SEMARAK ILMU PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage: https://semarakilmujournal.com.my/index.php/araset/index
ISSN: 2462-1943

A Comparative Study of Deep Neural Networks for Aedes Mosquito Species Acoustic Wingbeat Classification

Nur Faizah Hambali[1], Siti Salasiah Mokri[1,*], Asma' Abu Samah[1], Seri Mastura Mustaza[1], Asma' Amirah Nazarudin[1]

[1] Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Buit Environment, Universiti Kebangsaan Malaysia, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 13 May 2026
Received in revised form 30 May 2026
Accepted 1 June 2026
Available online 3 June 2026

ABSTRACT

In tropical regions, mosquito borne disease such as dengue and malaria remain major public health challenges. The common approaches to detect mosquito species have limitations in terms of large scale surveillance, the time taken and the need for specific taxonomic expertise. Thus, this study investigates identification approach using deep learning applied to mosquito wingbeat audio. Wingbeat signals were transformed into spectrogram images using the Short Time Fourier Transform which provides frequency representations as inputs for classification models. This paper investigates several deep learning architectures; convolutional neural networks (CNNs) based, such as ResNet, DenseNet and MobileNet as well as Capsule Networks and Vision Transformer. Experiments were conducted using the Kaggle mosquito wingbeat audio dataset with an 80/20 split for training and validation. Results showed that the Vision Transformer achieved the highest classification accuracy at 85% followed by DenseNet at 83.5%, CapsNet at 80%, MobileNet at 79.5% and ResNet at 78%. These findings show that deep learning based analysis of wingbeat spectrograms is a promising approach for mosquito species identification. Among the tested models, Vision Transformer achieved the highest validation accuracy in this experimental setup highlighting its potential to support automated vector surveillance and improved public health interventions.

Keywords:
Minimum three keywords; wingbeat classification, convolutional neural network, transfer learning, vision transformer, capsule network

1. Introduction

There are some disease borne mosquito species that cause malaria, dengue, Zika, yellow fever and chikungunya. It is estimated by The World Health Organization that over 700,00 deaths occur annually due to these deadly mosquitoes. Malaria accounts for an estimated 249 million clinical cases and more than 608,000 deaths each year [1]. In Malaysia, an estimation of 100,000 clinical cases of dengue are reported in 2024 [2]. The reasons for dengue spread are due to extensive urban expansion, ineffective waste disposal systems and the tropical climate. Thus, there is a need to control dengue mosquito populations through accurate species identification to help prevent the spread of dengue. Traditional mosquito species identification is done by an entomologists who

* *Corresponding author.*
E-mail address: siti1950@ukm.edu.my

https://doi.org/10.37934/araset.60.5.116

identifies the specimens under the microscope [3]. However, this method takes time, requires expertise and difficult for large scale surveillance [4]. In dengue endemic regions such as Malaysia, manual identification may create diagnostic bottlenecks when thousands of specimens are collected from many sites during routine monitoring or outbreak investigations [2]. In addition, accurate identification heavily depends on expert taxonomic skills especially to distinguish between Aedes aegypti and Aedes albopictus especially when the specimens are damaged or unclear. The collected specimens from the sites need to be transported to the laboratory also resulting in delay in the species identification [5]. Therefore, these limitations highlight the need for automated mosquito identification approaches. Acoustic based identification using mosquito wingbeat signals offers a promising alternative because wingbeat frequencies contain specific information that can be captured and analyzed computationally to determine the species. Alternatively, deep learning architectures for example Convolutional Neural Networks (CNNs) based transfer learning frameworks such as MobileNet, ResNet, and DenseNet as well as Capsule Network (CapsNet) can be used to automatically classify the mosquito species [6]. By delivering fast processing of mosquito wingbeat data, such models not only improve accuracy of species identification but indirectly allow for a strategic mosquito vector management [9].

Several studies have been reported to identify species specific mosquito audio wingbeat using deep learning. Wei et al. (2022) trained ResNet18 architecture on wingbeat recordings by extracting SpecAugment and Mel spectrogram. They achieved an accuracy of 89.9% on the WINGBEATS dataset and 98.9% on the ABUZZ collection [8]. Bilal et al. (2023) proposed a ResNet50 for mosquito species classification using acoustic features extracted from wingbeat sounds from Humbug dataset. They achieved a ROC-AUC score of 0.921 and a PR-AUC score of 0.885 [7]. Fanioudakis et al. (2018) used optical wingbeat recordings and showed that deep learning can extract discriminative acoustic features. This finding highlights the potential of deep learning methods for adapting signal based representations for mosquito identification [10].

Asgari et al. (2022) proposed an deep convolutional neural network for vector mosquito image classification by incorporating transfer learning and dropout layers to improve classification performance and reduce overfitting. Their model was validated on four publicly available mosquito datasets and achieved high classification accuracies of 98.82%, 98.92%, 94.66%, and 98.40% for each dataset [11]. Szekeres et al. (2023) adopted a ResNet 9 architecture to tackle multi class identification of mosquito species among other insect groups and achieved 95% classification accuracy [12].

In terms of network integration, Karuppaiah et al. (2023) integrated convolutional neural networks with Transformer blocks that achieved a classification accuracy of 81.27%. In this model, the convolutional layers extract spatial details and the Transformer captures temporal and spatial dependencies. The hybrid architecture maintained high performance even in acoustically cluttered settings [13]. Wang et al. (2024) applied ViT to environmental sound classification such as animal sounds such as cat, dog etc., and action sounds such as crying, laughing, coughing and recorded 93.75% accuracy on the test set. This indicates that ViT could substantially improve the characterization of mosquito wingbeat sounds, as compared to the classical convolutional networks in terms of robustness and accuracy [14].

In addition, Ong et al. used MobileNetV2 for mosquito species identification and deploying it for real application. Their study classified mosquito vectors such as Aedes aegypti, Aedes albopictus and Culex quinquefasciatus using mosquito image data collected in North Borneo, Malaysia [15]. Although this study was image based rather than acoustic based, it supports the practicality of lightweight MobileNet architectures for mosquito surveillance applications. Muhammad et al. (2023) analyzed the performance of CapsNet and CNN models on the detection of cough, breath and speech coswara dataset recordings. The audio was processed through feature extraction using MFCC.

Preprocessing involved class imbalance normalization and SMOTE oversampling. Results showed that CNN performed better with a higher accuracy of 94% as compared to CapsNet's (90%) with 10 fold cross validation. The study also showed that CapsNet performed poorly without preprocessing and CNN performed better showing CNN's robustness [16]. Li et al. (2022) proposed an enhanced version of CapsNet performed on weakly labeled polyphonic sound event detection (DCASE 2017 task 04) employing log Mel, multiresolution convolution, GRU layer and weight sharing routing with a reduction of parameters to 99%, obtaining F scores of 61. 1% for audio tagging and 50.5% for SED [17].

Most of the proposed deep learning frameworks for mosquito classification are based on conventional convolutional neural networks (CNNs). While CNNs have been broadly successful across diverse tasks, they struggle to maintain precise spatial localization due to missing critical positional information in the process. This limitation is significant when the spectrograms of mosquito wingbeats are used because precise spatial arrangement of spectrogram features directly influences classification outcomes. On the other hand, Vision Transformers (ViT) by Dosovitskiy et al. (2020) is based on self attention mechanism that recursively refining spatial hierarchies. This mechanism enable the ViTs to resolve multiscale patterns prevalent in spectrogram representations [18]. As a result, it is expected that ViT performs better than traditional CNNs in mosquito classification tasks.

The original contribution of this study the comparative evaluation of five different deep learning architectures: ResNet, DenseNet, MobileNet, CapsNet and ViT for Aedes mosquito wingbeat classification using STFT based spectrogram representations. Previous mosquito classification studies have commonly focused on CNN based models on handcrafted acoustic features or specific datasets such as HumBug, WINGBEATS, ABUZZ and smartphone based acoustic recordings. However, the application of Vision Transformer to mosquito acoustic classification remains relatively underexplored. Although ViT has demonstrated promising performance in environmental sound classification including animal and action sounds, its direct application to mosquito wingbeat spectrograms has not been systematically compared with established CNN based and alternative deep learning models. Therefore, this work contributes a comparative benchmark of ViT against ResNet, DenseNet, MobileNet and CapsNet for mosquito wingbeat classification.

2. Methodology

2.1. Data Acquisition

The dataset used in this analysis comprises mosquito wingbeat sounds which were collected from several open access bioacoustic databases on Kaggle. The dataset contains 4 mosquito species: Aedes aegypti, Aedes albopictus, Noise and Others. These recordings were preserved in controlled settings that best characterized the distinct frequency of wingbeats of each species. The sounds were recorded in a soundproof environment that eliminates any external noise interference.

All audio files in the dataset are appropriately labeled and stored in WAV format with a uniform length of 0.6 seconds and an approximate file size of 9.8 KB. Recordings were captured at a sampling frequency of 8 kHz. The dataset is divided into 80% for model training and 20% testing. The complete collection includes 1,628 samples. The number of recording audio for each class in "Wingbeats" database is summarized in Table 1.

Table 1
Wingbeat Database

Class	Number of recording audio
Ae. Aegypti	486
Ae. Albopictus	400
Noise	242
Others	500
Total	1628
Train (80% total)	1302
Test (20% total)	326

Figure 1 presents an example of an Aedes aegypti wingbeat signal acquired during flight. The waveform shows the signal amplitude over time. The oscillatory pattern reflects the periodic wing motion of the mosquito. Such wingbeat signals contain fundamental frequency and harmonic components that can be captured using optical or optoacoustic sensing techniques and used to characterize mosquito flight activity [19].

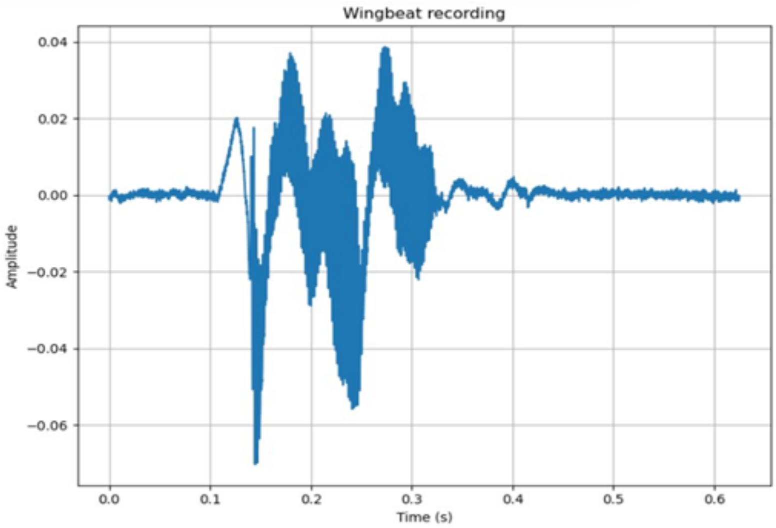


Fig. 1. Optical recording of an Aedes aegypti wingbeat

2.2. Data Preprocessing

The audio files were transformed into spectrogram representations. The spectrogram can accommodate 2D input to the deep learning architectures principally CNNs. Native audio waveforms are 1D signals and are incompatible with CNNs that are based on spatial characteristic of image data. By generating a spectrogram using Short Time Fourier Transform (STFT), it preserves temporal frequency content and at the same time it is a 2D grid that CNNs can process effectively. In the spectrogram generation algorithm, the audio stream is segmented into overlapping windows. Then, Fourier analysis is used to each segment to map the frequency content to temporal indices. This results in a 2D array where the abscissa is the temporal bins while the ordinate is the frequency bins. The pixel intensity represents the magnitude of the frequency component at the corresponding time. For the post processing, the spectrogram dataset was divided into training and testing subsets at an 80:20 ratio. This maximizes data utility for model training. Each class within the test subset is also equally represented.

2.3. Model Architecture

This study applies a wide range of deep learning architectures to evaluate their suitability for mosquito wingbeat spectrogram classification. ResNet, DenseNet and MobileNet were selected as representative of CNN based models. Each model offers different advantages in deep feature learning, feature reuse and computational efficiency. CapsNet was included to examine an alternative representation learning approach that preserves spatial relationships between spectrogram features. Finally, ViT was selected to explore the potential of self attention in capturing long range time frequency dependencies. By comparing these diverse architectures under the same experimental setting, this study provides a broader assessment of how different deep learning strategies perform for acoustic mosquito species classification from lightweight models to attention based approaches.

2.3.1 Transfer Learning Model

Transfer learning adopts a model trained on a large dataset to a distinct problem by avoiding scratch training. Transfer learning CNNs such as ResNet, DenseNet and MobileNet are designed to minimize training epochs and to generalize effectively across multiple data. Each architecture is initialized with pretrained weights derived from the ImageNet data. This procedure accelerates the optimization and enhances the final classification accuracy on the task of mosquito wingbeat sound. The configuration of the transfer learning model is shown in Figure 2 that summarizes the principal modules and their interconnections for the three architectures.

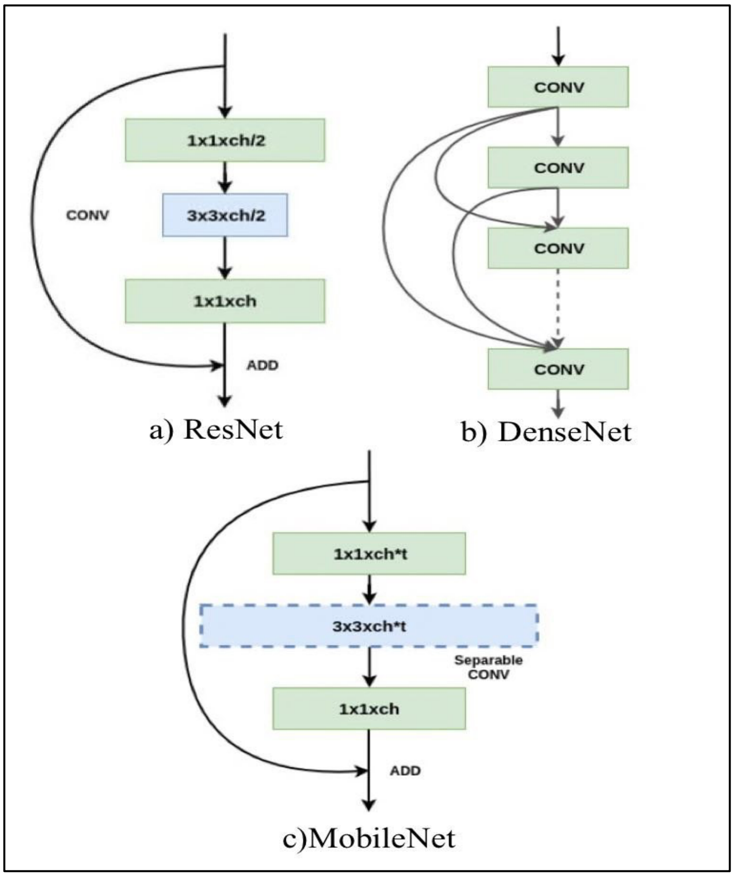


Fig. 2. Transfer Learning Model Architecture [17]

ResNet is based on residual learning framework with skip connections are used to facilitate the direct passage of activations from earlier to later layers and thus preserves gradient flow and accelerating convergence without excessive computational burden. This topology is suitable for intricate datasets such as the classification of mosquito wingbeat [21].

In DenseNet, each layer is connected to all of previous layers so that the previous information is maintained. This makes the model able to capture and process complex features. DenseNet enhances feature propagation and making it more efficient in terms of computation and memory. Therefore, DenseNet is particularly good for tasks involving complex data like spectrograms [20].

MobileNet is designed for efficient resource utilization. It is suitable for deployment on mobile and embedded hardware constrained by processing capability. The burden of computational load in MobileNet is low as the standard convolutions have been substituted by depthwise separable convolutions. This characteristic leads MobileNet well suited for real time classification of mosquito species on devices of minimum resources. Table 2 represents the hyperparameters selected for fine tuning the transfer learning model applied to wingbeat mosquito classification.

Following an exhaustive series of experiments and hyperparameter testing, the values in Table 2 are the most effective to achieve peak classification performance.

Table 2
Transfer Learning Model Parameter and Values

Parameter	**Value**
Epochs	50
Learning Rate	0.0001
Optimizer	Adam
Batch Size	32

2.3.2 Capsule Network

Capsule Network (CapsNet) is a neural network model introduced by Geoffrey Hinton and colleagues at the University of Toronto in 2017 [21].The primary objective of capsule networks is to overcome the limitations of standard Convolutional Neural Networks (CNNs) in terms of managing changes in the position, scale and rotation of the object in question. The main unit formed in CapsNet is called a capsule. A capsule is a collection of neurons which captures an object’s attributes such as pose, size and deformation where both inputs and outputs are in vector form [18]. Capsules are arranged in hierarchical layers. Higher level capsules correspond to more complex object components [22]. Figure 3 shows structure of the capsule network [23].

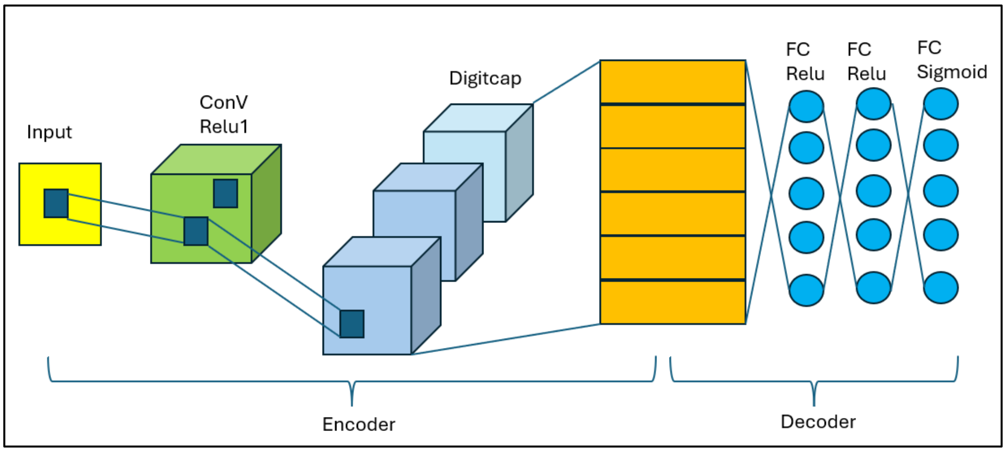


Fig. 3. Capsule Network Architecture

The convolutional layer stands in the first position of the network and is responsible for feature extraction. Output from the convolutional layer is fed as input to the primary capsule layer. The dynamic routing algorithm is responsible for primary to higher level capsule connections. It allows the network to learn to aggregate pose feature vectors and compute them into more abstract feature representations. The DigitCaps layer is a fully connected layer that consists of ten 16 dimensional capsules. Each of them is fully connected to all capsules of the previous layer [23]. Finally, the length of the output capsule is computed which is output from the previous layer. The training parameters used in the CapsNet model for the classification of mosquito species is shown in Table 3. After extensive testing and iterations, these parameters were identified as the optimal parameters that achieved high accuracy and consistent performance.

Table 3
CapsNet Model Parameters and Values

Parameter	Value
Epochs	50
Learning Rate	0.0001
Optimizer	Adam
Batch Size	32
Conv Kernel	128
Primary Capsules	32
Secondory Capsules	10
Primary Capsule Vector	8
Secondary Capsule Vector	32

2.3.3 Vision Transformer

Transformers were introduced by Vaswani et al. (2017) [24] in the context of neural machine translation initially for Natural Language Processing (NLP) applications. The Vision Transformer (ViT) is adopted form Transformer network for classification task and has demonstrated excellent performance by effectively modeling local and global correlations in image data in comparison to the established Convolutional Neural Networks (CNN) [18]. Figure 4 illustrates the model of the Vision Transformer (ViT) architecture. First, the input image is segmented into non overlapping patches (16 × 16 pixels). Each patch is subsequently reshaped into a one dimensional vector and subjected to a linear embedding layer. Following this embedding transformation, a positional embedding vector that corresponds to each patch's discrete location in the image grid, is incorporated to reintroduce spatial structure [18].

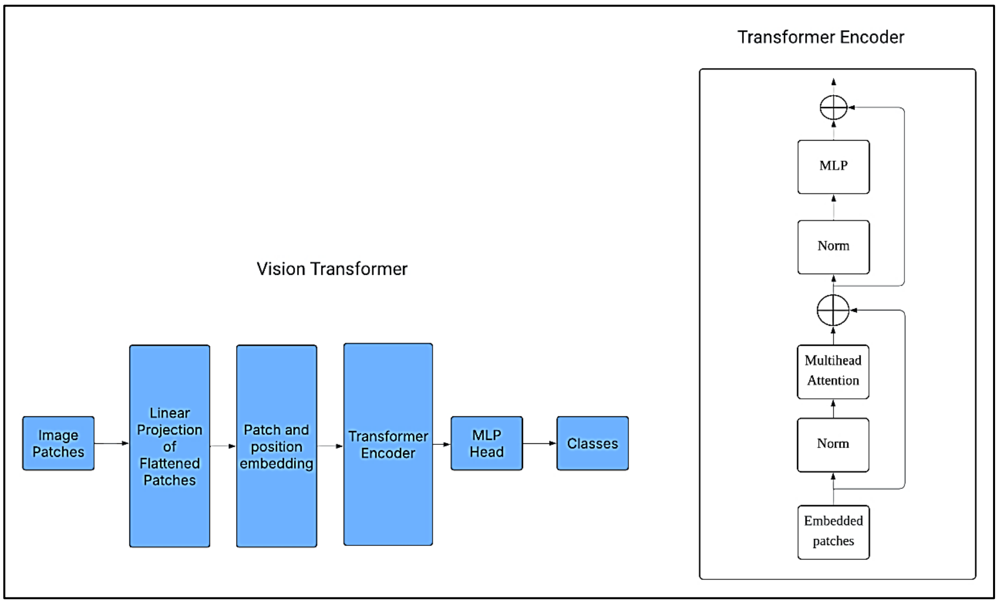


Fig. 4. Vision Transformer Architecture [15]

The integrated vector sequence comprises of positionally encoded patches is subsequently input to a multilayer Transformer encoder. This architecture is characterized by stacked self attention blocks to evaluate the relational structure among all patches concurrently. This approach acquires long range dependencies over the spatial dimensions of the image. To augment the representational capacity, the multihead attention framework executes several attention computations in parallel. Following attention aggregation, the feature tensor is forwarded through a stack of multilayer perceptrons that produces the final classification. The parameter configuration that yielded peak performance in the Vision Transformer for mosquito wingbeat classification is presented in Table 4.

Table 4
Vision Transformer Model Parameters and Values

Parameter	**Value**
Epochs	50
Learning Rate	0.0001
Optimizer	Adam
Batch Size	32
Image Size	128
Patch Size	16
Projection Dimension	64
Transformer Layers	8
MLP Head Units	[2048, 1024]

2.4. Model Training and Evaluation

The models were developed on a dataset divided according to a standard 80/20 ratio for training and testing. The optimization utilized Adam algorithm with a learning rate of 0.0001. Each architecture underwent 50 epochs of training and early stopping criteria was applied to avoid overfitting. Performance evaluation used are accuracy, precision, recall and the F1 score. Accuracy represents the fraction of correct classifications across the entire dataset. Precision isolates the ratio of true positive identifications against the total predicted positives while recall captures the ratio of true positives against the total actual positives. The F1 score is the mean of precision and recall. Accuracy is calculated according to Equation (1)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

where TP is True positive, FN is False Negative, FP is False Positive and TN is True Negative.

3. Results and Discussions

3.1. Spectrogram Transformation

In this investigation, spectrograms extracted from recorded mosquito wingbeat audio served as input for all deep learning architectures. The Short Time Fourier Transform (STFT) was applied to the unprocessed audio waveforms that produced frequency time distributions. The conversion of audio signals to spectrograms enables the models to use both spectral and temporal features that are important for classification. Following this process, the spectrograms were systematically presented to each classifier for species assignment.

In Figure 5, a spectrogram of the Aedes aegypti wingbeat is shown. This visual representation is crucial for understanding the variations in wingbeat patterns which are essential for accurate classification of mosquito species.

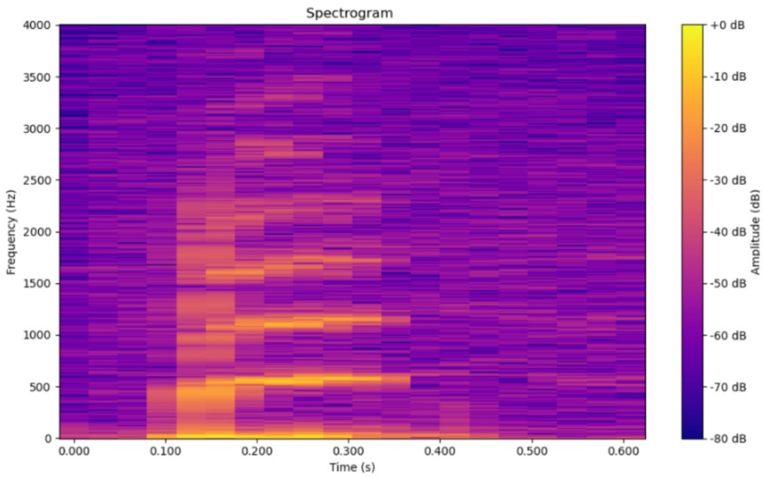
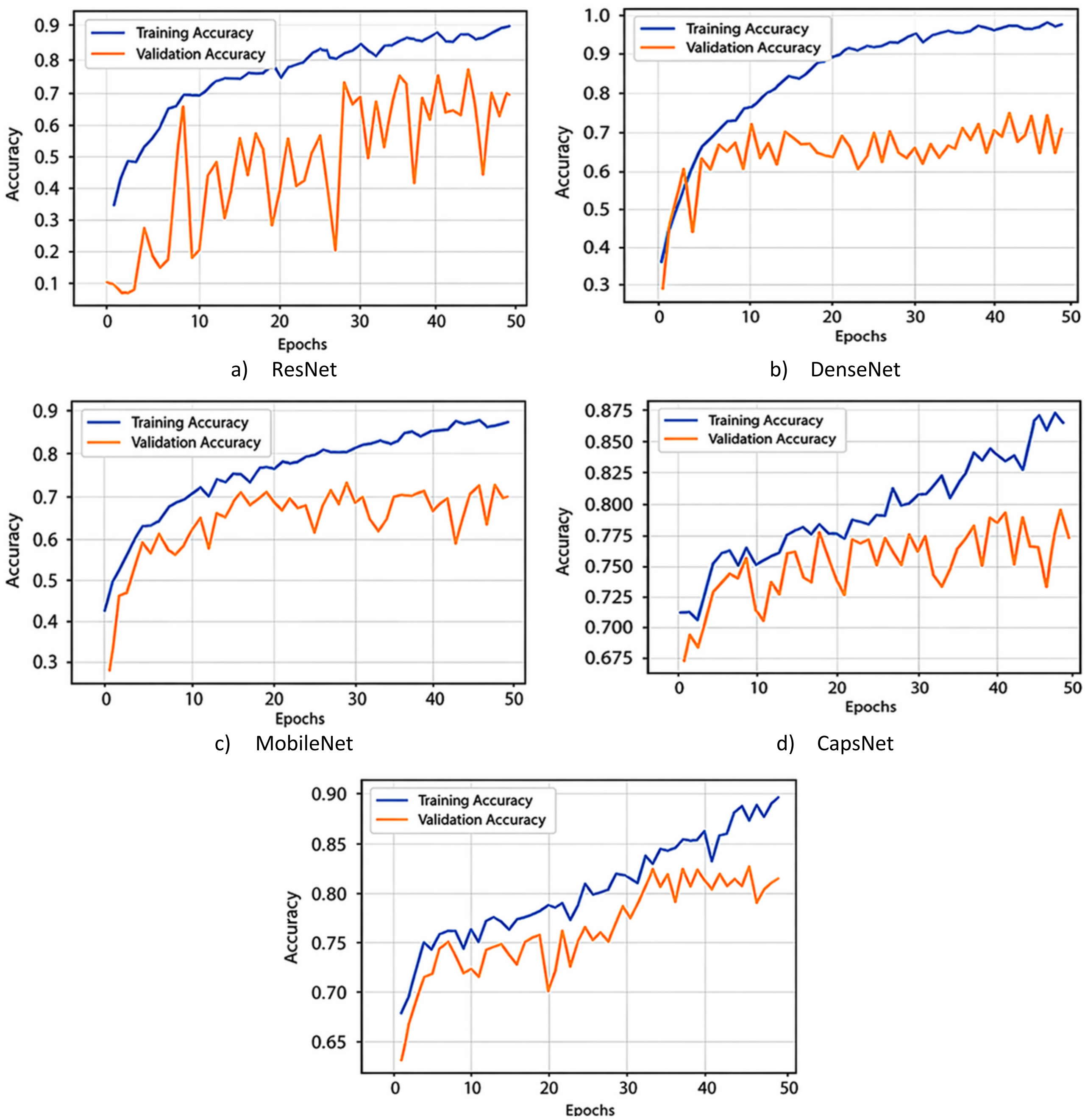


Fig. 5. Spectrogram for an Aedes Aegypti audio sample

3.2. Model Training Parameter

In this study, the training and testing accuracy curves reveal the learning behavior and performance of the models as training progresses. The training accuracy curve records the proportion of correctly classified mosquito wingbeat samples during training while the testing accuracy curve reflects performance on a distinct test dataset. These curves highlight overfitting and underfitting. Overfitting is a condition where the training accuracy is high but the testing accuracy is low. Conversely, underfitting is shown when the model performs poorly on both training and testing sets which indicates insufficient learning from the data to form reliable predictions [25].

In the experiments conducted, ResNet exhibits a gradual upward trend in training accuracy. However, the test accuracy varies widely showing possible overfitting. Figure 6a illustrates this behavior and the test accuracy plateaus at roughly 78% after 50 epochs. By contrast, DenseNet, depicted in Figure 6b, records a smoother trajectory with both training and validation accuracy rising steadily. By the 50th epoch, DenseNet achieves a validation accuracy of 83.5% reflecting its robust generalization capabilities.



a) ResNet
b) DenseNet
c) MobileNet
d) CapsNet
e) ViT

Fig. 6. Training and Validation Accuracy

MobileNet as shown in Figure 6c continues to produce reliable results throughout training. Training accuracy rises uniformly along with the validation accuracy and graphs at 79.5% by epoch 50. This performance indicates improved generalization over previous configurations. Next, as shown in Figure 6d, CapsNet exhibits significant increases in both training and validation accuracy to the 20th epoch. Following this, validation accuracy begins to exhibit some volatility. Overall, it continues to rise reaching 80% by the 50th epoch. This indicates a strong initial learning phase followed by some instability but the model still demonstrates solid performance. The Vision Transformer (ViT), depicted in Figure 6e is the best among all architectures. Its training and validation accuracy graphs move upward without fluctuation and by epoch 50, validation accuracy peaks at 85% making ViT as the leading model for classifying mosquito species.

Overall, Vision Transformer achieved the leading validation accuracy for this task. Second is DenseNet that achieved strong accuracy with good computational efficiency. CapsNet also performed well although some validation accuracy dip in the later epochs. MobileNet maintained consistent performance and balanced accuracy. However, ResNet seems to give inconsistent test accuracy.

3.3. Model Evaluation
3.3.1 Confusion Matrix

The confusion matrix shows the performance of each tested model to classify Aedes aegypti, Aedes albopictus, Noise and Other Mosquitoes. As shown in Figure 7, all models have successfully classified the Noise class accurately (49 samples). But in terms of differentiating between Aedes aegypti and Aedes albopictus, each model performed differently. For ResNet (Figure 7a), it correctly classified 66 out of 97 Aedes aegypti, incorrectly classified as Aedes albopictus and 22 as Other Mosquitoes. For Aedes albopictus, ResNet correctly classified 44 out of 80 samples while 28 were misclassified as Aedes aegypti and 8 as Other Mosquitoes. The results indicates that ResNet cannot easily differentiate these two close species.

DenseNet as shown in Figure 7b performed better compared to ResNet by correctly identifying 74 out of 97 Aedes aegypti samples and 55 out of 80 Aedes albopictus sample. MobileNet as shown in Figure 7c obtained the best classifier among the CNN based models with 79 out of 97 samples correctly identified. However, it performed least effectively for Aedes albopictus where 35 samples misclassified as Aedes aegypti. In Figure 7d, CapsNet produced a more balanced performance between the two Aedes species. It correctly identified 74 Aedes aegypti samples and 62 out of 80 Aedes albopictus samples. Finally, ViT as shown in Figure 7e achieved the best performance for Aedes aegypti. It correctly classified 82 out of 97 samples. It also correctly classified 51 out of 80 Aedes albopictus samples and maintained perfect classification for the Noise class.

Overall, the best classifier for Aedes aegypti and Aedes albopictus are ViT and CapsNet respectively. MobileNet showed the weakest performance for Aedes albopictus and ResNet showed the weakest performance for Aedes aegypti. These results indicate that the main classification challenge was to differentiate between Aedes aegypti and Aedes albopictus. However, Noise samples were accurately recognized by all models.

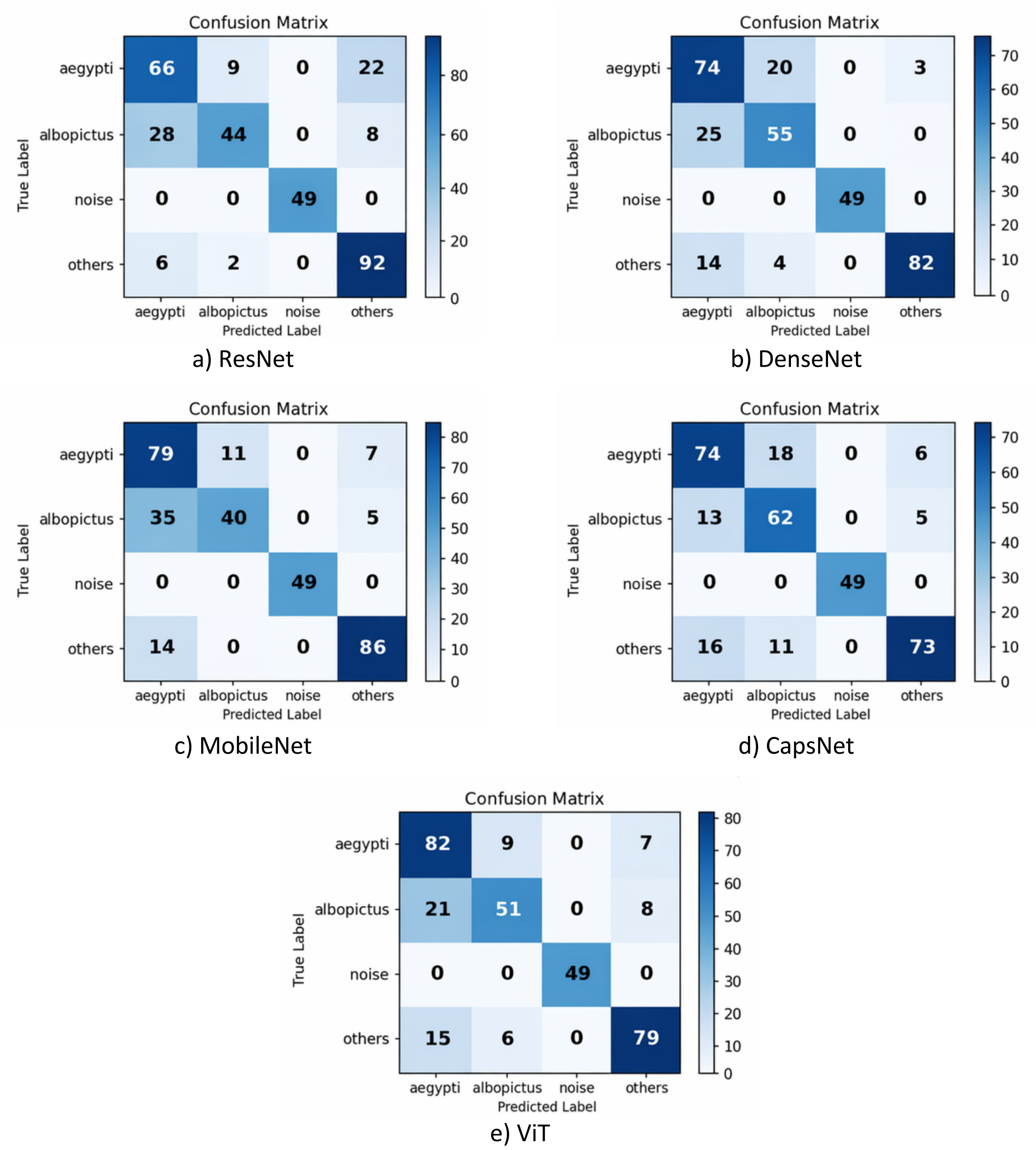


Fig. 7. Confusion Matrix for all models

3.3.2 Performance Metrics

Model performance is first examined through the confusion matrix. From this, accuracy, precision, recall and the F1 score are derived to provide a comprehensive view of each model's classification capability for mosquito species. Together, these numerical indicators reveal not only how well each species is correctly identified but also how the models behave in the presence of non target examples such as background noise and other mosquito species. As summarized in Table 5, the ResNet architecture achieved perfect classification of the Noise class (100%) and a precision of 92% for the Other Mosquitoes category. Nevertheless, performance deteriorated for the target species: Aedes aegypti obtaoned a precision of only 0.66 while Aedes albopictus recorded a recall of 0.55. These show that although ResNet is robust, it faces difficulty in differentiating between these closely related species.

DenseNet in Table 6 surpassed ResNet in interspecies differentiation. It obtained an accuracy of 0.81 for Aedes aegypti and 0.85 for Aedes albopictus. Furthermore, DenseNet exhibited good precision and recall for the 'Other Mosquitoes' category having a recall of 0.82. In comparison, MobileNet (Table 7) exhibited performance metrics that were comparable, recording a slight reduction in accuracy for Aedes aegypti (0.79) and Aedes albopictus (0.84). MobileNet excelled in the Noise class with 100% accuracy and the Other Mosquitoes classification was strong yielding an F1 score of 0.87. These characteristics render MobileNet a competitive candidate for implementation within resource constrained, real time mosquito classification systems.

CapsNet (Table 8) achieves a significant enhancement in the accuracy of Aedes aegypti (0.84) and Aedes albopictus (0.86) in comparison to the MobileNet and DenseNet models. These species' F1 scores are also quite higher than 0.74 and 0.73 which show a better balance of precision and recall. It achieves perfect scores for the noise class and still performs well with an F1 score of 0.79 for the Other Mosquitoes class. Meanwhile, Vision Transformer (ViT) in Table 9 superseded all preceding architectures with accuracy of 0.86 for Aedes aegypti and 0.87 for Aedes albopictus. The ViT model achieved good recall and F1 scores within the Noise class as well as Other Mosquitoes. Its capacity to resolve the most minute variations in mosquito wingbeat characteristics establishes ViT as the most effective architecture in the current investigation. While Vision Transformer (ViT) demonstrated the best overall performance particularly in distinguishing between similar mosquito species, both DenseNet and CapsNet proved to be highly consistent models with strong performance across all classes.

Post hoc analysis of the five deep learning architectures provides additional insight into their comparative performance for mosquito wingbeat classification. In the current experimental setup, ViT achieved the highest overall accuracy at 85% followed by DenseNet at 83.5%, CapsNet at 80%, MobileNet at 79.5% and ResNet at 78%. This indicates that ViT obtained a 7 percentage point improvement over ResNet and a smaller 1.5 percentage point improvement over DenseNet. At the class level, ViT achieved the highest reported accuracy for Aedes aegypti at 86% while CapsNet achieved the highest accuracy for Aedes albopictus at 86% suggesting that different architectures may have complementary strengths in distinguishing closely related mosquito species. All models achieved 100% classification accuracy for the Noise class indicating that the main classification challenge was not noise detection but differentiation between mosquito species. For the two Aedes classes, CapsNet and ViT produced comparable mean F1-scores. This suggests that their species level performance was relatively close despite ViT achieving the highest overall accuracy.

This study provides an initial proof of concept evaluation using a public Kaggle mosquito wingbeat dataset consisting of 1,628 samples and an 80:20 train testing split. This setup offers a useful starting point for comparing different deep learning architectures under a consistent experimental condition. However, the findings should be interpreted with consideration of several limitations. First, the dataset has an imbalanced class distribution across Aedes aegypti, Aedes albopictus, Noise and Other Mosquitoes which may influence model learning and contribute to differences in precision and recall across classes. Second, the audio recordings were obtained under controlled or soundproof conditions at an 8 kHz sampling frequency. Thus, the dataset may not fully represent real world surveillance environments where background noise such as rain, wind, human speech, vehicle sounds, air conditioning and other insect sounds can affect acoustic signals. Third, although early stopping was applied, some models showed fluctuations between training and validation performance suggesting possible overfitting risks.

Further validation using other mosquito wingbeat datasets such as HumBug, ABUZZ and WINGBEATS would also help determine whether the observed model performance is dataset specific or transferable to other mosquito acoustic recordings. In addition, future studies should examine model robustness under realistic field conditions by adding environmental noise during training and testing together with systematic signal to noise ratio analysis to identify which architecture is most reliable under noisy conditions. To improve the broader applicability of the model, the dataset should also be expanded to include more mosquito species particularly important disease vectors such as Anopheles and Culex. Finally, future studies should evaluate the feasibility of deploying these models on edge devices by measuring inference time, memory usage, model size, power consumption and the effects of model compression or quantization. Such evaluation would help identify models that are not only accurate but also practical for real time mosquito surveillance.

Table 5
ResNet Performance Metrics

Class	Accuracy	Precision	Recall	F1-Score
Aedes Aegypti	0.80	0.66	0.68	0.67
Aedes Albopictus	0.85	0.80	0.55	0.65
Noise	1.00	1.00	1.00	1.00
Other Mosquitos	0.88	0.75	0.92	0.83

Table 6
DenseNet Performance Metrics

Class	Accuracy	Precision	Recall	F1-Score
Aedes Aegypti	0.81	0.65	0.76	0.70
Aedes Albopictus	0.85	0.70	0.69	0.69
Noise	1.00	1.00	1.00	1.00
Other Mosquitos	0.94	0.96	0.82	0.87

Table 7
MobileNet Performance Metrics

Class	Accuracy	Precision	Recall	F1-Score
Aedes Aegypti	0.79	0.81	0.70	0.62
Aedes Albopictus	0.84	0.50	0.61	0.78
Noise	1.00	1.00	1.00	1.00
Other Mosquitos	0.92	0.86	0.87	0.88

Table 8
CapsNet Performance Metrics

Class	Accuracy	Precision	Recall	F1-Score
Aedes Aegypti	0.84	0.72	0.76	0.74
Aedes Albopictus	0.86	0.68	0.78	0.73
Noise	1.00	1.00	1.00	1.00
Other Mosquitos	0.88	0.87	0.73	0.79

Table 9
ViT Performance Metrics

Class	Accuracy	Precision	Recall	F1-Score
Aedes Aegypti	0.86	0.69	0.84	0.76
Aedes Albopictus	0.87	0.77	0.64	0.70
Noise	1.00	1.00	1.00	1.00
Other Mosquitos	0.91	0.84	0.79	0.81

4. Conclusion

This paper investigates the performance of Vision Transformer (ViT) compared with ResNet, DenseNet, MobileNet and CapsNet for classifying mosquito species based on their wingbeat frequencies. The models were tested to distinguish between Aedes aegypti, Aedes albopictus, Noise and Other Mosquitoes. In the present experimental setup, ViT achieved the highest accuracy of 85%, followed by DenseNet at 83.5%, CapsNet at 80%, MobileNet at 79.5% and ResNet at 78%. The performance of ViT may be attributed to its self attention mechanism which enables the model to capture important local and global patterns across spectrogram representations. DenseNet also performed strongly due to its dense connectivity pattern that supports feature reuse and efficient information propagation across layers. CapsNet provided a different feature representation approach through capsule based learning which may help preserve spatial relationships and hierarchical patterns in spectrogram data. Overall, the findings suggest that deep learning based spectrogram analysis is a promising approach for mosquito wingbeat classification.

Acknowledgement and Funding

The authors would like to acknowledge the Ministry of Higher Education Malaysia (MOHE) and the Universiti Kebangsaan Malaysia for the research grant: Fundamental Research Grant Scheme (FRGS) Project Code: FRGS/1/2024/TK07/UKM/02/6.

References

[1] World Health Organization. "Global Health Statistics." World Health Organization, 2024. https://www.who.int/statistics.

[2] Ministry of Health Malaysia. "Dengue Surveillance Report." Ministry of Health Malaysia, 2024. https://www.moh.gov.my.

[3] Lakyiere, A. B., R. Mensah, N. Y. Agbenor-Efunam, E. Yamba, and K. Badu. "Trends and Advances in Image-Based Mosquito Identification and Classification Using Machine Learning Models: A Systematic Review." *Computers in Biology and Medicine* (2025). https://doi.org/10.1016/j.compbiomed.2025.110373.

[4] Mandal, S., and P. Singh. "AI and ML Approaches for Automated Mosquito Species Identification: A Review." In *2026 5th International Conference on Innovative Practices in Technology and Management (ICIPTM)*. https://doi.org/10.1109/ICIPTM69057.2026.11465938.

[5] Nolte, K., F. Sauer, J. Baumbach, P. Kollmannsberger, C. Lins, and R. Lühken. "Robust Mosquito Species Identification from Diverse Body and Wing Images Using Deep Learning." *Parasites & Vectors* (2024). https://doi.org/10.1186/s13071-024-06459-3.

[6] Kiskin, Ivan, Bernardo Pérez Orozco, Theo Windebank, Davide Zilli, Marianne Sinka, Kathy Willis, and Stephen Roberts. "Mosquito Detection with Neural Networks: The Buzz of Deep Learning." *arXiv preprint* arXiv:1705.05180, 2017.

[7] Bilal, Muhammad, Abdul-U. Rehman, and Sajid Razzaq. "Mosquitoes Species Classification Using Acoustic Features of Wing Beats." In *2023 International Conference on Communication Technologies (ComTech)*. 2023.

[8] Wei, Xin, Md. Zahangir Hossain, and Kamal A. Ahmed. "A ResNet Attention Model for Classifying Mosquitoes from Wing-Beating Sounds." *Scientific Reports* 12 (2022): Article 10334.

[9] Siddiqui, B. R., and V. Jain. "Convolution Neural Network-Based Mosquito Classification System." In *Proceedings of the 3rd International Conference on Artificial Intelligence and Signal Processing (AISP 2022)*. 2023.

[10] Fanioudakis, Eleftherios, Matthias Geismar, and Ilyas Potamitis. "Mosquito Wingbeat Analysis and Classification Using Deep Learning." In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2410–2414. 2018.

[11] Asgari, Marjan, Amir Sadeghzadeh, Md. Baharul Islam, Luminita Rada, and John Bozeman. "Deep Learning-Based Vector Mosquitoes Classification for Preventing Infectious Diseases Transmission." *Image Analysis and Stereology* 41, no. 3 (2022): 203–215.

[12] Szekeres, Bence J., Márk Natabara Gyöngyössy, and János Botzheim. "A ResNet-9 Model for Insect Wingbeat Sound Classification." In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, 587–592. 2023.

[13] Shiva Charan, Oruganti, and Deepa Karuppaiah. "A Hybrid Network Combining CNN and Transformer Encoder to Classify Mosquitoes Based on Wing Beat Frequencies." *SSRN Electronic Journal*, 2023.

[14] Hong, H., and Y. Li. "Environmental Sound Classification Using Vision Transformer." In *International Conference on Electronic Information Technology and Smart Agriculture*, *Proceedings of SPIE* 12981, Article 129811Z. 2024.

[15] Ong, S. Q., N. H. A. R. Rahman, M. I. M. Yusoff, J. Jeffree, K. Y. Lum, and C. D. Chen. "Application of Computer Vision and Deep Learning Models to Automatically Classify Medically Important Mosquitoes in North Borneo, Malaysia." *Bulletin of Entomological Research* (2024).

[16] Muhammad, Abdulaziz, Muhammet Ali Arserim, and Ömer Türk. "Compare the Classification Performances of Convolutional Neural Networks and Capsule Networks on the Coswara Dataset." *Dicle University Journal of Engineering* 14, no. 2 (2023): 265–271.

[17] Li, Haitao, Shuguo Yang, and Wenwu Wang. "Improved Capsule Routing for Weakly Labeled Sound Event Detection." *EURASIP Journal on Audio, Speech, and Music Processing* 2022 (2022): Article 5.

[18] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *International Conference on Learning Representations (ICLR)*. 2021.

[19] Potamitis, Ilyas, and Iraklis Rigakis. "Measuring the Fundamental Frequency and the Harmonic Properties of the Wingbeat of a Large Number of Mosquitoes in Flight Using 2D Optoacoustic Sensors." *Applied Acoustics* 109 (2016): 54–60.

[20] Donthi Reddy, N. R. S., Arun Ross Rattani, and Reza Derakhshani. "Comparison of Deep Learning Models for Biometric-Based Mobile User Authentication." In *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 1–7. 2018.

[21] Patrick, Mensah Kwabena, Adebayo Felix Adekoya, Ayidzoe Abra Mighty, and Baagyire Y. Edward. "Capsule Networks – A Survey." *Journal of King Saud University - Computer and Information Sciences* 34, no. 1 (2022): 1295–1310.

[22] Khan, Asif Iqbal, Junaid Latief Shah, and Mohammad Mudasir Bhat. "CoroNet: A Deep Neural Network for Detection and Diagnosis of COVID-19 from Chest X-Ray Images." *Computer Methods and Programs in Biomedicine* 196 (2020): Article 105581.

[23] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic Routing Between Capsules." In *Advances in Neural Information Processing Systems 30*. 2017.

[24] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30*. 2017.

[25] Singla, M., K. S. Gill, M. Kumar, and R. Rawat. "Classification of Musical Genres Utilizing the CNN Sequential Model and Deep Learning Techniques." In *Proceedings of the IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems*, 1–5. 2024.