SiulMalaya: an annotated bird audio dataset of Malaysia lowland forest birds for passive acoustic monitoring

Nursuriati Jamil[1], Ahmad Nazem Norali[2], Muhammad Izzad Ramli[1], Ahmad Khusaini Mohd Kharip Shah[3], Ismail Mamat[3]
[1]College of Computing, Informatics, and Media, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]GBA Corporation Sdn Bhd, Wisma GBA, Petaling Jaya, Malaysia
[3]Department of Wildlife and National Parks, Agro-Biotechnology Institute (ABI), Kuala Lumpur, Malaysia

Article Info

Article history:

Received Nov 15, 2022
Revised Dec 21, 2022
Accepted Jan 16, 2023

Keywords:

Annotated audio dataset
Bird identification
Convolutional neural network
Lowland forest birds
Passive acoustic monitoring

ABSTRACT

The laborious point count method of conducting bird surveys is still a common practice in Malaysia. An alternative method known as passive acoustic monitoring (PAM) is deployed in many countries by placing sound recorders at surveying sites to collect bird sounds. Studies revealed that the number of bird densities counted by human observers was agreeable with those obtained using PAM. However, one of the most essential gaps in conducting PAM is the lack of expert-verified bird-call databases. Therefore, the aim of this study is to construct the first annotated Malaysia lowland forest bird sounds called *SiulMalaya* to be used as ground-truth datasets for PAM-related activities. The raw bird sounds dataset was downloaded from Macaulay Library using the eBird platform. Data pre-processing was done to produce annotated audio tracks that can be used as training datasets for bird classification. *SiulMalaya* dataset was further validated by two bird experts from the Department of Wildlife and National Parks, Malaysia. A bird identification experiment was carried out to assess and validate *SiulMalaya* dataset using a convolutional neural network (CNN) learning model. Even though the accuracy of bird identification is slightly above 50%, the annotated dataset is shown to be viable for PAM-related operations.

Corresponding Author:

Muhammad Izzad Ramli
College of Computing, Informatics, and Media, Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Email: izzadramli@uitm.edu.my

1. INTRODUCTION

The lowland dipterocarp forest in Malaysia generally refers to forests located below 300 meters [1], facing a greater threat of an estimated 45.3% total loss between 2000 and 2010 as reported by Miettinen *et al.* [2]. The constant deforestation and human-related activities have threatened the survival of lowland forest birds [3], thus a national red list for the birds of Malaysia was recommended by Lang *et al.* [4] to align the global International Union for Conservation of Nature (IUCN) red list with the conservation status of birds in Malaysia. Official work on bird conservation and surveys are collectively done by the bird conservation council (BBC) of the Malaysian nature society (MNS) and the Department of Wildlife and National Parks Peninsular Malaysia. Bird surveying and monitoring in Malaysia was and still is commonly done by volunteers or expert observations at the site of the survey. Terrestrial surveys were either conducted by foot [5]–[7] or using boats [5] and sometimes aerial surveys [5]. The most widely used method to estimate the number of birds is using the point count method where the volunteers and experts remain stationary at the point area for a certain period jotting down the species seen or heard. The bird counting process is often very laborious and requires the involvement of volunteers and researchers to be stationed at the different survey sites.

The use of acoustic sensors or passive acoustic monitoring (PAM) in biodiversity monitoring has shown an increase of fifteenfold in publications since 1992 [8]. PAM has several advantages over the traditional point count method for bird surveying and monitoring based on several factors [9] such as: i) sound recording of the birds can be done in the absence of the observer; ii) allowing monitoring to be done in areas that are difficult to access; and iii) vocally less active and nocturnal birds can be recorded using highly sensitive acoustic sensors. A study by Leach *et al.* [9] compared point count method and automated acoustic monitoring for detecting birds in a rainforest in Queensland, Australia. The results showed that each method detected different unique species, thus the authors recommended the adoption of both methods for future biodiversity assessments as they are complementary to each other. A more recent study by Pérez-Granados and Traba [10] compared 31 articles using autonomous recording devices and human observers to estimate bird densities. Twenty-six studies revealed that the number of bird densities counted by human observers was agreeable with those obtained using autonomous recording devices. Therefore, the use of PAM in bird monitoring is encouragingly positive and should be pursued to bring an impact on biodiversity research. However, one of the most essential gaps in conducting PAM is the lack of expert-verified bird-call databases [11]. Ground-truth databases are time-consuming, laborious, and costly as high-quality audio recordings are difficult to acquire in the wild. The two most common bird calls databases contributed by citizen sciences are the Macaulay library [12] by Cornell Lab of Ornithology and Xeno-Canto [13]. While the Macaulay library contains audio, photos, and videos of birds, amphibians, fishes, and mammals, the xeno-canto database comprises only bird audio. A search for bird sounds in Xeno-Canto shows 1,303 non-captive bird sounds in Malaysia. Xeno-canto welcomes all bird sounds recording regardless of the quality and contributors must ensure that the bird's sound is identified correctly before it is uploaded. On the other hand, there are more than 1.3 million bird sounds in the Macaulay library as of 7th May 2022. A total of 6,506 bird sounds found in Malaysia are retrieved from Macaulay library for this research using the eBird [14] platform. The bird sounds comprised 3,007 songs, 2,630 calls, 3 flight songs, 24 flight calls, 70 non-vocals, 30 dawn songs, and 85 duets. All the bird sounds are from non-captive birds and their identities are confirmed. Six unconfirmed bird sounds are discarded. As Macaulay library offers more bird sounds recorded in Malaysia and the quality of the recordings are strictly adhered to by the contributors, we opt to use these sound datasets in our study.

The bird sound files uploaded by the contributors are in raw forms with a lot of noise. Macaulay library provides steps for contributors to adhere to, before uploading their sound recordings. The contributors are to trim the ends of the recording and boost the volume. Furthermore, group recordings of the same bird are stored in one long audio file with 1-second intermittent between each bird. They are also encouraged to append voice announcements and avoid using filtering or cosmetic editing. Thus, some recordings are preceded by human voice announcements and other animal sounds. One recording even has a call for prayer and most have overlapping bird sounds. The noisy bird sounds posed no problem if the bird identification is done by a human expert. However, automated bird identification requires a clean and annotated training dataset, especially systems that utilize machine learning techniques. The largest annual bird species identification based on audio recording known as the bird cross-language evaluation forum (BirdCLEF) challenge [15] can be seen as a de facto standard for evaluating machine-learning bird identification techniques. Since BirdCLEF's conception in 2014, the bird audio datasets from the xeno-canto collaborative database were used for the bird identification task.

Table 1 shows a summary of the data collection used in the BirdCLEF challenge since 2014. The earlier data used was collected from South and Central American countries as was the original aim of the xeno-canto database. All training datasets were annotated by experts and the number of species to be identified in the BirdClef challenge increased from 2014 to 2017. In later years, the challenge is focused on specific data collection. When the identification of birds from mono-directional recordings (i.e., short recordings of individual birds) was established, the BirdCLEF challenge moved to identify birds in soundscape recordings (i.e., long recordings from PAM devices containing multiple species calling simultaneously) beginning 2019. Some of the lessons learned [16] from the BirdCLEF challenge are: i) deep neural network such as convolutional neural network (CNN) is commonly used for sound event recognition; ii) PAM is favorable for bird density estimation, monitoring, and habitat assessment; and iii) lack of suitable validation and test data hampers the development of reliable techniques. As can be seen in Table 1, more efforts are put into producing annotated datasets of certain countries. Thus, this paper attempts to address the last lesson learned from the BirdCLEF challenge, which is to construct the first annotated Malaysia lowland forest bird sounds to be used as training datasets for machine learning-based PAM-related activities such as bird identification. The contribution of this paper is twofold: i) to the best of our knowledge, this is the first study to develop annotated audio bird sound datasets of lowland forest birds in Malaysia using citizen science and ii) this study provides a detailed description of the development of validation and test datasets to be used by PAM activities.

Table 1. Data collection of BIrdCLEF challenge

	Country	Species	Audio recordings	Recording/species
2014	Brazil	501	14,027	Min 51, max 91
2015, 2016	Brazil, Colombia, Venezuela, Guyana, Suriname, and French Guiana	999	33,203	Min 14, max 200
2017, 2018	Brazil, Colombia, Venezuela, Guyana, Suriname, French Guiana, Bolivia, Ecuador, and Peru	1500	36,496	NA
2019	Ithaca, New York	NA	350 hours soundscapes	NA
2020	North, Central, South America, and Europe	1500-2000	80,000	
2021	North, Central, South America, Eastern, Western United States, Costa Rica, and Columbia	397	60,000 & soundscapes	NA
2022	Hawaii	152	15,000	NA

2. RELATED WORK

In Malaysia, the earliest work of using acoustics for the vocal understanding of birds was done in 2013 by [17] where vocalizations of swiftlets were studied to identify the acoustic features of swiftlets that attracted other swiftlets. Not many details on the recording of the swiftlets were provided. Another similar work by [18] went further by identifying the swiftlet calls and synthesizing more swiftlet calls. In this study, ten swiftlets' sounds were recorded from inside and outside of the bird's house using song meter (SM2). Both studies focused on developing signal processing algorithms for acoustic feature extraction. The first work of bird surveying using vocal calls was done by Chang *et al.* [19] to explain Sunda Scops-Owl's terrestrial calls. Seventy-five recordings were collected from 12 owls in lowland forests and oil palm smallholdings in Selangor within six months. Six frequency features and two temporals were extracted from the spectrograms of the calls, and classification was done using discriminant function analysis. The results showed an accuracy of 97.1% for a correct individual owl call. However, the sample size was too small to confirm the stability of the calls.

A group of researchers from public universities began showing keen interest in PAM beginning the year 2018. Research by Chang *et al.* [19] studied the vocal individuality of large-tailed nightjar in oil palm smallholdings and isolated forest patches in Selangor and recorded 22 individuals. Nine vocal parameters were extracted from the calls and discriminant function analysis showed correct classification of original grouped cases of above 94.5%. A comparison of point count and acoustic sampling was done by [20] to estimate the diversity index of resident bird species in a mangrove forest and oil palm plantations in Selangor. They recorded 115 species of birds in the two habitats and the results also showed no significant difference in the identification of bird species when using point counts or acoustic sampling. PAM was used in a study by [21] to observe birds' activity patterns in relation to the distance to the forest edge, microclimate factors, and different survey periods. Few statistical measurements were done, and the study showed that there were no significant differences in the number of species regardless of the distances from the forest edge. However, a significant difference was found in the number of birds for different survey periods and microclimate factors. Most acoustic activities of the birds were higher in the morning compared to the afternoon, which is when the light intensity and temperature are lower. In 2020, bird sound detection and classification were carried out by Saad [22], Hong [23], Musa [24], Liang [25], and Zabidi *et al.* [26]. Saad [22] used xeno-canto bird audio sounds to compare several CNN-based learning models for the classification of 10 classes of European, South America, and African birds. While the latter [23]–[26] used xeno-canto and Urban8K sound to train CNN-based models in discriminating between birds and non-bird sounds. Musa [24] developed Bulbul-CNN and compared it with MobileNet to classify Asian Koel.

A summary of the related work is presented in Table 2. As can be seen, there are few isolated studies of bird monitoring and surveying using PAM in Malaysia. In general, most studies that conducted their own data collection are done in Selangor, Malaysia. There is not much information on swiftlet's dataset and the data collection methods. The pre-processing and feature extractions of the study [18], [27] followed standard signal processing, and the purpose was to identify the best vocal to attract swiftlets. Three studies by [19]–[21] managed to collect acoustic recordings of 22 nightjar individuals, 5,686 individuals of 115 species, and 90 species of birds from various locations in Klang Valley, Malaysia. Their data were analyzed using statistical tests for different purposes. Chang *et al.* [19] used PAM to identify the vocal structure of the large-tailed nightjar's territorial calls, [21] used it to understand the activity patterns of birds related to the forest edge and the time of the survey. Another group of researchers from the engineering department focused on the classification methods to detect bird sounds against non-bird sounds. While one study [24] specifically stated Asian Koel species, others stated only the use of the xeno-canto database as their positive datasets of bird sounds and Urban8K as negative datasets of non-bird sounds.

Table 2. Summary of related work

Author	Data collection	Pre-processing	Acoustic features	Purpose
Zaini *et al.* [27]	Swiftlet	Pre-emphasis, framing, Hamming	Mel-frequency cepstral coefficient	Vocalization representation
Nematollahi *et al*. [20]	Swiftlet	Denoising	Linear predictive analysis	Vocalization synthesis
Chang *et al.* [19]	22 Large-tailed nightjar individuals	NA	Call length, interquartile bandwidth, low, high, average, center and peak frequencies, third quartile frequencies	Vocal individuality
Hamzah [20]	5,686 individuals	NA	NA	Point-cloud vs acoustic sampling
Shoon [21]	90 species	NA	NA	Acoustic activity pattern
Musa [24]	Asian Koel (xeno-canto)	NA	Spectrogram	Bird sound detection
Hong [23]	xeno-canto, Urban8K	NA	Spectrogram	Bird sound detection
Saad [22]	xeno-canto, Urban8K	1-sec audio segmentation	Spectrogram	Bird sound detection
Liang [25]	xeno-canto, Urban8K	1-sec audio segmentation	Spectrogram	Bird sound detection
Zabidi *et al*. [26]	xeno-canto, Urban8K	1-sec audio segmentation	Spectrogram	Bird sound detection

As stated earlier, one of the problems of PAM research is the lack of suitable validation and test data for the development of reliable and robust techniques. Based on the related work, a reasonably good collection of acoustic bird vocals was collected but no clear explanation or description of developing a validation and test data that can be used by PAM research is found. Even though established signal processing methods can be used for bird surveying and monitoring, the lack of validation and test data will slow the progress of utilizing PAM.

3. METHOD

The main aim of this paper is to develop an annotated bird sound dataset from raw audio recordings such as PAM that can be used as validation and test data for bird identification and monitoring. The first phase of the study was the data collection and data pre-processing of validated raw audio recordings contributed by citizen sciences on the eBird website. An annotated dataset called *SiulMalaya* was developed from the raw audio recordings in the second stage and the third stage involved the training and validation of a CNN model using *SiulMalaya.* The trained CNN model was then used to perform bird identification to evaluate the validity of *SiulMalaya* in the final stage of the study. Figure 1 illustrates the processes involved in this study. Each phase is described in detail in the following subsections.

3.1. Data collection

As this was the initial step of developing annotated bird sounds of lowland forest birds in Malaysia, we only focused on birds found in Kuala Krau Wildlife reserve, Pahang which is the biggest wildlife reserve in Peninsular Malaysia. Although the BirdLife international data zone [18] listed approximately 13 bird species in Kuala Krau Wildlife reserve, our study depends on the availability of the bird sounds on eBird website. The construction of the bird sounds dataset which is named *SiulMalaya* requires several steps as shown in Figure 2. The bird information was sought using its species' common name from the explore menu on the *ebird.org website*. The information was further filtered from the search results by location and data type, which in this case is audio data. After filtering, the result would only provide the information list of birds with audio data at a specified place given by the download URL list. The text pad with the download URLs was imported using simple mass downloader v0.831, a Google Chrome plugin. The extension was then executed to begin the download. As needed for this study, each step was repeated for each bird species. All the audio files in a total of 166 tracks were downloaded into specific folders according to the species' common name in moving pictures experts group-1 layer-3 (MP3) file format for pre-processing and the detailed list is shown in Table 3.

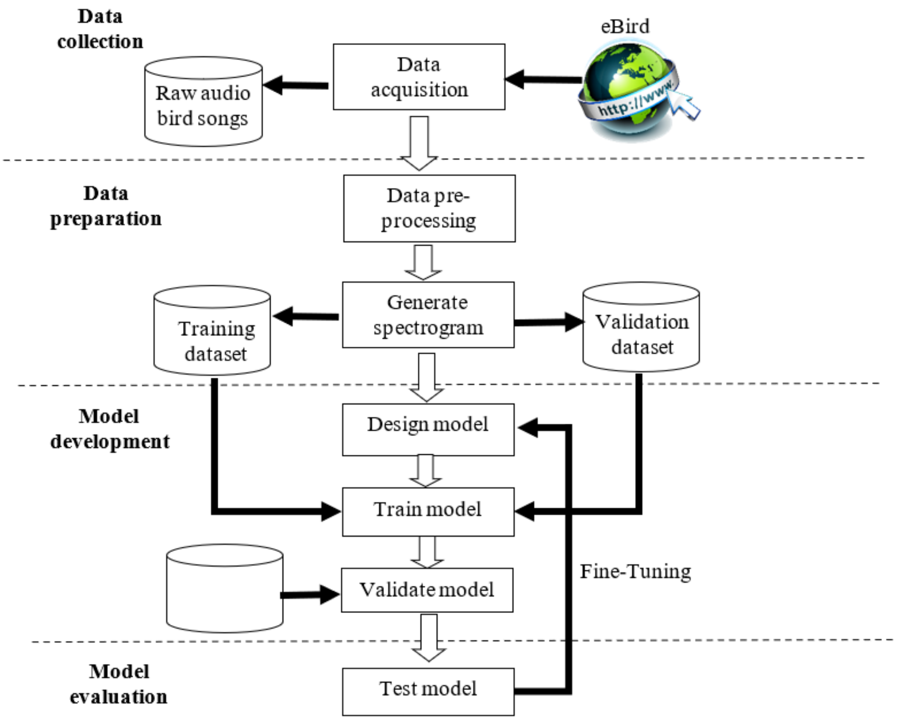


Figure 1. Process flow of *SiulMalaya* development and validation

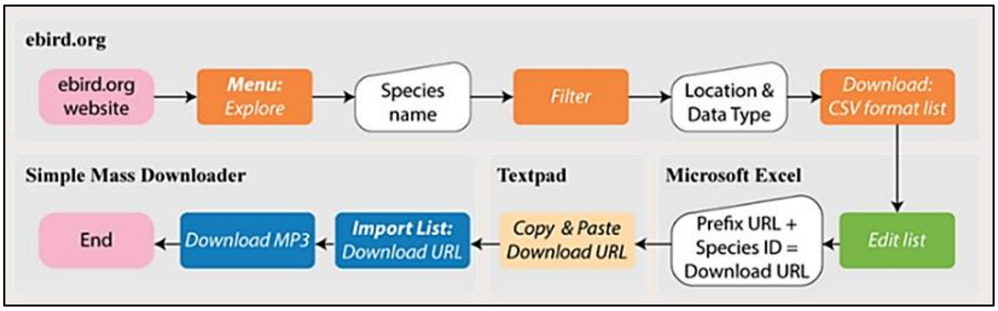


Figure 2. Data acquisition from eBird web portal

Table 3. Raw audio bird songs of Kuala Krau Wildlife reserve

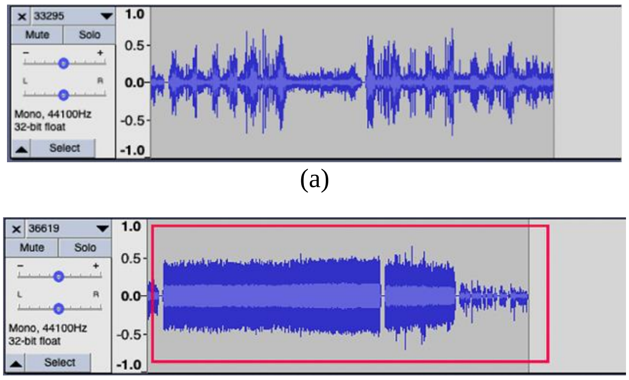
No.	Bird name	Total MP3 tracks	Duration (h:m:s)
1.	Black-capped Babbler	21	000:27:25
2.	White-chested Babbler	4	000:06:49
3.	Ferruginous Babbler	16	000:20:32
4.	Moustached Babbler	18	000:23:41
5.	Sooty-capped Babbler	13	000:35:42
6.	Scaly-crowned Babbler	8	000:08:30
7.	Rufous-crowned Babbler	15	000:34:12
8.	Black-throated Babbler	13	000:16:27
9.	Chestnut-rumped Babbler	13	000:27:06
10.	Chestnut-winged Babbler	17	000:26:03
11.	Fluffy-backed Tit-babbler	13	000:20:43
12.	Black-naped Monarch	8	000:08:00
13.	Rufous-winged Philentoma	7	000:09:23
	Total tracks	166	

3.2. Audio sampling

The raw audio birdsongs were resampled and converted from MP3 to waveform audio (WAV) using audacity v3.0. All audio tracks for a single species were loaded and converted from stereo to mono audio type before being resampled to 16,000 KHz. Finally, they were converted to linear pulse-code modulation (LPCM) WAV file format for the audio filtering process. The LPCM format was used because it uncompressed and preserved the information in the audio track which is crucial to the training of a deep learning model [28].

3.3. Audio filtering

Audio filtering was done to remove superfluous sounds such as animal sounds, environmental noises, and human voices from the sampled audio. Each audio track has unique noise profiles, therefore different noise reduction (dB), sensitivity, and frequency smoothing (bands) were applied on each audio track based on trial and error. The audio filtering was a delicate process and extremely time-consuming because excessive noise reduction will distort the bird's song. In the sections where the bird songs were non-existent, the audio tracks were reduced to 0 dB. Therefore, only the sections of the audio tracks with birds' sounds were left intact. In the case where the audio tracks were exceptionally noisy, they were discarded. An example of a very noisy audio track that was discarded and an audio track that went through filtering is shown in Figure 3. The Figure 3(a) is referred to raw audio birdsong with normal audio track. Meanwhile, Figure 3(b) showed the extremely noisy audio track. After the audio filtering, each audio track was reduced in duration and the results are shown in Table 4.



(a)

(b)

Figure 3. Two raw audio birdsong tracks (a) normal audio track and (b) extremely noisy audio track

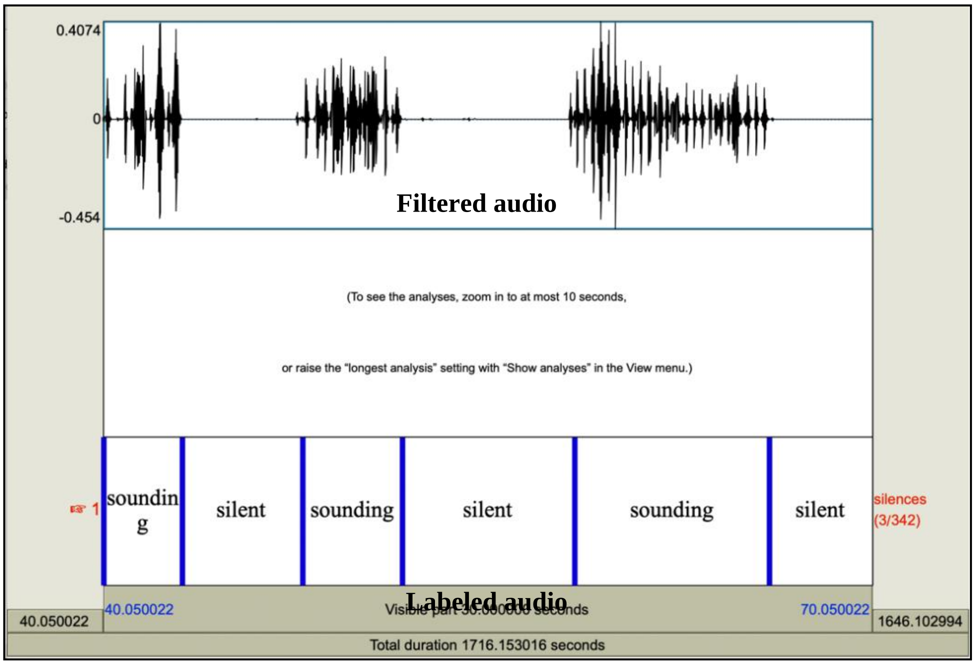
Table 4. Raw audio bird songs are slightly reduced in duration after audio filtering

No.	Bird name	Duration (before)	Duration (after)
1.	Black-capped Babbler	000:27:25	000:23:42
2.	White-chested Babbler	000:06:49	000:05:55
3.	Ferruginous Babbler	000:20:32	000:19:20
4.	Moustached Babbler	000:23:41	000:15:38
5.	Sooty-capped Babbler	000:35:42	000:28:36
6.	Scaly-crowned Babbler	000:08:30	000:05:53
7.	Rufous-crowned Babbler	000:34:12	000:26:06
8.	Black-throated Babbler	000:16:27	000:08:01
9.	Chestnut-rumped Babbler	000:27:06	000:17:08
10.	Chestnut-winged Babbler	000:26:03	000:16:36
11.	Fluffy-backed Tit-babbler	000:20:43	000:14:15
12.	Black-naped Monarch	000:08:00	000:06:12
13.	Rufous-winged Philentoma	000:09:23	000:05:59

3.4. Audio annotation and validation

Audio annotation is a method of making a speech or sounds more identifiable and understandable for deep learning [29]. In this study, the cleaned data was labeled as 'silent' and 'sound' to prepare the training and validation dataset of the deep learning model. The annotation was done using Praat v6.1.42. The challenge during annotation was to determine the most suitable pitch and time steps to segment the 'sound' and 'silence'. After several trials and errors using visual observation, the optimum setting that could produce the best annotation is -50.0 dB for the silence threshold, 2 s for the minimum silent interval, and 0.3 s for the minimum sounding interval. An example of the audio annotation can be seen in Figure 4. The filtered audio can be seen in the top part of Figure 4 and the annotation is seen at the bottom part labeled as 'sounding' and 'silent'.

After audio annotation, the audio tracks were segmented into smaller audio tracks of bird songs totaling 1,665 tracks. The highest number of tracks was from the black-capped Babbler species amounting to 265 tracks. The details are shown in Table 5. The audio tracks in .wav format were then sent to two bird experts from the Department of Wildlife and National Parks, Malaysia. They were given two months to complete the validation of the segmented annotated tracks to ensure all the pre-processing was done correctly and only one dominant bird call was annotated in each segment. The expert listened to the segmented audio tracks and confirmed whether the assigned label was correct or otherwise. If in doubt, a discussion will be held among all parties involved in the annotation process.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

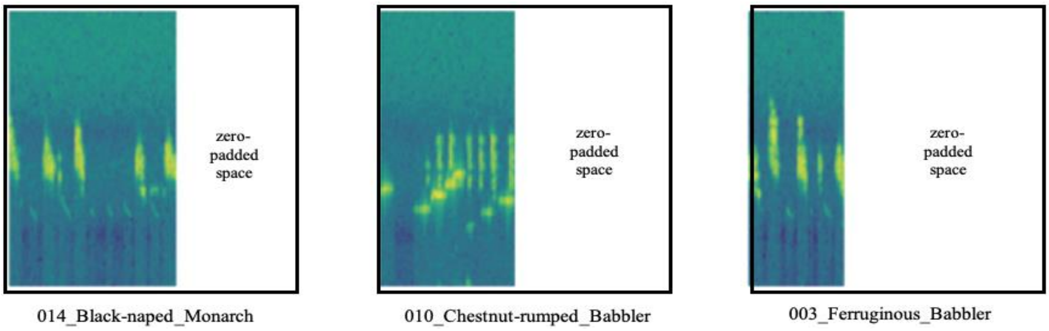
Figure 4. An annotated, noise-filtered audio track

Table 5. Segmented birdsong tracks by species

No.	Bird name	No. of tracks
1.	Black-capped Babbler	265
2.	White-chested Babbler	79
3.	Ferruginous Babbler	124
4.	Moustached Babbler	154
5.	Sooty-capped Babbler	157
6.	Scaly-crowned Babbler	64
7.	Rufous-crowned Babbler	179
8.	Black-throated Babbler	49
9.	Chestnut-rumped Babbler	148
10.	Chestnut-winged Babbler	150
11.	Fluffy-backed Tit-babbler	122
12.	Black-naped Monarch	75
13.	Rufous-winged Philentoma	99
	Total tracks	1,665

3.5. Spectrogram generation

The next step was to represent the annotated audio tracks using spectrograms, which was a representation of audio in a two-dimensional (2D) image based on frequency variations over time [30]. The spectrogram was created by applying a short-time fourier transform (STFT) to the audio tracks to extract spectro-temporal information, followed by normalization. The 1,665 annotated tracks were set to have a consistent length of 1 second, thus the audio tracks that did not meet these requirements are padded with zeros to ensure they have the same dimensions when transformed into spectrograms. The zero-padded spectrograms are shown in Figure 5. As can be seen, the amount of zero-padded differs from each track depending on the duration of the filtered audio track itself. STFT generated an array of complex integers that indicate magnitude and phase. However, for this study, the magnitude would suffice.



014_Black-naped_Monarch 010_Chestnut-rumped_Babbler 003_Ferruginous_Babbler

Figure 5. Spectrograms of three audio tracks of birdsongs

3.5. Model development

As stated earlier, the basic CNN model was chosen in this study as the main aim is to validate the annotated bird sounds of lowland forest birds. The pre-processed birds' audio tracks were stored in different folders for each species indicating the classes for the training of the CNN model. The CNN model was developed using Python and libraries such as NumPy, Seaborn, Matplotlib, Keras, and TensorFlow. The seed for the model was then configured. A seed was used to save the state of a random function, allowing it to create the same random numbers on many executions of the code on the same or other computers. The design, training, validation, and fine-tuning of the CNN model are further described in the following subsections.

3.5.1. Model design

The CNN model was designed to have 10 layers of networks consisting of the input layer, several convolutions, max pooling, fully connected, and output layer. The first layer was the input layer of the resized spectrogram image of 32×32 dimensions to allow the model to train quickly. A batch normalization layer was added to normalize each pixel in the image based on its mean and standard deviation. The normalized spectrogram was further convolved with 32 filters of size 3×3 resulting in dimensions 30×30×32. The convolution layers were configured with the rectified linear unit (ReLU) activation function. The second convolution was learned from a total of 64 filters of size 3×3 producing output dimensions of 28×28×64. The fifth layer was the pooling operation using max-pooling to reduce the spatial dimensions of the output volume. Pooling was done using filter size 2×2 and stride of 2 reducing the image dimension to 14×14×64. A dropout layer was added after pooling to drop a few neurons to avoid overfitting the CNN during the training process. In this study, 25% of the neurons were dropped. The dimension of the image was then flattened to a vector of 12,544 pixels. A single fully connected layer with 128 nodes was appended to CNN and another dropout of 50% was done. The last layer was a ReLU output layer of 13 nodes which was the number of bird classes in our dataset. Table 6 summarized the CNN architecture and the layers.

Table 6. CNN architecture

Layer type	Layer dimensions	Filter size	Stride	Filters	Output dimensions
Input	32×32×1				
Batch normalization	32×32×1				32×32×1
Convolutional	32×32×32	3×3	1	32	30×30×32
Convolutional	30×30×64	3×3	1	64	28×28×64
Max pooling	28×28×64	2×2	2		14×14×64
Dropout					14×14×64
Flatten					12,544
Fully connected	1×128			128	1×128
Dropout					1×13
Output	1×1×3			3	1×13

3.5.2. Model fine-tuning

The CNN model was fine-tuned to improve its performance by modifying the network layers. An additional convolutional layer with a size of 16 and a depth of 3 with ReLU activation function and pooling layer was added to the CNN model. One dropout layer was also removed. The modified CNN architecture is depicted in Table 7.

Table 7. Modified CNN architecture

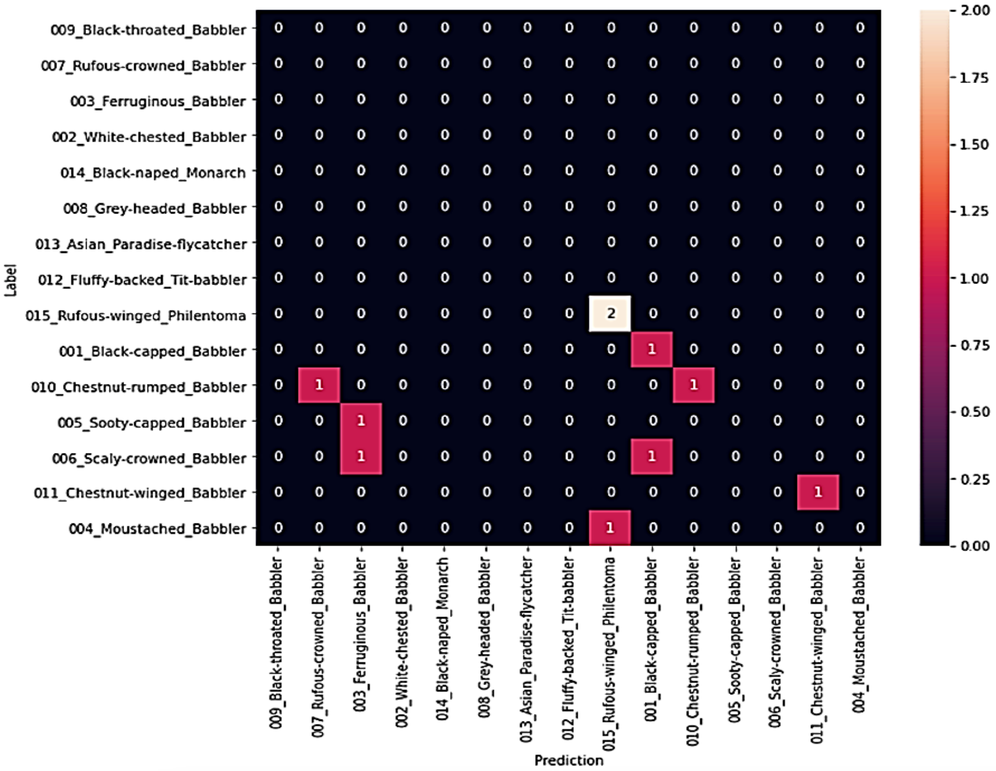
Layer type	Layer dimensions	Filter size	Stride	Filters	Output dimensions
Input	32×32×1				
Batch normalization	32×32×1				32×32×1
Convolutional	32×32×16	3×3	1	16	32×32×16
Max pooling	32×32×16	2×2	2		16×16×16
Convolutional	16×16×32	3×3	1	32	16×16×32
Max pooling	16×16×32	2×2	2		8×8×32
Convolutional	8×8×64	3×3	1	64	8×8×64
Max pooling	8×8×64	2×2	2		4×4×64
Flatten					1,024
Fully connected	1×128			128	1×128
Dropout					1×128
Output	1×1×3			3	1×13

4. RESULTS AND DISCUSSION

The training, validation and testing are the standard process for the creation and evaluation of the CNN model. Each processes used difference datasets with a certain percentage. In this section, the results of training, validation, and testing are presented in different sections.

4.1. Training and validation

The training and validation of the CNN model were evaluated using an accuracy and confusion matrix. The confusion matrix does not show encouraging results. There are very few true positives as shown in Figure 6. Black-throated Babbler only achieved a single positive classification from 49 audio tracks and Rufous-crowned Babbler acquired two correct classifications out of 179 audio tracks. Meanwhile, the zero values imply the non-classification of the bird species. Figure 7 shows the accuracy and loss rate for the training and validation of the CNN model. The training performance was very low, only achieving an accuracy of 47% for training and 10% for validation. From Figure 7, the accuracy observed was 0.47 and 0.10 for both the training and validation, respectively. The margin difference between the training and validation dataset was too large. The training accuracy increased linearly over time, but validation accuracy peaked at around 10% during the training phase. This disparity in accuracy between training and validation accuracy is a clear indication of overfitting. Overfitting occurred because the training data is rather small. Each class has limited samples and the classes were also imbalanced.



	009_Black-throated_Babbler	007_Rufous-crowned_Babbler	003_Ferruginous_Babbler	002_White-chested_Babbler	014_Black-naped_Monarch	008_Grey-headed_Babbler	013_Asian_Paradise-flycatcher	012_Fluffy-backed_Tit-babbler	015_Rufous-winged_Philentoma	001_Black-capped_Babbler	010_Chestnut-rumped_Babbler	005_Sooty-capped_Babbler	006_Scaly-crowned_Babbler	011_Chestnut-winged_Babbler	004_Moustached_Babbler
009_Black-throated_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
007_Rufous-crowned_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
003_Ferruginous_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
002_White-chested_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
014_Black-naped_Monarch	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
008_Grey-headed_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
013_Asian_Paradise-flycatcher	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
012_Fluffy-backed_Tit-babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
015_Rufous-winged_Philentoma	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
001_Black-capped_Babbler	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
010_Chestnut-rumped_Babbler	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
005_Sooty-capped_Babbler	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
006_Scaly-crowned_Babbler	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
011_Chestnut-winged_Babbler	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
004_Moustached_Babbler	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Figure 6. Confusion matrix of the CNN training

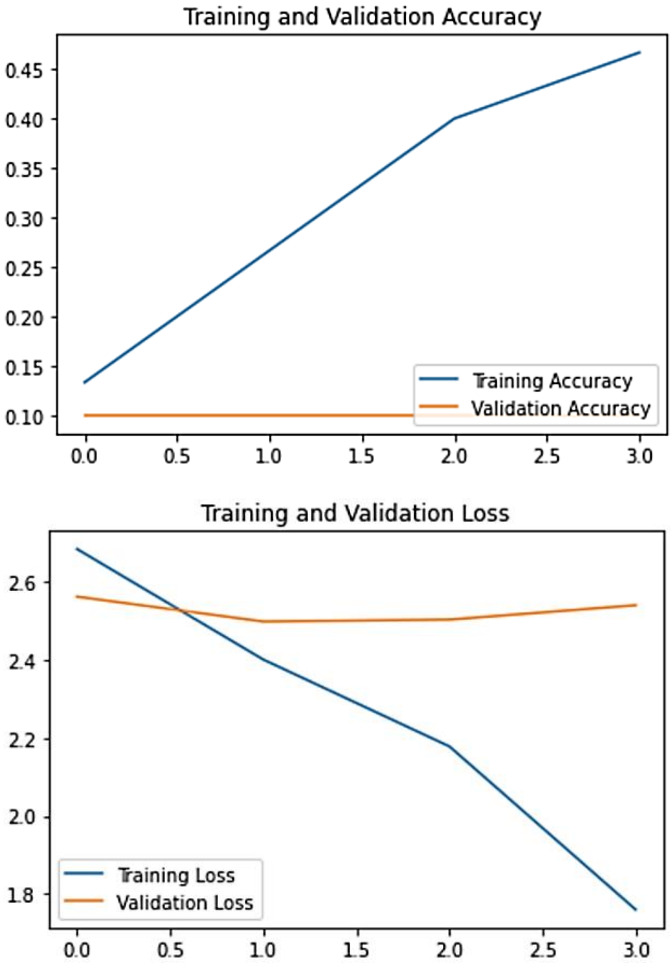


Figure 7. Spectrograms of the audio tracks of birdsongs

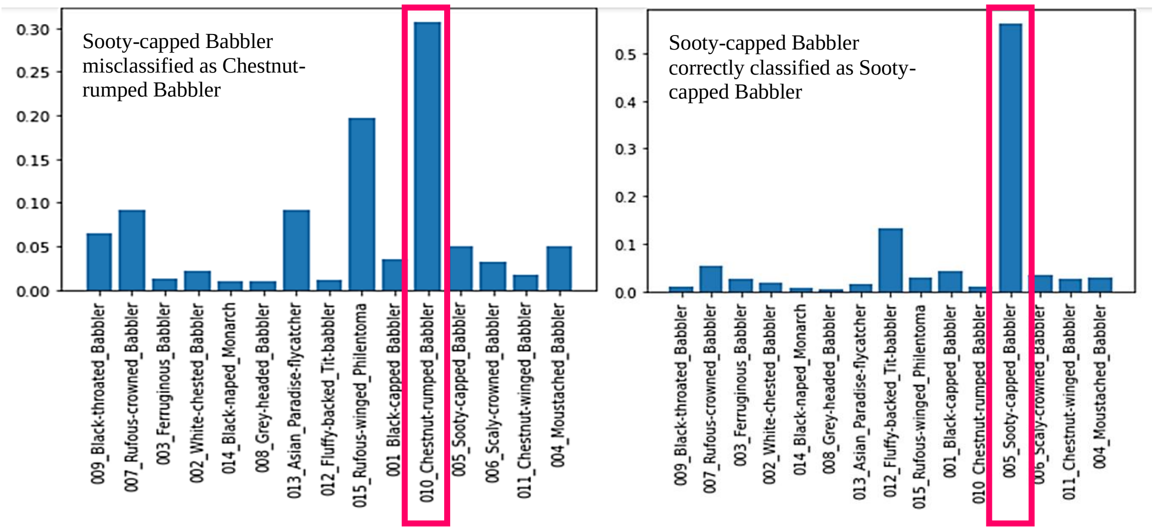
4.2. Testing

Table 8 shows the accuracy and confidence level of classification for both CNN and modified CNN. Each model was executed three times with two different epochs (i.e., 10 and 30) and the average accuracy and confidence level were computed. The highest accuracy was achieved by the modified CNN at an average of 53.33% and a confidence level of 70.37% using 10 epochs. For both learning models, the accuracy was decreased when 30 epochs were used.

We tested the CNN and Modified CNN model with an audio track of a species named Sooty-capped Babbler with a duration of 7 seconds. The CNN model misclassified the test data as Chestnut-rumped Babbler with a confidence level of 30.78% see Figure 8(a). After the modification of the CNN model as stated in Table 7, the test data was correctly classified as Sooty-capped Babbler with a confidence level of 56.30% see Figure 8(b). As the modified CNN has fewer layers, thus fewer trainable parameters. As stated by [31] the lesser the trainable parameters, the harder it is for the model to remember the target class for each training sample, which is optimal for generalizing to new data.

Table 8. Accuracy and confidence level of CNN and modified CNN

Model	Epoch	Number of experiments executed						Average	
		1		2		3			
		Accuracy (%)	Confidence (%)	Accuracy (%)	Confidence (%)	Accuracy (%)	Confidence (%)	Accuracy (%)	Confidence (%)
Modified	10	40	64.04	60	96.58	60	50.49	53.33	70.37
CNN	30	50	99.69	40	48.98	70	21.21	53.33	56.63
CNN	10	30	27.44	30	24.53	20	38.60	26.67	30.19
	30	50	68.13	40	42.64	50	57.31	46.67	56.03



(a) (b)

Figure 8. Test results of Sooty-capped Babbler classified by (a) CNN and (b) modified CNN

5. CONCLUSION

An annotated bird sound audio dataset named *SiulMalaya* for lowland forest birds of Malaysia was developed in this paper. The dataset was verified by two bird experts from the Department of Wildlife and National Parks, Malaysia and it was validated for the purpose of bird identification using a CNN method. Even though the accuracy was only slightly above 50%, the dataset is shown to be suitable for PAM purposes. As stated earlier, there are many established signal processing and machine learning techniques that can be used for PAM-related operations. However, the absence of validation and test data will hinder the adoption of PAM for bird monitoring and surveying in Malaysia. Detailed steps of creating *SiulMalaya* were presented in this paper so that it can serve as guidelines for other researchers to add more bird sound audio tracks to *SiulMalaya*. Therefore, future work should involve parties from different disciplines and governmental bodies to create a bird sound audio dataset at a bigger scale for a more successful effort of conserving the birds of Malaysia.

ACKNOWLEDGEMENTS

Due acknowledgments are accorded to Research Management Centre, Universiti Teknologi MARA for the research funding received through LESTARI research grant scheme no. 600-RMC/LESTARI SDG-T 5/3 (139/2019).

REFERENCES

[1] Forestry Department of Peninsular Malaysia Headquarters, “Forest types,” *Forestry Department of Peninsular Malaysia Headquarters*, 2022. https://www.forestry.gov.my/en/2016-06-07-02-31-39/2016-06-07-02-35-17/forest-type (accessed May 21, 2022).

[2] J. Miettinen, C. Shi, and S. C. Liew, “Deforestation rates in insular Southeast Asia between 2000 and 2010,” *Global Change Biology*, vol. 17, no. 7, pp. 2261–2270, Jul. 2011, doi: 10.1111/j.1365-2486.2011.02398.x.

[3] E. N. Hashim and R. Ramli, “Comparative study of understorey birds diversity inhabiting lowland rainforest virgin jungle reserve and regenerated forest,” *The Scientific World Journal*, pp. 1–7, 2013, doi: 10.1155/2013/676507.

[4] D. A. I. Lang, D. Bakewel, and M. Mohamed, “A national red list for the birds of Malaysia,” *Journal of Wildlife and Parks*, vol. 28, no. 9, pp. 41–49, 2014.

[5] A. Hamza, H. Mamat, and M. T. Abdullah, “Results of a seabird survey at the southern seribuat archipelago, Johor, Malaysia,” *Marine Ornithology*, vol. 47, pp. 49–53, 2019.

[6] C. L. Puan, K. L. Yeong, K. W. Ong, M. I. A. Fauzi, M. S. Yahya, and S. S. Khoo, “Influence of landscape matrix on urban bird abundance: evidence from Malaysian citizen science data,” *Journal of Asia-Pacific Biodiversity*, vol. 12, no. 3, pp. 369–375, Sep. 2019, doi: 10.1016/j.japb.2019.03.008.

[7] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro, and D. Llusia, “Terrestrial passive acoustic monitoring: review and perspectives,” *BioScience*, vol. 69, no. 1, pp. 15–25, Jan. 2019, doi: 10.1093/biosci/biy147.

[8] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, Sep. 2010, doi: 10.1016/j.patrec.2009.09.014.

[9] E. C. Leach, C. J. Burwell, L. A. Ashton, D. N. Jones, and R. L. Kitching, “Comparison of point counts and automated acoustic monitoring: detecting birds in a rainforest biodiversity survey,” *Emu - Austral Ornithology*, vol. 116, no. 3, pp. 305–309, Sep. 2016, doi: 10.1071/MU15097.

[10] C. Pérez-Granados and J. Traba, “Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research,” *Ibis*, vol. 163, no. 3, pp. 765–783, Jul. 2021, doi: 10.1111/ibi.12944.

[11] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, “Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring,” *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, Feb. 2019, doi: 10.1111/2041-210X.13101.

[12] Macaulay Library, “The cornell lab of ornithology Macauley Library,” *The Cornell Lab Merlin*, 2020. https://www.macaulaylibrary.org/ (accessed May 03, 2022).

[13] “Cornell Lab Annual Report,” 2022. Accessed: May 03, 2022. [Online]. Available: https://www.birds.cornell.edu/home/annual-report/.

[14] ImageCLEF, “ImageCLEF/LifeCLEF-multimedia retrieval in CLEF,” *ImageCLEF*, 2022. https://www.imageclef.org/LifeCLEF2022 (accessed May 03, 2022).

[15] A. Joly *et al.*, “Overview of lifeclef 2021: an evaluation of machine-learning based species identification and species distribution prediction,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021*, 2021, pp. 371–393, doi: 10.1007/978-3-030-85251-1_24.

[16] “BirdLife International (2022) Important Bird Areas factsheet: Krau Wildlife Reserve,” BirdLife International, 2022. http://datazone.birdlife.org/site/factsheet/krau-wildlife-reserve-iba-malaysia (accessed May 03, 2022).

[17] S. Ravindran, “Birding expert to talk on forest species,” *Star Media Group Berhad*, 2019. https://www.thestar.com.my/metro/metro-news/2019/05/03/birding-expert-to-talk-on-forest-species/ (accessed May 03, 2019).

[18] M. A. Nematollahi, S. A. R. Al-Haddad, A. R. Ramli, A. Kassim, and S. J. Hashim, “Frequency domain processing for artificial synthesis of Swiftlet’s sound waves,” *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 1–3, pp. 89–93, 2017.

[19] P. K. Chang, C. L. Puan, S. A. Yee, and J. Abu, “Vocal individuality of large-tailed nightjar (Caprimulgus macrurus) in Peninsular Malaysia,” *Bioacoustics*, vol. 27, no. 2, pp. 131–144, Apr. 2018, doi: 10.1080/09524622.2017.1292408.

[20] A. S. M. A. Hamzah, “Comparison of resident bird communities in mangrove forest and oil palm plantations in Selangor, Malaysia,” Ph.D. dissertation, Dept. Philosophy, Universiti Putra Malaysia, Selangor, Malaysia, 2018.

[21] K. L. K. Shoon, “Acoustic activity pattern of birds in Ayer Hitam forest reserve, Selangor,” M.S. thesis, Dept. Forestry Sci. Univ. Putra Malaysia, Selangor, Malaysia, 2018.

[22] A. Saad, “Bird species identification using spectrograms and convolutional neural networks,” Ph.D. dissertation, Dept. Comput. and Microelectronic Syst., Univ. Teknologi Malaysia, Johor, Malaysia, 2020.

[23] T. Y. Hong, “1-D and 2-D convolutional neural network for bird sound detection,” Ph.D. dissertation, Dept. Comput. and Microelectronic Syst., Univ. Teknologi Malaysia, Johor, Malaysia, 2020.

[24] N. B. Musa, “2D convolutional neural network for the detection of Asian Koel (Eudynamys Scolopaceus) vocalizations,” Ph.D. dissertation, Dept. Comput. and Microelectronic Syst., Univ. Teknologi Malaysia, Johor, Malaysia, 2020.

[25] W. K. Liang, “Acoustic event detection with binarized neural network,” Ph.D. dissertation, Dept. Comput. and Microelectronic Syst., Univ. Teknologi Malaysia, Johor, Malaysia, 2020.

[26] M. M. Zabidi, K. L. Wong, U. U. Sheikh, S. S. A. Manan, and M. A. N. Hamzah, “Bird sound detection with binarized neural networks,” *ELEKTRIKA- Journal of Electrical Engineering*, vol. 21, no. 1, pp. 48–53, Apr. 2022, doi: 10.11113/elektrika.v21n1.349.

[27] S. N. Z. H. Zaini, Sunardi, K. H. Ghazali, and S. N. Tajuddin, “Application of speech recognition for swiftlet,” in *Artificial Intelligence in Computer Science and ICT (AICS)*, 2013, vol. 25–26, no. November, pp. 260–264.

[28] I. Siegert, A. Requardt, L. Duong, and A. Wendemuth, “Measuring the impact of audio compression on the spectral quality of speech data,” *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 229–236, 2016.

[29] P. Guyot *et al.*, “Audiovisual annotation procedure for multi-view field recordings,” in *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece*, 2019, pp. 399–410, doi: 10.1007/978-3-030-05710-7_33.

[30] K. Venkataramanan and H. R. Rajamohan, “Emotion recognition from speech,” Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.10458

[31] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

BIOGRAPHIES OF AUTHORS

Nursuriati Jamil is a Professor at College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM), Malaysia. She is currently heading the Digital Image, Audio and Speech Technology Research Group and is the Director of National Autism Analytics Centre in UiTM. She has authored 2 books and published over 100 scientific papers on speech synthesis and speech recognition of Malay language; biometrics; image segmentation and recognition in agriculture and medical domain; gait analysis of autism children; and image retrieval. She can be contacted at email: liza@tmsk.uitm.edu.my.

Ahmad Nazem Norali graduated with M.Sc. Computer Science from Universiti Teknologi MARA (UiTM) and obtained his Bachelor in Science (Conservation Biology) from Universiti Malaysia Sabah (UMS). He is currently a Full Stack Developer in developing web application and system. He can be contacted at email: ahmadnazem2010@gmail.com.

Muhammad Izzad Ramli obtained his Bachelor of Computer Science (Hons) (Multimedia Computing) in 2011, M.Sc. Computer Science in 2013 and Ph.D in Computer Science in 2018 from Universiti Teknologi MARA. He is currently a senior lecturer in College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM), Malaysia specializing in speech processing. He is a member of Digital Image and Speech Technology (DIAST) and Computational Intelligence Group (CIG) Research Group. He can be contacted at email: izzadramli@uitm.edu.my.

Ahmad Khusaini Mohd Kharip Shah is an assistant director in Department of Wildlife and National Park, Malaysia. He is very expert in conservation specifically in birds (ecology). He obtained Bachelor in Biological Sciences in 2009 from Universiti Malaysia Terengganu. In 2009 to 2014, he was a research officer in Agro-Biotechnology Institute (ABI) and starting from 2014 until now, he one of the wildlife officers at Department of Wildlife and National Parks Peninsular Malaysia. He can be contacted at email: khusaini@wildlife.gov.my.

Ismail Mamat is an assistant Wildlife Officer in Department of Wildlife and National Park, Malaysia. He obtained diploma in forestry in 1987 from Universiti Putra Malaysia. He can be contacted at email: ismailmat@wildlife.gov.my.