Sound Classification for Javanese Eagle Based on Improved Mel-Frequency Cepstral Coefficients and Deep Convolutional Neural Network

Silvester Dian Handy Permana[1], T.K. Abdul Rahman[2]
Informatics Engineering, Universitas Trilogi, Jakarta, Indonesia[1]
School of Science and Technology, Asia e University, Shah Alam, Malaysia[2]

***Abstract*—The Javanese Eagle is a rare and protected animal in Indonesia. These animals only live in a few species and are threatened with extinction. These birds need to be bred to avoid extinction. One form of communication between the Javanese eagles and each other is the sound of their tweets. These tweets can be studied and classified to conserve endangered animals. This study will classify the sound of the Javanese Eagles for the benefit of animal conservation. Data in the form of voice tweets will be classified. This classification uses algorithms from improved MFCC (Mel-Frequency Cepstral Coefficients) and Deep Convolutional Neural Network. The result of this study was to classify the sound of the Javanese Eagle from the lack of food or drink, the normal tweets state of the bird, and to find out the Javanese Eagle in finding a partner. This research has been carried out by comparing the CNN architecture with AlexNet and VGGNet models and various combinations of training, validation, and test data. The best model dataset underwent division into 80% for training, 10% for validation, and 10% for testing. Training and testing of both IMFCC and VGGNet models occurred using the same dataset. During training, VGGNet achieved 100% accuracy, while testing yielded 99%. ROC Curve: 'Normal' AUC 0.996, 'Looking for Partner' AUC 1.000, 'Looking for Food' AUC 0.996. This study aids Javanese Eagle conservation, crucial for preventing extinction at conservation sites.**

Keywords*—*Improved MFCC; deep convolutional neural network; Javanese eagle sound; sound classification

I. INTRODUCTION

The Javanese Eagle (*Nisaetus Bartelsi*) is a rare and protected animal in Indonesia. The Javanese Eagle existence is increasingly rare because of the eruption of Mount Merapi, which caused the death of many Javanese Eagles and it only lays 1 to 2 eggs per year. In addition, many illegal hunters hunt these birds to sell and make a profit [1]. Even though in 1990, eagles were protected by the government, there are still many who trade eagles illegally [2]. These animals only live a few species globally and are threatened with extinction [3, 4]. The Javanese Eagle is one of the animals that are conserved in zoos and nature reserves. These birds need to be bred to avoid extinction [5]. Especially in zoos, caretakers need to pay attention to the needs of these birds, especially in maintaining a balance nutrition. Because, balanced nutrition will keep Javanese eagles to survive. However, sometimes they cannot understand the Javanese Eagle's needs quickly. Javanese eagles usually use their tweet to code their environment to find food or before eating other animals. From the tweet sound, it can be identified the conditions and needs of the Javanese Eagles. The voice of this tweet is very distinctive and very specific which can be heard [6, 7].

In helping to preserve the Javanese Eagle, research is needed to identify the needs of the Javanese Eagles. The chirping sound of this Javanese Eagle can be studied and classified to help in the conservation of endangered animals. With the tweets studied by the proposed technique and verified by experts, can know the basic needs of the bird especially in searching for prey. This research will develop a Javanese Eagle's sound classification technique that will classify the sound of the Javanese Eagle into lack of food or drink, knowing the Javanese Eagle in search of partner, and normal state of bird tweets through combination of algorithms from Mel-Frequency Cepstral Coefficients (MFCC) [8] and Deep Conventional Neutral Network [9, 10, 11, 12, 13]. The data from this study were taken from zoos and nature reserves in Indonesia such as the Ragunan Zoo, PSSEJ, and the Bogor Botanical Garden. The sound that was taken and use as a data are validated by experts. Data in the form of voice tweets will be classified.

MFCC is a feature extraction that produces features in the form of cepstral coefficient parameters. Feature extraction Mel Frequency Cepstral Coefficient (MFCC) converts sound waves into several types of parameters such as the cepstral coefficient which represent the audio file. In addition, Improved MFCC generates feature vectors that convert voice signal into several vectors for speech recognition. This signal is known as spectrogram. Convolutional Neural Network (CNN) is the development of Multilayer Perceptron (MLP) which is designed to process two-dimensional data. CNN is included in the type of Deep Neural Network because of its high network depth and widely applied to image data. The sound image formed from the MFCC model can provide a specific picture so that CNN can train properly so as to produce an accurate model. Before being processed using CNN, the Javanese eagle's sound needs to be converted to a spectrogram first. The existence of this spectrogram provides a significant difference from the presence of audio in a tweet. Spectrograms provide higher accuracy in training that audio signals trained in digital form. The Improved MFCC and followed by CNN in deep learning architecture was designed to classify the Javanese Eagle's voice. Javanese eagle sound classification was used to identify whether the Javanese eagle is lacking of food or drink,

finding a partner, or it is a normal tweet of bird. The result of this research was used to help the bird's caretaker to better understand the basic needs of the Javanese Eagle.

The combination of signal processing and deep learning has the potential to reveal the complex layers of meaning hidden within the vocal repertoire of the Javanese Eagle. By combining different methodologies, the objective was to classify the tweets of the Javanese eagle into specific categories: regular sounds representing their daily activities, sound indicating their hunger and search for food and sound expressing their desire for a mate, reflecting the intricate social dynamics of these magnificent birds of prey.

The output of this research has helped to protect Javanese eagle birds from extinction. In this research, the needs of Javanese eagle can be identified from the sound of their chirping. This research created an application that can differentiate the sound of the tweets of the Javanese eagle. This research is expected to have a major impact in the caring for the existence of the Javanese eagle by providing what is needed quickly. By providing caretakers and conservationists with a tool to swiftly identify and respond to the needs of the Javanese Eagle, the study contributes to the ongoing efforts to protect this species from extinction. Developing an application capable of distinguishing between different tweet sounds holds significant promise for enhancing the management and conservation of the Javanese Eagle population. Ultimately, by leveraging advancements in signal processing and deep learning, this research underscores the importance of interdisciplinary collaboration in safeguarding Indonesia's biodiversity and preserving the ecological balance of its natural habitats.

II. Literature Review

In this section discussed about the Javanese eagle, animal sound and the voice of the Javanese eagle's tweet. And a quick overview of literature review on previous research about classification of animal's sounds, MFCC, and CNN deep learning.

A. Javanese Eagle

The Javanese Eagle, scientifically classified as *Spizaetus bartelsi*, represents a distinctive species within the medium-sized bird category. Endemic to the lush landscapes of Java Island, this avian species thrives in the verdant forest that adorn the highlands and mountain slopes. Regrettably, the very existence of the Javanese eagle faces a precarious future, imperiled by the relentless march of time and insidious encroachment of illegal deforestation, which threaten to drive this rare species to the brink of extinction [14]. Historically, the Javanese eagle was a prominent inhabitant of Java Island's forests and mountainous terrains. Yet, in the wake of anthropogenic activities and environmental transformations, the once-thriving population has witnessed a distressing decline. This medium-sized eagle exhibited a body ranging from 60 to 70 cm, measured from the beak to the tip of the tail [15]. The Javanese eagle eats various types of small birds and other poultry, small mammals such as mice, squirrels, rabbits, to medium-sized one such as monkeys. This bird also eats various types of small reptiles such as lizards, monitor lizards, and snakes. This bird lives on the slopes of mountains and hills. Now its existence is only in the rain forest alone. This animal is endemic to the island of Java [16].

B. Animal Sound

Sound is a form of energy that always propagates in all directions in the form of longitudinal waves. Sound can be heard if there is a sound source, medium or intermediary to propagate, as well as an object to listen to / which is used to capture the sound signal. A signal is a variation of variables such as the pressure wave of sound, the color of an image, the depth of a surface, the temperature of a body, the voltage or current of a conductor or biological system, light, radio electromagnetic signals, the price of goods or the volume and weight of an object. It can be said that a signal is a medium to carry information about the past, present and future state of a variable [17].

C. Classification of Sounds

Environmental Sound Classification (ESC) is one of the most challenging tasks in forensic digital signal processing and machine learning. Many methods have been proposed to perform ESC, one of which is self-supervised learning (SSL) for ESC. SSL is a model used to study unsupervised representations by completing pretext tasks and using them to perform downstream task such as classification or regression [18]. This study uses the ESC-10 and DCASE 2019 Task-1(A) datasets. The first dataset used is the ESC-10 containing 400 signals from 10 different types of environmental sounds. 10 types of sounds from the ESC-10 dataset: Dog Barking, Baby Cry, Clock Tick, Fire Cracking, Helicopter, Person Sneezing, Rain, Rooster and Sea Waves. The test signal contains 10 different types of environmental sounds including Airport, Bus, Metro Station, Metro, Park, Public Square, Shopping Mall, Street Pedestrian, Street Traffic and Trams [19]. In this study the model developed uses a spectrogram image as its input, in the early stages of extracting the spectrogram signal. This research discusses the evolution of object detection, highlighting the shift from traditional methods reliant on handcrafted features to deep learning approaches. It emphasizes the advantages of deep learning in learning semantic, high-level features and explores various architectures, training strategies, and optimization functions. The paper provides a comprehensive review of deep learning-based object detection frameworks, covering both generic and specific detection tasks such as salient object detection, face detection, and pedestrian detection. Experimental analyses are conducted to compare methods and suggest future directions for research in object detection and neural network-based learning systems [20]. In this research discusses about an audio extraction technique using MFCC, LPC and DTW and use CNN methods for training and classification process. It develop a hardware module to collect the audio data and sent to the server and used data from online source. The hardware module can function well to classify and send audio signal to the cloud server and store. The audio signal in the cloud server can be reused for the training phase. The highest accuracy obtained reached 91.3% [21].

D. Classification of Animal's Sounds

Research on the classification of animal sounds were carried out by [22, 23]. Reference [22] took the theme of classifying animal sounds using the Convolutional Neural Network method. In this study, the problem of classifying animal sounds using deep learning was investigated and a system based on a convolutional neural network was proposed. The result obtained from this study are unsatisfactory where the highest accuracy is only up to 75% which in the confusion matrix gives the result 4 bird sound are misclassified as cats. While reference [23] took the theme of an IoT-based sound classification system. In this study a well-known feature extraction technique called Mel Frequency Cepstral Coefficient (MFCC) is used to extract features from a given audio clip, send it to the CNN architecture. The result obtained from the research after the dataset test was run on the model, the best accuracy was obtained by AdaDelta, Gradient Descent, and RMSProp optimizer, which was 91.3%, and the worst accuracy was obtained by Momentum optimizer which was 82.6%. Research conducted by [24], with the theme Workflow for automatic identification of animal sounds. In this study, the development and application of a convolutional neural network for the automatic detection of 14 birds and mammals adapted in the forest by classifying spectrogram images generated from short audio clips were explained. The result of this study was stored after Epoch 100 with a training loss of 0.0182, validation loss of 0.0139, training accuracy of 0.9954, and validation accuracy of 0.9969. Research conducted by [25] with the theme of differences between MFCC and IMFCC for the classification of bird sounds. In this study a comparison between MFCC and IMFCC features for automatic bird species recognition systems was carried out with the aim of validating the use of IMFCC features as features that can also be extracted for bird species recognition and the method of bird sound classification system used in this study using Hidden Markov Models (HMM) for data training needs. To compare the efficiency of the features of MFCC and IMFCC with the proposed algorithm using the TAR and TRR performance. Based on the TAR and TRR performance on *Automolus rubiginosus*, *Synallaxis erythrothorax*, *Cardinalis*, *Cercomacra Tyrannina*, and *Myiozetetes Similis* the result of the IMFCC method provide a percentage increase than MFCC. Finally, study conducted by [26], with the theme of a forest fire early warning system with the sound of birds. In this study, the bird data used in the form of recordings of bird sounds from four bird species. The four bird species used in this study were Cipoh, Prenjak, Merbah Cerucuk and Pleci. The result of this study obtained a test accuracy value of 96.45% based on the results of the testing process carried out on the experience of the system program. At this stage it can be concluded that the proposed method is able to classify bird sounds based on the condition of the two birds with an accuracy of 96.45%.

E. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a method that is quite good and the most widely used in the field of speech recognition. MFCC is a feature extraction that produces in cepstral coefficient parameters. Feature extraction MFCC converts sound waves into several types of parameters, such as the cepstral coefficient, which represent the audio file. In addition, MFCC produces feature vectors that convert voice signals into several vectors for speech feature recognition [27]. MFCC has seven stages, namely pre-emphasis, frame blocking, windowing, Fast Fourier Transform (FFT), Mel Frequency Wrapping (MFW), Discrete Cosine Transform (DCT) and cepstral lifting, which produces parameters like features, namely frames and cepstral coefficient. The final result of MFCC method will improves the quality of the speech recognition, it can be seen using plot from cepstral lifting stage [28].

F. Convolutional Neural Network (CNN) Deep Learning

Convolutional Neural Network (CNN) is one of the deep learning algorithms included in the feed forward class method that is inspired by the visual cortex of the brain. However, to be able to predict well, CNN must be designed with a more complicated architecture. As a result, CNN training is very computationally expensive and has implementation because of its slow speed. CNN has several models in the training process where the method has a different architecture according to the problem. The training model is used in accordance with the state of the object of identification because CNN has several layers implemented at the training stage. The modern CNN discovered by LeChun has seven layer structures (not including the input layer) namely LeNet-5 which has the following structures C1, S2, C3, S4, C5, F6 output [29]. Example of case studies in image recognition on CNN have three stages, namely the input, CNN, and output stages [20]. Input layer is the stage for inserting images into the program and further be processed by changing the image into a binary form so that it can be process at the CNN stage [30]. The CNN algorithm develops multi-layer perceptron (MLP) to process data, one of which is two-dimensional image data, for example images. This CNN algorithm classifies labeled data using the supervised learning method, namely, targeted data and the appropriate variables. Convolutional Neural Network (CNN) has five stages, namely (a) convolution, (b) ReLu layer, (c) max pooling, (d) flattening, and (e) full connection.

III. METHODOLOGY

It is not an easy task to extract sound features, recognize them and classify them from various short audio clip. This was due to background noise, very short sound intervals, and fast clip changes. These things interfere with the recognition process carried out by artificial intelligence. Therefore, the recognition of the input voice is very important and affects the result of the voice classification. The system must distinguish the sound that it wants to process and which sound it does not want to process. As a result, the voice data obtained needs to be processed first before entering the voice classification model for processing. The research flow can be seen in Fig. 1.

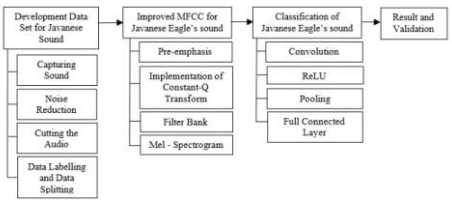


Fig. 1. Research flows.

A. Develop a Data Set of Sounds of Javanese Eagle

The data used in this study is data in the form of recordings of the sound of birds chirping which is captured by recording on the spot or looking for it in the datasets that are already available. The dataset was divided into two consisting of a true dataset containing the sound of an eagle and a noise dataset. The sounds of the birds used are the sounds of the Javanese eagle so that the sounds of other birds were included in the noise dataset. These two types of data sets were further divided into two, namely the dataset used in the training process and the dataset used in the testing process. The self-collected bird sound clips are short audio clips containing only one tweet. The steps of the Development Data Set for Javanese Sound flow are shown in Fig. 2 below.

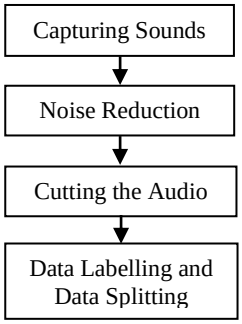


Fig. 2. Development data set for Javanese sound flow.

The preprocessing phase plays a crucial role in the successful identification of Javanese Eagle's primary needs using the hybrid improved Mel-Frequency Cepstral Coefficients (MFCCs) and Deep Convolutional Neural Network (CNN). This phase involves two key steps: Capturing Sounds, noise reduction and cutting audio, and Data Labelling and Data Splitting. The data for Capturing sounds phase are gathered from three specific locations: Taman Safari, PSSEJ and Ragunan Zoo with a total 300 samples for the dataset. After that, to enhancing the signal component the unwanted background noise from the audio signal were attenuate or eliminate using the noise reduction phase [31]. Then the audio data is dissected into individual syllable segment to separate each distinct syllable inside cutting audio phase. After that, the data is assigned with labels and split into three distinct subsets known as training data, validation data and testing data.

B. Develop an Improved Mel Frequency Cepstral Coefficients (MFCC) Technique for Converting the Javanese Eagle Tweet into Spectogram

At this stage, will convert the sound image into the form of a spectrogram using the Improved Mel Frequency Cepstral Coefficients (IMFCC), which will extract the sound features using an artificial intelligence model. The sound spectrogram example is shown in Fig. 3 below.

In order to obtain an effective feature representation for the sound detection of the Javanese eagle, in this study, an improved MFCC-based feature extraction algorithm is proposed. Improved MFCC using Constant-Q (CQT) and Mel Spectrogram. Constant-Q Transform (CQT) is an algorithm that can efficiently compute the Fourier transform. The improved MFCC has two stages before using MFCC itself. The first step is Implementation of Constant-Q Transform (CQT) and the second is Implementation of Mel-Spectrogram. The steps of the improved MFCC flow are shown in Fig. 4 below.

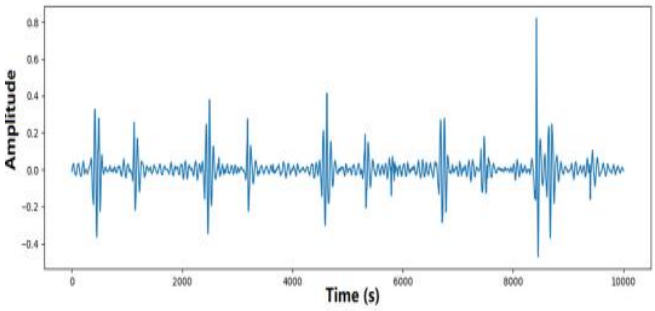


Fig. 3. Sound spectrogram.

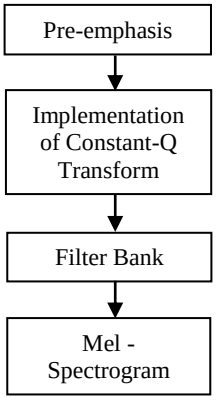


Fig. 4. Improved MFCC.

The audio signal is one-dimensional data, so to convert the audio data it is necessary to convert the audio signal to a log scale time frequency using Constant-Q Transform (CQT). The CQT transformation can be seen in Eq. (1).

$$X[k,n] = \sum_{q=n-\lfloor Nk/2 \rfloor}^{n+\lfloor Nk/2 \rfloor} x(q) a_k^* \left(q - N + \frac{Nk}{2}\right) \quad (1)$$

where, k = 1,2 …, K indexes the catch-frequency coefficient of CQT. Shows a_k^*(n). The basic function of a_k^*(n) is a waveform with complex values. Then in the atomic Eq. (2) the time frequency is defined as follows:

$$a_k(n) = \frac{1}{Nk} \omega \left(\frac{1}{Nk}\right) exp\left(-j2\pi \frac{f_k}{f_s}\right) \quad (2)$$

The value fk is the storage centre frequency k, and fs denotes the sampling rate, and w (t) is a continuous window function sampled at the point determined by t. Nk is the length of the window which is inversely proportional to fk in order to achieve the same Q-factor for all k containers. The center frequency fk is defined as Eq. (3) follows:

$$f_k = f_1 \, 2^{\frac{k-1}{B}} \quad (3)$$

where, f_1 is the middle frequency of the lowest frequency container, and B is the coefficient to determine the number of containers per octave. In the process, B is an important parameter to make choices when using CQT, because it determines the time-frequency resolution considerations of CQT. The IMFCC consist of four stages, namely: pre-

emphasis, implementation of Constant-Q Transform, filter bank and mel- Spectrogram.

C. CNN for Classification of Eagle's Tweet

In this study, the classification model used is Convolutional Neural Network (CNN) which consists of three stages, namely: convolution, Rectified Linear Unit (ReLu), and pooling. The CNN training itself will be carried out by repeating the convolution and ReLu stages using the training dataset that has been prepared. CNN itself is used because it has the characteristics of sparse interaction, parameter sharing, and equivalent representation. The steps of the CNN Model are shown in Fig. 5 below.

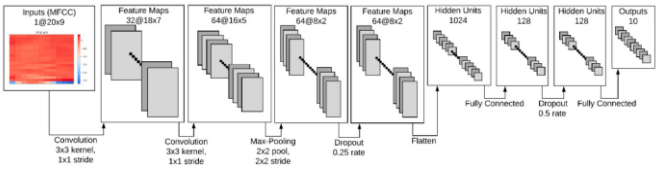


Fig. 5. CNN model.

Introducing a Comparative Analysis: AlexNet and VGGNet for Sound Recognition. Two such seminal architectures, and VGGNet and AlexNet, have etched their names in the annals of deep learning, each characterized by its unique design and contributions. To compare AlexNet and VGGNet effectively, several factors need to be considered, such as model complexity, adaptability to temporal data, and computational efficiency. AlexNet's deep architecture and local response normalization can be modified to accommodate the temporal characteristics of sound, while VGGNet's uniform structure may be better able to capture diverse temporal features. A nuanced approach is necessary to evaluate these architectures, with adaptations and optimizations being crucial to realizing the full potential of both AlexNet and VGGNet in sound classification tasks. AlexNet's architecture is characterized by its depth, large convolutional filters, ReLU activation, and hierarchical feature extraction. The utilization of convolutional layers, max-pooling, and fully connected layers in the architecture of AlexNet facilitated the attainment of cutting-edge performance in picture classification tasks. While, the VGGNet architecture is characterized by its simplicity, regularity, and deep stack of convolutional layers. This design philosophy allows the network to learn hierarchical features of increasing complexity from the input image, making is effective for image classification tasks. The AlexNet CNN and VGGNet Model Architecture are shown in Fig. 6 and Fig. 7 below:

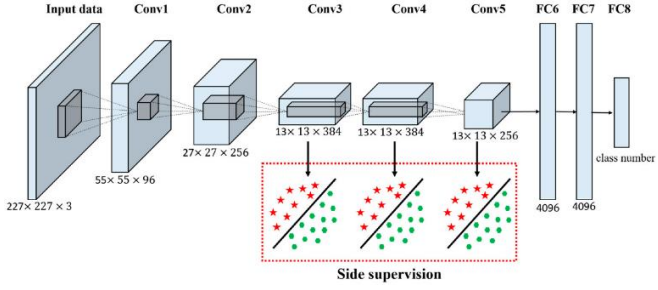


Fig. 6. AlexNet CNN model architecture.

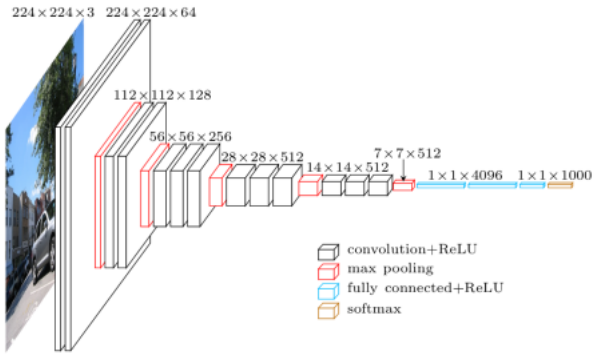


Fig. 7. VGGNet CNN model architecture [32].

1) Convolutional: CNN is a neural network model designed to process data in a grid-like structure. Therefore, the MFCC input in the form of these images will be entered into a matrix that contain the pixel values in the existing image. After that the image map will be multiplied by a matrix that we call the kernel model, for each pixel value of the image matrix. It is this kernel model that the model will study and create. This stage uses mathematical equations such as Eq. (4)

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (4)$$

At the end of the image matrix itself will be filled by an auxiliary variabel with a value of 0. The kernel model will also move by shifting it according to a predefined value, which can call stride. These three parameters are very important at this stage.

2) Rectified Linear Units (ReLu): The result of the image matrix that have been multiplied by this kernel model will be normalized so that negative values in the image matrix must be removed. On CNN, a thresholding process will be carried out, which changes the negative value to zero, using ReLu. ReLu itself has similarities as in equation 3.19 whose activities can be seen in Fig. 10 below.

3) Pooling: After the pixel values have been convoluted and denormalized, the CNN matrix will be reduced in size so that the calculation and classification process become faster and more precise using the pooling method [33]. There are many pooling methods, in this study using the max-pooling method which is very commonly used in CNN and also easy. Basically max-pooling works by grouping the CNN matrix into small matrices and then taking the highest value from the small matrix.

4) Fully connected layer: The CNN matrix values that have gone through the three stages will then proceed to the last stage, namely Fully Connected Layer. At this stage the data obtained from the CNN matrix will be combined and flattened into a one-dimensional layer containing the values from the CNN matrix. This process is often called the flatten process [33]. After completing the flattening stage, the spectrogram image value data that has been generated by CNN will be classified using the SoftMax classifier function.

D. Analysis of Result

To determine the success or failure of this study, it is necessary to Testing data or test data. Tests on learning outcomes are carried out to assess the level of success they have. The success in the test is presented in the form of a percentage, such as the following equation. The results of the test are decisions that were the result from the classification carried out. Testing can be done by using the Correlation coefficient (R), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE). The equation for the correlation coefficient (R) uses the Eq. (5) as follows:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{j=1}^{N}(y_i - \bar{y})^2}} \quad (5)$$

Furthermore, to calculate the Mean Absolute Error (MAE) can use the Eq. (6) follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \quad (6)$$

Then to calculate the Mean Absolute Percentage Error (MAPE) you can use the Eq. (7) follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{X^i}\,.100\% \quad (7)$$

And lastly, to calculate the Mean Square Error (MSE) you can use the Eq. (8) follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \quad (8)$$

where, x; y is the target value and predicted value, and N is the number of sample data.

The Confusion Matrix is an indispensable tool for understanding the strengths and weaknesses of a classification model. It allows researchers and practitioners to pinpoint specific areas of improvement, assess the model's robustness, and make informed decisions about model refinement or optimization based on real-world implications. The matrix comprises four essential components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) represents instances correctly classified as positive by the model. True Negative (TN) denotes instances accurately identified as harmful. False Positive (FP) signifies instances wrongly classified as positive, and False Negative (FN) includes instances incorrectly identified as harmful.

When the model accurately identifies positive audio samples associated with Javanese eagles, it is denoted as True Positive (TP). On the contrary, True Negative (TN) arises when the model correctly rejects audio samples unrelated to Javanese eagles. False Positive (FP) instances occur when the model incorrectly identifies negative audio samples as positive for Javanese eagles. Finally, False Negative (FN) incidents happen when audio samples relevant to Javanese eagles are incorrectly classified as irrelevant by the model.

Comprehending these concepts aids in evaluating the reliability of the model or system in predicting Javanese eagle sounds. The quantification of TP, TN, FP, and FN counts facilitates the calculation of evaluation metrics such as precision, recall, and F-Score. Precision is the ratio of true positives to the sum of true and false positives, reflecting the accuracy of optimistic predictions. Recall, also known as sensitivity or true positive rate, is the ratio of true positives to the sum of true positives and false negatives, indicating the model's ability to identify all relevant instances. F1-score, a harmonic mean of precision and recall, provides a balanced assessment of the model's overall performance. A confusion matrix is typically presented in a Table I format with rows and columns corresponding to the actual and predicted classes, respectively. Here's a simple representation:

TABLE I. REPRESENTATION OF CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

E. Validation on Expert Verification

Validation through expert verification is a crucial component of this research, ensuring the robustness and reliability of the developed sound classification system for Javanese eagles. Pusat Suaka Satwa Elang Jawa (PSSEJ) serves as the primary domain for this validation process. The expertise of caretakers at PSSEJ plays a pivotal role in assessing the accuracy and effectiveness of the developed system. Their extensive experience caring for Javanese eagles equips them with a nuanced understanding of the eagles' vocal expressions corresponding to distinct needs and providing a real-world evaluation that aligns with the practical aspects of eagle care and conservation efforts. Their validation ensures that the developed system aligns with scientific precision and integrates seamlessly into the context of caretaking at PSSEJ.

IV. RESULT AND DISCUSSION

The outcomes and discussions of the study on sound classification for identifying the primary needs of the Javanese Eagle are presented, employing a hybrid approach of improved Mel-Frequency Cepstral Coefficients (MFCC) and Deep Convolutional Neural Network (DCNN). The section commences with a concise overview of the research objectives and methodology, followed by an elaborate account of the utilized dataset, encompassing the recording process, data pre-processing, and feature extraction.

A. Development of Data Set for Sounds of Javanese Eagle

Data on Javanese eagle sounds were gathered from three separate sites. A bidirectional microphone was used to record audio, which was then saved in uncompressed WAV format with a sampling rate of 44.1 kHz and a resolution of 16 bits. Our machine learning algorithm was trained using the training set, and its performance was assessed using the testing set. Eighty percent of the dataset (240 sound samples) was used for training, and the remaining twenty percent (60 sound samples) was used for testing. The sound clips were pre-processed to remove any background noise, normalize the amplitude, and cut them to a set length of a syllable. The steps in this phase are:

1) Capturing sound: The data-gathering process involved recording the sounds produced by Javanese Eagles in their

natural habitat within the park. The recordings were made using a high-quality microphone and a digital audio recorder and conducted during the daytime when the Javanese Eagles were most active and vocal. This study used a bidirectional microphone to record the sounds of Javanese eagles. A bidirectional microphone, also known as a figure-eight microphone, captures sound from two opposing directions while denying sound from other approaches. Using a bidirectional microphone, the sounds of Javanese eagles can be accurately recorded.

2) Noise reduction: The sound data collected has gone through pre-processing once the Javanese eagle's sound was recorded. Pre-processing was the process of cleaning and filtering recorded audio to eliminate distracting elements like background noise. A crucial first step in employing sound classification to determine the main requirements of Javanese Eagles has been data pre-processing. To achieve accurate sound categorization findings, pre-processing has consisted of multiple phases aimed at cleaning and filtering the collected audio data.

This study has made use of spectral subtraction to separate the desired eagle vocalizations from the unwanted background noise present in the audio files. By applying noise reduction to the audio files, undesired background noise has been successfully removed, increasing the sound categorization process' accuracy and precision. Prior to conducting spectral reduction, it has been necessary to apply a High Filter to the sound data. The High filter has served to discriminate between low-frequency sounds and high-frequency sounds.

3) Cutting the audio: Efficient techniques for audio translation were crucial in comprehensively studying and evaluating the noises emitted by Javanese eagles, thereby enhancing our understanding of their behavior and communication. The aim of this method has been to analyze Javanese eagle audio recordings by segmenting them into syllabic units, known as "syllables" in the field of phonetics, with a minimum duration of 0.5 seconds. Additionally, to achieve the best possible outcome, the audio cutting process has utilized the noise-free audio sound.

B. Converting Javanese Eagle Sound into a Spectogram Using Basic Mel Frequency Cepstral Coefficients (MFCC) Technique

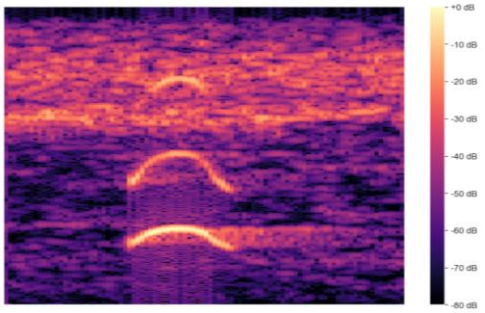
Converting Javanese eagle sounds into spectrograms involves using the basic technique of Mel Frequency Cepstral Coefficients (MFCC) as the primary method. The MFCC method serves as the foundation for generating spectrograms from Javanese eagle sounds, enabling a detailed visual representation of the frequency and energy components of each sound segment.

The MFCC technique involves several crucial steps. First, the sound signal is divided into small time intervals known as frames. Each frame is then analyzed to extract important features encompassing frequency and sound energy information. The next step involves applying Fourier transformation to each frame to convert it from the time domain to the frequency domain. In the MFCC technique, this transformation is typically done using the Short-Time Fourier Transform (STFT). After the transformation, a filter bank is applied to the frequency spectrum to capture more relevant information in the audio spectrum.

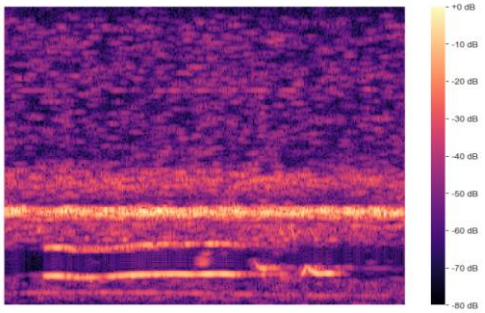
1) Pre-emphasis of audio signal: Pre-emphasis is an audio signal processing technique that enhances sound characteristics by emphasizing high frequencies in the signal. This results in a clearer representation of acoustic details and an overall improvement in sound quality.

Implementing pre-emphasis on each syllable could enhance comprehension of the acoustic properties of that sound. Additionally, pre-emphasis could mitigate the presence of noise in a speech signal, thereby yielding a more pristine rendition of the sound. The pre-emphasis was implemented on sound signals at the syllable level.

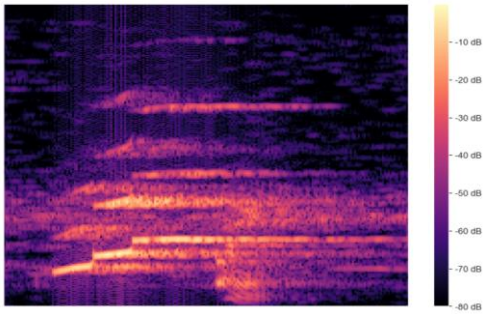
2) Implementation of Short-Time Fourier Transform (STFT): STFT is used to analyze how the frequency of an audio signal changes over time. This technique divides the audio signal into small segments called frames and then performs Fourier transformations on each of these frames. The STFT process enables us to observe how the frequency in an audio signal changes over time, thereby generating a spectral representation of the signal. STFT is employed after obtaining the audio signal that has undergone pre-emphasis. The implementation of STFT in this study, the configuration used includes a NFFT of 512, a window length of 256, a hop length of 128, and the use of a 'hamming' window. The effect of applying the Short-Time Fourier Transform (STFT) on syllables explained in Fig. 8. Fig. 8(a), (b), and (c) depict the STFT representations.



(a)



(b)



(c)

Fig. 8. (a) STFT on normal tweets, (b) STFT on looking for food tweets, (c) STFT on looking for partner tweets.

3) Filter bank implementation for STFT representation: This Filter bank operates by segregating the frequency signal into smaller sub-bands. By applying Filter bank to STFT, frequency information from the audio signal can be more effectively separated, allowing for a more detailed analysis of each frequency sub-band. Each filter within the Filter bank responds to specific frequency ranges. The fusion of STFT and Filter bank has enabled the creation of Mel Spectrogram, facilitating the analysis of Javanese eagle vocalizations within a frequency range similar to human auditory perception and sound processing. This aids in understanding significant changes in the acoustic characteristics of Javanese eagle sound. The resulting Mel Spectrogram serves as a visualization of frequency representation in the form of a spectrogram derived from employing Filter bank techniques on the STFT output.

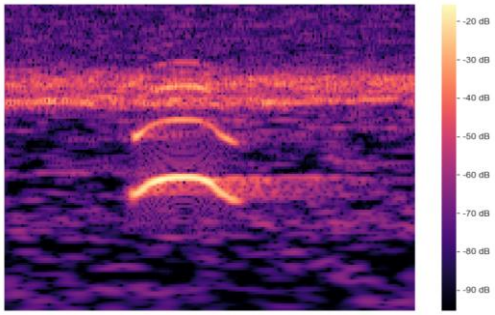
4) Converting a normalized Mel Spectrogram to an image representing the audio: The representing Javanese eagle sound were transformed into a visual image from a standard Mel Spectrogram. The aim is to illustrate audio data in an easily comprehensible visual format, aiding in a simpler understanding of the vocal attributes of the Javanese eagle. The initial step in this study involves normalizing the Mel Spectrogram. Normalization is a crucial procedure to convert spectral data into a standardized format suitable for analysis. The normalization of the Mel spectrogram function balanced the spectrum and increased the Signal-to-Noise Ratio (SNR).

C. Converting Javanese Eagle Sound into a Spectrogram Using Improved Mel Frequncy Cepstral Coefficients (IMFCC) Technique

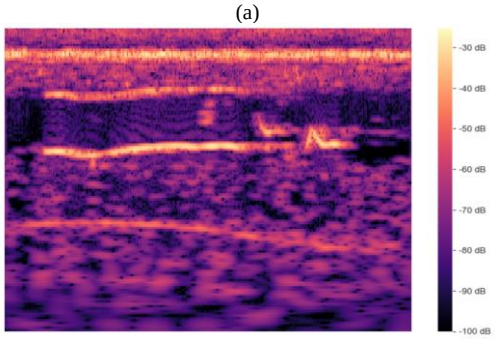
The objective of this study was to provide a novel method for assessing the noises produced by Javanese eagles by transforming them into a visually informative representation known as a spectrogram. The Improved Mel Frequency Cepstral Coefficient (Improved MFCC) technique was employed, involving the substitution of the Short-Time Fourier Transform (STFT) with the Constant-Q Transform (CQT). This research utilizing Javanese eagle vocalizations that had been divided into syllabic segments. Segmentation facilitated the detection of variations in the vocalization of the Javanese eagle, such as alterations in pitch or rhythm that might have been present within each syllable.

1) Pre-emphasis of audio signal: Applying pre-emphasis to each syllabic segment holds the potential to enhance understanding of the acoustic characteristics of that sound. This process aids in distinguishing sounds with high intensity at high frequencies from those with a more balanced frequency range. Additionally, pre-emphasis can reduce noise in speech signals, resulting in a cleaner and clearer representation of the sound.

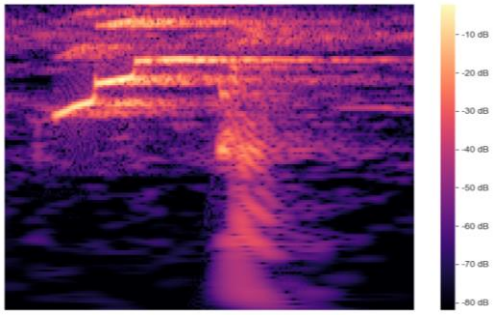
2) Implementation of Constant-Q transform: The CQT setup we used to comprise multiple essential settings. The hop length was configured at 128, and the transformation process employed 'hamming' windows. By utilizing CQT with this particular configuration, it was expected that a more accurate and informative spectral representation of the Javanese eagle's vocalizations could be obtained at the syllable level. The effects of applying the Constant-Q Transform (CQT) on syllables that had undergone pre-emphasis in the preceding sub-chapter are depicted in Fig. 9 below. Fig. 9(a), (b), and (c) depicted a CQT representation, which provided information about energy levels across different frequencies during a certain period of time.



(a)



(b)



(c)

Fig. 9. (a) CQT on normal tweets, (b) CQT on looking for food tweets, (c) CQT on looking for partner tweets.

3) Filter bank implementation for CQT representation: This study involved the creation of a novel technique for evaluating the sounds of Javanese eagles. It achieved this by combining Filter bank with the Constant-Q Transform (CQT) output to generate a more detailed spectral representation called the Mel Spectrogram. The fusion of Constant-Q Transform (CQT) and Filter bank has enabled the generation of a Mel Spectrogram, which facilitated the examination of the vocalizations of the Javanese eagle in a frequency domain that closely aligns with human auditory perception and sound processing. This has aided in comprehending significant alterations in the acoustic properties of sound. The outcome of applying Filter bank to the CQT output has yielded a Mel Spectrogram, as seen in Fig. 10 below. Fig. 10(a), (b), and (c) depicted a Mel Spectogram representation.

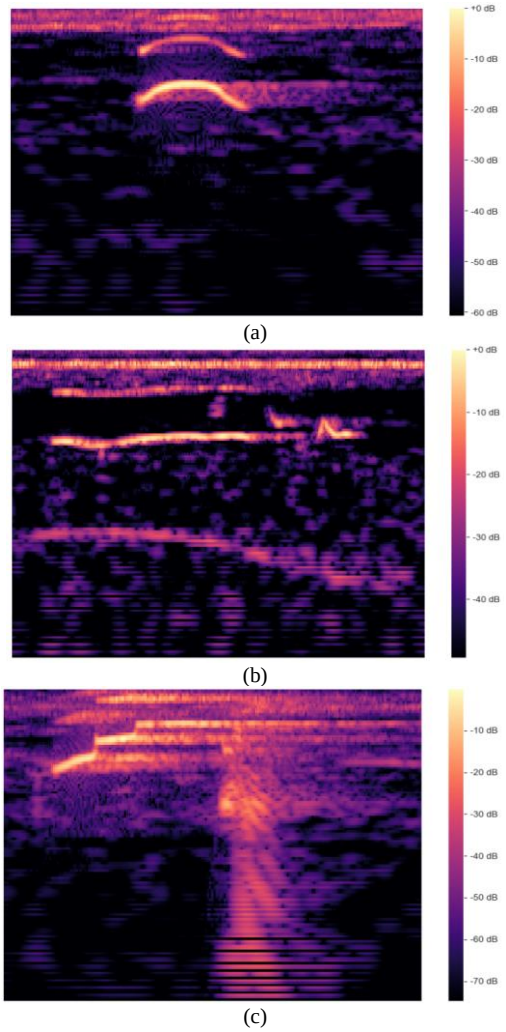


Fig. 10. (a) Normal tweets, (b) Looking for food tweets, (c) Looking for partner tweets Mel spectrogram.

4) Converting a normalized Mel Spectrogram to an image representing the audio: The representing Javanese eagle sound were transformed into a visual image. The first phase in this investigation involved normalizing the Mel Spectrogram. Normalization was a crucial procedure for transforming spectral data into a standardized format suitable for analysis using mean normalization approach to normalize the data, resulting in a spectral average of zero and decreased data variability. The normalization of the Mel spectrogram function balanced the spectrum and increased the Signal-to-Noise Ratio (SNR).

D. Convolutional Neural Network (CNN) Deep Learning for Classifying the Tweets of Javanese Eagle Spectrogram

Convolutional Neural Networks (CNNs) have been a potent category of deep learning models that have demonstrated remarkable efficiency in many picture and signal processing applications, including the classification of spectrograms, such as those depicting the tweets of the Javanese Eagle. Spectrograms are graphical depictions of audio signals, illustrating the variations in frequency components as they evolve over time. By utilizing a Convolutional Neural Network (CNN) and providing it with a dataset of categorized Javanese Eagle tweets, the network has acquired the ability to identify and differentiate unique auditory patterns linked to other categories, such as diverse calls or behaviors exhibited by the eagles.

1) Data Validation and Image Pre-processing: Data validation was crucial in the context of training Convolutional Neural Networks (CNN) on Javanese Eagle sound spectrograms to ensure the dataset's trustworthiness. Spectrogram picture data was considered valid only if it had a file extension of png, jpg, or bmp. This approach guaranteed that only image data adhering to the required format was used for training, preventing any compatibility problems that might have arisen and led to errors during the preprocessing and model training stages.

After completing the data validation stage, the subsequent step involved image preprocessing. This procedure involved modifying the dimensions of the spectrogram image to fulfill the specifications of the CNN model. the spectrogram image underwent a conversion process to either a size of 227 x 227 or 224 x 224. Therefore, the picture preprocessing procedure readied the dataset in a suitable format and ensured the data was prepared for utilization in CNN training for the interpretation of Javanese Eagle sound spectrograms.

2) Data splitting and data labelling: Data labeling and data splitting were two crucial stages in data preprocessing for training and assessing Convolutional Neural Network (CNN) models. Data labeling involved assigning a specific label or category to each individual data instance inside a dataset. Labels were crucial for CNN models to acquire knowledge and comprehend the correlation between characteristics (attributes) and intended outputs. In an image classification task, each image had to be assigned a label that indicated the object or category seen in the image.

Data splitting entailed the partitioning of a dataset into three distinct subsets, commonly known as training data, validation data, and testing data. The objective was to partition a subset of data for the purpose of training a Convolutional Neural Network (CNN) model, referred to as the training data.

Subsequently, data that had not been previously encountered during training was employed to evaluate the CNN model's ability to generate precise predictions, known as the test data. Validation data was employed throughout the training process to oversee and enhance the model. The data splitting are shown in Table II below.

TABLE II. DATA SPLITTING

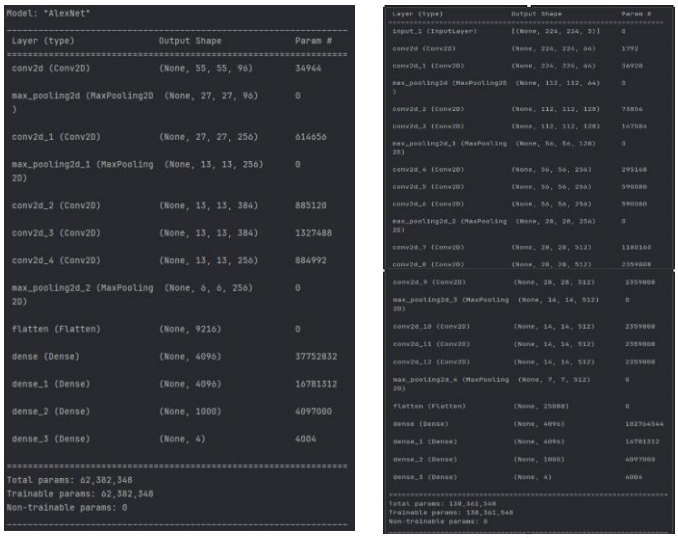
Experiment	Data	Training	Validation	Testing
Experiment I	900	630 (70%)	90 (10%)	180 (20%)
Experiment II	900	630 (70%)	180 (20%)	90 (10%)
Experiment III	900	720 (80%)	90 (10%)	90 (10%)

3) CNN Architecture Model: The Convolutional Neural Network (CNN) architecture has served as the fundamental framework in image processing and has consisted of multiple layers that collaborated to achieve a profound comprehension of picture data. The method commences with convolution layers, where tiny filters or kernels are employed to extract distinctive features like edges, textures, and patterns from the image. The outcome is a feature map that accentuates significant details within the image.

Max Pooling layers have typically succeeded convolution layers, serving to reduce data dimensionality and computational complexity by selecting the highest value within an overlapping region. Subsequently, a Flatten layer is employed to transform the three-dimensional feature map into a singular, one-dimensional vector. A Dense layer is a type of neural network layer characterized by having every neuron connected to every neuron from the previous layer. This means that there is a full and direct connection between all neurons in the layer and the neurons in the previous layer. The ReLU (Rectified Linear Unit) activation function is applied to each neuron in the Dense and convolution layers to introduce non-linearity to the model, aiding in acquiring more intricate features. The Softmax activation function computes the probabilities for each distinct class, and the class with the highest probability is selected as the final prediction of the model.

Apart from that, two architectures have been employed, namely Alex Net and VGGNet-16. The architectures were used in the training, testing and evaluation process for experiments 1, 2, and 3 that have been carried out. Model summary of Alex Net and VGGNet-16 were displayed in Fig. 11.

4) Training and testing CNN model: Assessing the performance of a Convolutional Neural Network (CNN) model during training is crucial. This involves tracking loss function values and accuracy to gauge the model's success in categorizing the training data. Choosing the appropriate number of epochs is important, with 25 epochs being ideal to ensure the model converges well. Evaluating the model's performance is done through experiments on the dataset, with results providing valuable insights into the impact of validation and testing dataset sizes and the proportion of training data. After training, the model is tested on never-before-seen testing data to evaluate its ability to generalize learned information. Evaluation metrics such as accuracy, precision, recall, F1-score, Confusion Matrix, and AUC-ROC curve offer valuable insights into the model's effectiveness in classifying previously unseen new data.



(a) (b)

Fig. 11. Model summary of (a) Alex net, (b) VGGNet-16.

Experiment 1: The trained MFCC with AlexNet model achieved 97% accuracy in identifying Javanese Eagle's sound patterns. However, the testing showed a slightly lower accuracy rate of 95% with some identification failures in specific categories, indicating the need for further improvement.

Experiment 2: The trained MFCC with VGGNet Model showed stable accuracy despite a significant increase in loss. Testing showed an accuracy rate of 94% with some identification failures in specific categories. The early stopping method was effective in optimizing the training process.

Experiment 3: The trained MFCC with AlexNet Model achieved 97% accuracy in identifying Javanese Eagle's sound patterns. However, the testing showed a slightly lower accuracy rate of 93% with some identification failures in specific categories, indicating the need for further improvement.

Experiment 4: The trained MFCC with VGGNet model on a 70:20:10 dataset split achieved 97% accuracy during training. However, there was a significant increase in loss values during epochs 15-20. The early stopping method was implemented at epoch 20 and proved successful. The model's ability to classify spectrogram images remained highly reliable during testing.

Experiment 5: The trained MFCC with AlexNet model achieved 98% accuracy during training, showcasing exceptional ability to capture the patterns and characteristics within the spectrogram data. The model's performance during testing remained highly accurate.

Experiment 6: The trained MFCC with VGGNet model achieved 98% accuracy during training, showcasing exceptional ability to capture the patterns and characteristics

within the spectrogram data. The model's performance during testing remained highly accurate.

Experiment 7: A CNN with AlexNet architecture was trained to learn the IMFCC spectrogram of Javanese Eagles. The model achieved an accuracy rate of 99% during training. In testing, it achieved an accuracy rate of 98% with some identification failures.

Experiment 8: A CNN with VGGNet architecture was trained to learn the IMFCC spectrogram of Javanese Eagles. The model achieved an accuracy rate of 97% during training. In testing, it achieved an accuracy rate of 97% with some identification failures.

Experiment 9: A CNN with AlexNet architecture was trained to learn the IMFCC spectrogram of Javanese Eagles. The model achieved an accuracy rate of 97% during training. In testing, it achieved an accuracy rate of 97% with some identification failures.

Experiment 10: The results of training a CNN with VGGNet architecture to learn the IMFCC spectrogram of Javanese Eagles are not provided in the given text.

Experiment 11: The IMFCC and AlexNet models were trained and tested for the 80:10:10 dataset. The AlexNet architecture recognized and learned the sound patterns of Javanese Eagles with high precision, achieving an accuracy rate of 98% during training. Testing involved classification analysis, Confusion Matrix, and ROC Curve, with an overall accuracy rate of 97%.

Experiment 12: The IMFCC and VGGNet models were trained and tested for the same dataset. The VGGNet architecture achieved a peak accuracy rate of 100% during training and an accuracy rate of 99% during testing. The ROC Curve showed an area under the curve of 0.996 for the 'Normal' category, 1.000 for 'Looking For Partner', and 0.996 for 'Looking For Food'.

Improved Mel Frequency Cepstral Coefficient (IMFCC). The CQT approach in IMFCC was found to be more precise that the Short-Time Fourier Transform in MFCC. In experiment 12, the CNN model based on the VGGNet-16 architecture achieved 99% accuracy. The resulting AUC value shows an Average Recall of 0.89, an Average Specificity of 0.965, and an Area under Curve (AUC) value of 0.93. The MAPE value for the LookingForFood class is 0.006; for the LookingForPartner class, it's 0.315; and for the Normal class, it's 0.045. The average MAPE value can be calculated by adding the MAPE values for each class and dividing the result by the number of classes. In this case, the average MAPE value is (0.006 + 0.315 + 0.454) / 3 = 0.141. Precision for LookingForFood was 0.93, for LookingForPartner it was 1.00, and for Normal, it was 0.95. The F-Score values for the classes were 0.96 for LookingForFood, 0.81 for LookingForPartner, and 0.97 for Normal. These values calculate the average F-Score as (0.96 + 0.81 + 0.97) / 3 = 0.91. This results show that the purposed method is suitable for detecting real-time Javan Hawk-Eagle sounds.

5) CNN model validation: The evaluation process begins by considering the complexity of analyzing real-time sound recordings of the Javan hawk-eagle. The CNN model, utilizing the VGGNet-16 architecture, undergoes rigorous testing, analyzing the intricate vocalizations of these birds within audio frames. The raw dataset, consisting of the model's predictions, acts as a preliminary insight into the model's performance. Instances where the Javan hawk-eagle sounds are correctly identified with confidence reflect the model's accuracy. The VGGNet-16 architecture-based CNN model achieved 99% accuracy in experiment 12, making it suitable for detecting real-time Javan hawk-eagle sounds.

Metrics like the Area under Curve (AUC), Mean Absolute Percentage Error (MAPE), and F-Score are pivotal in this evaluation to measure the model's performance. The evaluation csv file contains authentic sound recordings, labels, and tweet counts directly obtained from the Javanese Eagle conservation site. It provides a deep understanding of the Javan hawk-eagle sound detection balance between precision and recall, aiding in refining and enhancing the accuracy and reliability of the model for better practical applications in sound detection in natural environments.

The prediction results were processed to compare the predicted data with the actual data, which can be observed in the actual and predicted tweet sounds image. The Confusion Matrix generated from the prediction outcomes for each audio file in Evaluation.csv provides a detailed breakdown of the model's performance in classifying different sound categories from the audio files in the evaluation dataset. Lastly, the metrics TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) derived from the Confusion Matrix present a detailed account of the model's performance in accurately identifying each class and its corresponding errors or correct identifications.

The Area under Curve (AUC) is a performance metric used to measure the reliability of classification models in distinguishing between different classes. AUC measures how well the model recognizes instances of a class and separates that class from others. The average Recall and Specificity values for each class are then computed to calculate the AUC, which ranges from 0 to 1. A high AUC value signifies that the model exhibits proficiency in distinguishing between behaviors.

The Mean Absolute Percentage Error (MAPE) is a metric used to measure the relative error in a model's predictions. It is crucial in evaluating the accuracy of the model's prediction for each behavior class of Javanese eagles. MAPE is computed by averaging the absolute percentage errors between the actual and predicted values. Lower MAPE values indicate the model's ability to accurately predict eagle behavior, while higher values signify a greater level of inaccuracy.

The F-Score is an essential metric for evaluating the performance of classification models in identifying Javan Hawk-Eagle behaviors. It combines Precision and Recall to provide an overview of how well the model can differentiate between LookingForFood, LookingForPartner, and Normal behaviors based on the observed sounds. Precision measures the accuracy of the model's predictions for a specific class,

while Recall evaluates how well the model can identify all instances of that class.

V. Conclusion and Future Works

In conclusion, the research presented a novel approach, combining Improved Mel Frequency Cepstral Coefficients (IMFCC) with Deep Convolutional Neural Networks (CNN), to decode the intricate vocalizations of the Javanese Eagle. The research aimed to decode the Javanese Eagle's vocalizations by combining Improved Mel Frequency Cepstral Coefficients (MFCC) with Deep Convolutional Neural Networks (CNN), promising enhanced comprehension and species conservation. Using IMFCC, experiment 12 achieved 99% accuracy with VGGNet-16 architecture, showing significant promise in real-time sound detection. The method achieved high accuracy rates, demonstrating its suitability for real-time sound detection. With an Average Recall of 0.89, an Average Specificity of 0.965, and an AUC value of 0.93, the study's findings underscore the effectiveness of the proposed technique in understanding and classifying Javanese Eagle sounds.

Moving forward, future research avenues include exploring alternative machine learning models such as Recurrent Neural Networks (RNNs) and Convolutional Recurrent Neural Networks (CRNNs) to improve temporal analysis. Additionally, investigating advanced feature extraction techniques beyond IMFCCs, like Hybrid Cepstrum Analysis (HCA) or Mel-Spectrograms, could provide deeper insights. Further exploration of evaluation metrics focusing on temporal accuracy and ethical considerations surrounding model deployment in natural environments will enhance the robustness and applicability of sound detection and recognition systems.

References

[1] S. N. Utami, "Usaha Untuk Melestarikan Elang Jawa," [Online]. Available: www.kompas.com, 2021. [Accessed: August 15, 2021].

[2] E. P. Putra, “Habitat Elang Jawa Diambang Kepunahan,” [Online]. Available: www.replubika.co.id, 2015. [Accessed: Aug. 15, 2021].

[3] A. Karpyn, G. Sawyer-Morris, S. Grajeda, K. Tilley, and H. Wolgast, “Impact of Animal Characters at a Zoo Concession Stand on Healthy Food Sales,” *Journal of Nutrition Education and Behavior*, vol. 52, np. 1, pp. 80-86, 2020, doi: 10.1016/j.jneb.2019.09.013.

[4] P. Lindhout and G. Reniers, “Reflecting on the safety zoo: Developing an integrated pandemics barrier model using early lessons from the Covid-19 pandemic,” *Safety Science*, col. 130, 104907, 2020. doi: 10.1016/j.ssci.2020.104907.

[5] P. E. Rose, S. M. Nash, and L. M. Riley, “To pace or not to pace? A review of what abnormal repetitive behaviour tell us about zoo animal management,” *Journal of Veterinary Behaviour: Clinical Applications and Research*, vol. 20, pp. 11-21, 2017, doi: 10.1016/j.jveb.2017.02.007.

[6] F. Berger, W. Freillinger, P. Primus, and W. Reisinger, “Bird audio detection-dcase 2018,” DCASE2018 Challenge, Tech. Rep., 2018.

[7] L. Kettler and C. E. Carr, "Neuroethology of Sound Localization in Birds," in Encyclopedia of Animal Behavior, 2nd ed., 2019, doi: 10.1016/B978-0-12-809633-8.01274-7.

[8] S. Paul, A. X. Glittas, ad L. Gopalakrishnan, “A low latency modular level deeply integrated MFCC feature extraction architecture for speech recognition,” *Integration*, col. 76, pp. 69-75, Dec. 2019, doi: 10.1016/j.vlsi.2020.09.002.

[9] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, “Application of deep larning technique for heartbeats detection using ECG signals-analysis and review,” *Computers in Biology and Medicine*, vol. 120, Mar. 2020, doi: 10.1016/j.compbiomed.2020.103726.

[10] J. Niemi, and J. T. Tanttu, “Deep learning case study for automatic bird identification,” *Applied Sciences*, vol. 8, no. 11, pp. 1-15, Nov. 2018, doi: 10.3390/app8112089.

[11] W. Zhang and L. Guoxin, "The Research of Feature Extraction Based on MFCC for Speaker Recognition," 2013.

[12] J. Song and S. Li, "Bird sound detection based on binarized convolutional neural networks," *Lecture Notes in Electrical Engineering*, vol. 568, pp. 63–71, 2019. DOI: 10.1007/978-981-13-8707-4_6.

[13] J. Xie and M. Zhu, "Ecological Informatics Handcrafted features and late fusion with deep learning for bird sound classification," *Ecological Informatics*, vol. 52, pp. 74–81, May 2019, doi: 10.1016/j.ecoinf.2019.05.007.

[14] Ferlazafitri, Syartinilia, and Y. A. Mulyani, “Habitat patch connectivity of Javanese Eagle (Nisaetus bartelsi) in Eastern Part of Java, Indonesia,” *IOP Conference Series: Earth and Environmental Science*, vol. 590, no. 1, 012003, 2020, doi: 10.1088/1755-1315/590/1/012003.

[15] I. Fahmi and Syartinilia, “Habitat preferences of current record of JHE (Nisaetus bartelsi) in lowland forest in Ujung Kulon National Park,” *IOP Conference Series: Earth and Environmental Science*, vol. 590, no. 1, 012004, 2020, doi: 10.1088/1755-1315/590/1/012004.

[16] R. A. Suyitno and Syartinilia, "Assessing potential habitat of Javanese Eagle (Nisaetus bartelsi) based on landscape characteristic in Banten Province," *IOP Conference Series: Earth and Environmental Science*, vol. 590, no. 1, 012001, 2020. DOI: 10.1088/1755-1315/590/1/.

[17] A. M. Tripathi and A. Mishra, "Self-supervised learning for Environmental Sound Classification," *Applied Acoustics*, vol. 182, 2021, 108183, doi: 10.1016/j.apacoust.2021.108183.

[18] T. Virtanen, M. D. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events," *Computational Analysis of Sound Scenes and Events*, pp. 1–422, 2017, doi: 10.1007/978-3-319-63450-0.

[19] H. Zhenyi and J. Dacan, "Acoustic scene classification based on deep convolutional neural network with spatial-temporal attention pooling," pp. 2–6, 2019.

[20] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," pp. 1–21. [Online]. Available: https://arxiv.org/pdf/1807.05511.pdf, 2019.

[21] K. Nagy, T. Cinkler, C. Simon and R. Vida, "Internet of Birds (IoB): Song Based Bird Sensing via Machine Learning in the Cloud : How to sense, identify, classify birds based on their songs?," *2020 IEEE SENSORS*, Rotterdam, Netherlands, 2020, pp. 1-4, doi: 10.1109/SENSORS47125.2020.9278714.

[22] E. Sasmaz and F. B. Tek, "Animal Sound Classification Using A Convolutional Neural Network," in *UBMK 2018 - 3rd International Conference on Computer Science and Engineering*, 2018. DOI: 10.1109/UBMK.2018.8566449.

[23] L. G. C. Vithakshana and W. G. D. M. Samankula, "IoT based animal classification system using convolutional neural network," *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka, 2020, pp. 90-95, doi: 10.1109/SCSE49731.2020.9313018.

[24] Z. J. Ruff, et al., "Workflow and convolutional neural network for automated identification of animal sounds," *Ecological Indicators*, vol. 124, p. 107419, 2021.

[25] A. D. P. Ramirez, J. I. de la Rosa Vargas, R. R. Valdez and A. Becerra, "A comparative between Mel Frequency Cepstral Coefficients (MFCC) and Inverse Mel Frequency Cepstral Coefficients (IMFCC) features for an Automatic Bird Species Recognition System," *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, Gudalajara, Mexico, 2018, pp. 1-4, doi: 10.1109/LA-CCI.2018.8625230.

[26] S. D. H. Permana, et al., "Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4345-4357, 2022.

[27] S. D. H. Permana, and K. B. Y. Bintoro, “Implementation of Constant-Q Transform (CQT) and Mel Spectrogram to converting Bird’s Sound,” in *2021 IEEE International Conference on Communication, Networks and*

Satellite (COMNETSAT), Jul. 2021, pp. 52-56, doi: 10.1109/COMNETSAT53738.2021.9539187.

[28] B. H. Juang, L. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no.7, pp. 847-954, July 1987, doi: 10.1109/TASSP.1987.1165237.

[29] V. Tavakkoli, K. Mohsenzadegan, and K. Kyamakya, "A visual sensing concept for robustly classifying house types through a convolutional neural network architecture involving a multi-channel features extraction," *Sensors (Switzerland)*, vol. 20, no. 19, pp. 1–16, 2020, doi: 10.3390/s20195672.

[30] W. Caesarendra, T. Triwiyanto, V. Pandiyan, A. Glowacz, S. D. H. Permana, and T. Tjahjowidodo, "A CNN prediction method for belt grinding tool wear in a polishing process utilizing 3-axes force and vibration data," *Electronics*, vol. 10, no.12, p. 1429, 2021. doi: 10.3390/electronics10121429.

[31] P. C. Loizou, "Spectral-Subtractive Algorithms," in *Speech Enhancement: Theory and Practice*, 2nd ed., CRC Press, 2017, ISBN 9781138075573.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] D. Anggraeni, W. S. M. Sanjaya, M. Munawwaroh, M. Y. S. Nurasyidiek, and I. P. Santika, "Control of robot arm based on speech recognition using Mel-Freuency Cepstrum Coefficient (MFCC) and K-Nearest Neighbors (KNN) method," *2017 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA)*, Surabaya, Indonesia, 2017, pp. 217-222, doi: 10.1109/ICAMIMIA.2017.8387590.