Journal of Theoretical and Applied Information Technology
15th March 2022. Vol.100. No 5
2022 Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYSIS ON VOWEL /E/ IN MALAY LANGUAGE RECOGNITION VIA CONVOLUTION NEURAL NETWORK (CNN)

NIK MOHD ZARIFIE HASHIM [1], NIK ADILAH HANIN ZAHRI [2], MOHD JUZAILA ABD. LATIF [3,4], ROSTAM AFFENDI HAMZAH [5], NIK FARIZAL HASHIM [6], MAISARAH KAMAL [7], MAHMUD DWI SULISTIYO [8] and AFIQAH IYLIA KAMARUDDIN [9]

[1] Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka, Malaysia

[2] Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis, Malaysia

[3] Fakulti Kejuruteraan Mekanikal, Universiti Teknikal Malaysia Melaka, Malaysia

[4] Advanced Manufacturing Centre, Universiti Teknikal Malaysia Melaka, Malaysia

[5] Fakulti Teknologi Kejuruteraan Elektrik dan Elektronik, Universiti Teknikal Malaysia Melaka, Malaysia

[6] Kolej Kemahiran Tinggi Mara, Pasir Mas, Kelantan, Malaysia

[7] Centre for Foundation Studies in Science, Universiti Malaya, Malaysia

[8] School of Computing, Telkom University, Indonesia

[9] Pusat Rehabilitasi PERKESO, Melaka, Malaysia

E-mail: [1]nikzarifie@utem.edu.my

ABSTRACT

In recent years, the silent killer disease, defined as a non-communicable disease, has become a frequent topic discussed in many academic discussions. Although this disease is not transferable from one to another, starting from 1990, the increment trend was annually published by the world statistic data for this disease, e.g., heart attack and stroke. The more significant consequence of these two diseases is to disable one or more human capabilities. One of the stroke disease effects is becoming disabled from hearing. Speech disabilities are the focus of this proposed study in this paper. Since the person diagnosed as a stroke patient requires attending the recovery session or rehabilitation session, the rehabilitation center must prepare and provide a sound module and system to help the patient regain their capability. Rehabilitation is an alternative path to gradually giving routine practice to the patient to improve their capability back. For this purpose, the rehab center requires a quantity of time to provide the patient to attend the training session. The training, however, is conducted in two ways, physically and virtually. For the Malaysia stroke patient, the training for pronouncing the vowel in the Malay language is crucial in getting back the speaking capability. Since the Malay language has 6 types of vowels, which are /a/, /e/, /ê/, /i/, /u/, and /o/. Here, there is a limitation to smartly recognizing the difference between the two /e/ vowels. Malay's /e/ vowel is crucial as the similar spelling vocabulary conveys two different meanings. This study analyzed the differences in recognizing the two /e/ vowels using Convolution Neural Network (CNN) with the help of the existing sound-image dataset.

Keywords: *Convolutional Neural Network (CNN), /e/ vowel, Malay language, Non-Communicable Disease (NCD), Recognition, Rehabilitation, Stroke patient*

1. INTRODUCTION

An illness that affects a person and a condition that prevents the human body or mind from working defines a disease. Here, the disease is categorized into congenital and acquired diseases. In acquired disease then is grouped into two types, communicable and non-communicable disease. Non-communicable disease (NCD) is a common disease that transpires during a period of time. NCD could be caused by several bases, including the genetic transfer and combination of the genome, a frequent habit of the human body in daily life, environmental conditions, and behavioral

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

occurrences. NCD killed nearly forty-one million people in one year, equal to seventy-one percent of worldwide human death percentages. According to the global death numbers, a total of seventy-seven percent is from NCD deaths happening in the developing and low-income region, e.g., South-East Asia, which includes Malaysia. Cardiovascular disease became the most death number from NCD, 17.9 million people annually, where the stroke patients' numbers also counted in this statistical numbers. Daily human life, related to the harmful habit, smoking, reluctance to do physical activity, and over drink alcohol, boost the probability risk of this NCD being infected in their life. Therefore, there are still steps to be taken as a precaution from the NCD, which are periodically health screening, quick NCD treatment, and intensive care are important elements as to response the NCDs.

Due to the increment trend of NCD patients in Malaysia, this event puts the same increment in the number of diagnosed people as disabled people with NCD. Government and Non-government organizations (NGOs) collectively provide awareness programs via physical programs, virtual events, distributing posters, and many more to keep people understanding the effect of NCD. The most comprehensive action provided by the authority is providing a center for the diagnosed people with NCD to gain back the lost capability in the schedule of time. As the provided treatment requires consistent commitment from the rehab staff and patient, the rehab session's time frame is crucial for ensuring the patient's ability can be obtained back. One example of a disability that is nearly possible for training and learning back is speech ability. For this reason, in the proposed paper, we could conduct a study on speech ability for helping stroke patients to have a systematic rehab training system.

As rehabilitation activities are essential for gaining back the speech disability, an efficient and reliable training program must be designed and provided. Most of the existing Malaysian rehabilitation center has recently provided a manual rehab program for speech-disabled people. It combined the direct assessment from the staff to the patient and partially indirect assessment, which used the modern tool for voice recording. Later, the rehab expertise will examine the patient's performance. The two types of assessment conducted in the rehab center requires them plenty of time to assess and examine every session manually. This manual assessment, "listen and evaluate," also demands several rehab centers staff to assess the patient periodically, especially when the rehabilitation has many patients to be treated at one time. The two factors that could influence the effectiveness of the rehab activities are time and human resources; however, they could affect a long period to complete the rehabilitation, which later could extend and prolong the whole recovery time for one patient. Likewise, the COVID-19 is still surrounding us; this pandemic would contaminate the patients' restorative treatment.

Dealing with the stroke patient in the rehab center is not an easy task for ensuring complete training for the patient in a dedicated time. Despite using the manual evaluating procedure, which could cater to the time-consuming and less staff in the rehab center, an intelligent system for the training could be the choice. An intelligent training system for these stroke patients was proposed and conducted by existing research using sound features and the help of the computer. Although the human-computer interface was promoted to be used widely in rehab, the cost to provide this service is expensive to come across the globe. Due to broad global coverage for stroke patients, the intelligent system could not perform reliably as we expected. Thus, the manufacturer or the researchers should initiate the regional dataset images of sounds. Nevertheless, considering the region sharing a similar historically language such as Malaysia, Indonesia, Brunei, Singapore and South of Thailand, the dataset could be possibly shared among them for research purposes [1]. Malaysia, which is a country that uses the Malay language as their national and first language. The language consists of six types of vowels /a/, /e/, /i/, /o/, /u/ and the second vowel of /e/ cap or /e/ hat makes them different from the English. As the wide regional coverage of the language and dissimilarity with English, the paper studies and analyses the differences of the vowel individually and the two /e/vowels in the Malay language dedicated to the NCD patient, a stroke patient.

Signal was originally constructed by an object or hard material vibration, which later produced a wave signal. In the scientific outlook, these transmitted wave signals from the sound source to our ear denote as speech waveforms. In the proposed paper, we utilized the traditional speech waveform to another perspective: an image of the sound properties. Generally, the sound is defined as a combination of two-axis graph components: time in the x-axis and amplitude in the y-axis. In bringing the sound study with an intelligent approach in human-computer interaction, these images of sound properties will be manipulated and analyzed using the Convolution Neural Network (CNN). Since the CNN only

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

occupies the images with pixel properties, we utilized sound to image conversion in the middle of study preparation. However, this sound to image conversion could be possible with various methods; one of them is converting the sound into frequency domain which happens in Short Time-Frequency Transformation or Wavelet Transformation. Here, for analyzing the vowels for the Malay language, in this paper, we look at the wave signal as a spectrogram image which later delivers adequate knowledge about these six Malay language vowels.

Malay is recognized as Malaysia's national language under Article 152 of the Malaysian Constitution, and it became the sole official language in Peninsular Malaysia in 1968 and Borneo Island as East Malaysia gradually from 1974. Before the year, this Malay language had already become lingua franca since the 14 centuries when the Malacca Sultanate time. Ever since, the Malay language has been nearly 33 000 000 people spoken as their daily conversation language in several countries around the southeast of Asia, especially Malaysia, South of Thailand, Singapore, and Brunei Darussalam [2]. Based on the large numbers of spoken people in South-East Asia, the study of vowels classification for the stroke patient is considered a crucial sub-area in the biomedical engineering field, which is crucial for rehab activities [3 - 6].

Since the previous study was conducted for five vowels for the Malay language [8], in this paper, we filled the gap in this field area for the study on the full vowels available. The Malay language's two types of /e/ vowels produce a similar spelling but not pronunciation. The Malay speaker is the only one who can generally differentiate them by hearing the voice and distinguishing them in the usage context. An example of this is when we use "kepak" and "k *ê* pak," which means “wings” and “to open something”. These two words would confuse the new Malay language user as they read or pronounce it during the conversation. This pronunciation will affect the performance of the speech rehabilitation among the stroke patient.

We employed six vowels in the Malay language as our main analysis subject, /a/, /e/, /i/, /o/, /u/ and including the second vowel /e/ for the analysis. CNN is used as the main classification tool to classify the six vowels, as a widely used tool in diverse research for classification. It also produces a stable and relative classification result. For presenting the classification performance, VGG16 [7] is one of the straightforward and universal CNN networks, is utilized as the primary model for vowel classification in the experimental work.

Our contribution can be summarized as follows:

• we compare the performance of two types /e/ vowels classification information by utilizing the spectrogram image,

• we show the effect of analytic study in vowels classification via CNN using two, five, and six classes.

Note that this paper is an extended version of the previous conference paper [8]. We add the study on the difference of /e/ vowels classification performance in Malay compared to the previous paper. The evaluation also is conducted using six classes of vowels. This paper delivers Malay language vowel recognition for stroke patients, utilizing Convolution Neural Network (CNN). The method introduced spectrogram images to substitute the traditional sound file in CNN to recognize all the six vowels for the Malay language. The proposed paper is divided into several sections: Section II presents the related works; Section III discusses the proposed methods; Section IV explains the dataset arrangement. Section IV is later followed by Section V, which addresses the results of the proposed method's conducted experiments. Then, in Section VI, we conclude the paper and suggest several future works.

2. RELATED WORKS

Much research works on helping disabled people, specifically NCD patients, recover in getting the previous sound or speech capability. Although the study was not directly dedicated to the stroke patient, the conducted study aimed to help the patient who is diagnosed with the articulation disorder [9]. For this reason, a manual approach for the speech therapy activity was firstly introduced with a systematical treatment. By definition, articulation disorder is described as speech confusion. This speech confusion included difficulty differentiating the various types of sounds, hardship on interchanging sound with other sounds, speech mumbling, or it could also be intermediate silenced speech. Most of the early treatment and proposed methods focused on the error sound's phonetic placement. The motor skills considered a significant part of the training session could be trained by producing the sound appropriately during physical therapy. This treatment session could be conducted by utilizing flashcards and gesture training.

Apart from the speech capability improvement, a hierarchy method to help the diagnosed children demonstrate the right and correct way to sound the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

vowels without any help of smart methodology. The proposed research work for this purposed was focused on the severity of speech disorders among the child who has mild level, intermediate level, intense level, and inability to speak level. Here, the traditional approach is still applied to help the children pronounce the vowel correctly based on the individual treatment. This technique, treating the severity of speech disorder patients, was among the earliest speech rehabilitation methods conducted by the professional standard [10]. Then, using the articulation drills and motor learning, Van Riper proposed training involving the movement of the patient's tongue. He also suggested the steps for the remaining articulators, a manual rehab approach, including lips and jaw [11].

The dialect element of voice and speech recognition has been the subject of numerous studies. A deep learning-based system for identifying shrieking by [12], is claimed to be the first to use CNN. The new proposed idea of using MFCC characteristics to convert audio to an image form. Audio recordings are extracted and then provided to the DL classifier to detect the screaming sound. Other researchers, meanwhile, focused on the Gaussian Mixture Model (GMM) and the subsequent classification approach, Support Vector Machine (SVM), and compared their performance with CNN in 2D and 3D pattern signals [12 - 17].

Machine Learning (ML) was one of the first ways to improve traditional methods. With their proposed risk-stratification models created by self-reports from the baseline patient, the ML method results in clinically meaningful major disorder disease (MDD) [18]. The CNN is used to construct a fruits classification based on a fruit control classification system on the Convolution Neural Network (CNN) project published in 2018. Nine hundred seventy-one images from 30 different classes account for 94% of the classification's accuracy. In addition, the classification of fruits using CNN implementation is proposed. [10] has also done work on a control system by building an automated vision-based system using a computer vision technique. [5-6] uses image processing to describe the vocal realm using magnetic resonance imaging (MRI). The work of Husni T. et al. demonstrates the application of four different types of frequencies to investigate the continued Malay children who are between seven to twelve years old. The most recent work conducted by Nur Syakirah et al. focused on the five vowels in the Malay language. The proposed study and analyze was in the preliminary stage of analyzing the class of vowels.

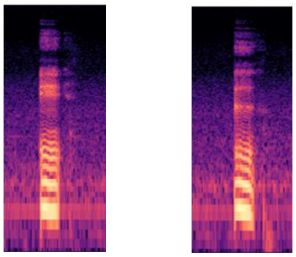


Figure 1: Two types of /e/ vowel in the Malay Language which impose a similar appearance in spectrogram images

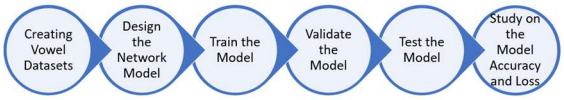


Figure 2: Proposed workflow

They presented only five vowels comparison using the sound-image dataset. Although the comparison study showed a good result, but still, there is a need to foresee the performance of the actual six types of vowels in Malay, which reflected to the original five vowels from their work plus the additional type of /e/ vowel. For this reason, the additional /e/ vowel for the classification is a crucial task to be studied as illustrated in Figure 1.

This paper provides a comprehensive study and analysis on vowel classification via CNN for speech rehabilitation. Since the limited research conducted for the specific /e/ vowel in the Malay language initiated from past time, we presented the effect of several analysis settings in the experimental work. The classification group numbers, batch size selection, and epoch size selection are among the analysis settings to understand the vowel /e/ classification specifically and the other five vowels generally. The existing dataset sound-image for six vowels in the Malay language will be used to input the whole vowel classification in this paper.

3. PROPOSED METHODS

In this study, we offer a basic CNN model network used to create a comparison network model for the newly proposed six Malay language vowel kinds. This section describes the process of obtaining an image from an original audio signal file. We transformed the appearance of a vowel sound into a new type of information, which is in the image format. A comprehensive dataset of images comprising several dimensions is required to develop a new intelligent classification system for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

stroke patients. This dataset is accomplished by carefully selecting the wave signal captured from the patient under various wave circumstances.

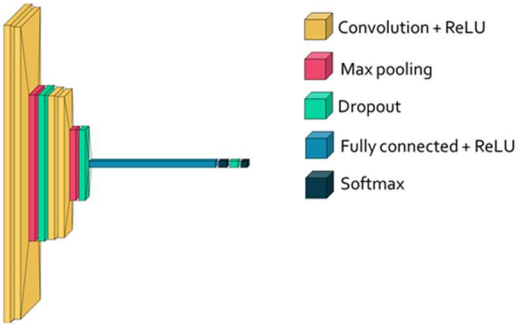


Figure 3: The proposed network architecture

Converting the audio input to an image is critical for CNN's study on vowel recognition. Amplification of the audio stream results in the generation of spectrogram images. We combined existing network models such as VGG16 and VGG 19 with our

suggested six-class network model. The overall approach for this article is depicted in Figure 2.

2.1 The proposed network

For the vowel recognition problem, we developed a basic network model. A convolutional layer, a pooling layer, a fully connected layer, a dropout layer, and activation functions comprise the proposed network based on the CNN design. This project chose a resolution of 240 × 55 pixels for the input image based on the images from the six vowels collection. Before beginning the convolution phase, these input images are set to 240, 55, and 3 in total. The first convolution layer, dubbed Conv1, is followed by Conv1-1, which employs 32 filters and considers the dimension of the image to be 238 × 53. The convolution layer with 64 filters and max-pooling dimensions 236 × 51 for the images which are used in this case. Conv2 is the second convolution layer, consisting of 32, 64, or 128 filters.

Following that is the Conv2-1, which employs 128 filters and maximum pooling with an image resolution of 116 × 23. Simultaneously, the Conv2 2 employs 128 filters with a maximum pooling size of 9 × 56. Following that, we create a dense layer with 1024 units of a dense layer and five units of a dense SoftMax layer. Figure 3 illustrates the image size decrease for each layer. The model was constructed using a CNN with an input image dimension of 240 × 55, an ADAM classifier as the optimizer, and SoftMax as the activation function.

Following that, the model is tested by running it with

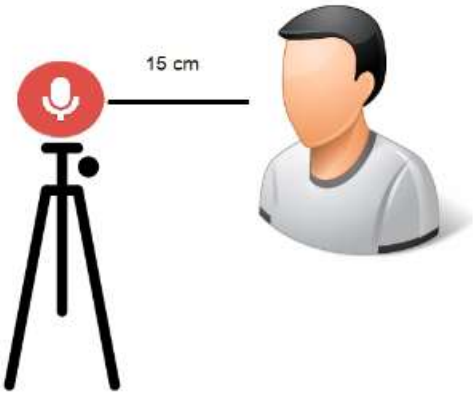


Figure 4: The recording arrangement schematic

various of vowel class setting, selection of batch sizes and selection of epoch sizes for the study comparison purposes.

4. DATA ARRANGEMENT

This section describes how the photos for the datasets used in the article were acquired and organized. We transformed the audio signal wave file to a spectrogram image to get information about the vowel. We will address audio recording, wave signal to an image, and dataset organization.

2.2 Audio Recording

The vowel audio signals are obtained from nine male and female subjects aged twenty to twenty-four years. Recognizing that a single dataset layout should contain a range of audio vowel kinds, we conducted a recording session using audio signals with varying durations: short period, medium period, and long period. Three distinct periods of audio signal are captured for vowel /a/, /e/, / ê /, /i/, /u/, , and /o/.

To minimize background noise during the recording, we put the recorded voice 15 centimeters away from each participant. When recording a brief audio vowel signal, the individual must speak /a/ in less than a second. In comparison, the subsequent two-period audio vowel signal, middle-period signal, and long-period signal are captured at 2- and 3-second lengths, respectively. We employ a voice recorder, the REMAX RP1 8GB Digital Audio

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Voice Recorder, to capture the voices of the nine sample participants in a manner similar to the recording arrangement schematic shown in Figure 4.

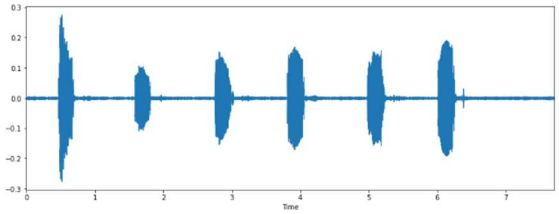


Figure 5: Sample of vowel audio signal in time domain for the six vowels /a/, /i/, /u/, /e/, /ê/, and /o/ in Malay Language

```
plt.figure(figsize=(14, 5))
librosa.display.specshow(Xdb, sr=sr, x_axis='time', y_axis='log')
plt.colorbar()
```

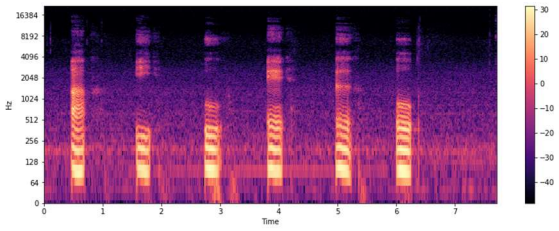


Figure 6: The Python code for converting an audio vowel signal for vowel /a/, /i/, /u/, /e/, /ê/, and /o/ to a single spectrogram image, as well as an illustration of a spectrogram image at the bottom

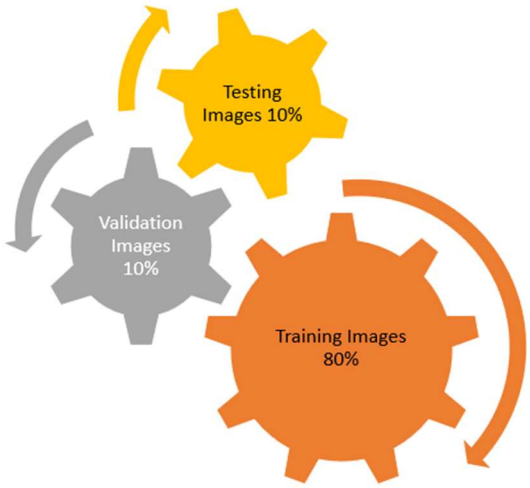
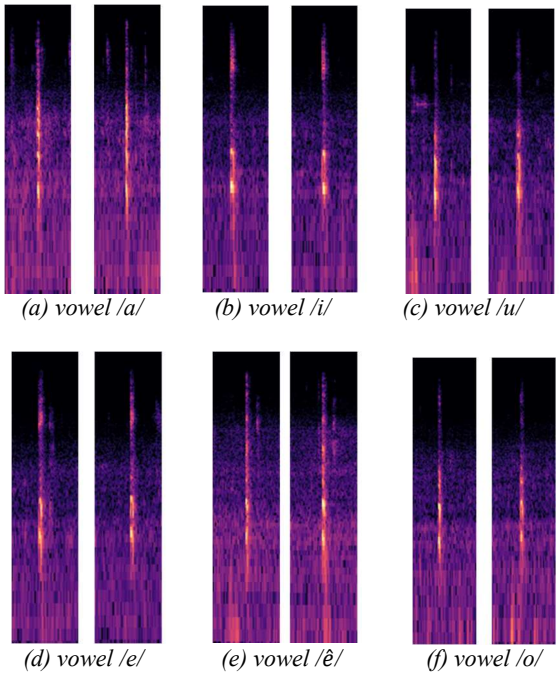


Figure 7: The proposed dataset images in percentage based on the total images

2.1 Conversion of a Wave Signal to an Image

The captured wave signal is eventually transformed to an image format known as a spectrogram image. We adjusted the y-axis to log frequency to provide a more detailed view of the spectrogram's top section. The modified spectrogram images are then manually cropped to a precise size of 240 × 55 pixels to create a single vowel image from /a/ to /o/ as in Figure 5.



(a) vowel /a/ (b) vowel /i/ (c) vowel /u/

(d) vowel /e/ (e) vowel /ê/ (f) vowel /o/

Figure 8: The example spectrogram for six vowels /a/, /i/, /u/, /e/, /ê/, and /o/ in the Malay language

2.2 Dataset Images

To gain the images for each vowel in Figure 5, we manually crop the images from Figure 6 with 240 × 55 (for the height × width). Then, the dataset is divided into three segments: training, validation, and testing images with a ratio of 8:1:1, as presented in

Figure 7. These three segments are divided for gathering and investigating the relevant network for stroke patient use. The total number for the newly introduced dataset consists of 4860 images. These introduced new dataset images were not intended to be used only for rehabilitation even to the extent of other related field.

4. EXPERIMENTAL RESULTS

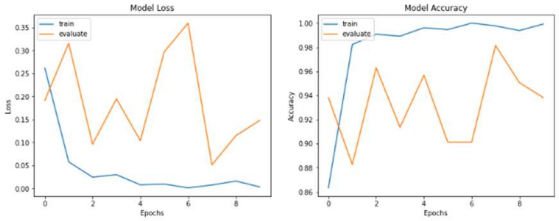
We conducted experiments based on several settings to present the model's effectiveness for recognizing the Malay language vowel for stroke patients using all vowels /a/ to /o/ with the emphasis analysis and study on the two vowels /e/ with many

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

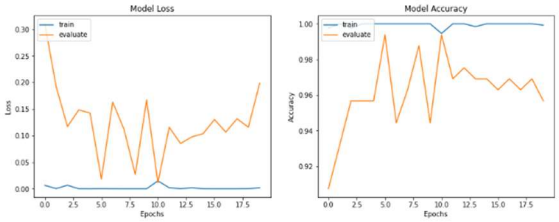
various experimental settings. We divided the analysis and experimental setting into three primary analyses within the classes' comparison approach. By looking into the two, five, and six classes, we compared the classification between the /e/ vowel in Malay with other types of the existing vowels. Various batch sizes, 3,6, and 9, and epoch sizes, 10, 20, and 50, are presented in the 2-class vowel /e/ classification study. However, in the 5- and 6-class classification studies, we set the similar batch size and epoch size to study the effectiveness during the small epoch size, which is reliable to the actual application. Comparing 2-, 5-, and 6-class vowel classification provides a comprehensive study for difference vowel /e/s in Malay. The proposed network model is compared to the comparative network models, VGG16 and VGG19, to show the classification performance. All the conducted experiments using the Google Colab by Google Inc with a free account require plenty of time to finalize the result.

A. Comparison analysis among the types of /e/ vowel (2 vowel classes)

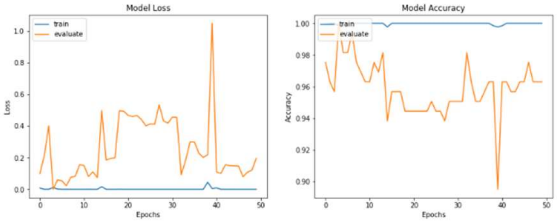
Since vowel /e/ in Malay is categorized into two types, in this study, the 2-class classification is conducted to deeply compare them individually using three types of batch size, four types of epoch size, and three types of network models. We provided the detail of the experiment result for all these settings. Using the proposed network model, in Figure 9 the classification accuracy for batch size = 3 is at 93.83%, 95.68%, 96.30%, and 95.68% for each epoch sizes.



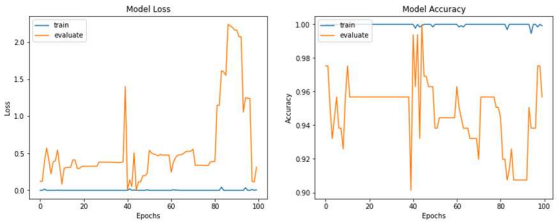
(a) Epoch size = 10



(b) Epoch size = 20

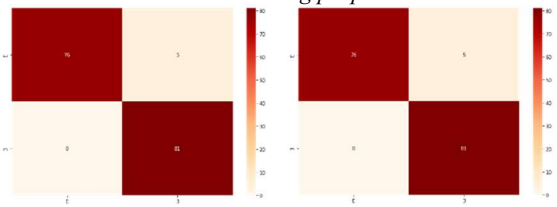


(c) Epoch size = 50

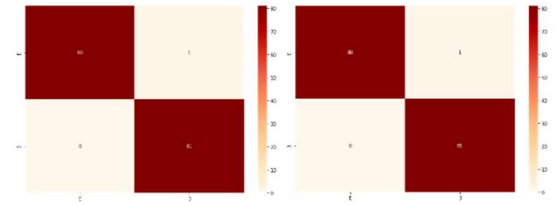


(d) Epoch size = 100

Figure 9: Model Loss and Accuracy performance when the batch size 3 is selected using proposed network model



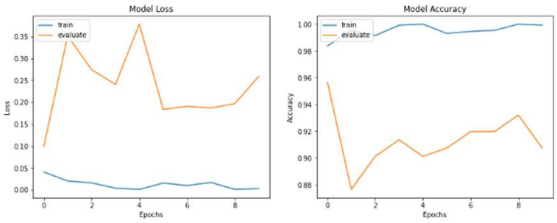
(a) Epoch size = 10 *(b) Epoch size = 20*



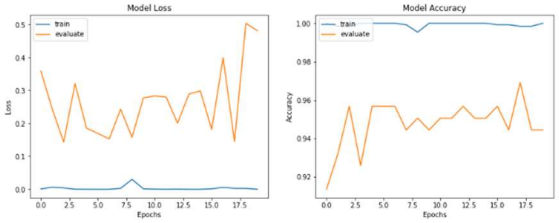
(c) Epoch size = 50 *(d) Epoch size = 100*

Figure 10: Confusion Matrix when the batch size 3 is selected using proposed network model

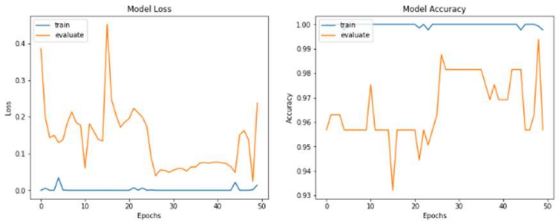
Figure 10 shows the 2-class classification managed to classify vowel /e/s with the batch size = 3. Via this confusion matrix, we understand that when epoch size = 10 and 20, similar results are attained in the testing with 93.83%. The vowel /ê/ performs better than the vowel /e/for the all-epoch sizes.



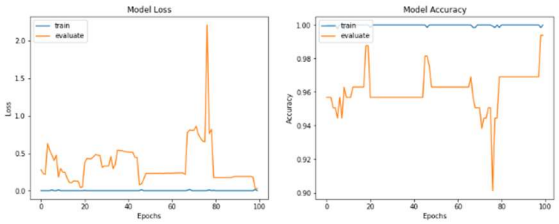
(a) Epoch size = 10



(b) Epoch size = 20



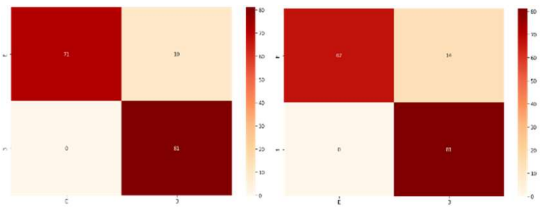
(c) Epoch size = 50



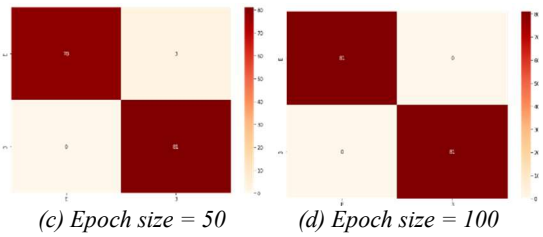
(d) Epoch size = 100

Figure 11: Model Loss and Accuracy performance when the batch size 6 is selected using proposed network model

In Figure 11 the classification accuracy for batch size = 6 is at 90.74%, 94.44%, 95.68%, and 99.38% for each epoch sizes.



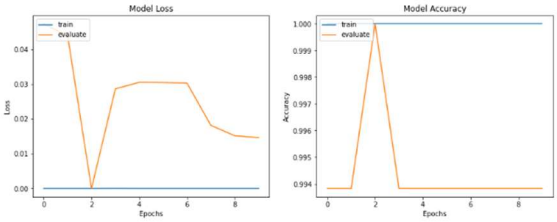
(a) Epoch size = 10 *(b) Epoch size = 20*



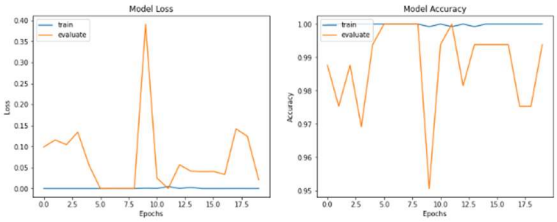
(c) Epoch size = 50 *(d) Epoch size = 100*

Figure 12: Confusion Matrix when the batch size 6 is selected using proposed network model

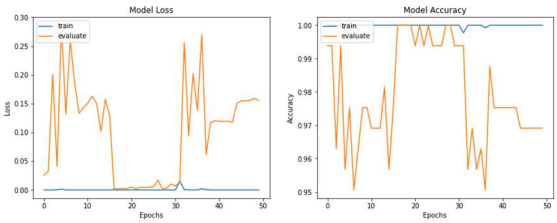
Figure 12 shows the 2-class classification managed to classify vowel /e/s with the batch size = 6. Via this confusion matrix, we understand that epoch size = 10, 20, and 50 infers a difference in the vowel /e/ classification. For the all-epoch sizes, the vowel /ê/ performs better than the vowel /e/.



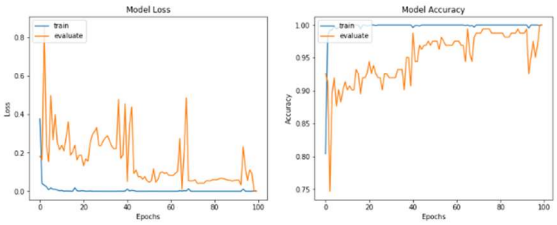
(a) Epoch size = 10



(b) Epoch size = 20



(c) Epoch size = 50

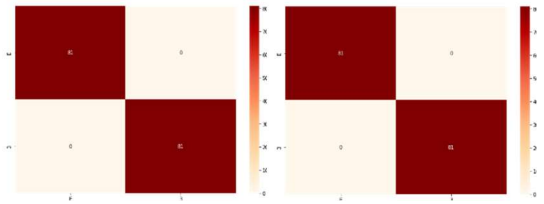


(d) Epoch size = 100

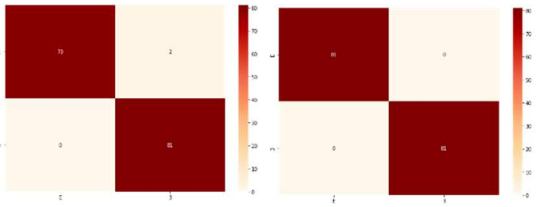
Figure 13: Model Loss and Accuracy performance when the batch size 9 is selected using proposed network model

In Figure 13 the classification accuracy for batch size = 9 is at 99.38%, 99.38%, 96.91%, and 99.38% for each epoch sizes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) Epoch size = 10 (b) Epoch size = 20

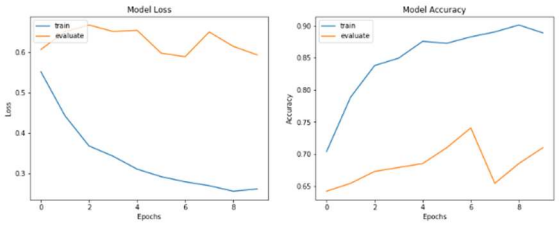


(c) Epoch size = 50 (d) Epoch size = 100

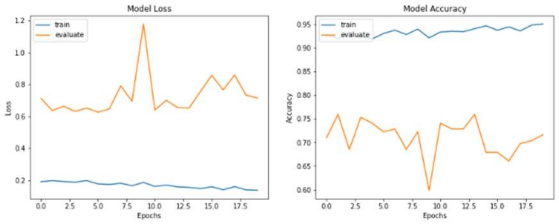
Figure 14: Confusion Matrix when the batch size 9 is selected using proposed network model

Figure 14 shows the 2-class classification managed to classify vowel /e/s with the batch size = 9. Via this confusion matrix, we understand that all epoch sizes except epoch size = 50 provided good performance for the vowel /e/ classification in testing. The vowel /ê/ shows a good performance in overall epoch sizes for the all-epoch sizes.

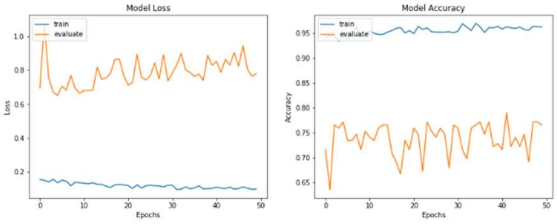
Using the VGG16 network model as the first comparative method, in Figure 15, the classification accuracy for batch size = 3 is at 70.99%, 71.60%, 76.54%, and 74.69% for each epoch size.



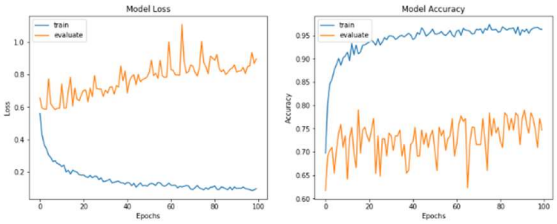
(a) Epoch size = 10



(b) Epoch size = 20

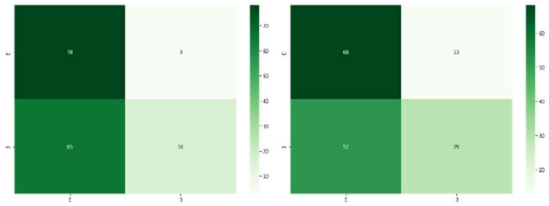


(c) Epoch size = 50

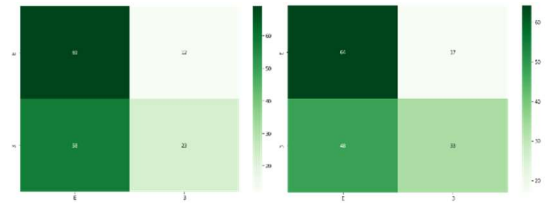


(d) Epoch size = 100

Figure 15: Model Loss and Accuracy performance when the batch size 3 is selected using VGG16 network model



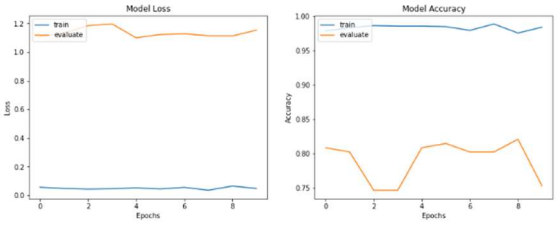
(a) Epoch size = 10 (b) Epoch size = 20



(c) Epoch size = 50 (d) Epoch size = 100

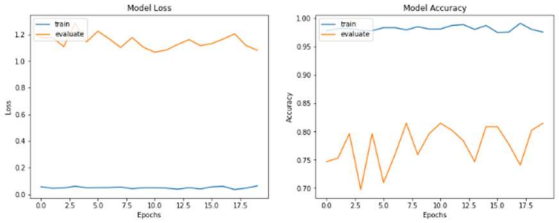
Figure 16: Confusion Matrix when the batch size 3 is selected using VGG16 network model

Figure 16 shows the 2-class classification for the true two vowel /e/s the exact prediction as vowel /e/ in all epoch sizes. The vowel /ê/ presents less performance than vowel /e/for the all-epoch sizes.

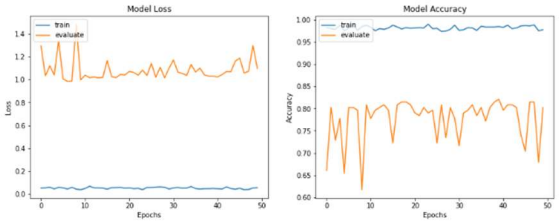


(a) Epoch size = 10

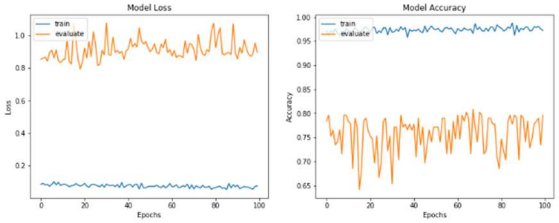
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(b) Epoch size = 20



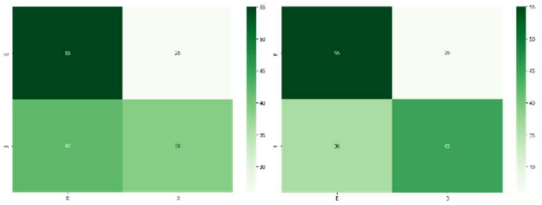
(c) Epoch size = 50



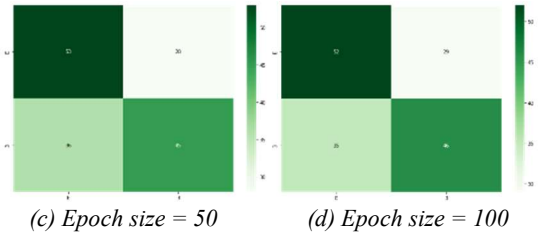
(d) Epoch size = 100

Figure 17: Model Loss and Accuracy performance when the batch size 6 is selected using VGG16 network model

In Figure 17 the classification accuracy for batch size = 6 is at 75.31%, 81.48%, 80.25%, and 79.63% for each epoch sizes.



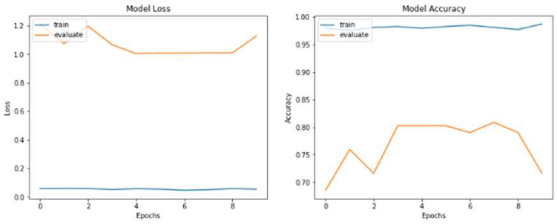
(a) Epoch size = 10 *(b) Epoch size = 20*



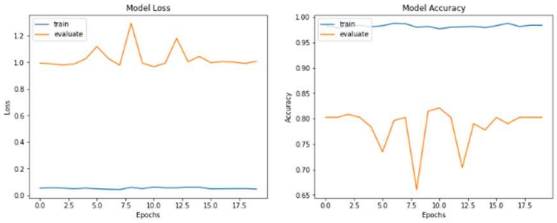
(c) Epoch size = 50 *(d) Epoch size = 100*

Figure 18: Confusion Matrix when the batch size 6 is selected using VGG16 network model

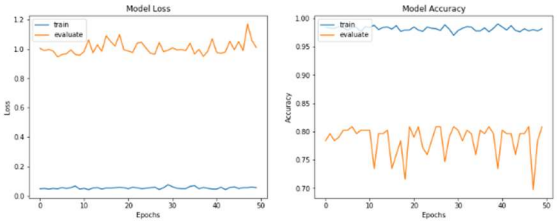
Figure 18 shows the 2-class classification shows the classification for the true two vowel /e/ and /ê/ are slightly same prediction in epoch size = 20, 50, and 100. For the epoch size = 10, the vowel /ê/ presents less performance than vowel /e/.



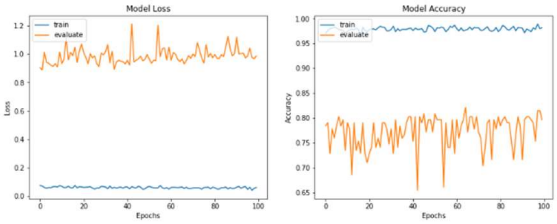
(a) Epoch size = 10



(b) Epoch size = 20



(c) Epoch size = 50

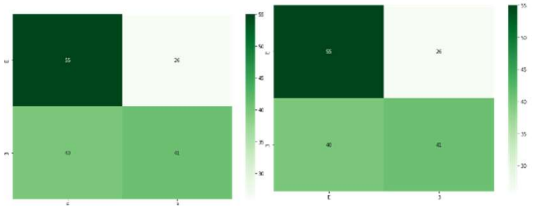


(d) Epoch size = 100

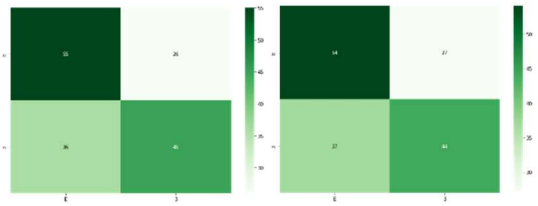
Figure 19: Model Loss and Accuracy performance when the batch size 9 is selected using VGG16 network model

In Figure 19 the classification accuracy for batch size = 9 is at 71.60%, 80.25%, 80.86%, and 79.63% for each epoch sizes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) Epoch size = 10 (b) Epoch size = 20

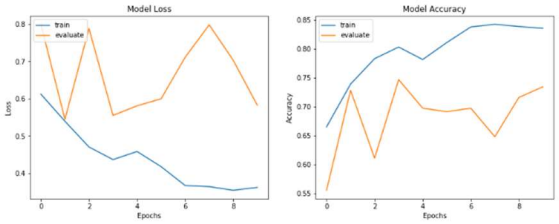


(c) Epoch size = 50 (d) Epoch size = 100

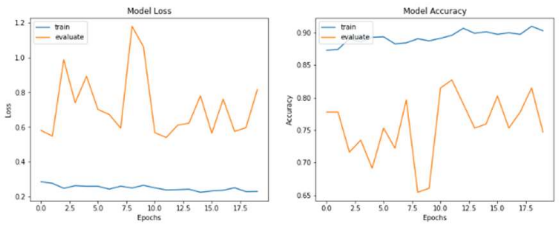
Figure 20: Confusion Matrix when the batch size 9 is selected using VGG16 network model

Figure 20 shows the 2-class classification shows the classification for the true two vowel /e/ and /ê/ are slightly same prediction in epoch size = 50 and 100. For the epoch size = 10 and 20, the vowel /ê/ presents less performance than vowel /e/.

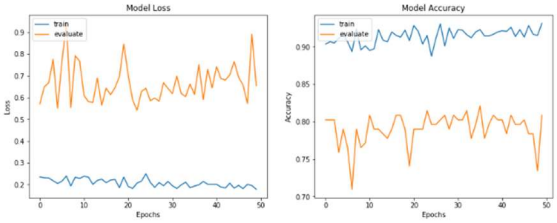
Using the VGG19 network model as the first comparative method, in Figure 21, the classification accuracy for batch size = 3 is at 73.46%, 74.69%, 80.31%, and 75.31% for each epoch size.



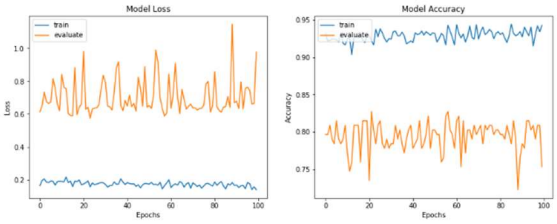
(a) Epoch size = 10



(b) Epoch size = 20

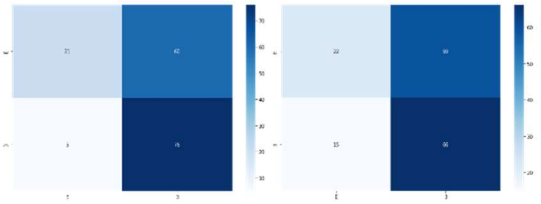


(c) Epoch size = 50

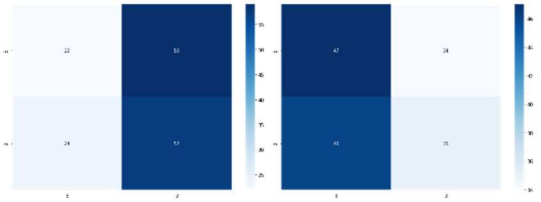


(d) Epoch size = 100

Figure 21: Model Loss and Accuracy performance when the batch size 3 is selected using VGG19 network model



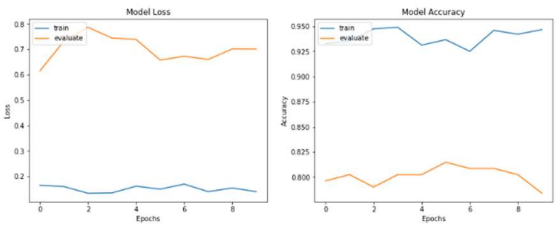
(a) Epoch size = 10 (b) Epoch size = 20



(c) Epoch size = 50 (d) Epoch size = 100

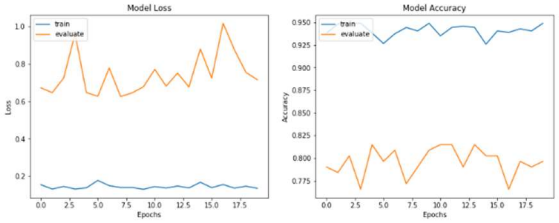
Figure 22: Confusion Matrix when the batch size 3 is selected using VGG19 network model

Figure 22 shows the 2-class classification shows the classification for the true two vowel /e/ and /ê/ are oppositely predicted for epoch size = 10 and 20 as vowel /ê/, and 50 and 100. For the epoch size = 10 and 20 as vowel /e/.

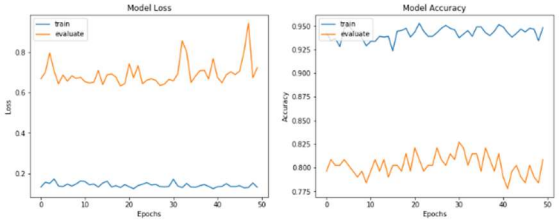


(a) Epoch size = 10

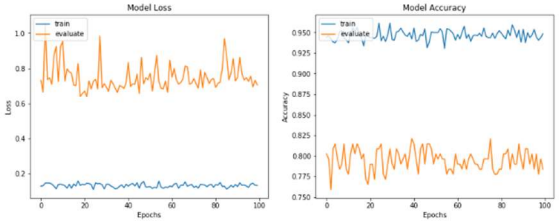
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(b) Epoch size = 20



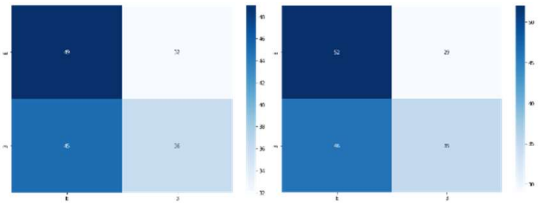
(c) Epoch size = 50



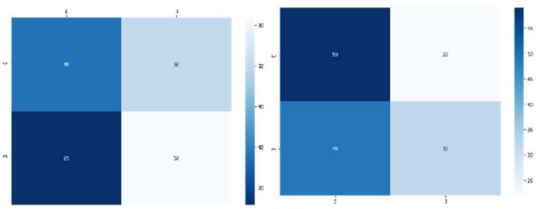
(d) Epoch size = 100

Figure 23: Model Loss and Accuracy performance when the batch size 6 is selected using VGG19 network model

In Figure 23 the classification accuracy for batch size = 6 is at 78.40%, 79.63%, 80.86%, and 78.40% for each epoch sizes.



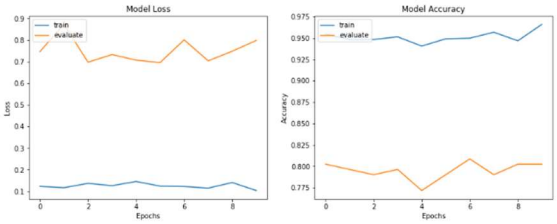
(a) Epoch size = 10 *(b) Epoch size = 20*



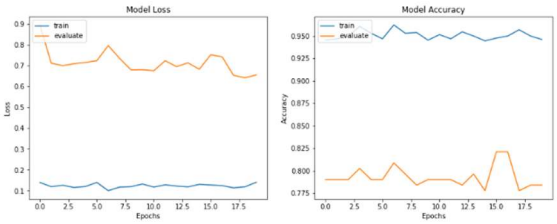
(c) Epoch size = 50 *(d) Epoch size = 100*

Figure 24: Confusion Matrix when the batch size 6 is selected using VGG19 network model

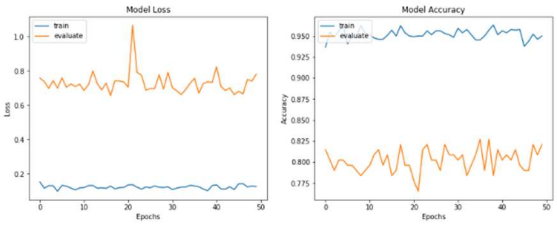
Figure 24 shows the 2-class classification shows the classification for both two vowel /e/ and /ê/ are oppositely predicted as vowel /e/ for all epoch sizes.



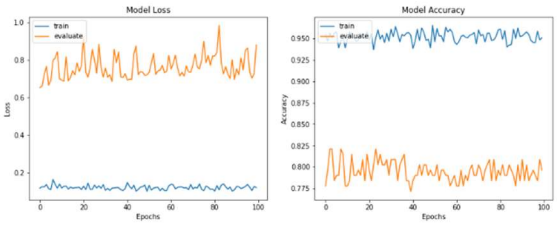
(a) Epoch size = 10



(b) Epoch size = 20



(c) Epoch size = 50

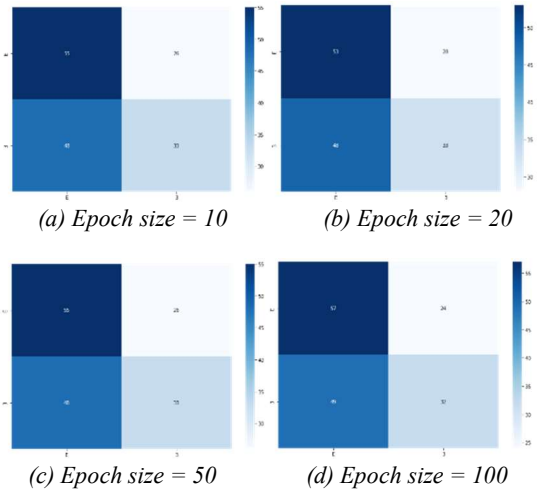


(d) Epoch size = 100

Figure 25: Model Loss and Accuracy performance when the batch size 9 is selected using VGG19 network model

In Figure 25 the classification accuracy for batch size = 9 is at 80.25%, 78.40%, 82.10%, and 78.40% for each epoch sizes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) Epoch size = 10 *(b) Epoch size = 20*

(c) Epoch size = 50 *(d) Epoch size = 100*

Figure 26: Confusion Matrix when the batch size 9 is selected using VGG19 network model

Figure 26 shows the 2-class classification shows the classification for both two vowel /e/ and /ê/ are oppositely predicted as vowel /e/ not for vowel /ê/ for all epoch sizes.

Table 1: Precision, Recall for the /e/ vowel classes, and the test accuracy (epoch size 10)

Batch Size		3		6		9	
Net.	Vow.	Prec	Rec	Prec	Rec	Prec	Rec
Acc.		**0.96**		**0.94**		**1.00**	
Pro	/e/	1.00	0.91	1.00	0.88	1.00	1.00
Pro	/ê/	0.92	1.00	0.89	1.00	1.00	1.00
Acc.		0.58		0.58		0.59	
VGG 16	/e/	0.55	0.96	0.57	0.68	0.58	0.68
VGG 16	/ê/	0.84	0.20	0.60	0.48	0.61	0.51
Acc.		0.58		0.58		0.59	
VGG 19	/e/	0.81	0.26	0.52	0.60	0.53	0.68
VGG 19	/ê/	0.56	0.94	0.53	0.44	0.56	0.41

Table 2: Precision, Recall for the /e/ vowel classes, and the test accuracy (epoch size 20)

Batch Size		3		6		9	
Net.	Vow.	Prec	Rec	Prec	Rec	Prec	Rec
Acc.		**0.97**		**0.91**		**1.00**	
Pro	/e/	1.00	0.94	1.00	0.83	1.00	1.00
Pro	/ê/	0.94	1.00	0.85	1.00	1.00	1.00
Acc.		0.60		0.62		0.59	
VGG 16	/e/	0.57	0.84	0.60	0.68	0.58	0.68
VGG 16	/ê/	0.69	0.36	0.63	0.56	0.61	0.51
Acc.		0.60		0.62		0.59	
VGG 19	/e/	0.59	0.27	0.53	0.64	0.53	0.68
VGG 19	/ê/	0.53	0.81	0.55	0.43	0.56	0.41

From the batch sizes and epoch sizes based on the several network models, both vowels/e/ and /ê/ are classified correctly using the proposed network. Using the VGG16 showed the prediction result for vowel /e/ for both vowel /e/ and /ê/ with batch size 3. In other batch sizes, 6 and 9, the classification

Table 3: Precision, Recall for the /e/ vowel classes, and the test accuracy (epoch size 50)

Batch Size		3		6		9	
Net.	Vow.	Prec	Rec	Prec	Rec	Prec	Rec
Acc.		**0.99**		**0.98**		**0.99**	
Pro	/e/	1.00	0.99	1.0	0.96	1.00	0.98
Pro	/ê/	0.99	1.00	0.96	1.00	0.95	1.00
Acc.		0.57		0.60		0.62	
VGG 16	/e/	0.54	0.85	0.60	0.65	0.60	0.68
VGG 16	/ê/	0.66	0.28	0.62	0.56	0.63	0.56
Acc.		0.57		0.60		0.62	
VGG 19	/e/	0.48	0.27	0.53	0.64	0.52	0.65
VGG 19	/ê/	0.49	0.70	0.55	0.43	0.54	0.41

Table 4: Precision, Recall for the /e/ vowel classes, and the test accuracy (epoch size 100)

Batch Size		3		6		9	
Net.	Vow.	Prec	Rec	Prec	Rec	Prec	Rec
Acc.		**0.99**		**0.98**		**1.00**	
Pro	/e/	1.00	0.99	1.0	0.96	1.00	1.00
Pro	/ê/	0.99	1.00	0.96	1.00	1.00	1.00
Acc.		0.60		0.60		0.60	
VGG 16	/e/	0.57	0.79	0.60	0.64	0.59	0.67
VGG 16	/ê/	0.66	0.41	0.61	0.57	0.62	0.54
Acc.		0.60		0.60		0.60	
VGG 19	/e/	0.51	0.58	0.55	0.73	0.54	0.73
VGG 19	/ê/	0.51	0.43	0.59	0.40	0.57	0.40

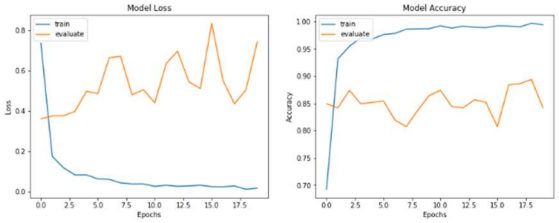
performance is comparable to another network model. Using VGG19 at batch size 3 shows the prediction result as vowel /ê/ for both vowels. However, when the batch size is enlarged, most of the images' predicted class is vowel /e/. This shows unstable classification for the VGG16 and VGG19 compared to the proposed network model. The epoch size analysis shows that the epoch size 20 is suitable for most of the experimental cases. Therefore, we set epoch 20 for the upcoming analysis.

Table 1 – 4 can understand the analysis by changing the epoch sizes into 10, 20, 50 and 100. Using the CNN as the basis of the network, the proposed network model outperformed other

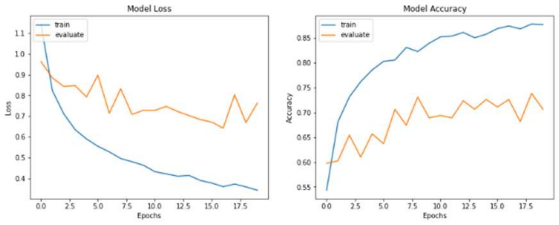
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

network models VGG16 and VGG16 in all settings, batch sizes, and epoch sizes. This experiment shows that all the test accuracy for the proposed method is reliable in classifying 2-class vowel /e/ in Malay.

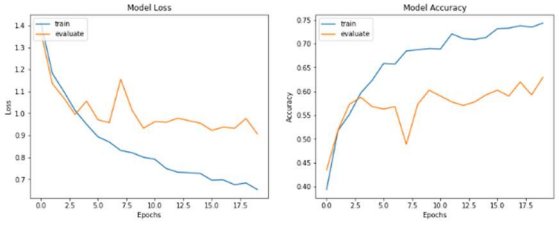
B. Comparison analysis between vowel /e/ and other vowels (5 vowel classes)



(a) Proposed Network

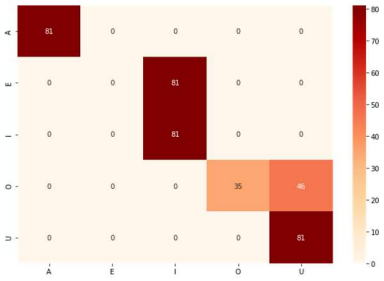


(b) VGG16

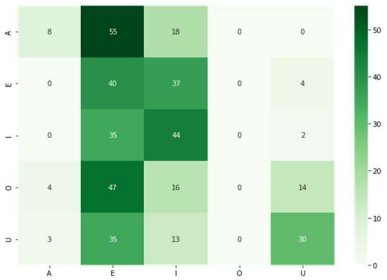


(c) VGG19

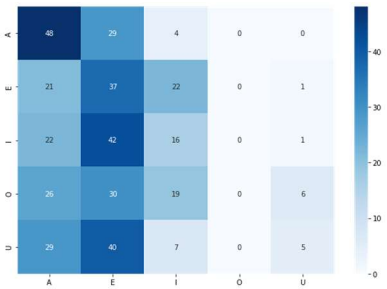
Figure 27: Model Loss and Accuracy performance for the /e/ vowel with /a/, /i/, /u/ and /o/ classes, when the batch size 4and epoch size 20



(a) proposed network



(b) VGG16



(c) VGG19

Figure 28: Confusion Matrix when the batch size 4 and epoch size 20, /e/ vowel with /a/, /i/, /u/ and /o/ classes

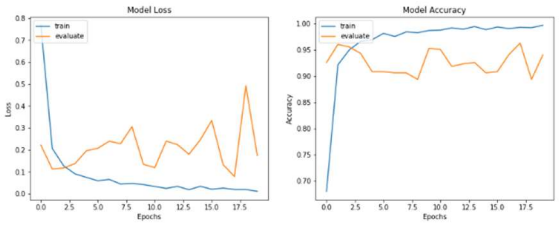
Table 5: Precision, Recall for the /e/ vowel with /a/, /i/, /u/ and /o/ classes, and the test accuracy (When the batch size 4 and epoch size 20)

Network	Vowel	Prec	Rec
Accuracy		**0.69**	
Proposed Network	/a/	1.00	1.00
	/e/	0.00	0.00
	/i/	0.00	1.00
	/u/	1.00	0.43
	/o/	0.64	1.00
Accuracy		0.30	
VGG16	/a/	0.53	0.10
	/e/	0.19	0.49
	/i/	0.34	0.54
	/u/	0.00	0.00
	/o/	0.60	0.46
Accuracy		0.26	
VGG19	/a/	0.33	0.59
	/e/	0.21	0.46
	/i/	0.24	0.20
	/u/	0.00	0.00
	/o/	0.38	0.06

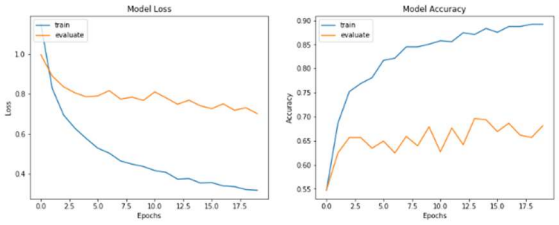
Figure 27 and Table 5 show that the proposed network model outperforms all the comparative network models with 0.69 for testing accuracy. Even though the results still presented a good performance classification, the vowel /e/ is not classified correctly by the proposed network model compared to other VGG16 and VGG19 in Figure 28. Vowel /a/, /i/, and

/o/ are correctly predicted by the proposed model. Vowel /o/ presented the test prediction as /o/ and /u/ which reflected to the class number 4 (vowel /o/) as 0.43 in percentage. For this vowel /o/, on the other hand for VGG16 and VGG19 were disabled to classify this vowel correctly.

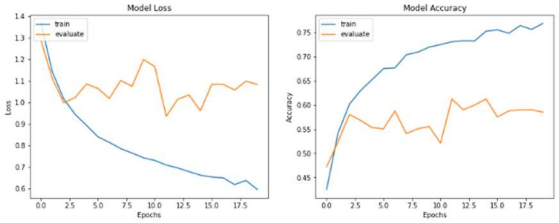
C. Comparison analysis between /ê/ vowel and other vowels (5 vowel classes)



(a) Proposed Network

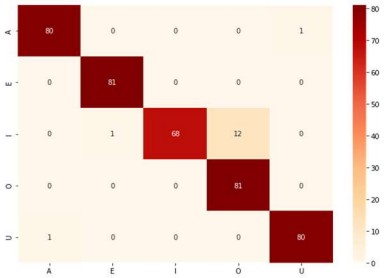


(b) VGG16

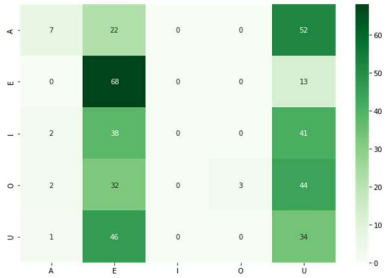


(c) VGG19

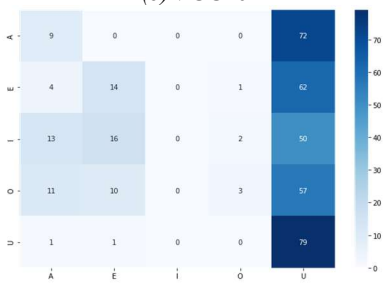
Figure 29: Model Loss and Accuracy performance for the /ê/ vowel with /a/, /i/, /u/ and /o/ classes when the batch size 4 and epoch size 20



(a) proposed network



(b) VGG16



(c) VGG19

Figure 30: Confusion Matrix when the batch size 4 and epoch size 20, /e/ vowel with /a/, /i/, /u/ and /o/ classes

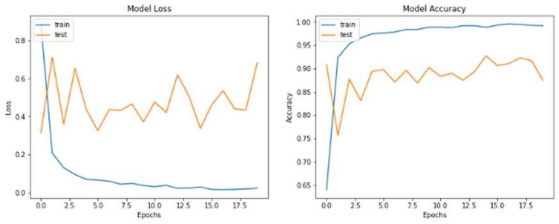
Table 6: Precision, Recall for the /ê/ vowel with /a/, /i/, /u/ and /o/ classes, and the test accuracy (When the batch size 4 and epoch size 20)

Network	Vowel	Prec	Rec
Accuracy		**0.96**	
Proposed Network	/a/	0.99	0.99
	/e/	0.99	1.00
	/i/	1.00	0.84
	/u/	0.87	1.00
	/o/	0.99	0.99
Accuracy		0.28	
VGG16	/a/	0.58	0.09
	/e/	0.33	0.84
	/i/	0.00	0.00
	/u/	1.00	0.04
	/o/	0.18	0.42
Accuracy		0.26	
VGG19	/a/	0.24	0.11
	/e/	0.34	0.17
	/i/	0.00	0.00
	/u/	0.50	0.04
	/o/	0.25	0.98

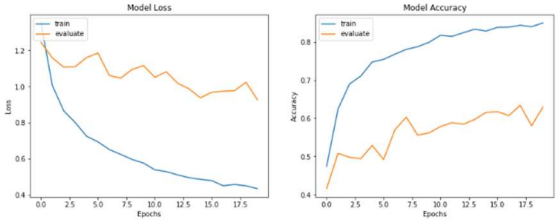
Figure 29 and Table 6 show that the proposed network model outperforms all the comparative network models with 0.96 for testing accuracy. This performance shows that the proposed network model successfully differentiates the vowel /ê/ correctly compared to the vowel /e/ in the previous experiment in Section 3 B. The vowels in Malay are classified correctly by the proposed network model compared to other VGG16 and VGG19 in Figure 30. However, the proposed network model shows a 0.84

percentage for vowel /i/'s test accuracy. Vowel /e/ and /o/ are correctly predicted by the proposed model. Vowel /a/ and /o/ also presented a good classification in the test prediction. For this vowel /a/, /i/ /o/, on the other hand for VGG16 and vowel /a/, /e/, /i/ /o/ VGG19 were disabled to classify this vowel correctly.

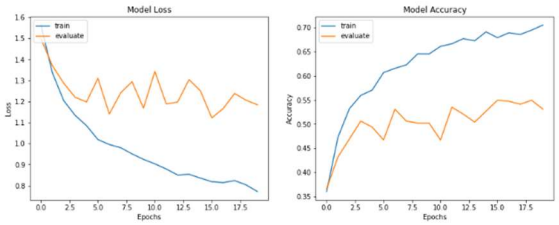
D. Comparison analysis using various of batch size (6 vowel classes)



(a) Proposed Network

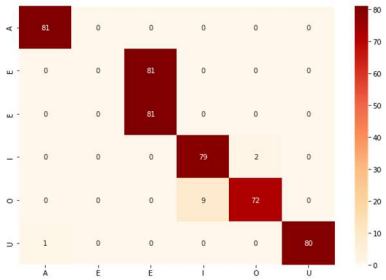


(b) VGG16

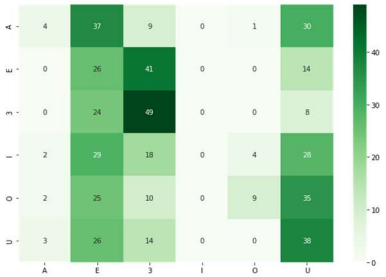


(c) VGG19

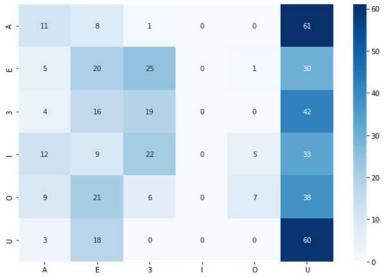
Figure 31: Model Loss and Accuracy performance for all Malay vowels when the batch size 4 and epoch size 20



(a) proposed network



(b) VGG16



(c) VGG19

Figure 32: Confusion Matrix when the batch size 4 and epoch size 20, /a/, /e/, /ê/, /i/, /u/, and /o/ classes

Table 7: Precision, Recall for all Malay vowels and the test accuracy (When the batch size 4 and epoch size 20)

Network	Vowel	Prec	Rec
Accuracy		**0.81**	
Proposed Network	/a/	0.99	1.00
	/e/	0.00	0.00
	/ê/	0.50	1.00
	/i/	0.90	0.98
	/u/	0.97	0.89
	/o/	1.00	0.99
Accuracy		0.26	
VGG16	/a/	0.36	0.05
	/e/	0.13	0.32
	/ê/	0.35	0.60
	/i/	0.00	0.00
	/u/	0.64	0.11
	/o/	0.25	0.47
Accuracy		0.24	
VGG19	/a/	0.25	0.15
	/e/	0.22	0.25
	/ê/	0.26	0.23
	/i/	0.00	0.00
	/u/	0.54	0.09
	/o/	0.23	0.74

Figure 31 and Table 7 show that the proposed network model outperforms all the comparative network models with 0.81 for testing accuracy. This performance shows that the proposed network model successfully differentiates the vowel /ê/ correctly from the vowel /e/ in the previous experiment in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Sections 3 A and B. The vowels in Malay are classified correctly by the proposed network model compared to other VGG16 and VGG19 in Figure 32. However, the proposed network model shows a null percentage for vowel /e/'s test accuracy. Vowel /a/, /e/, and /o/ are nearly correctly predicted by the proposed model. Vowel /i/ and /o/ also still presented a good classification in the test prediction. For this vowel /a/, /i/ /o/, on the other hand for VGG16 and vowel /a/, /e/, /i/ /o/ VGG19 were disabled to classify this vowel correctly.

Based on the four experiments conducted utilizing four types of class numbers, vowel /ê/ provided an excellent analysis result via the proposed network model compared to the vowel /e/. Within the 2-class analysis, vowel /e/ and /ê/, the evaluation and testing accuracy are highly accurate. The study with 2-class classification is conducted to find the suitable batch size and epoch size for the details analysis with 5-class and 6-class. We understand that the epoch size here is reliable when this number is set as 20. Overall, the proposed network model performs better than VGG and VGG19 in the testing phase by comparing it with confusion matrix 2 x 2 and the precision and recall value for both values.

Analysis for 5-class vowels, which for each /e/ and /ê/ with other vowels in Malay, shows the variety of results depending on the three network models were used in the experiment. In this experimental setting, the differentiation of classification effectiveness of these two vowels could be described and understood. Vowel /ê/ shows that it could be classified successfully if the vowel /e/ is not available in the classification task. Here, it also means that vowel /ê/ is performed well when the vowel /e/ is not classified together. On the other hand, the vowel /e/during the experimental analysis shows that the prediction is not successfully ensued and is even most likely null. The proposed network model in the paper has the limitation in comparing the differentiation between the two vowels/e/ in Malay. Conceivably, the proposed network model's structure and setting were looking to be relooked and redesigned to ensure the classification is done successfully. Since the convolution layer in the CNN used in the proposed layer is still considered a basic one, to distinguish the vowel /e/, which looks similar in image behavior, deeper and more layers could be helpful to classify both vowels.

Nevertheless, in the 6-class of vowel classification, the performance of vowel /e/ is still struggling to achieve a good prediction. In overall testing prediction percentage, the number still considered good as the average of the testing performance managed to gain 81%. This is not enough to analyze the vowel /e/deeper on an excellent classifier in the actual application. Limitation on the spectrogram images could provide difficulties in classifying these two vowels, especially the difference between the two in image knowledge at the higher frequency part. The less distinction from the images could be the less testing accuracy percentage in experiments conducted in this paper. The more significant number of images provided in the training phase could also be considered in the new evaluation for looking for a better classifier for these six vowels in Malay.

4. CONCLUSION AND FUTURE WORKS

We presented four Malay language analyses for the vowel /e/using the proposed network with two comparative network models, VGG and VGG19. The paper contributes an overall analysis on batch size, epoch sizes, classes variety and various network models for giving a good understanding of the differentiation of vowel /e/s. We trained, evaluated, and tested the images for the six Malay language vowels with the proposed dataset images by utilizing a simple CNN model. Overall, in the four analyses, we understand that all vowels in Malay were classified successfully in an average manner. It is still difficult to classify the vowel /e/ when classing the vowel together. Therefore, we are looking to study the effectiveness of changing the sound to another image format, such as scalogram and STFT images. This could be helping us to outlook the effectiveness of having another image format in the analysis. Perhaps, the more complex network model also could be designed and tested in an overall and versatile experimental setting in the future. The vowel could also be implemented only as a raw audio file than changing the audio o image format.

ACKNOWLEDGEMENT

The authors would like to thank Perkeso Rehab Center, Melaka for the Research Project Collaboration, and Universiti Teknikal Malaysia Melaka for the financial support through PJP/2020/FKEKK/PP/S01792.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] Indirawati Z. and Mardian S. O., Phonetics and phonology in Malay. Kuala Lumpur: PTS Professional, 2006.

[2] Farid M. O. Aspects of Malay phonology and morphology: A generative approach. Universiti Kebangsaan Malaysia, 1980.

[3] Mark D., Malay as a mirror of austronesian: Voice development and voice variation," *Lingua*, 118-10, 2008, pp1470–1499.

[4] Peter C., Gabriella H., Yanti, Voice in Malay/Indonesian," *Lingua*, 118-10, 2008, pp 1500–1553.

[5] Husni T., Gunawan A., Irma Y., Wawan J. P. , Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia, *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 17-5, 2019, pp 2321-2326.

[6] Hua-Nong T., Alireza Z., See-Yan C., Boon-Fei Y., Badrulzaman A. H., Formant frequencies of Malay vowels produced by Malay children aged between 7 and 12 years, *Journal of Voice*, 26-5,2012, pp 664.e1-664.e6.

[7] Karen S. and Andrew Z., Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2015.

[8] N. S. M. Zamri, N. M. Z. Hashim, A. S. Ja'afar, A. M. Darsono, M. J. A. Latif and P. Rajaandra, A Preliminary Study on Vowel Recognition via CNN for Disorder People in Malay Language, 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2021, pp. 312-316, doi: 10.1109/ISMSIT52890.2021.9604589.

[9] Charles V. R. and Robert L. E, *Speech correction : an introduction to speech pathology and audiology*, 9th Edition. p. cm. Needham Heights, MA: A Simon Schuster Company, 1995.

[10] Sara R. and Patti S., *Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program*, Washington (DC): National Academies Press (US), 2016.

[11] Md Z. Z., Jin Y. K., Hyoung-Gook K., and Seung Y. N., A preliminary study on deep-learning based screaming sound detection, *5th Int. Conf. IT Converg. Secur. ICITCS*. 2015.

[12] Cristhian P., Saman P., Asif R., Bryan C., Ensemble of Feature:based and Deep learning:based Classifiers for Detection of Abnormal Heart Sounds, *Comput. Cardiol. Conf*. 43, 2017, pp 621–624.

[13] Juncheng L., Wei D., Florian M., Shuhui Q., and Samarjit D., A Comparison Of Deep Learning Methods For Environmental Sound Detection, *IEEE Int. Conf. Acoust. Speech, Signal Process,* 2017, pp 126–130.

[14] Rahmawati R. and Dessi L., Java and Sunda dialect recognition from Indonesian speech using GMM and I-Vector, *Proceeding 11th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA.* 2018-Janua, 2017, pp 1–5.

[15] Pronaya P. Das, Shaikh M. A., Ruhul A., Zahida R., Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model, *Proc. 8th Int. Conf. Adv. Comput. Intell. ICACI*. 2016, pp 359–364.

[16] Jacqueline I. and Dessi L., Classification and clustering to identify spoken dialects in Indonesian, *Proc. 2017 Int. Conf. Data Softw. Eng. ICoDSE 2017*, 2018-Janua, 2017, pp 1–6.

[17] Du G., Wang X., Wang G., D. G. Yan, Xia W., Guangyan W,Yan Z. and Dan L. Speech recognition based on convolutional neural networks, *2016 IEEE Int. Conf. Signal Image Process. ICSIP 2016*, 2016, pp 708–711.

[18] Khaing Y. H. P. H., Min N. Z., Development of control system for fruit classification based on convolutional neural network, *Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018,* 2018-January-10, 2018, pp 1805–1807.

[19] Chamberlain D., Kodgule R., Ganelin D., Miglani V., and Einstein A., Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports, *Molecular Psychiatry*, 21-10, 2016, pp 1366–1371.