INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 14, Issue 10, (2024) E-ISSN: 2222-6990

Decode Malay Syllables: CNN's Key to Malay Language Understanding

Haleeda Norelham Harun[1], Nik Mohd Zarifie Hashim[1], Nursyahmina Ahmad Azhar[1], Azahari Salleh[1], Mahmud Dwi Sulistiyo[2], Afiqah Ilya Kamarudin[3]

[1]Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Durian Tunggal, Melaka, Malaysia, [2]School of Computing, Telkom University, Bandung, Indonesia, [3]Pusat Rehabilitasi PERKESO, Bandar Hijau, Hang Tuah Jaya, Melaka, Malaysia

Corresponding Author Email: nikzarifie@utem.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v14-i10/23118 DOI:10.6007/IJARBSS/v14-i10/23118

Published Date: 05 October 2024

Abstract

Malay serves as the language of knowledge and is instrumental in educational settings, as exemplified by its significance in the Education Act of 1961. Furthermore, to support the government's initiatives in advancing the quality of education toward achieving global standards, there is a growing demand for innovative pedagogical methods in teaching the Malay language. In this context, numerous researchers have concentrated their efforts on developing speaker-independent systems, which find applications in language training, articulation therapy, and aiding language learners in mastering the intricacies of Malay phonetics, particularly focusing on vowels. Hence, the principal aim of this paper is primarily dedicated to the recognition of intelligently pronounced Malay syllables by distinct male and female groups, employing Neural Network technology. The primary objective of this research paper is to develop a robust system for the accurate recognition of Malay language syllables, which play a pivotal role in the context of the Malay language, widely used as the primary medium of communication in Malaysia. The implementation of this system leverages the Python programming language, known for its versatility and adaptability to various applications. The paper's primary focus lies in the careful observation and analysis of specific syllable components, particularly those involving the pronunciation of /a/, /e/, /i/, /o/, and /u/. These segments of the language pose particular challenges in terms of pronunciation, and the paper seeks to develop a comprehensive solution for their accurate recognition.

Keywords: Audio Signal Processing, Convolutional Neural Networks (CNN), Natural Language Processing (NLP), Speech Recognition, and Spoken Keyword Spotting

Introduction

Sound processing and AI have become increasingly important in recent years, with the development of new technologies and techniques that enable machines to understand and interpret sound signals. The use of artificial intelligence (AI) in audio processing has shown great promise in various applications, such as speech recognition, music analysis, and sound detection. AI has the potential to revolutionize the field of audio processing by enabling machines to understand and interpret sound signals in a way that was previously impossible. With the help of deep learning algorithms, machines can now recognize patterns in sound data and make predictions based on that data.

The use of AI in audio processing has led to the development of new tools and techniques for analyzing and manipulating sound signals. For example, AI-powered audio mixing tools can detect instruments and suggest fitting presets to the user, while AI-based sound recognition solutions can extract insights from speech, voices, snoring, music, industrial and traffic noise, and other types of acoustic signals. One of the key advantages of AI in sound processing is its ability to handle large amounts of data and learn from that data over time. This makes it possible to develop more accurate and efficient models for recognizing and classifying sound signals, which can be applied to various applications such as speech recognition, language learning, and natural language processing.

However, there are also challenges and limitations to the use of AI in sound processing. For example, the quality of the data used to train the AI models can have a significant impact on their accuracy and performance. In addition, there are ethical, and privacy concerns related to the use of AI in audio processing, such as the potential for misuse or abuse of the technology. Despite these challenges, the potential benefits of AI in sound processing are significant, and the field is likely to continue to grow and evolve in the coming years. As new technologies and techniques are developed, we can expect to see even more innovative applications of AI in audio processing, with the potential to transform the way we interact with sound signals in our daily lives.

From there, focusing on Malaysia main language, this proposed paper come in by introducing a research area that focuses on developing models and techniques for accurately identifying and classifying syllables in the Malay language. The Malay language is a syllable-friendly language, which makes it an ideal candidate for speech recognition systems. The use of Convolutional Neural Networks (CNN) in speech recognition has shown promising results in achieving high accuracy in recognizing Malay language syllables, especially vowels. Speech recognition is an essential technology that enables machines to understand human speech. The development of speech recognition systems for the Malay language is crucial for various applications, such as education, healthcare, and communication. However, the Malay language presents unique challenges for speech recognition systems due to its complex syllable structure and the presence of similar sounding vowels. The main objective of this proposed paper is to develop accurate and efficient models and techniques for identifying and classifying syllables in the Malay language. The proposed models and techniques should be able to overcome the challenges posed by the Malay language, such as recognizing similar sounding vowels and multi-syllable commands.

The research in this proposed paper involves the use of a different domain in looking into audio into another perspective with the use of spectrogram property of a recorded audio. The proposed methods aim to improve the accuracy and efficiency of Malay language syllables recognition using CNN. These papers provide a comprehensive overview of the current state of research in Malay Language syllables recognition using CNN. The proposed models and techniques have shown promising results in accurately identifying and classifying syllables in the Malay language. However, there are still challenges to be addressed, such as recognizing multi-syllable commands and differentiating between similar vowels. The proposed models and techniques have shown promising results in overcoming the challenges posed by the Malay language. However, further research is needed to develop more accurate and efficient models for Malay language syllables recognition using CNN.

Research Background and Motivation

Malay language syllables recognition using Convolutional Neural Networks (CNN) has gained significant attention in recent years. The use of CNN in speech recognition has shown promising results in achieving high accuracy in recognizing Malay language syllables, especially vowels. One study proposed a systematic approach for Malay language dialect identification using CNN (Sulaiman et al., 2021). The authors used Mel Frequency Cepstral Coefficients (MFCC) as features for Malay dialect identification. The CNN was trained on MFCC features extracted from sound files of native dialect speakers on selected vocabulary. The high accuracy across all dialects suggests that the dialects and words were classified with minimal errors. The proposed method can be useful for improving the accuracy of Malay language syllables recognition using CNN. Another study proposed a smart vowel recognition system for Malay language using CNN for people with speech disorders (Zamri et al., 2021). The authors used CNN for vowel recognition in Malay language. The proposed system achieved high accuracy in recognizing Malay language vowels. The proposed system can be useful for improving the accuracy of speech recognition for people with speech disorders. However, the paper does not specifically focus on CNN-based techniques for Malay syllables recognition.

A comprehensive review of Malay speech recognition and audio-visual speech recognition techniques was presented in a paper (Hariharan et al., 2012). The authors discussed various aspects of Malay speech recognition, including digit syllable structure, Malay speech corpus, end-point detection processing, feature extraction, and classification methods. The review can be a valuable resource for researchers and practitioners working on Malay language syllables recognition using CNN. Another paper analyzed the differences in recognizing the /e/ vowel in Malay language using CNN (Hashim et al., 2021). The authors discussed the challenges in recognizing the /e/ vowel, which is crucial as similar spelling vocabulary conveys two different meanings. The analysis provides insights into improving the accuracy of Malay language syllables recognition using CNN.

A paper proposed a method for Malay language vowel recognition using image outlook via CNN (Hashim et al., 2022). The authors discussed the challenges in recognizing Malay language vowels and proposed a novel approach using image outlook. The experimental results show that the proposed method achieves high accuracy in recognizing Malay language vowels. The proposed method can be useful for improving the accuracy of Malay language

syllables recognition, especially for vowels. However, the paper does not specifically focus on CNN-based techniques for Malay syllables recognition.

A critical analysis of speech recognition for Tamil and Malay languages through Artificial Neural Networks (ANN) was presented in a paper (Ponniah, 2021). The author discussed the syllable structure of Tamil and Malay languages and the challenges in speech recognition for these languages. The analysis can be useful for understanding the challenges and potential solutions in developing accurate speech recognition systems for these languages. However, the paper does not specifically focus on CNN-based techniques for syllables recognition. In another paper, the authors explored speech recognition performance for Malay language with multi-accents from speakers of different origins or ethnicities. The paper provides insights into the challenges in recognizing Malay language syllables for speakers of different origins or ethnicities. The analysis can be useful for developing more accurate models for Malay language syllables recognition using CNN. A paper proposed a method for Malay language syllables recognition using a hybrid model of CNN and Hidden Markov Model (HMM). The authors used MFCC as features for Malay syllables recognition. The proposed hybrid model achieved high accuracy in recognizing Malay language syllables. The proposed method can be useful for improving the accuracy of Malay language syllables recognition using CNN.

Another paper proposed a method for Malay language syllables recognition using a hybrid model of CNN and Recurrent Neural Network (RNN) (Juan et al., 2012). The authors used MFCC as features for Malay syllables recognition. The proposed hybrid model achieved high accuracy in recognizing Malay language syllables. The proposed method can be useful for improving the accuracy of Malay language syllables recognition using CNN. A paper proposed a method for Malay language syllables recognition using a hybrid model of CNN and Support Vector Machine (SVM) (Ong et al., 2011). The authors used MFCC as features for Malay syllables recognition. The proposed hybrid model achieved high accuracy in recognizing Malay language syllables. The proposed method can be useful for improving the accuracy of Malay language syllables recognition using CNN. In conclusion, the use of CNN in speech recognition has shown promising results in achieving high accuracy in recognizing Malay language syllables, especially vowels. The proposed methods in the reviewed papers can be useful for improving the accuracy of Malay language syllables recognition using CNN. However, further research is needed to develop more accurate and efficient models for Malay language syllables recognition using CNN.

Table 1 shows all the journals choose the syllables according to the stops and plosive sound for the Malay Language. The stops are the sound that flow of air which is active in creating the sound is completely blocked and it is called plosive is a common type of stop sound whereby air in the lungs is briefly blocked from flowing out through the mouth and nose, and it is also pressure that builds up behind the blockage (Nong et al., 2001, Ting et al., 2001, Ting et al., 2003). According to Ting Hua Nong, Jasmy Yunus and Sheikh Hussain Shaikh Salleh the Malay Language the list of syllables that includes stops and plosive sound contains 16 syllables such as [ba], [bi], [bu], [da], [du], [gi], [ka], [ki], [ko], [ku], [pa], [pi], [pu], [ta], [te] and [ti]. These 16 syllables contain 5 vowels.

Table 1
Syllables in the Previous Work

Work (Year)	Syllables
Classification of Malay Speech Sounds Based on Place of Articulation and Voicing Places of Articulation using Neural Networks (Nong et al., 2001)	- In Malay language, there are a total of 33 phonemes. - The pure Malay phonemes consist of 18 consonants (/p, t, k, b, d, g, c, j, s, h, 1, r, m, n, ŋ, ɳ, w, y/) and 6 vowels (/a, e, ə, i, o, u/). - The vocabulary of the system is comprised of 16 Malay syllables which are initialized with Plosives or Stops and followed by succeeding vowels. - The selected Malay syllables are /ba/, /bi/, /bu/, /da/, /du/, /gi/, /ka/, /ki/, /ko/, /ku/, /pa/, /pi/, /pu/, /ta/, /tə/, and /ti/.
Malay Syllable Recognition Based On Multilayer Perceptron And Dynamic Time Warping (August Ting et al., 2001)	- The vocabulary of the systems consists of 16 Malay syllables which are initialized with Stops or Plosives (h, d, g, p, t, k/) and followed by 5 vowels (**/a, e, i, o, u/).** - The selected syllables are **/ba/, /bi/,** /bu/, /da/, /du/, /gi/, /ka/, /k/l, /ko/, /ku/, /pa/, /pi/, /pu/, /ta/, /te/, and /ti/.
Computer-Based Malay Articulation Training For Malay Plosives At Isolated, Syllable and Word Level (Ting et al., 2003)	- The list of isolated plosive sounds includes [b], [d], [g], [p], [t] and [f]. - The list of plosive syllable sounds contains 16 syllables **[ba], [bi], [bu],** [da], [du], [gi], [ka], [ki], [ko], [ku], [pa], [pi], [pu], [ta], [te] and [ti].

Methodology

Conventional machine learning techniques for speech recognition and natural language processing often rely on complex feature engineering to extract relevant patterns from audio signals. Convolutional neural networks (CNNs), inspired by biological visual processing, have emerged as a powerful alternative that can automatically learn hierarchical representations directly from raw input data (Ketkar et al., 2021). In this study, we propose applying CNNs for the analysis of Malay syllables involving the five core vowels, by leveraging their strength in visual pattern recognition. Specifically, we intend to record samples of syllables containing each vowel, then preprocess the audio by converting spectrograms - visual representations of the audio's spectral characteristics over time - into 2D images. By cropping these spectrogram images to focus on the regions containing the vowels, we aim to transform the problem into an image classification task compatible with CNNs. Through training a CNN model on large amounts of these spectrogram images (Wyse et al., 2017), we expect it to learn the most discriminative audio patterns between vowels directly from the input, without

reliance on handcrafted features. This end-to-end deep learning approach holds promise for accurate Malay vowel identification.

Data Collection

The data used in this study was collected by recording sounds produced by ten individuals who were classified as "normal." Importantly, these data were acquired in a controlled environment where the only recorded sounds were the human voice articulating the selected syllabus. As elucidated in Table 1, the research focused on a specific set of syllabus items, comprising a total of 16, namely /ba/, /bi/, /bu/, /pa/, /pi/, /pu/, /da/, /du/, /ta/, /pi/, /pu/, /gi/, /ka/, /ki/, /ko/, and /ku/. However, in the paper, a deliberate decision was made to narrow down the scope by selecting only eight distinct syllabus groups, specifically /ba/, /bi/, /bu/, /da/, /du/, /gi/, /ka/, and /ki/. This selection was primarily influenced by practical constraints, particularly the limited timeframe available for the comprehensive integration of all 16 syllabus items into the Convolutional Neural Network (CNN) framework.

In essence, the data collection process ensured that the recorded sounds were exclusively representative of the human voice articulating the chosen syllabus, and this was supplemented by the literature review's rationale for the selection of the specific syllabus items. Furthermore, the restriction to only eight syllabus groups was necessitated by time constraints, preventing the incorporation of the full set of 16 syllabus items into the CNN. These details encapsulate the technical aspects of the data collection and syllabus selection process, providing the foundation for subsequent analyses and discussions in the research study.

Data Recording

The recording was carried out within a room where it was imperative that silence be maintained, with no extraneous sounds, aside from the voices of the participants designated as "normal individuals." The cohort of individuals contributing to this recording comprised a total of ten participants, equally divided into five males and five females. For recording, a microphone and a voice recorder were employed. The positioning of both the microphone and the voice recorder in relation to the participants' mouths was consistently maintained at a distance of 15 cm. The rationale behind this precise measurement of 15 cm stemmed from its capacity to facilitate the clear capture of the participants' voices while minimizing the recording of any incidental breathing sounds.

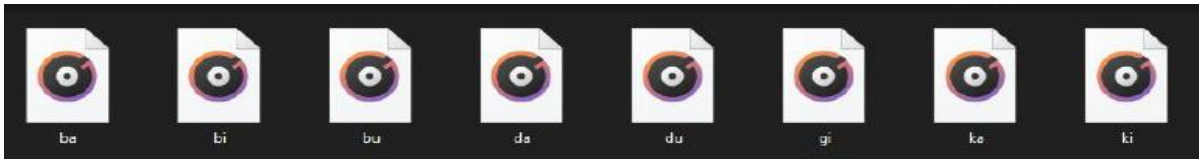

Fig. 1 The sample of the recorded syllabus types audio files

In terms of the duration of the recording sessions, meticulous documentation was upheld. The standard recording session, characterized by an absence of mispronunciations, was systematically timed to span approximately 30 minutes. On the other hand, recording sessions that included instances of mispronunciation extended to an hour. These measures were put in place to ensure the integrity and consistency of the recorded data, allowing for a comprehensive and accurate analysis of the voice recordings, particularly in the context of

the study's objectives. In this manner, the passive voice was applied consistently throughout the description, emphasizing the procedural and methodological aspects of the recording process, the role of the participants, the equipment used, the rationale behind the 15 cm distance, and the deliberate timing protocols employed during the recording sessions. This passive construction adheres to scholarly conventions and prioritizes objectivity and clarity in the research documentation.

Upon completion of the recording process, the audio data undergoes transformation into waveforms or images, facilitated through the utilization of Python IDE, Google Colab. After this conversion, each image is systematically cropped, a step designed to aid in the subsequent processing of the Convolutional Neural Network (CNN) in the subsequent phases of the paper. The process commences with the audio data being subjected to conversion into either waveform representations or images. Once this conversion process is executed, the audio data is effectively transmuted into a format that can be readily processed and analysed by the subsequent stages of the paper. Following the conversion, a critical post-processing step involves the systematic cropping of each individual image. The purpose of this cropping operation is to segment the images into distinct, manageable units, with each unit corresponding to a specific syllabus item. This deliberate segmentation facilitates the subsequent operations of the Convolutional Neural Network (CNN) by presenting the network with a structured dataset that is amenable to systematic analysis.

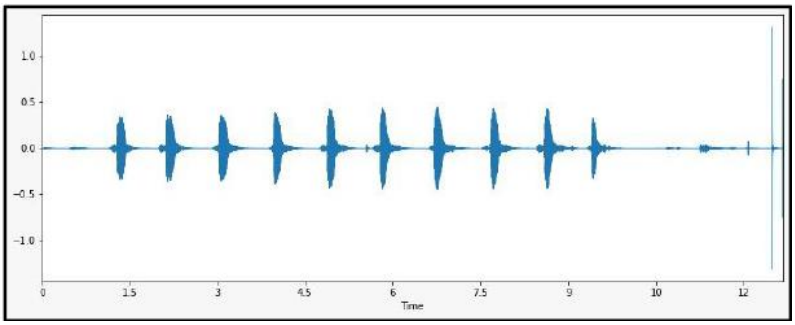

Fig. 2 The sample of recorded audio waveform

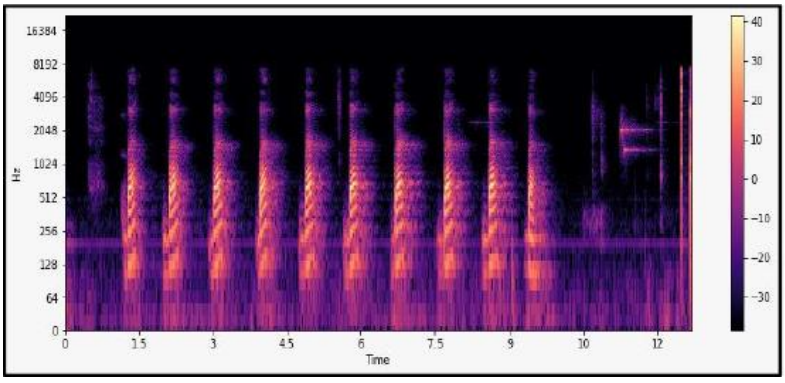

Fig. 3 The sample of recorded audio in spectrogram waveform

Data Arrangement for Analysis

The recorded data comprises two distinct groups, consisting of five male and five female participants. These two groups were recorded individually and at separate times to ensure data integrity. The conversion of the recorded audio into image representations, specifically

focusing on the syllabus /ba/, revealed that data from recordings L1 through L3 exhibited high-quality spectrogram images. Conversely, data from recordings L4 to L5, included in the appendix, exhibited noticeable noise in the waveforms, particularly in the representations from the male group. A similar pattern was observed in the female group, with the most favourable images arising from recordings P1 to P2 and noise artifacts evident in the remaining representations from P3 to P5. In the case of the syllabus /bi/, a distinct pattern emerged where the images generated from recordings L1 to L5 and P5 exhibited no noise artifacts.

However, the converse was true for recordings P1 to P4, which displayed noise interference. The conversion of the syllabus /bu/ produced images with detectable noise in recordings L2, P1, P2, P3, and P4, while noise-free representations were obtained from recordings L1, L3, L4, L5, and P5. For the /da/ syllabus conversion, noise artifacts were identifiable in images from recordings L2 to L5 and P1 to P5. In contrast, noise-free images were evident in recordings L1 to L3 and P2. Regarding the syllabus /du/, noise was detected in images from recordings L1, P1, and P3 to P4, while noise-free images were discernible in recordings L2 to L4, P2, and P5. In the context of the syllabus /gi/, noise was evident in images from recordings L1 to L3, P1, and P3 to P5, while noise-free images were only identifiable in recordings L5 and P2. The conversion of images for the syllabus /ka/ revealed noise artifacts in recordings L1, P1, and P3 to P5, with noise-free images present in recordings L2 to L5 and P2. Finally, in the case of the syllabus /ki/, noise was detectable in images from recordings L1, P1, and P3 to P4, whereas noise-free images were obtained from recordings L2 to L5, P2, and P5.

Data Preparation before Processing via Deep Learning

The cropping process was initiated after the conversion of audio-recorded data into image representations. The dimensions chosen for cropping the spectrogram images were set at 55 pixels in width and 230 pixels in height. It is imperative to note that this cropping operation was executed through an online platform, and the undertaking spanned a period of one week to complete the cropping task for each individual and syllabus image. The purpose underlying this meticulous cropping process was to facilitate the Convolutional Neural Network (CNN) in its capacity to selectively concentrate on the specific waveform patterns representative of the syllabus in question.

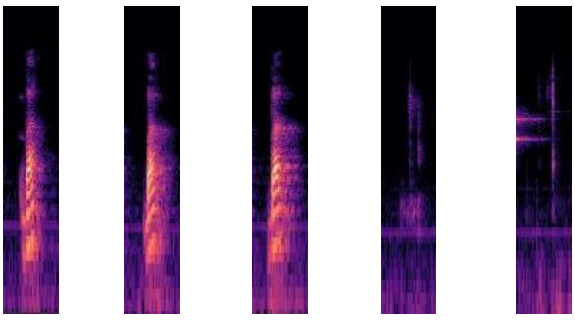

Fig. 4 The Data Cropping Process

Finding

In this section, the results will be discussed accordingly, with consideration given to the network type, epoch number, and batch size. Several networks have been chosen for evaluation, specifically VGG16 and VGG19. The primary objective is to determine which network performs optimally for the Malay Language Syllabus through the utilization of Convolutional Neural Networks (CNN). The selection of the most suitable network is contingent upon the inherent design features tailored to the Malay Language Syllabus. Furthermore, the batch size has been chosen to be either 8 or 16. The batch size signifies the number of samples processed in each iteration of the CNN training process. In the context of this proposed paper, the decision to use 8 or 16 as the batch size is underpinned by specific considerations. Notably, the dataset comprises 640 images categorized into 8 distinct syllabus classes, and the total number of syllabus classes is 16. Hence, the rationale for selecting 8 and 16 as batch sizes is twofold. Firstly, 16 is the actual number of syllabus classes, and 16 divided by 2 yields 8, aligning with the dataset structure. Secondly, each image is characterized by a bit depth of 32 bits and dividing 32 by 2 results in 16.

Subsequently, the epoch number is varied between 10, 20, and 50. The epoch number represents the quantity of times the algorithm iterates through the entire training dataset. The rationale behind this selection is rooted in the need to train the CNN to effectively recognize the syllabus depicted in the images. Commencing with a small number of epochs and incrementally progressing to a larger number serves the purpose of aiding the CNN in learning and reinforcing its ability to discern the syllabus.

Table 2
Comparison of proposed network model classification accuracy (%) with other comparative models using different epoch numbers and network type for batch size = 8

Batch Size / Epoch	10	20	50
VGG16	*37.50%*	*43.75%*	*42.50%*
VGG19	26.25%	26.25%	30.00%
Proposed	**45.00%**	**57.50%**	**53.75%**

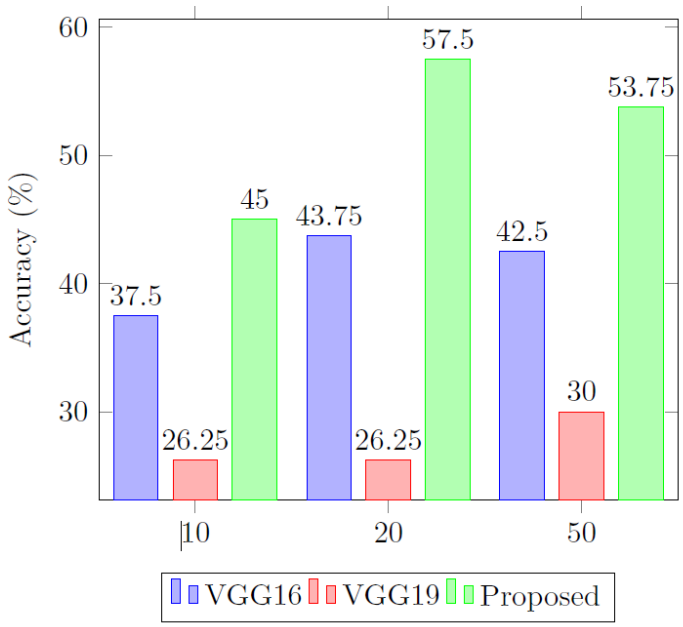

Fig. 5 The bar graph of batch size 8, and epoch numbers are 10, 20 and 50

Table 2 presents a comprehensive depiction of the outcomes associated with a batch size of 8. The results are categorized according to different epoch values and network types (CNN) to facilitate a meticulous analysis aimed at determining the most suitable CNN architecture for the paper. The data from Table 3 elucidate that the highest percentage accuracy is achieved at epoch 20. Conversely, deploying epoch 50 results in a decrease in the percentage accuracy compared to epoch 20. It is important to emphasize that the primary function of varying the epoch values is to enable the CNN to adapt and consolidate its comprehension of the syllabus images through repetitive exposure to the training dataset.

Table 3 Comparison of proposed network model classification accuracy (%) with other comparative models using different epoch numbers and network type for batch size = 16

Batch Size / Epoch	10	20	50
VGG16	31.25%	*36.25%*	*41.25%*
VGG19	*38.75%*	27.50%	30.00%
Proposed	**40.00%**	**60.00%**	**68.75%**

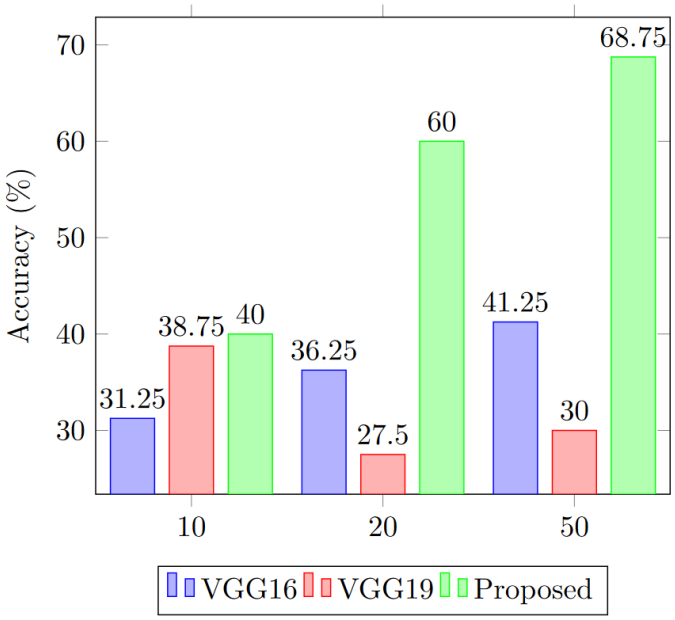

Fig. 6 The bar graph of batch size 16, and epoch numbers are 10, 20 and 50

Table 3, conversely, offers a comprehensive presentation of the outcomes linked to a batch size of 16. These outcomes are organized based on different epoch values and the utilization of distinct convolutional neural network architectures (CNN). The data in Table 3 reveal that the highest percentage accuracy is attained at epoch 50. In contrast, adopting epoch values of 20 and 10 leads to a reduction in the percentage accuracy. The motivation for this exploration resides in the quest to ascertain the most effective CNN configuration for the proposed paper.

Conclusion and Future Work

The paper has yielded a proposed Malay language syllable recognition system designed to categorize Malay syllables based on their places of articulation and voicing. Various critical aspects of the neural network architecture have undergone rigorous investigation, including considerations of time and environmental factors. The data collection process involved the recording of two distinct gender groups, denoted as L1 - L5, representing males, and P1 - P5, representing females. Each recording was conducted individually within a controlled, noise-free environment employing essential recording equipment such as voice recorders and microphones. Ensuring the quality and consistency of these recordings was paramount, as they were essential during the subsequent stages of the proposed paper. The core methodology employed in this paper involved the utilization of Convolutional Neural Networks (CNN), which was organized into three primary phases: testing, evaluation, and training. The testing and evaluation phases involved the analysis of 10 images of individual waveforms from random subjects in the case of testing, and 80 images for the evaluation phase. This division into separate stages facilitated the CNN in recognizing syllables effectively through the processing of cropped waveform images. The system achieved an average accuracy rate of 60%. Looking ahead to future developments in this paper, there is potential for expansion into diverse languages. Such an expansion holds the promise of assisting stroke patients in communicating across various local languages worldwide, utilizing the power of Convolutional Neural Networks (CNN). Furthermore, the proposed paper can be enhanced by

incorporating mouth movement detection technology, contributing to the advancement of Industry 4.0. Additionally, it has the potential to aid individuals with specific needs and can be invaluable in recognizing the speech of disabled individuals, particularly stroke patients. However, it is imperative to emphasize that comprehensive planning is crucial for the successful execution of this paper. The recording, cropping, and simulation phases, involving a significant number of participants, may entail several months to complete. Nevertheless, this paper holds the potential to make significant contributions to the field of bio-medicine, offering a promising avenue for research and application in this specialized domain.

Acknowledgement

The authors would like to thank PERKESO Rehabilitation Centre, Melaka for the Research Project Collaboration and Fakulti Teknologi dan Kejuruteraan Elektronik dan Kejuruteraan, Universiti Teknikal Malaysia Melaka for providing the facilities to do the experimental analysis and work towards completing this research paper.

References

Sulaiman, M. A. H., Abd Aziz, N., Zabidi, A., Jantan, Z., Mohd Yassin, I., Megat Ali, M. S. A., & Eskandari, F. (2021). A systematic approach for Malay language dialect identification by using CNN. *Journal of Electrical and Electronic Systems Research (JEESR)*, *19*, 25-37.

Zamri, N. S. M., Hashim, N. M. Z., Latif, M. J. A., & Rajaandra, P. (2021, October). A Preliminary Study on Vowel Recognition via CNN for Disorder People in Malay Language. In *2021 5 th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 312-316). IEEE.

Hariharan, M., Yaacob, S., & Adom, A. (2012). A review: Malay speech recognition and audio visual speech recognition In Biomedical Engineering (ICoBE). In *International Conference*.

Hashim, N. M. Z., Zahri, N. A. H., Abd, M. J., Sulistiyo, M. D., & Kamaruddin, A. I. (2022). Analysis on Vowel /E/ in Malay Language Recognition via Convolution Neural Network (CNN). *Journal of Theoretical and Applied Information Technology*, *100*(5).

Hashim, N. M. Z., Zahri, N. A. H., Sulistiyo, M. D., Latif, M. J. A., Darsono, A. M., & Rajaandra, P. (2021, October). Malay Language Vowel Recognition using Image Outlook via Convolution Neural Network (CNN). In *International Conference on Emotions and Multidisciplinary Approaches-ICEMA* (Vol. 2021, p. 588).

Ponniah, K., Sivanathan, I., & Kumar, M. (2021). An Critical Analysis of Speech Recognition of Tamil and Malay Language Through Artificial Neural Network.

Juan, S. S., Besacier, L., & Tan, T. P. (2012, November). Analysis of malay speech recognition for different speaker origins. In *2012 International Conference on Asian Language Processing* (pp. 229-232). IEEE.

Ong, H. F., & Ahmad, A. M. (2011). Malay language speech recogniser with hybrid hidden markov model and artificial neural network (HMM/ANN). *International Journal of Information and Education Technology*, *1*(2), 114.

Nong, T. H., Yunus, J., & Salleh, S. H. S. (2001, August). Classification of Malay speech sounds based on place of articulation and voicing using neural networks. In *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001 (Cat. No. 01CH37239)* (Vol. 1, pp. 170-173). IEEE.

Ting, H. N., Jasmy, Y., Hussain, S. S., & Cheah, E. L. (2001, August). Malay syllable recognition based on multilayer perceptron and dynamic time warping. In *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat. No. 01EX467)* (Vol. 2, pp. 743-744). IEEE.

Ting, H. N., Yunus, J., Vandort, S., & Wong, L. C. (2003, December). Computer-based Malay articulation training for Malay plosives at isolated, syllable and word level. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint* (Vol. 3, pp. 143-1426). IEEE.

Ketkar, N., Moolayil, J., Ketkar, N., & Moolayil, J. (2021). Convolutional neural networks. *Deep learning with Python: learn best practices of deep learning models with PyTorch*, 197-242.

Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*.