VOWEL'S CLASSIFICATION FOR STROKE PATIENTS THROUGH REHABILITATION PERFORMANCE VIA IMAGE-PROFILED SOUND DATA

Nur Syahmina Ahmad Azhar[1], Nik Mohd Zarifie Hashim[1], Afiqah Iylia Kamaruddin[2], Nik Adilah Hanin Zahri[3] and Mahmud Dwi Sulistiyo[4]
[1]Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Pusat Rehabilitasi PERKESO Tun Abdul Razak, Melaka, Malaysia
[3]Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis, Perlis, Malaysia
[4]School of Computing, Telkom University, West Java, Indonesia
E-Mail: nikzarifie@utem.edu.my

ABSTRACT

In terms of medicine, a disorder is a disturbance of the mind or body's normal functioning. Communication issues may result from stroke because it damages the parts of the brain that control language. The capacity to talk, read, write, and comprehend speeches can all be affected after a stroke. The rehabilitation and treatment of patients generally take a long time and include constant medication, exercise, and rehabilitation training. However, most rehab facilities throughout the world still manually carry out this rehabilitation process. Machine learning and deep learning have been introduced to this medical field to aid rehabilitation using the new technology due to computer vision's impact on this field. A reliable Convolution Neural Network with a graphical user interface is introduced in this study to support and enhance rehabilitation efforts. The spectrogram in the image-profiled sound is used to provide the optimum outcome and accuracy. This project aims to develop a neural network that can distinguish vowels between a normal person's and stroke patients' voices. In this proposed paper's result, an intelligent Convolution Neural Network system for Malay language vowel detection with high-performance accuracy is demonstrated with maximum accuracy was 92.96% with 20 epoch numbers and 6 batch size. This outcome showed that the proposed method, even using a simple network design, is still competitive compared with other methods. The proposed method is ideal for training and validating vowel recognition accuracy especially for stroke patients.

Keywords: convolutional neural network (CNN), image-profiled sound, rehabilitation, spectrogram image, stroke patients, vowel recognition.

Manuscript Received 26 October 2022; Revised 13 August 2023; Published 30 August 2023

INTRODUCTION

Genetic, medical, or traumatic factors can contribute to mental illness and mental health issues. Mental illnesses include depression, anxiety disorders, schizophrenia, eating problems, and addictive behaviors. Mental illnesses, physical disorders, genetic diseases, emotional and behavioral disorders, and functional disorders are several types of medical problems [1]. An available anomaly or disruption is referred to as a disorder. In some contexts, the term disorder is the preferred language over sickness or illness since it is frequently thought to be more value-neutral and less stigmatizing. However, the term disorder is also used in many other fields of medicine, primarily to identify physical disorders that are not caused by infectious organisms, such as metabolic disorders. In mental health, mental disorder acknowledges the complex interaction of biological, social, and psychological factors in psychiatric conditions. Disorder patients experience physical problems that affect their disposition, thinking, and actions [2].

Speaking disorders are just one of several issues that make communication challenging. The communication range of a patient with a condition is slightly different from that of a healthy individual. Communication issues may result from stroke because it damages the parts of the brain that control language. The capacity to talk, read, write, and comprehend speeches can all be affected after a stroke. Some diseases, like aphasia, make communicating difficult in real life. Damage to the language centers of the brain results in aphasia and causes a complex impairment of language and communication. This harm can be brought on by a stroke, a head injury, brain tumors, or other neurological conditions.

About a third of stroke survivors get aphasia [2]. Next, people with apraxia know what they want to express but struggle to put it into words. Reading, writing, swallowing, and other motor skills may be complex for them [4]. On the other side, dyspraxia is a neurological ailment and brain-based motor condition that impairs a person's coordination, movement, and planning abilities. Dyspraxia affects children's ability to produce sounds, repeat sounds or sentences, and maintain regular intonation patterns. Most of these patients won't be able to convey their needs and wants since they won't be able to communicate [5]. More speech disorders like articulation disorders, cognitive-communication disorders, dysarthria, expressive disorders, fluency disorders, sensory disorders and resonance disorders need help to improve their voice and sound production, comprehension, fluency, clarity, and expression.

The term speech rehabilitation refers to medical therapy focusing on communication problems that most frequently affect those unable to communicate effectively. A practical and trustworthy training program must be created and offered because rehabilitation activities are crucial for recovering from the speech impediment. To help people, overcome communication hurdles, many therapy centers and rehab facilities now provide rehabilitation treatments. With the development of technology, there has been a lot of research and development in speech therapy and rehabilitation. Vowel identification was the initiative's main goal to improve communication for those with communication disorders.

The documentation of the speech pronunciations of standard Malay using vowel charts will serve as the fundamental framework for further research on speech synthesis, speech education, speech rehabilitation, and speech reproduction, even though these works were significant and valuable in furthering our understanding of spoken Malay [6]. Additionally, earlier research [7] focused solely on the generation of the six conventional Malay vowels which are /a/, /e/, /ê/, /i/, /o/ and /u/. Figure 1 shows a spectrogram for the vowel /a/ as an example. In a wideband spectrogram, the formants appear as black bands. It is generally known that the first two formants, F1 and F2, are the primary bearers of the data required for vowel recognition [7]. Although vowel reduction can take many forms, it is lexical in certain languages. For instance, in English, the centralized vowel "schwa" is used to sound vowels that are distant from the stressed syllable.

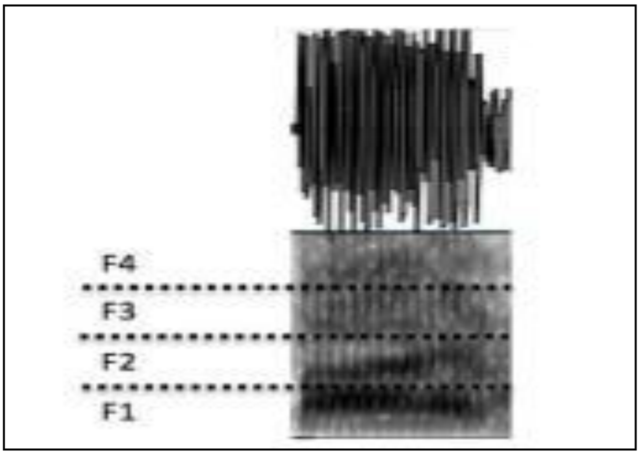


Figure-1. Dark Region of Spectrogram [7].

Despite adopting a manual evaluation process that might accommodate the rehab center's time constraints and staff shortage, an intelligent system for training may be the better option. The existing research was used to suggest and carry out an intelligent training system for these stroke patients using sound characteristics. These systems will help them communicate since they will concentrate on vowel recognition by collecting sound data from healthy and disordered patients. To assess and validate the recorded data, it will compare the sound recordings of disordered patients with healthy individuals using neural networks [7]. This initiative will concentrate on vowel recognition for people with behavioral issues. The sound/voice recording data for this study is collected from two groups: disorder patients and normal persons. Voices and sounds will be recorded for the project, which will later be kept in the data collection. Then, by comparing the sound recordings of disordered people to normal people, it will train, evaluate, and test the recorded data using the proposed network model, the Neural Network. Journals and papers pertinent to the project are used as references. This study offers two contributions:

a) We compare the performance of two group directories between the normal person and disorder patients using a new approach to see sound in the image that looks like a spectrogram image, and

b) We designed a new simple network model for evaluating the training and validation accuracy performance for Malay vowel classification tasks.

Note that the research work [7] is expanded in this study. In addition to the previous work, here, we conducted a new study on the performance of a disordered patient, and six kinds of vowels are also used in the evaluation. This study uses a convolution neural network to provide stroke patients with Malay language vowel recognition (CNN). The technique used spectrogram visuals in place of CNN's conventional sound file to distinguish all six vowels in the Malay language. Five models of networks will be employed in this paper's experiment, which will be conducted via a network.

RELATED WORKS

The term "speech rehabilitation" refers to medical therapy focusing on communication problems that most frequently affect those unable to communicate effectively. To help people overcome communication hurdles, many therapy centers and rehab facilities now provide rehabilitation treatments. With the development of technology, there has been a lot of research and development in speech therapy and rehabilitation. Vowel identification was the initiative's primary goal to improve communication for those with communication disorders. Before the existence of deep learning, there were several methods of rehabilitation that focused on speech disorders. According to Van Riper's Speech Correction in his book, speech and motor abilities are related. The studies of young children that demonstrate a parallel development path for speech acquisition and essential motor skills and the development of delayed speech might be impacted by the delay in motor skill acquisition, especially in young children. Therefore, the articulation drill and the motor learning approach concentrated on the motor practice of the tongue movement and synchronization of the other articulators, such as the lips and jaw, intending to make the people who lack communication have precise pronunciation [9].

In addition to motor learning and articulation drills, phonological/lexical interventions are a technique used with words and phrases. This approach takes into account voice sound concerning word creation. These interventions aim to increase productivity at the word level [9]. This approach differs from the Articulation Drill and Motor Learning approaches, which emphasize the placement and production of each speech sound using the body. Recently, there have been methods of speech rehabilitation using machine learning. Deep learning is currently one of the most popular machine-learning (ML) based techniques, and CNN is the major deep learning (DL) architecture used for image processing. The research paper on speech recognition using machine learning created a system that can translate between Hindi and English and detect human speech and audio samples. They offered options to convert audio from one language to another, with the output in text form. The architecture of neural machine translation consists of two recurrent neural networks that work together to build an encoder-decoder structure [10].

Another recent method of speech therapy in rehabilitation is deep learning applications in telerehabilitation speech therapy scenarios. Using a speaker-dependent approach, a convolutional neural network has been trained to detect a handful of terms inside the unusual speech. They concentrated on isolated word identification for native Italian speakers with dysarthria. They used an already-existing mobile app to gather audio data from users with speech disorders as they engaged in articulation exercises to benefit speech treatment [11].

PROPOSED WORKS

The proposed approach in this study covers the full procedure of enhancing training and validation accuracy. To create a new intelligent classification system for stroke patients, a large dataset of images covered for normal person and patients is necessary. The recording, conversion, and cropping procedures create datasets for healthy individuals and stroke patients. The complete dataset will then be trained using a convolutional neural network.

The Arrangement of Dataset

This section explains the process used to collect and arrange the images for the datasets used in the article. To learn more about the vowel, we converted the real-time audio in *wav file* format to a spectrogram image. In this study, we used spectrogram images to produce the result of amplification of the audio stream.

Audio Recording

The vowel sound /a/, /e/, /E/, /i/, /o/, and /u/ should be recorded first. A voice recorder was used to capture the vowels /a/, /e/, /E/, /i/, /o/, and /u/ on healthy individuals and stroke patients. Everybody records in the same way, with a 15 cm gap between their mouths and the voice recorder shown in Figure-2. A normal and healthy person must pronounce every vowel in three sections which are short, middle, and long. The short-period, middle-period, and long-period signals are recorded at lengths of 2, 3, and 4 seconds, respectively. When stroke sufferers recorded their voices for this component, it was designed with their requirements in mind. A voice recorder, REMAX RP1 8GB Digital Audio was used to capture the vowels /a/, /e/, /E/, /i/, /o/, and /u/ on healthy individuals and stroke patients.

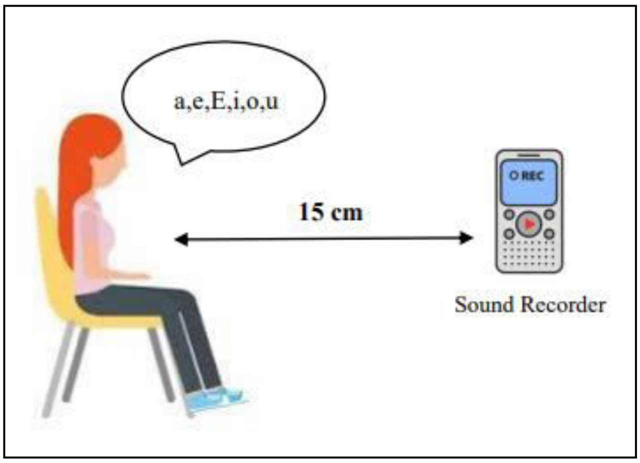


Figure-2. The proposed recording session setting.

Conversion from *wav file* to Spectrogram Images

The conversion from wav file to spectrogram images is then carried out, as seen in Figure-3 and Figure-4. We used spectrogram image-profile besides of real-time image as spectrogram is a graphic representation of the strength of a signal over time at different frequencies that make up a waveform. A spectrogram's vertical axis displays frequency-based information. The highest frequency is shown at the top, while the lowest frequency is shown at the bottom. Spectrograms can be either three-dimensional graphs with a fourth color-based variable or two-dimensional graphs with a third variable that is based on color. Figure-4 displays the spectrogram image of 10 repeatedly uttered vowels from a normal person. Each vowel in the spectrogram image has a unique set of qualities.

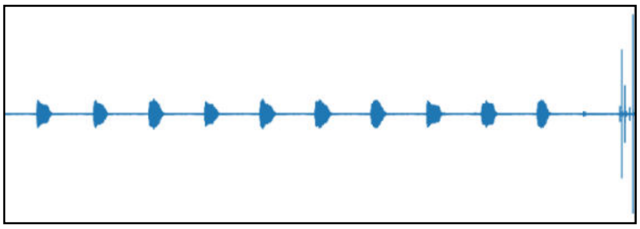


Figure-3. Real-time Audio in *wav.file*

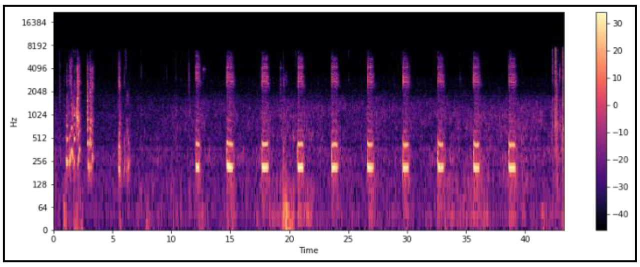


Figure-4. Spectrogram images

The cropping procedure will start after the conversion is finished, with each vowel's image having a size of 240 x 55 pixels and a bit depth of 32 bits. Figure-5 shows the cropped images of normal person and disordered patients in every class.

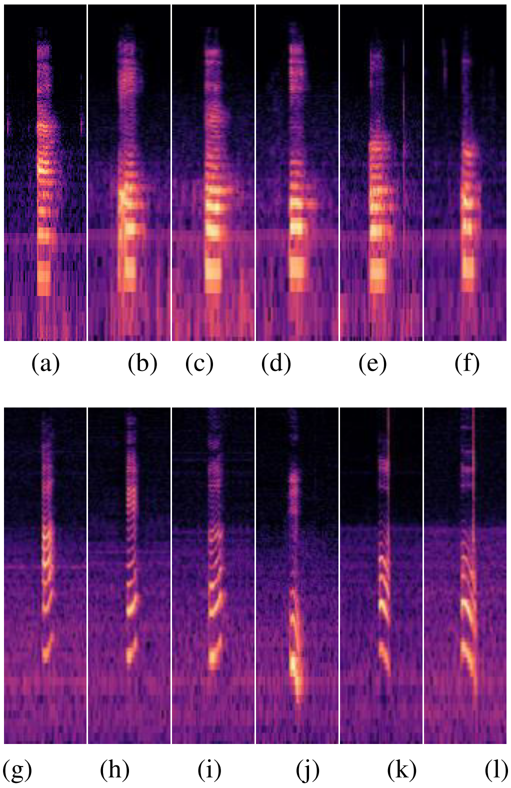


Figure-5. Spectrogram images for vowels (a)-(f) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for normal person and vowels (g)-(l) for vowel /a/, /e/, /E/, /i/, /o/ and /u/ for disordered patient.

Proposed Network

The dataset of conversion spectrogram images has been put to the test using convolutional neural network architecture. We integrated our proposed network model with already-existing network models like VGG16, VGG19, Inception and AlexNet. The five layers of CNN are the convolutional layer, the pooling layer, the fully connected layer, the dropout layer, and the activation layer [10].

The project's input consists of spectrogram images with a size of 55 by 240 pixels and a depth of 32 bits. The convolutional layer will control the output of neurons connected to certain regions of the input. The pooling layer will then down sample the input after that. The dropout layer, which removes a small number of neurons from the neural network during training, is then used to learn and approximately represent any kind of continuous and complex relationship between network variables using the activation functions. We have split the model training, evaluation, and testing into 80% training data, 10% validation data and 10% test data as in Figure-6. There are no hyper-parameters and optimal split percentage to split the machine learning data. The model and dataset split ratio are both determined by the number of samples in the dataset. The amount of data that was gathered for the experiment that was undertaken for this study is large. When choosing the best split, there are two main considerations. The machine learning model will display large variation in training if there is little training data. The model evaluation/model performance measure will have more variance if there are fewer testing data/validation data. We have constructed an ideal optimum split that meets the demands of the dataset and model.

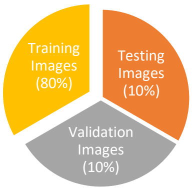


Figure-6. The percentage of optimal split dataset.

In the designed model, we have designed the network for the better accuracy. Conv1 is the first convolution layer. Conv1-1, the second convolution layer, has 32 filters and assumes the picture is 238 by 53 in size. Conv1 is the first convolution layer. The convolution layer utilized in this instance has 32 filters and a max-pooling dimension of 236 x 51 for the pictures. With 32 filters, Conv2 is the second convolution layer. Next is the Conv2-1, which has 32 filters and maximum pooling with an image resolution of 116 x 23.

The Conv2-2 uses 64 filters simultaneously, with a maximum pooling size of 55 x 8. Then, we combine 1024 units of a dense layer with 6 units of a dense SoftMax layer to generate a dense layer. The reduction in image size for each layer is seen in Figure-7. A CNN with an input picture dimension of 240 x 55 was used to build the model, together with an ADAM classifier as the optimizer and SoftMax as the activation function.

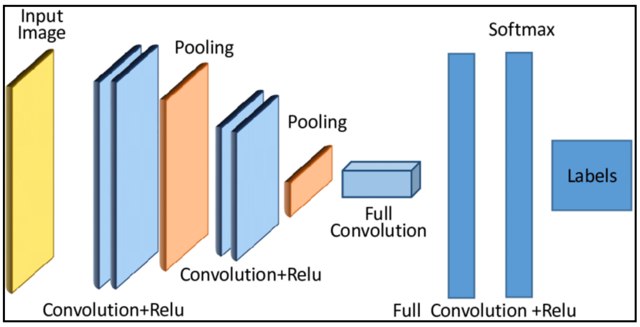


Figure-7. Designed model architecture.

One of the simple and common CNN networks, VGG16, is used as the main model for vowel classification besides the designed model in the experimental work to provide classification results. A deep convolutional neural network with 16 layers is called VGG16. It repeatedly combines 3x3 convolutional layers and 2x2 pooling layers. VGG16 is simpler to use and gives a better capacity for feature learning than AlexNet.

Additionally, it provides better picture classification results than Inception. The number of separable convolutions and convolutions in the new model almost exactly match the number of convolutions and separable convolutions in VGG16 and VGG-19. To evaluate how changing the architecture layers of the AlexNet Convolution Neural Network can affect and improve the classification result, we suggested a new CNN model in this study.

We chose the AlexNet architecture because it has the simplest structure of all the CNN systems and requires the most time and efficiency during the learning phase, especially since we are utilizing a standard personal computer with a standard CPU-based system [11]. High-performance training that is quick and uses few epochs is another incentive to choose this design. The development of CNN-based classifiers reached a turning point with the introduction of the Inception module in as a component of the GooLeNet architecture [12].

EXPERIMENTAL RESULTS

The training and validation results of study 1, study 2 and study 3 have been noted and recorded. The studies use five network models with the same epoch value and batch size setting. Each study's training process for the dataset image satisfied all of the specifications. We conducted studies in various settings to demonstrate the model's efficiency in detecting stroke patients' vowels. We divided the analytic and experimental settings into three main analyses to compare the classes. We compared the accuracy of normal person results and disorder patient results, and we observed the effectiveness and efficiency of the results for the usage of those with disabilities.

To explore the effectiveness during the tiny epoch size, which is reliable to the actual application, we set the batch size and epoch size to be similar in the normal person, disorder patients and mixed group (normal person + disorder patient) classification investigations. A research paper by Hashim *et al.* has shown the various experiments of different epochs and batch sizes of six vowels [7]. The larger the batch size, the faster the model per epoch during training but the higher number of batch sizes will lead to poor generalization. The number of epoch have no optimal number, and the dataset is sufficient for epoch 20 throughout the training phase without an overfitting graph in every study. Every dataset uses a variable number of epochs, and the diagram depicts the training process when the model is learning the training and validation datasets.

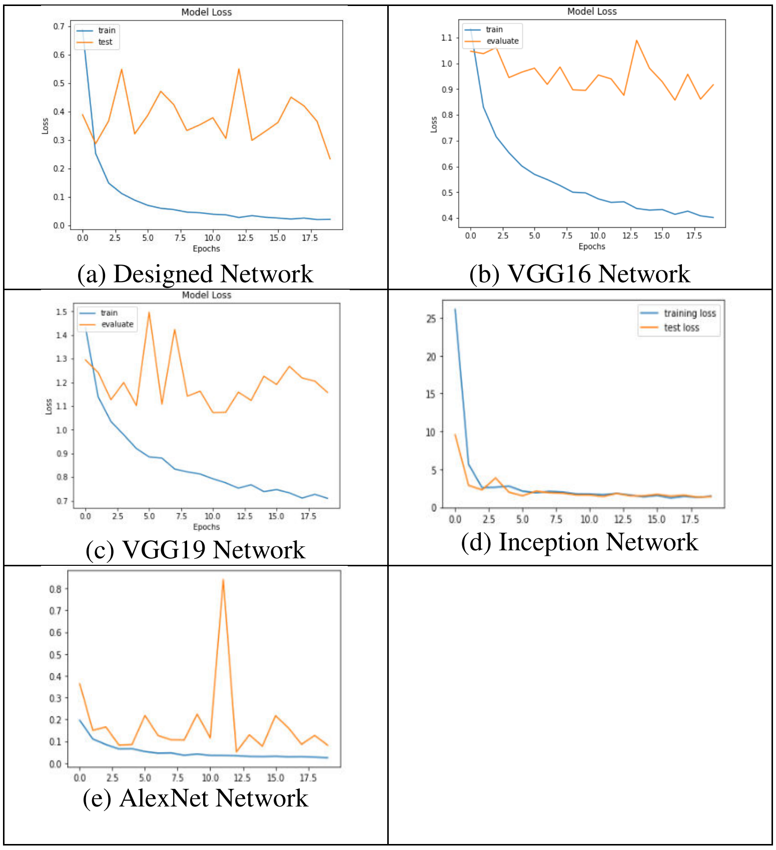
A thorough examination of Malay vowels is provided by comparing normal person, disorder patient, and mixed group (normal person + disorder patient) vowel classification. To demonstrate the classification performance, the designed network model is compared with the designed network models, VGG16 model, VGG19 model, Inception model and AlexNet model. In the initial analysis, no stroke patients were involved, and only 20 healthy participants were included. A total of 10800 spectrogram images were collected for the first comparison. A dataset with nine stroke patients was used in the second analysis, and no dataset from the normal group was included. From the second study, we collected a total of 1620 spectrogram images. A total of 12420 spectrogram images were used in the third study, whereas 9936 images from 20 healthy subjects and 1242 images from 9 stroke patients were used for the training process and evaluation process in the third study.

Model comparison of normal person classification (one group)

With the aid of five different network model, which is designed model, VGG16, VGG19, Inception and AlexNet, these models are carried out to compare each of them thoroughly. For each of these conditions, we detailed the experiment's findings. The classification accuracy for batch size = 6 and epoch 20 using the designed network model shown in Figure 9 is 92.96%. Table-1 compares each model's validation accuracy. From the study, we can observe that the accuracy of the designed network is the highest compared to the other network.

Table-1. Result of training and validation accuracy for different model in first study.

Epoch	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**99.95**	**92.96**
		VGG16	87.88	68.52
		VGG19	76.30	56.02
		Inception	33.70	32.30
		AlexNet	94.17	87.11



(a) Designed Network

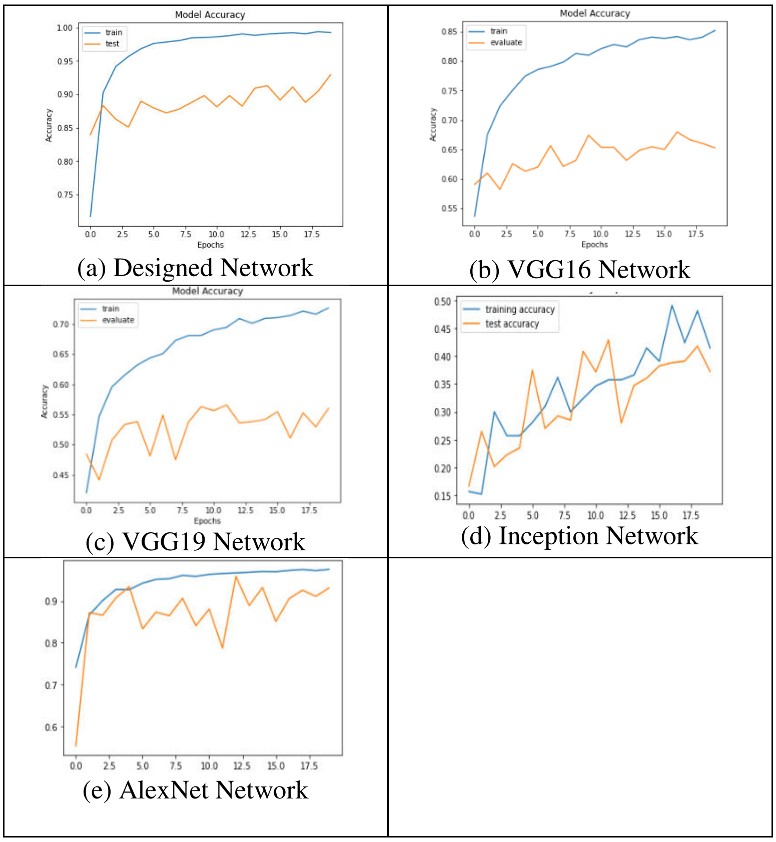
(b) VGG16 Network

(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-8. Model Loss performance in the first study.



(a) Designed Network

(b) VGG16 Network

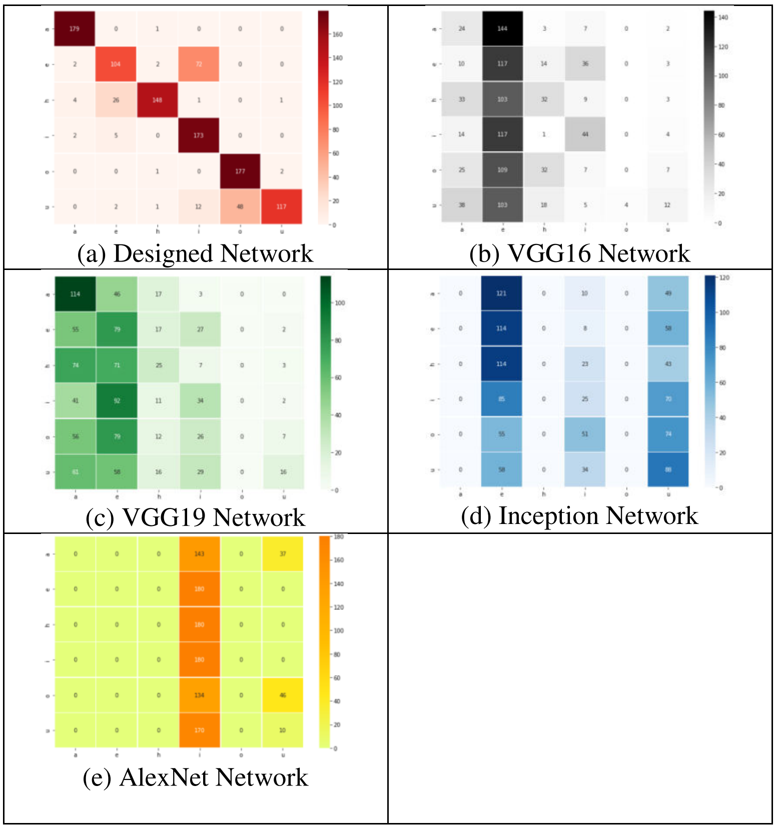
(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-9. Model accuracy performance in the first study.

Between the five model networks conducted, the designed model has the highest validation accuracy for epoch 20 and batch size 6, with 92.96%. Compared to Hashim et al., their accuracy is 81% for the designed network and epoch 20. For this research experiment, we have successfully developed a better accuracy network with a validation accuracy of 92.96%. In this research, the dataset collected for normal and healthy persons is more than the experiments conducted using Hashim's dataset.



(a) Designed Network

(b) VGG16 Network

(c) VGG19 Network

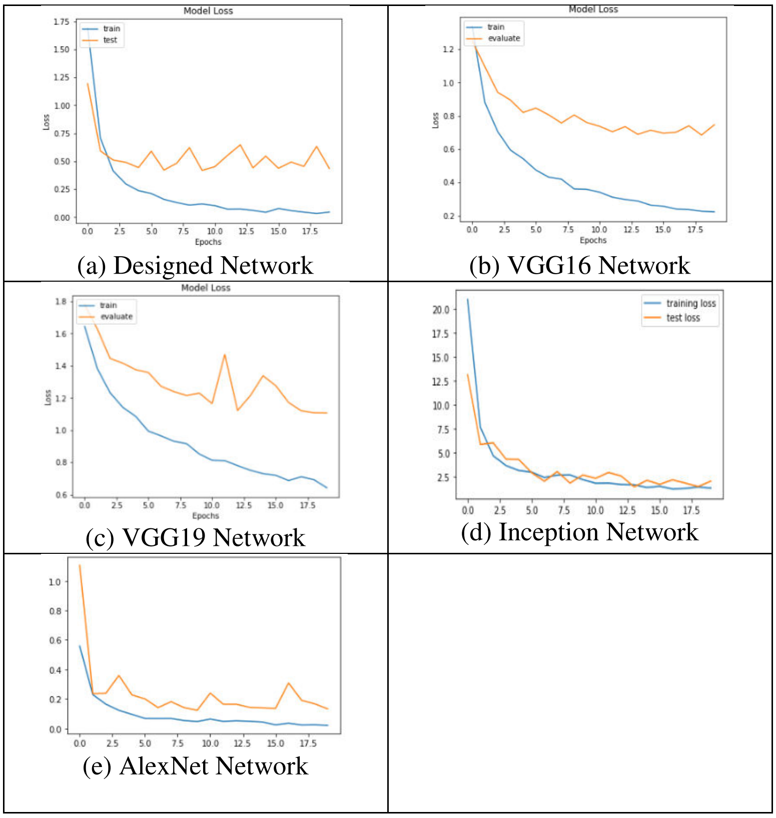
(d) Inception Network

(e) AlexNet Network

Figure-10. Confusion matrix of network proposed.

Model comparison of disorder patients' classification (one group)

The second study is conducted to observe the performance of the dataset among the disorder patient. From the experiment shown in study 1, we can conclude that the designed model has the highest accuracy compared to others. The carried method of training is the same as the first study with an epoch of 20 and batch size of 6. The second study contains 1620 spectrogram images from nine stroke patients. The 80% training process had 1296 spectrogram images of six classes of vowels, while the 10% validation process took 162 images of six classes of vowels. The remaining 10% of the testing process also contained 162 spectrogram images.



(a) Designed Network

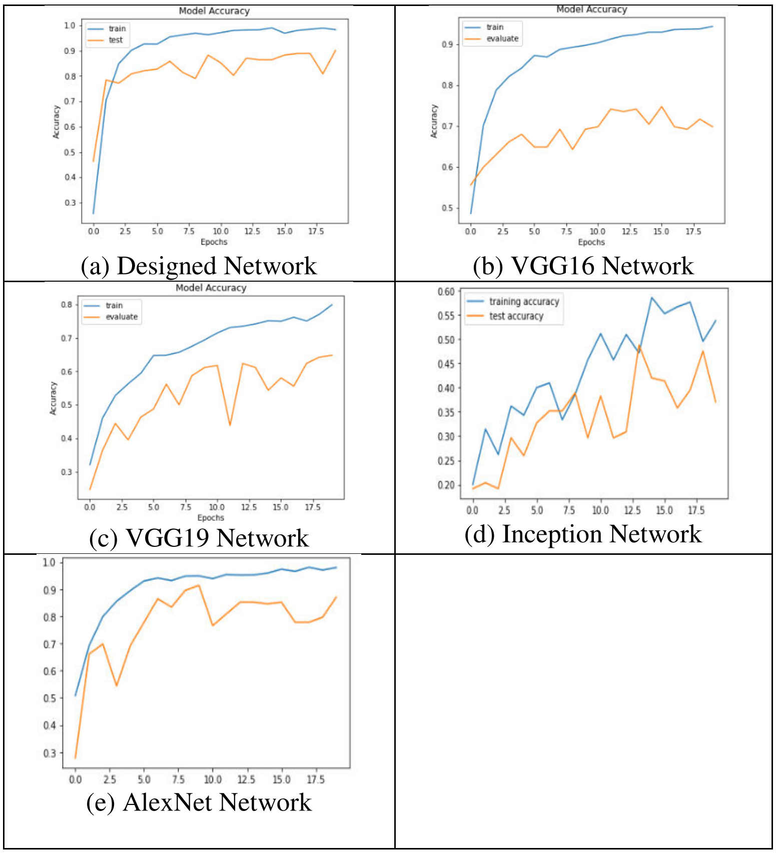
(b) VGG16 Network

(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-11. Model loss performance in the second study



(a) Designed Network

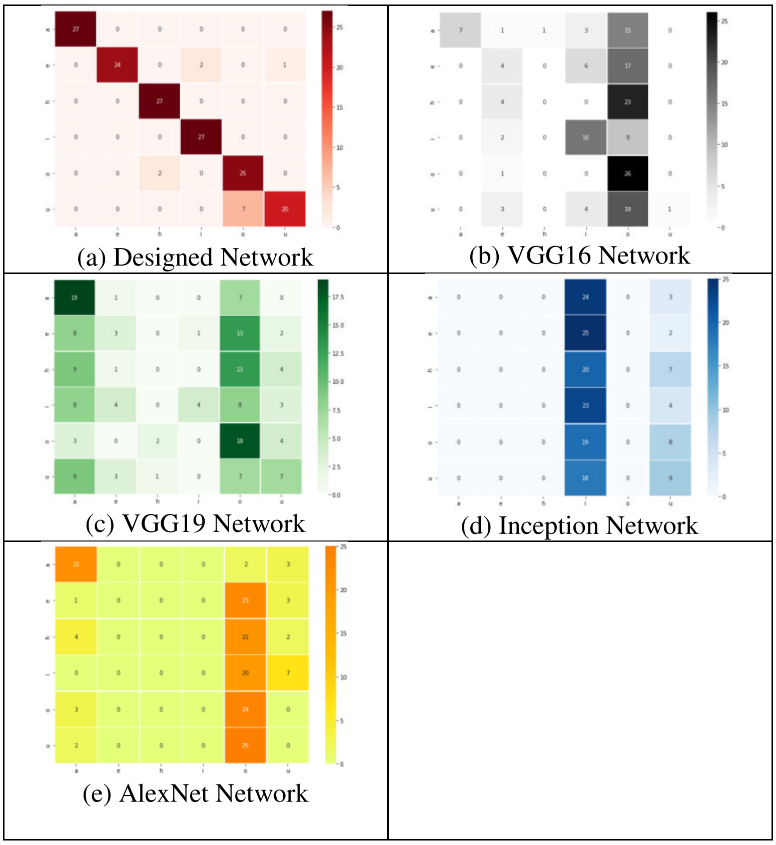
(b) VGG16 Network

(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-12. Model accuracy performance in the second study.


(a) Designed Network

(b) VGG16 Network

(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-13. Confusion matrix of network proposed.

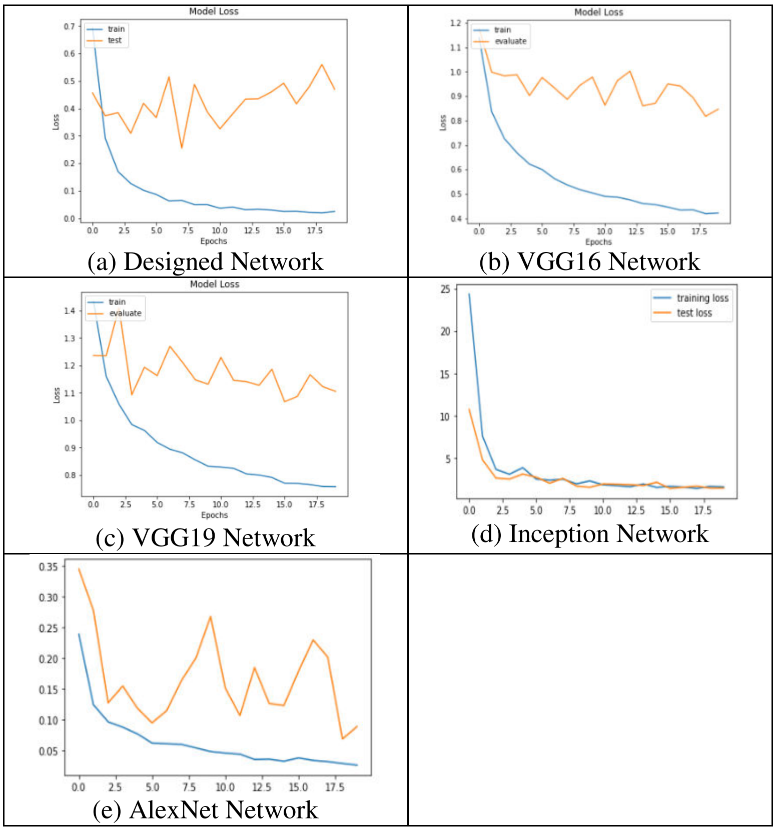
Table-2. Result of training and validation accuracy for different model in second study.

EPOCH	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**99.77**	**90.12**
		VGG16	87.88	68.52
		VGG19	80.71	64.81
		Inception	41.37	33.83
		AlexNet	90.33	76.76

In the second study, which contained a dataset of disorder patients, the validation accuracy for the designed network was lower compared to the first study shown in Table-2. This is because the dataset of the first study only included data from normal and healthy persons. Stroke patient validation accuracy is poorer than healthy individuals because they have difficulty pronouncing vowels throughout the recording process. Some of them have trouble pronouncing particular vowel families. Vowels /e/ and /i/ frequently have comparable sounds, making training data sets difficult. Both healthy individuals and stroke patients are included in the dataset for the second and third analyses. During the training process, we may see the model accuracy of the designed model and the other networks' models. The designed training had higher accuracy and required fewer layers through each layer than the other networks. A comparison of the designed and other models' training and validation accuracy graphs is shown in Figure 12.

Model comparison of normal person mixed with disorder patients classification (two groups)

The third study combines the dataset of a normal person and a disordered patient. With a total of 12420 spectrogram images, the process took 80% of training with 9936 spectrogram images and 10% of evaluation with 1242 images. The third study is conducted to observe the performance of the combination dataset in a normal person and a disordered patient. From the experiment shown in Study 1 and Study 2, we can conclude that the designed model has the highest accuracy compared to others. The carried method of training is the same as the first study with an epoch of 20 and batch size of 6.



(a) Designed Network

(b) VGG16 Network

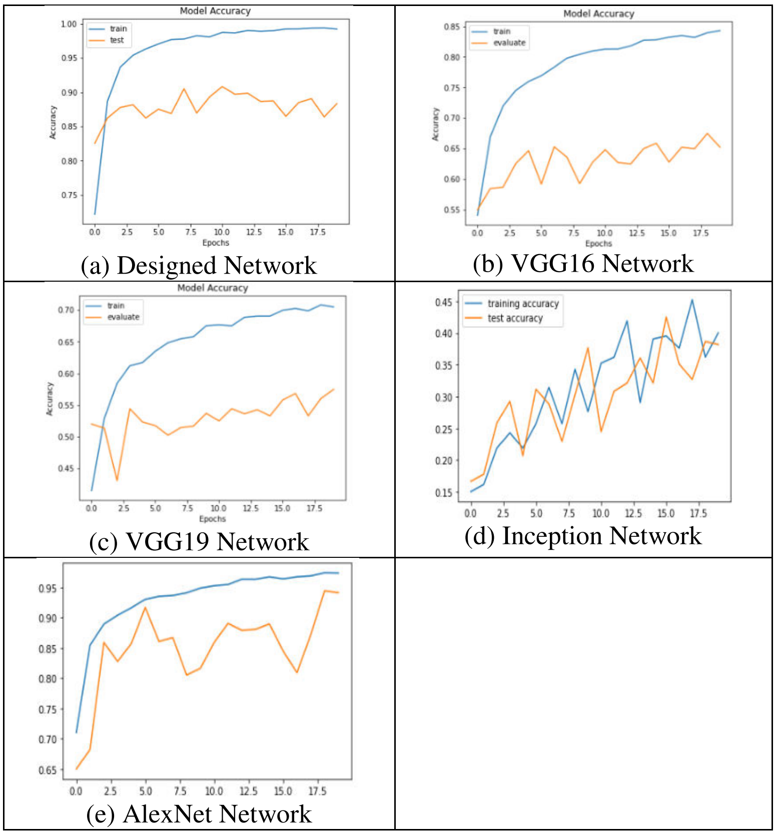
(c) VGG19 Network

(d) Inception Network

(e) AlexNet Network

Figure-14. Model loss performance in the third study

The third study shows the accuracy for the designed model is the highest at 88.33%, followed by the AlexNet model with 84.77%. The Inception model has the lowest accuracy carried out by all the studies, with 30.19%. Deep neural networks that perform in Inception must be immense. A neural network requires numerous additional network layers and units inside those layers to be categorized as significant.

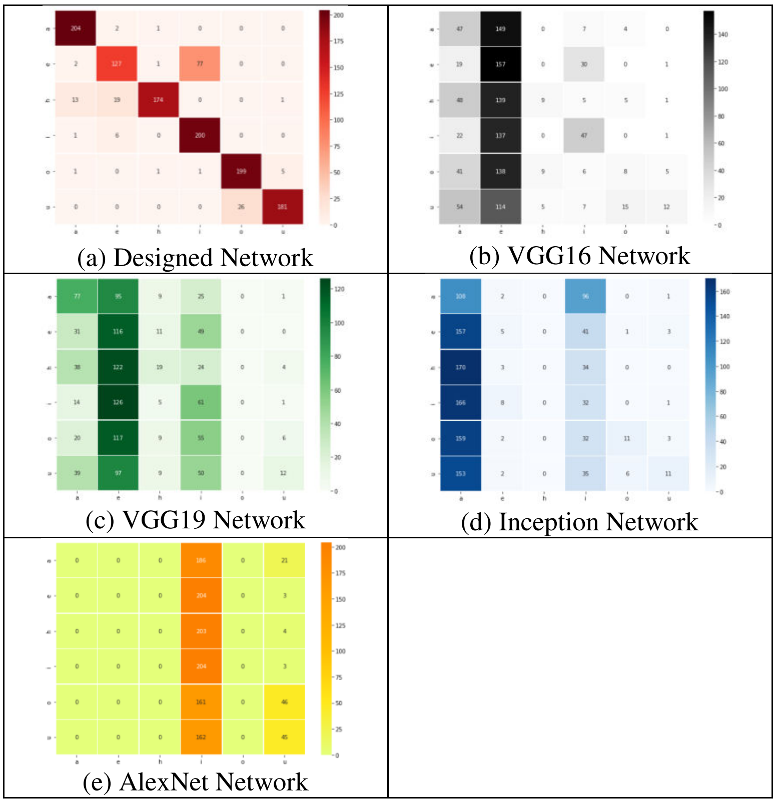


(a) Designed Network (b) VGG16 Network

(c) VGG19 Network (d) Inception Network

(e) AlexNet Network

Figure-15. Model accuracy performance in the third study.



(a) Designed Network (b) VGG16 Network

(c) VGG19 Network (d) Inception Network

(e) AlexNet Network

Figure-16. Confusion matrix of network proposed in the third study.

Table-3. Result of training and validation accuracy for different models in third study.

EPOCH	Batch Size	Model	Accuracy Percentage	
			Training (%)	Validation (%)
20	6	Designed	**99.97**	**88.33**
		VGG16	84.33	65.22
		VGG19	72.93	57.49
		Inception	31.20	30.19
		AlexNet	93.09	84.77

CONCLUSIONS

In conclusion, choosing suitable deep models is crucial since they must fit every data set and determine how well each analysis performs. The developed model must be subjected to data augmentation, dropout, and fine-tuning to prevent over-fitting. If the suitable model is selected and the proper adjustments are performed, deep models can fit a relatively small amount of data. The project has chosen to use the designed model as its model. It verifies its accuracy in all studied data sets and has the most outstanding validation accuracy. Using the suggested network and four comparison network models, VGG16, VGG19, AlexNet and Inception, we presented various models with different accuracy. The article provides a comprehensive analysis of batch size, period sizes, a variety of classes, and differences for a thorough understanding of the distinction of the vowel, especially for disorder patients. We trained, validated, and tested for Malay language vowels' spectrogram images using a straightforward CNN model.

All the outcomes from the first study, which contained datasets from normal person, the disorder patient dataset of the second study and the third study with the mixed dataset of the normal and patients have been documented, and a comparison of the outcomes shows that the designed model is more precise than another model conducted in every study. However, every analysis shows good accuracy and the data set of a normal person and stroke patient is accurate. The goal has, in general, been accomplished. In the future, we plan to investigate the efficacy of converting the sound to another image profile, the Mel-Frequency Cepstral Coefficient (MFCC) and observe how it works. This may aid us in gauging the value of using another image format for the study. Upcoming, creating and testing the more complicated network model in a comprehensive and varied experimental environment is possible.

ACKNOWLEDGEMENTS

We would also want to express our gratitude to the Perkeso Rehab Center in Ayer Keroh, Melaka and Universiti Teknikal Malaysia Melaka for their provided funding and assistance in making this initiative a success for this research.

REFERENCES

[1] R. D. Hayes, A. Begum, David T. 2012. Functional Status and All-Cause Mortality in Serious Mental Illness. PLoS ONE. 7(9).

[2] S. Ayers, Carrie D. Liewellyn, Cambridge Handbook of Psychology, Health and Medicine: Third edition, Cambridge: Cambridge Handbooks in Psychology, 2019.

[3] S. Wortman-Jutt, D. Edwards. 2019. Poststroke Aphasia Rehabilitation: Why All Talk and No Action? Neurorehabilitation and Neural Repair. 33(4): 235-244.

[4] C. Code. 2021. Contemporary Issues in Apraxia of Speech. Aphasiology. 35(4): 391-396.

[5] Waber D. P., Boiselle E. C., Yakut A. D., Peek C. P., Strand K. E. & Bernstein J. H. 2021. Developmental Dyspraxia in Children with Learning Disorders: Four-Year Experience in a Referred Sample. Journal of Child Neurology. 36(3): 210-221.

[6] N. Narasimhan. 2019. Vowel Space Area in Speech of Children with Hearing Impairment. International Journal of Health Sciences & Research. 9(8): 97-102.

[7] Hashim N. M. Z., Zahri N.A.H., Latif M.J.A., Hamzah R.A., Hashim N.F., Kamal M., 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. 5(1301-1318): 100.

[8] N. Amir, O. Tzenker, O. Amir, J. Rosenhouse. 2012. Quantifying Vowel Characteristics in Hebrew and Arabic. Afeka Conference for Speech Processing.

[9] S. P. Rosenbaum 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.

[10] S. Hershey, Shawn Chaudhuri. 2017. CNN Architectures for Large-scale Audio Classification. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 131-135.

[11] N. Milosevic. 2020. Introduction to Convolutional Neural Networks. Introduction to Convolutional Neural Networks. pp. 1-31.

[12] X. Liang, Ming Hu. 2015. Recurrent Convolutional Neural Network for Object Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015(1): 3367-3375.