Speech Emotion Recognition for Autism Spectrum Disorder using Deep Learning

MD Gunathilake[1] and GAI Uwanthika[1]

[1]*Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka*

36-cs-0002@kdu.ac.lk

Abstract:*Children who belong to autism spectrum disorder have difficulty identifying emotions and expressing their emotions. Because it is hard to identify the emotions like anger, disgust, fear, happiness, neutral, sad, and surprise in other people and themselves. This can be even more severe when it could not be found at the beginning and may lead to impairment of social communication of the child. Through the proposed systematic methodology child can identify their basic emotions and try to express them. This evolved methodology was developed using python language. For emotion recognition used a deep machine learning model like Recurrent Neural Network (RNN) using Keras with a TensorFlow backend. RNN consists of four layers with two long short-term memory (LSTM) layers. To optimize the performance of the model used Adam optimizer. For the training and testing of the model used online available data. For the classification of the emotion's valuable features of the audio signal like Zero Crossing Rate (ZCR), Chroma STFT, Mel- Frequency Cepstral Coefficient (MFCC), Root Mean Square (RMS) value, and Mel spectrogram are extracted using the python libROSA library. Due to the lack of the data amount and GPU requirements model's performance can be decreased. This model performed well with the TESS data corpus with 91% test accuracy.*

Keywords: *speech emotion recognition, autism spectrum disorder, ZCR, Chroma STFT, MFCC, RMS, Mel spectrogram*

1. Introduction

Autism Spectrum Disorder is a neurological disorder that mainly affects children's development. Children may show less communication and social interaction, restricted interest, repetitive behaviours, and mannerisms due to this disorder ("Psychiatry.org - What Is Autism Spectrum Disorder?"). It has been widely known that the connection between the brain and behaviour neurological disorder is affected in those who have autism spectrum disorder. The disorder may affect the brain's structure and function, which may lead to the symptoms. There is no cure for autism spectrum disorder, but there are treatments to help manage and improve the symptoms ("Autism spectrum disorders,"). Historically, autism was thought to be caused by bad parenting, but today it is founded that autism is a complex disorder. It is not one thing, but a group of symptoms that can only be understood in the context of the person who experiences them. Over the past few decades, discovered a lot about autism and have begun to understand that it is not a single disorder, but a group of related disorders that share symptoms, diagnosis, and treatment. One of the primary causes of autism spectrum disorder is the environment in which a child is raised. It is a complex condition that requires multiple interventions to help improve a child's social and communication skills. Because of the complexity of ASD, effective treatment requires a multi-component approach. It is estimated that 1 in 68 children in the United States has autism spectrum disorder, making it one of the most common disorders among children ("Autism spectrum disorders,"). Many parents and educators have become concerned about the rise of autism spectrum disorder, especially among boys, and have called for greater research on the causes and potential ways to prevent it.

Once a child is diagnosed with ASD, parents are faced with the difficult task of finding the best ways to help their children. The first step to helping a child with ASD is to find the right specialist. This can often be frustrating because there is a wide variety of specialists that are available to help. Special people in this field like paediatricians and speech therapy pathologists help people with ASD to overcome challenges in their communication skills, while others specialize in behaviour modification, helping individuals with ASD to manage their behaviour so that they can function in a classroom or work environment. When considering a country like Sri Lanka, where there is a high prevalence of autism spectrum disorders, it is crucial to have a system in place that is designed to help individuals with ASD. One of the most difficult things in our country is that there is a lack of services for individuals with ASD ("Autism spectrum disorders (ASD) in Sri Lanka,"), making it difficult for those with the disorder to get help and support that they need. This often causes families to take a do-it-yourself approach to help their children, which can be ineffective and cause more harm than good. One of the most effective ways to help an individual with ASD is to provide them with the professional support that they need.

Today, some researchers turn to advanced technology, like machine learning, to detect ASD in children because some symptoms are hard to detect by the parents especially in young children's because they may not have obvious symptoms, making it difficult to know when to refer the child to a specialist. Therefore, new technology, like

machine learning, can help diagnose ASD earlier, improving access to care for those with the disorder. Machine learning has been used to detect ASD in children with high sensitivity and specificity, but to do so, the algorithm needs a large amount of data.

This paper is structured as follows: Section 2 describes the related works which cover the research topic, section 3 presents the methodology, section 4 presents the conclusion, and section 5 presents the future work.

2. Related Works

Most of the previous works are based on speech emotion recognition (SER) but some are based upon automatic speech recognition (ASR) to detect the autistic symptoms of the child's speech. Works based on the ASR are still in a developmental stage and there are still many challenges to be solved in this direction. The main challenges are, firstly, automatic speech recognition is still a new field of study and there is still much to be learned from the existing works. Secondly, such kind of speech recognition can only be performed when the software has enough background information about the language and the speaker. In this paper, speech emotion recognition is addressed.

The researchers (Rouhi et al., 2019) have developed a website to identify the emotions such as happiness, sadness, anger, and neutrality of ASD children with two. The first part is for learning and the second part is for checking the child's performance. Then the child must express the assigned emotion using the tone of voice. To classify this emotion first audio is extracted to avoid the noise contained in the audio. Then extracted the MFCC (Mel-frequency Cepstral Coefficient). Then those extracted features are passed to the random forest classifier to classify the emotion in the audio. This model achieved an average accuracy of 72% over five independent runs. The specialty of this model is that can predict emotions in more than one language. Another study (Matin and Valles, 2020) developed a speech recognition model which identifies emotions in social interactions. Python LibROSA library s used to extract the features from the recordings. After extracting MFCC Zero Crossing Rate (ZCR) extracted because it increases the test accuracy of the utterance when noise is contained in audio. Used grid search for the SVM model and 48,000 Hz frequency as the native sampling rate to process the utterances. Achieved test accuracy of 77%. But overfitting due to the lack of audio recording used to train the model. Another study developed a tool to identify a youngster's proper emotion while the child is speaking (Welarathna et al., 2021). The children's input audio stream was normalized into a specified range, sub-framed into 2s duration for language-independent, noise reduction, and age independence features, and the most effective 40 audio characteristics were extracted. Even in an uncontrolled setting, the CNN-based model distinguishes eight different emotions; sad, disgust, surprise, neutral, happy, calm, fear, and furious with an accuracy matrix of F1 score of 0.90. The trained model is capable of handling tiny frequency fluctuations in categorizing emotions.

The game developed by researchers (Heni and Hamam, 2016) is a different kind of research because it identified emotion via facial emotion recognition by analysing the facial expressions of the user. After that voice recognition is used. This educational game "Worlds of Kids" identify mobile users' emotions via facial detection to extract the best appropriate and favourable game. These game developers paid more attention to the user interface because these designs are for ASD children. The game evolves with emotion recognition, ASR which integrates the deviation in emotional sounds with three distinctive degrees of intensity. After the notion of synchronization presents the emotional speech, facial emotion recognition. This game can measure emotions with an average accuracy of 87%.

However, the current works use advanced ASR software and machine learning techniques to detect the symptoms of the child with ASD. The researchers use deep learning methods, such as the convolutional neural network, to achieve the detection of speech emotion and fidgeting gestures. The biggest challenge in this area is to detect autism while the child is communicating with the other person in a normal way. Some use machine learning techniques to detect autism while the child is communicating in the normal way. Following research used ASR to identify the symptoms of autism.

A two-stage system was developed to identify articulation disorders and learning assistive systems for acoustic children (G Pillai and Sherly, 2017). Articulation disorder is based on the 14 Malayalam vowels. Deep encoder used for pre-training with several features including MFCC, ZCR, etc. LASAC is used to provide speech training, speech analysis, and articulation tutorial. This system achieved 65% of accuracy with the autistic dataset. Tianyan Zhou et al., have developed an automated assessment framework to assist clinicians in quantifying typical prosody related to ASD. Used SVM to extract utterances level large dimensional acoustic features using OpenSmile toolkit. Then used a deep neural network to model the typical prosody label from the speech spectrogram. This system can predict the atypical prosody score for young children with the severity of ASD. Another study developed (Pawar et al., 2017) Automatic analysis of the LENA recordings system which can classify the recording of the people who are suffering from autism which are recorded in controlled home and clinical environments to child and adult vocalization. For pre-processing of the recordings Hamming windowing method is used. MFCC low-level features like deltas, double-deltas, pitch statistics, and a few additional features are extracted in the feature extraction. For classification Support Vector Machine (SVM) was used. This vocalization detector can detect both child and adult vocalizations at high precision and recall. To identify the speech deficiency of ASD children another study

developed a machine learning-based automatic speech analysis tool for the Sinhala language (Wijesinghe et al., 2019). The first stage of this system utilizes thresholding for silence detection and vocal activity detection. The recurrent algorithm is used to segment long audio files. Vocal filtration is applied to the segment labels which are not silent to relabel vocal audios as vocal and non-vocal audios as noise. After the first stage to classify the utterance into seven categories used CNN which gave 79% training accuracy and 78% testing accuracy. Also, provide high precision accuracy of 86%. Then these classified utterances send to another CNN network to check the presence of the autism traits. This gave 90% of training accuracy, and 72% testing accuracy but the average precision nearby 58% due to the limited training dataset. A special Cognitive-based intelligent learning assistant was developed to provide suitable courseware by identifying the child's specialty by analysing the behavioural patterns of the child (Vijayan et al., 2018). To analyse a child's specialty the author not only considers the child's real-time responses but also behavioural, medical, genotype, and phenotype data with brain and facial images. This system is based on the deep learning and prediction method. In the input, a layer chatbot is used to interact with the child either directly or in speech. Then using ASR is used to convert this speech to text and NLP is used to process text data. For images, classification can be done with high accuracy using a Convolutional neural network. Therefore, the brain and facial images which are taken using visual aid are classified using CNN and R-CNN (Regional Convolutional Neural Network) respectively. To suggest the courseware for the child author used reinforcement learning and a deep Learning algorithm. Predicting the autism severity through speech recordings is another accomplishment achieved in this area. The researchers (Eni et al., 2020) used voice recordings of Hebrew-speaking children who completed an Autistic Observation Schedule (ADOS) evaluation to extract arrange of prosodic, acoustic, and conversational aspects. The recordings of 72 youngsters yielded sixty characteristics, 21 of which were strongly linked with the children's ADOS scores. This was developed using multiple DNN methods to predict ADOS scores based on these variables and compared their performance to linear regression and support vector regression models. Among them, the convolutional algorithm produces the best results. When trained and evaluated on several subsamples of available data, this method predicts ADOS scores with a mean RMSE of 4.65 and a mean correlation of 0.72 with genuine ADOS scores. Automated algorithms that can reliably and sensitively predict ASD severity have the potential to revolutionize early ASD detection, symptom severity measurement, and treatment efficacy evaluation.

3. Methodology

This research involves developing a speech emotion recognition model to detect Autism by using children's emotional speeches. By taking the real-time recording of the child who is capable of verbal communication for emotion according to their preferences analyse whether they can express the feelings correctly within the given time.

A. Dataset

For the training and testing purpose of the model used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Toronto Emotional Speech Set (TESS). RAVDESS contains 1,440 audio files in the carrier phase “Kids are talking by the door,” and “Dogs are sitting by the door” of 24 actors (12 male actors and 12 female actors) and 60 trials from each actor. Phrases are pronounced in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgusted expressions. There are two emotional intensity levels (normal and strong) and one neutral expression created for each expression. The audio was captured in WAV format at a sample rate of 48,000 Hz and a bit depth of 16 bits. TESS contains a set of 200 target words in the carrier phrase "Say the word _," and recordings of the set evoking each of the seven emotions were created (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are a total of 2800 data points (audio files). All the audio files are in WAV format.

B. Pre-processing and Feature extraction Pre-process the data set using data augmentation methods such as noise injection and changing the pith of the data. The objective is to make our model resistant to these perturbations and increase its generalizability. The label from the initial training sample must be preserved when adding the perturbations for this to work. When implementing this work into a real-world scenario the audio file may contain some Gaussian noise which should be removed from the audio file. Therefore, framing is essential for audio pre-processing. The speech signal is continuously changing over time. Splitting the audio signal allows the audio signal to become static. Usually, this audio frame is 20-30ms long, but this can depend on the author of the system. These frames are adjacent to each other to ensure no loss. This online available data set is already framed using necessary algorithms. Therefore, we only apply noise injection and change the pitch of the data as pre-processing steps.

After performing the data augmentation feature extraction took place. This model is not able to process audio files directly therefore instead of the whole bunch of the audio files it extracted some valuable features of the audio signal and then trained the model. Features can be identified through the different relationships of the audio files. We extracted features like Zero Crossing Rate (ZCR), Chroma STFT, Mel- Frequency Cepstral Coefficient (MFCC), Root Mean Square (RMS) value, and Mel spectrogram. for the extraction of these features used the python libROSA library. ZCR is the speed at which a signal switches from one sign to another, from positive to negative, or vice versa. Both voice recognition and the retrieval of music information have made extensive use of this characteristic. Typically, highly percussionist sounds, like those found in metal and rock, have higher values (Doshi, 2019). MelFrequency Based on the linear cosine transform of the

log power spectrum on a nonlinear Mel scale frequency, cepstrum is a representation of the short-term power spectrum of a sound. MFCC is a valuable feature in ASR implementations (Paulin et al., n.d.). Chroma STFT The Chroma esteem of a sound essentially speaks to the escalation of the twelve pitch classes that are utilized to ponder music. They can be utilized within the separation of the pitch course profiles between sound signals. Mel spectrograms visualize the visual representation of the frequency of the given signal.

C. Classification

For the classification of each emotion, we used Recurrent Neural Network (RNN) model which is effective for long sequence data or natural language. This model is a fully connected four-layer neural network. All the extracted features are saved and after that normalized and split the data set for training and testing. Speech is classified into seven categories such as anger, disgust, fear, happiness, neutral, sad, and surprise. The model used a fully connected four layers neural network. Added two LSTM (Long Short Term Memory) layers for the first layer and the first hidden layer. Based on accessible runtime equipment and limitations, this layer will select diverse executions (cuDNN-based or pure-TensorFlow) to maximize the execution. On the off chance that a GPU is accessible and all the contentions to the layer meet the prerequisite of the cuDNN bit, the layer will utilize a quick cuDNN usage. There are two hidden layers which consist of 128 neurons and one layer used the ReLu activation function. For the output, the layer user has seven neurons because emotions categorize into seven categories. Therefore, we used SoftMax as the activation function of the output layer. There is a Dropout layer with 0.3 dropouts in between the last hidden layer and the output layer. Dropout could be a regularization technique for neural organize models. It may be a strategy where haphazardly chosen neurons are overlooked amid preparing. They are "dropped out" haphazardly. This implies that their commitment to the enactment of downstream neurons is transiently evacuated on the forward pass and any weight upgrades are not connected to the neuron on the reverse pass. For the optimization used Adam optimizer for the model. Adam is an optimization calculation that can be utilized rather than the classical stochastic slope plummet strategy to overhaul organized weights iterative based on training information.

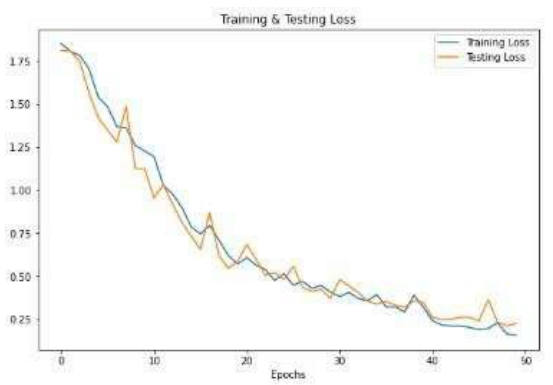

Figure 1: Model loss for Tess training and testing data

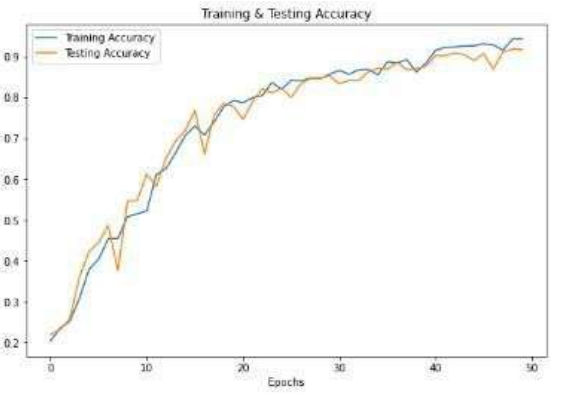

Figure 2: Model accuracy for Tess training and testing data

D. Results

After setting up and training the model when it tested for the testing data the accuracy of it is for the TESS data corpus is 91% with a 0.25 loss percentage after fifty epochs.

4. Discussion

When we applied the same model to the RAVDESS data corpus the accuracy of the model decreased by 50%. Then we go with the CNN model then the accuracy increased but it is not satisfactory. These arose because of the amount of the corpus as well as the performance of the model. Therefore, we only consider the TESS data corpus. For the testing of the model, we took the random amount of data set from the data set, because of the lack of data in our country and the difficulty in collecting data. This proposed methodology applies to any autistic child who can verbally communicate.

5. Conclusion

Autistic children have poor communication skills. Some are unable to speak or understand the language. Others have difficulty reading social cues and expressing themselves. Many cannot interact with other people or follow a conversation. They often miss cues and do not understand others' emotions or intentions. Therefore, the proposed system used emotional intelligence to help autistic children communicate. Through this model, we can help autistic children to get a basic understanding of their emotions and how to express them correctly. The model can get the 91% of accuracy for emotion recognition. When we consider the CNN model for the classification the accuracy of the classification is not efficient and for the RAVDESS data set it gets decreased which is not efficient. But this model should train using more data to identify some other emotions and to optimize the performance of the model.

6. Future Works

As for the future works, we hoped to get the real-time recording of the child and then identifying articulation

disorder like blabbering, neologism, and echolalia of it by CNN based model and then developed a speech recognition bases therapy system to give speech therapy to according to the level of the autism, to overcome these disorders.

References

Autism spectrum disorders (ASD) in Sri Lanka (no date) *Sri Lanka Foundation*. Available at: https://www.srilankafoundation.org/newsfeed/autism-spectrumhttps://www.srilankafoundation.org/newsfeed/autism-spectrum-disorders-asd-in-sri-lanka/disorders-asd-in-sri-lanka/ (Accessed: 14 January 2022).

Autism spectrum disorders (no date). Available at: https://www.who.int/news-room/fact-sheets/detail/autismhttps://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disordersspectrum-disorders (Accessed: 14 January 2022). Brownlee, J. (2016) 'Dropout Regularization in Deep Learning Models With Keras', *Machine Learning Mastery*, 19 June. Available at: https://machinelearningmastery.com/dropouthttps://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/regularization-deep-learning-models-keras/ (Accessed: 9 July 2022).

Brownlee, J. (2017) 'Gentle Introduction to the Adam Optimization Algorithm for Deep Learning', *Machine Learning Mastery*, 2 July. Available at: https://machinelearningmastery.com/adam-optimizationhttps://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/algorithm-for-deep-learning/ (Accessed: 9 July 2022).

Doshi, S. (2019) *Extract features of Music*, *Medium*. Available at: https://towardsdatascience.com/extract-features-of-musichttps://towardsdatascience.com/extract-features-of-music-75a3f9bc265d75a3f9bc265d (Accessed: 8 July 2022).

Eni, M. *et al.* (2020) 'Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network', *IEEE Access*, 8, pp. 139489–139500.

G Pillai, L. and Sherly, E. (2017) 'A Deep Learning Based Evaluation of Articulation Disorder and Learning Assistive System for Autistic Children', *International Journal on Natural Language Computing*, 6(5), pp. 19–36.

Heni, N. and Hamam, H. (2016) 'Design of emotional educational system mobile games for autistic children', in *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, Monastir, Tunisia: IEEE, pp. 631–637.

Matin, R. and Valles, D. (2020) 'A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions', in *2020 Intermountain Engineering, Technology and Computing (IETC)*. *2020 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA: IEEE, pp. 1–6.

Paulin, H., Milton, D. R. S. and Janakiraman, D. S. (no date) 'Efficient Pre Processing of Audio and Video signal dataset for building an efficient Automatic Speech Recognition System', p. 8.

Pawar, R. *et al.* (2017) 'Automatic analysis of LENA recordings for language assessment in children aged five to fourteen years with application to individuals with autism', in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Orland, FL, USA: IEEE, pp. 245–248.

Psychiatry.org - What Is Autism Spectrum Disorder? (no date). Available at: https://psychiatry.org:443/patientshttps://psychiatry.org/patients-families/autism/what-is-autism-spectrum-disorderfamilies/autism/what-is-autism-spectrum-disorder (Accessed: 26 April 2022).

Rouhi, A. *et al.* (2019) 'Emotify: emotional game for children with autism spectrum disorder based-on machine learning', in *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. *IUI '19: 24th International Conference on Intelligent User Interfaces*, Marina del Ray California: ACM, pp. 31–32.

Team, K. (no date) *Keras documentation: LSTM layer*. Available at: https://keras.io/api/layers/recurrent_layers/lstm/ (Accessed: 9 July 2022).

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" by Livingstone & Russo is licensed under CC BY-NA-SC 4.0

Velayudham, V. (2020) 'Audio Data Processing— Feature Extraction — Science & Concepts behind them', *Analytics Vidhya*, 2 May. Available at: https://medium.com/analyticshttps://medium.com/analytics-vidhya/audio-data-processing-feature-extraction-science-concepts-behind-them-be97fbd587d8vidhya/audio-data-processing-feature-extraction-sciencehttps://medium.com/analytics-vidhya/audio-data-processing-feature-extraction-science-concepts-behind-them-be97fbd587d8concepts-behind-them-be97fbd587d8 (Accessed: 8 July 2022).

Vijayan, A. *et al.* (2018) 'A Framework for Intelligent Learning Assistant Platform Based on Cognitive Computing for Children with Autism Spectrum Disorder', in *2018 International CET Conference on Control, Communication, and Computing (IC4)*. *2018 International CET Conference on Control, Communication, and Computing (IC4)*, Thiruvananthapuram: IEEE, pp. 361–365.

Welarathna, K. T. *et al.* (2021) 'Automated Sinhala Speech Emotions Analysis Tool for Autism Children', in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, Negombo, Sri Lanka: IEEE, pp. 500–505.

Wijesinghe, A. *et al.* (2019) 'Machine Learning Based Automated Speech Dialog Analysis Of Autistic Children', in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam: IEEE, pp. 1–5.

Wijesinghe, A. *et al.* (2019) 'Machine Learning Based Automated

Speech Dialog Analysis Of Autistic Children', in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam: IEEE, pp. 1–5.

Zhou, T. *et al.* (no date) 'An Automated Assessment Framework for Speech Abnormalities related to Autism Spectrum Disorder', p. 5.

Authors Biography

MD Gunathilake is pursuing a BSc (Hons) in Computer Science at General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka. Her primary areas of research interest are Machine Learning, Natural Language Processing, and Image Processing.

GAI Uwanthika received a BSc(sp) in Computer Science and Technology degree from Uva Wellassa University and MSc in Computer Science degree at University of Peradeniya. She has been awarded a merit scholarship for the performance in course work of the MSc in Computer Science 2017/ 2018. Her research interests include Bioinformatics, Deep Learning and Digital Image Processing.