VOL. 20, NO. 2-3, 2021, 1-6
www.elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

Speech-Based Depression Detection for Bahasa Malaysia Female Speakers Using Deep Learning

Mugahed Al-Ezzi Ahmed Ezzi [1], **Nik Nur Wahidah Nik Hashim** [1*], **Nadzirah Ahmad Basri** [2] and **Siti Fauziah Toha** [1]

[1] Department of Mechatronics, Faculty of Engineering, International Islamic University Malaysia, Selangor, Malaysia.
[2]Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, Malaysia.

[*]Corresponding author: niknurwahidah@iium.edu.my, Tel: 603-6421 6564

Abstract: Depression is a mental disorder of high prevalence, leading to a negative effect on individuals, society, and the economy. Traditional clinical diagnosis methods are subjective and require extensive participation of experts. Furthermore, the severe shortage in psychiatrists' ratio per population in Malaysia imposes patients' delay in seeking treatment and poor compliance to follow-up. Besides, the social stigma of visiting psychiatric clinics also prevents patients from seeking early treatment. Automatic depression detection using speech signals is a promising depression biometric because it is fast, convenient, and non-invasive. This research attempts to develop an end-to-end deep learning model to classify depression from female Bahasa Malaysia speech using our dataset. Depression status was identified by the Patient Health Questionnaire 9, the Malay Beck Depression Inventory-II, and subjects' declaration of Major Depressive Disorder diagnosis by a trained clinician. The dataset consists of 110 female participants. We provided a detailed implementation of deep learning models using raw audio input. Multiple combinations of speech types were analyzed using various deep neural network models. After performing hyperparameters tunning, raw audio input from female read and spontaneous speech combination using AttCRNN model achieved an accuracy of 91%.

Keywords: Artificial intelligence, Deep learning, Depression detection, Mental health, Speech analysis.

Article History: received 25 May 2021; accepted 12 June 2021; published 15 October 2021

1. INTRODUCTION

One of the most common mental illnesses is depression. More than 300 million people of all ages are thought to suffer from depression worldwide [1]. As a result, the affected individual can perform poorly at work, school, and within the family. Depression can lead to suicide in the worst-case scenario. In Malaysia, depression is the most common mental disorder [2]. From 10.7% in 1996 to 11.2% in 2006 to 29.2% in 2015, the prevalence of mental health disorders has gradually risen [3]. According to the World Health Organization's Global Health Observatory data repository, Malaysia had just 1.05 psychiatrists per 100,000 people in 2016 [4]. According to a more recent survey, Malaysia had 410 registered psychiatrists in 2018, or 1.27 psychiatrists per 100,000 people. WHO recommends a psychiatrist-to-population ratio of 1:10,000 in Malaysia. The current ratio, however, is just 1:80,000 [5]. The country's severe shortage of psychiatrists may cause a slew of problems for those suffering from mental illnesses. Delays in seeking help, lengthy wait times for medical consultations, low-quality outpatient mental health services, inadequate follow-up and treatment compliance, increased substance abuse and addiction prosecutions, a rise in suicide rates, unemployment, and homelessness are just a few of these problems.

Early intervention to delay the onset of clinical depression may be a powerful tool for lowering the disease's burden. However, the number of diagnostic methods available for detecting depression is currently extremely limited. Patient self-reporting and professional judgement are used almost exclusively in assessment processes, which can lead to a number of subjective biases. Hence, it's important to seek out new objective approaches that can assist physicians in diagnosing and monitoring clinical depression [6].

Scientific studies have proven the relationship between depression and human speech. Prosodic irregularities are connected with a person's mental state of depression. Depressive voice can be described as low, slow, hesitative, monotonous, stuttering, and whispering. Phonological loop and neuromuscular motor control impairments associated with depression cause phonation and speech errors, resulting in these abnormalities. As a result, speech is a responsive output system; even minor physiological and cognitive changes can result in audible changes. [7][8][9] [10].

The utilization of speech signal processing for creating various tools is an essential aspect for cognitive info-

communication [11]. This fascinating research field of developing non-invasive diagnostic biomarkers using speech signals makes automatic depression detection possible. Speech-based depression detection is fast and easy to access by a broader range of people.

In recent years, more researchers adopted deep learning for automatic speech-based depression detection. Le Yang applied Deep Convolutional Neural Network (DCNN) [12]. The DCNN is trained using the input audio features and ground-truth labels of PHQ-8 and trained on distress analysis interview corpus (DAIC) [13]. The DAIC dataset contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post- traumatic stress. The dataset consists of 108 recordings. Each recording is labeled with the PHQ-8 value. Yang et al. model reported promising results as it shows a Mean Absolute Error (MAE) of 4.484. Chlasta achieved 77% accuracy in depression detection by training a CNN on the DAIC dataset using speech spectrograms [14]. Multiple pieces of research had used the DAIC dataset for training deep neural networks such as CNN and Long Short-Term Memory (LSTM) neural network [15]–[18]. Besides the DAIC dataset, Seneviratne used a different dataset consist of 200 subjects (124 female and 76 male) to train a CNN model [19]. Table 1 shows the summary of the most recent previous works using deep learning. The summary includes dataset, input type, classification techniques, and performance.

Using convolutional and recurrent neural networks, this paper proposes a novel approach for automated speech-based depression detection. We constructed multiple models that accept raw audio input data. To demonstrate the ability of our classification system and test the results obtained, we present extensive experiments on our dataset.

2. DATASET

2.1 Dataset Collection

An online form has been published to collect the dataset. In this form, participants are asked to answer some personal questions, submit the BDI-II questionnaire (Malay version), PHQ9 questionnaire, a voice recording of them reading a given passage, and voice recordings of them answering three general questions. Additionally, participants are asked to report if they have any psychiatric clinical records related to depression.

The submitted BDI-II and PHQ9 scores of participants with clinical reports will be used as a label for annotating each subject with their depression classification. Moreover, we have adopted two speech types in the collected voice recordings: read speech and spontaneous speech. Read speech fixes the syntax and vocabulary where spontaneous speech leaves the speaker free to choose their own syntax and vocabulary. The effectiveness of these speech types will be investigated. The detail of the number of the collected data is shown in Table 2.

As shown in Table 2 we could not gather many samples from male participants. Thus, we will be focusing on experimenting with female data only. There are two types of female speech data: female read data (FR) and Female Spontaneous data (FS). This research investigated the robustness of depression detection from these two speech types.

2.2 Raw Audio Input Pre-processing

Firstly, speech signals are split into 10 seconds clips. Each clip is to be used as one sample. Secondly, the voice signals amplitude is scaled between -1 and 1. Afterward, audio normalization is performed. Audio normalization is also an essential step used to reduce the recording variability without losing the audio's discriminative strength. By using audio normalization, the neural network model's ability of generalization is increased. There are different normalization types; however, the most widely used normalization method is the z-normalization (standard score). If μ is the mean of x audio signal and σ is the standard deviation, z-normalization can be expressed as in Equation (1).

Table 1. Summary of The Collected Dataset

Gender	Type of speech	Depression Status		Total
		Depressed	Normal	
Female	Read	42	68	110
	Spontaneous	42	69	111
Male	Read	11	9	20
	Spontaneous	11	10	21

Table 2. Summary of Recent Works on Speech Depression Detection Using Deep Learning

Author	Year	Dataset			Input Features	Model	Performance
		Size	Type	Language			
Le Yang [12]	2017	189	Spontaneous	English	238 LLDs	CNN	MAE 4.484
Karol Chlasta [14]	2019	107	Spontaneous	English	Spectrogram	CNN	Accuracy 77%
Emna Rejaibi [15]	2020	189	Spontaneous	English	MFCC	RNN	MAE 4.97
Ziping Zhao [18]	2020	189	Spontaneous	English	Spectrogram	RNN	RMSE 5.51 MAE 4.20
Le Yang [17]	2020	189	Spontaneous	English	238 LLDs	CNN	RMSE 5.520 MAE 4.634
Srimadhur N.S [16]	2020	189	Spontaneous	English	Raw spectrograms	CNN	F-Score 78%
Nadee Seneviratne [19]	2021	200	Spontaneous	English	Vocal Tract Variables MFCC	CNN	Accuracy 91.84%

$$z_{\text{normalization}} = \frac{x-\mu}{\sigma} \quad (1)$$

Lastly, the dataset is divided into 80% for training and 20% for validation. For the validation, we adopted the accuracy matrix. After the model is trained, we run predictions on the 20% of the dataset which were not included in the training and calculated the total number of correct predictions over the total number of the dataset.

3. METHOD

This section will illustrate the different neural network architectures that we implemented in this research study. Each model is trained and evaluated to perform a comparison between multiple types of DNNs. The best model architecture will be further tuned to reach the optimal results. Figure 1 shows the overall workflow procedures of speech depression classification.

3.1 Fully Connected Neural Network Model

Fully connected (FC) neural networks is a feedforward network, and it can be defined as a collection of nodes. This collection of nodes starts with the input layer and terminates at the output layer. Furthermore, they can be one or more hidden layers containing multiple nodes within each layer. Our model is built with five fully connected layers: the input layer, one as output layer, and three as hidden layers. Additionally, we use the dropout layer for all layers to prevent overfitting

3.2 Convolutional Neural Network Model

In our experiment, we used four convolutional (CONV) layers and four global average pooling layers. The average pooling layer reduces each feature map into one float by averaging the activation across the temporal dimension. Additionally, we adopted another type of layer called batch normalization (BN). BN reduces the problem of unbalanced gradients, a common issue in optimizing deep neural network architectures. BN normalizes the output of CONV layer, so the gradients are well-balanced. We applied BN layer to the output of each CONV layer before applying the ReLU activation function.

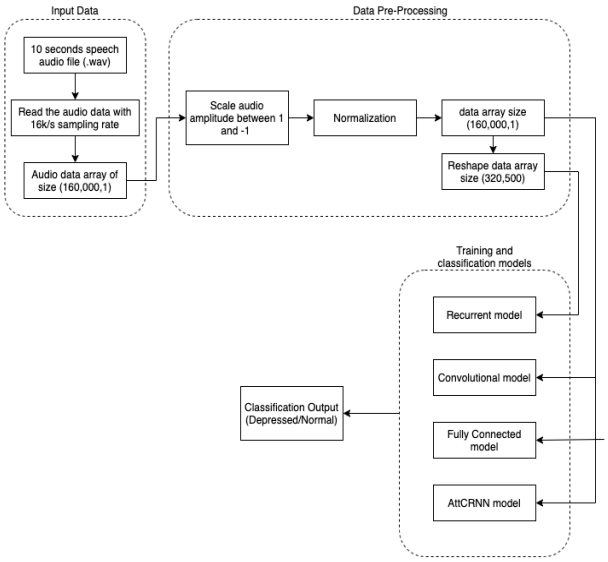

Figure 1. Overall workflow procedures of speech depression classification.

3.3 Recurrent Neural Network Model

In this model we constructed four LSTM layers and two FC layers at the end. Moreover, dropout and BN are applied to the output of each LSTM layer.

3.4 Attention Convolutional Recurrent Neural Network

In neural networks, attention takes two vectors and turns them into a matrix where the values of one vector form the columns and values of another vector form the rows, identifying relative context, which means understanding how some parts of these vectors are related to others. Hence, the neural network would learn to discard the noises and focus on what is relevant and look at the totality of the vector to make connections between any particular point and its relevant context.

Attention Convolutional Recurrent Neural Network (AttCRNN) model is adopted initially by Andrade et al., which is implemented for speech command recognition [20]. However, we are going to modify the adopted network to suit our depression detection objective. The model accepts raw audio data as an input. The following layer is non-trainable layer for Mel-spectrogram extraction. A couple of 2D CONV layers are added after the Mel-spectrogram is computed followed by two bidirectional LSTM layers. A normalization layer is added after each CONV layer. Afterwards, one of the last LSTM layer's output vectors is extracted, projected with a dense layer, and used as the query vector. Finally, for classification, the weighted average of the LSTM output is fed into three FC layers. Figure 2 summarizes the architecture.

4. EXPERIMENTS AND RESULTS

This section discussion includes speech depression detection evaluation results for raw audio input and acoustic features input. Moreover, the impact of the speech type (read and spontaneous) on the classification results was further investigated. We trained the four deep learning

models using read speech and spontaneous speech separately and combined them in a third test. For each test, the best validation accuracy model was saved, and the training stopped if no improvements were made in 30 consecutive epochs. The initial learning rate was set to 0.001 and decay of 0.5 if the accuracy does not increase over ten epochs. The batch size used is eight. Additionally, we have trained all models again after removing the normalization stage to investigate how significant its impact on the classification results.

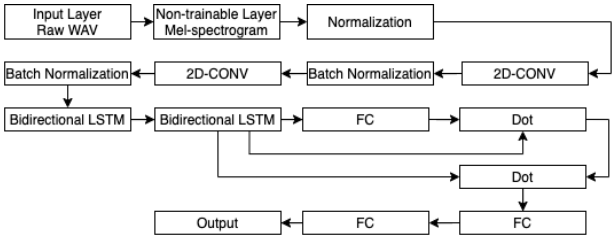

Figure 2: AttCRNN architecture with attention

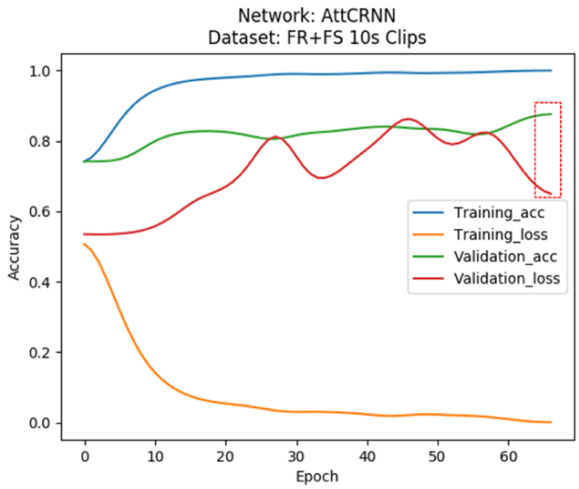

Figure 3. AttCRNN training graph using FR and FS data

Table 3. Training Result Summary

Model / Dataset	Normalization	FC	CNN	RNN	AttCRNN
		Validation Acc.	Validation Acc.	Validation Acc.	Validation Acc.
FR	Yes	0.63	0.83	0.62	0.53
FR	No	0.53	0.66	0.61	0.86
FS	Yes	**0.66**	**0.84**	**0.68**	0.64
FS	No	0.55	0.73	0.65	0.76
FR+FS	Yes	0.53	0.83	0.58	**0.89**
FR+FS	No	0.53	0.71	0.60	0.84

4.1 Raw Audio Input Results

Table 3 shows the training results summary of raw audio input data to the deep learning models. We can see from the table that good performance with high accuracy is achieved in the AttCRNN model with 89%, followed by the CNN model with 84%. In general, speech data normalization has a considerable impact on the overall classification accuracy. Moreover, spontaneous speech has a better indication for depression detection using raw input data. All models, except AttCRNN, have their best accuracies with normalized spontaneous speech. Nevertheless, AttCRNN achieved its best accuracy by combining spontaneous and read speech together after normalization is applied.

However, when the normalization is not applied, read speech has the best accuracy at 86%. Figure 3 and Figure 4 show the training graphs detail for AttCRNN and CNN models, respectively. All graphs are Gaussian smoothed with a factor of three. The optimal point is where the validation loss is minimized and the validation accuracy is maximized. Since the AttCRNN model has the best accuracy, it will be further optimized with hyperparameter tuning.

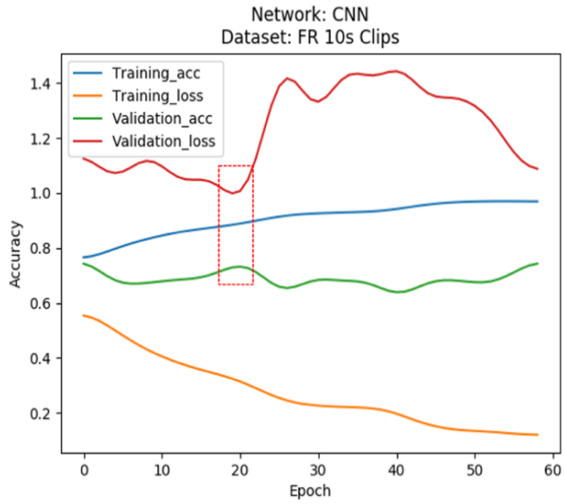

Figure 4. CNN training graph using FR and FS data

4.2 Hyperparameters Tuning

Hyperparameters are the variables that control the training process itself. For example, deciding how many hidden layers to use between the input layer and the output layer, the type of these layers, how many nodes each layer should use, and what learning rate leads to better output result. These variables are not directly associated with the training

data. Meaning that, unlike parameters that change during the training process, hyperparameters are configured before starting the training process, and they remain constant throughout the process. In hyperparameters tuning, we will be running the whole training process, observing the output accuracy, and adjusting the hyperparameters to find the best combination that works the best for depression detection.

We performed hyperparameter tuning only on the model with the best accuracy. Thus, the AttCRNN model with normalized FR and FS data is selected. The hyperparameters we are going to are number of LSTM layers (2, 3, 4), LSTM layer size (64, 128, 256) and dropout rate (0, 0.1, 0.2, 0.3). Hence, 36 combinations are going to be trained, and their result will be discussed.

Table 4. AttCRNN Model Hyperparameter Tuning Results

Model	Speech Type	Number of LSTM Layers	LSTM Size	Dropout	Validation Acc.
AttCRNN	FR+FS	2	64	0	0.76
		2	64	0.1	0.77
		2	64	0.2	0.84
		2	64	0.3	0.73
		2	128	0	0.77
		2	128	0.1	0.81
		2	128	0.2	0.88
		2	128	0.3	0.86
		2	256	0	0.77
		2	256	0.1	0.78
		2	256	0.2	0.77
		2	256	0.3	0.69
		3	**64**	**0**	**0.91**
		3	64	0.1	0.85
		3	64	0.2	0.82
		3	64	0.3	0.85
		3	128	0	0.84
		3	128	0.1	0.81
		3	128	0.2	0.86
		3	128	0.3	0.71
		3	256	0	0.69
		3	256	0.1	0.75
		3	256	0.2	0.75
		3	256	0.3	0.75
		4	64	0	0.79
		4	64	0.1	0.69
		4	64	0.2	0.68
		4	64	0.3	0.78
		4	128	0	0.88
		4	128	0.1	0.73
		4	128	0.2	0.75
		4	128	0.3	0.73
		4	256	0	0.72
		4	256	0.1	0.78
		4	256	0.2	0.72
		4	256	0.3	0.74

Table 4 shows the results of all hyperparameters combinations for the AttCRNN model with normalized FR and FS data input. The combination of three LSTM layers, 64 LSTM size, and zero dropout (AttCRNN_3_64_0) has achieved the highest accuracy at 91%. This is a promising result. It overcomes the trained model before hyperparameter tunning by around 2%. Although it may seem like it is not a significant improvement, hyperparameter tunning proved its effectiveness by showing the considerable variation between the lowest and highest accuracies. From the lowest accuracy at 68 percent to the highest at 91%, a range of improvement by 23% is achieved by only manipulating the neural network hyperparameters. This indicates the importance of hyperparameters tuning. Figure 5 shows the training graphs for AttCRNN_3_64_0 model.

It can be seen from the validation loss graph that the overall training is not stable. That is due to the amount of the dataset is small. Hence, overfitting is happening in an early stage at around epoch 28. The validation loss fluctuation also indicates that the learning rate might be a little too large for this model. However, the training loss and training accuracy show healthy trends, which implies that the model can handle the problem splendidly and learn efficiently. However, when it comes to generalizing the learning to the test dataset, it seems the model tends to overfit.

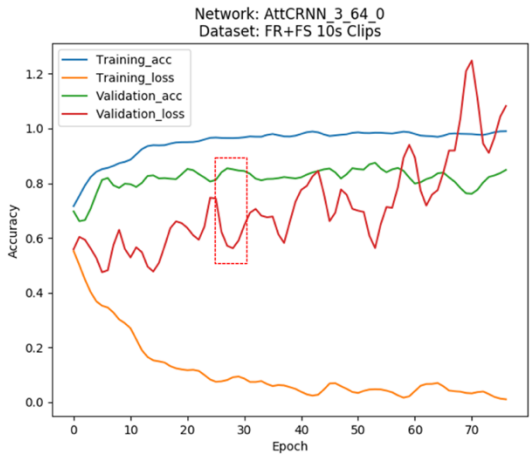

Figure 5. AttCRNN_3_64_0 training graphs using FR and FS data

5. CONCLUSION

This research study investigated the potential of speech-based depression detection for Bahasa Malaysia speakers using different neural network models. This study showed the high potential of using raw input audio data as an input to the system that can accurately detect the depression status. With the utilization of deep learning methods, this is no longer a dream. That means if we put more effort into this research field to achieve a higher performance and robust results, we would no longer need to go through the tedious and intensive acoustic features analysis for different languages or speaker gender. The acoustic features approach has proven its effectiveness in this field; however, it is a time-consuming process. There are vast

variations and types of acoustic features. Extensive analysis needs to be carried out to find which features have a better indication for the speaker's mental state of depression.

However, raw audio input data requires high specification hardware such as a powerful Graphics Processing Unit (GPU) to train the neural network model due to the large size of the input data compared to features input. Furthermore, raw audio input needs a more complex deep neural network architecture that can work as an end-to-end audio features extraction and classification model. Thus, more researches should be carried out to investigate the effectiveness of more DNNs with raw audio input.

Speech type investigations showed that generally, spontaneous speech has a better indication of the mental state of depression in speakers. This is aligned with the theory behind speech production and its relationship with the speaker's mental state. Moreover, it is the closest form of speech to the way psychiatrists use it in their clinics where the patients are interviewed.

For future work, more depression speech data should be collected for this research field. This will allow other researchers to perform benchmarking and improve depression detection further. Moreover, the developed model should be tested against male and female data, and possibilities of developing gender independent and even language-independent models should be further studied. Lastly, DNN models for raw input data should be studied further to develop a better neural network architecture to achieve even higher accuracy.

ACKNOWLEDGEMENT

This work was supported by funding from the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme
(FRGS/1/2018/TK04/UIAM/02/7).

REFERENCES

[1] World Health Organization, "Depression and other common mental disorders: global health estimates," World Health Organization, 2017.

[2] F. Mukhtar and T. P. S. Oei, "A Review on the Prevalence of Depression in Malaysia," *CPSR*, vol. 7, no. 3, pp. 234–238, Aug. 2011, doi: 10.2174/157340011797183201.

[3] Institute for Public Health, *National Health and Morbidity Survey 2015 (NHMS 2015)*, vol. 2. Ministry of Health Malaysia Kuala Lumpur, 2015.

[4] World Health Organization, "GHO | Human resources - Data by country," *World Health Organization*, 2019. https://apps.who.int/gho/data/view.main.MHHRv (accessed Jan. 25, 2021).

[5] N. C. Guan, T. C. Lee, B. Francis, and T. S. Yen, "Psychiatrists in Malaysia: The ratio and distribution," *Malaysian journal of psychiatry*, vol. 27, no. 1, pp. 4–12, 2018.

[6] H. Jiang *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, pp. 39–46, Jun. 2017, doi: 10.1016/j.specom.2017.04.001.

[7] E. Kraepelin, "Manic Depressive Insanity and Paranoia," *The Journal of Nervous and Mental Disease*, vol. 53, no. 4, p. 350, Apr. 1921.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015, doi: 10.1016/j.specom.2015.03.004.

[9] B. Stasak, J. Epps, and R. Goecke, "Elicitation Design for Acoustic Depression Classification: An Investigation of Articulation Effort, Linguistic Complexity, and Word Affect," in *Interspeech 2017*, Aug. 2017, pp. 834–838, doi: 10.21437/Interspeech.2017-1223.

[10] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of Voice Quality Features in Detecting Depression," in *Interspeech 2018*, Sep. 2018, pp. 1676–1680, doi: 10.21437/Interspeech.2018-1399.

[11] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive infocommunications (CogInfoCom)*. 2015.

[12] L. Yang, D. Jiang, W. Han, and H. Sahli, *DCNN and DNN based multi-modal depression recognition*. 2017, p. 489.

[13] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *LREC*, 2014, pp. 3123–3128.

[14] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, Jan. 2019, doi: 10.1016/j.procs.2019.12.228.

[15] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech," *arXiv:1909.07208 [cs, eess]*, Mar. 2020, Accessed: Jan. 25, 2021. [Online]. Available: http://arxiv.org/abs/1909.07208.

[16] N. S. Srimadhur and S. Lalitha, "An End-to-End Model for Detection and Assessment of Depression Levels using Speech," *Procedia Computer Science*, vol. 171, pp. 12–21, Jan. 2020, doi: 10.1016/j.procs.2020.04.003.

[17] L. Yang, D. Jiang, and H. Sahli, "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals," *IEEE Access*, vol. 8, pp. 24033–24045, 2020, doi: 10.1109/ACCESS.2020.2970496.

[18] Z. Zhao *et al.*, "Automatic Assessment of Depression From Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, Feb. 2020, doi: 10.1109/JSTSP.2019.2955012.

[19] N. Seneviratne and C. Espy-Wilson, "Deep Learning Based Generalized Models for Depression Classification," *arXiv:2011.06739 [cs, eess]*, Feb. 2021, Accessed: Feb. 21, 2021. [Online]. Available: http://arxiv.org/abs/2011.06739.

[20] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," *arXiv:1808.08929 [cs,*

eess], Aug. 2018, Accessed: Jan. 30, 2021. [Online].
Available: http://arxiv.org/abs/1808.08929.