Classification of Depression Audio Data by Deep Learning

Phanomkorn Homsiang
Department of Biomedical Engineering
School of Engineering, King Mongkut's Institute of Technology Ladkrabang, 10520
Bangkok, Thailand
63601218@kmitl.ac.th

Treesukon Treebupachatsakul
Department of Biomedical Engineering
School of Engineering, King Mongkut's Institute of Technology Ladkrabang, 10520
Bangkok, Thailand
treesukon.tr@kmitl.ac.th

Komsan Kiatrungrit
Department of Psychiatry
Faculty of Medicine Ramathibodi Hospital, Mahidol University, 10400
Bangkok, Thailand
komsan.kia@mahidol.ac.th

Suvit Poomrittigul
Department of Software Engineering and Information System
Faculty of Science and Technology
Pathumwan Institute of Technology, 10330
Bangkok, Thailand
suvit@pit.ac.th

2022 14th Biomedical Engineering International Conference (BMEiCON) | 978-1-6654-8903-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/BMEiCON56653.2022.10012102

Abstract—. **Due to many factors such as anxiety from contracting the disease and concern about the socio-economic impacts, Thai people have accumulated stress and are at risk of depression. The diagnosis of depression can be primarily assessed by testing the assessments such as PHQ-8, PHQ-9, and CES-D. The applied deep learning technology in medicine has received research interest and has been developing. In this research, we tried the classification of depression and non-depression audio datasets with the implementation of 4 model architectures: 1D CNN, 2D CNN, LSTM, and GRU. By converting wave audio format (WAV) of Daic-woz database to the Mel-frequency cepstrum (MFC). We have done the training and evaluated the 4 model architectures and compared the results between non-augmented and augmented datasets. The highest accuracy was obtained from 1D CNN with a non-data augmentation of 95%, and a 2D CNN with a data augmentation of 75%. These results confirm that human voices can differentiate between depression and non-depression.**

Keywords **Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Mel-Frequency Cepstrum (MFC), depressive order classification**

I. Introduction

The current situation of contracting the disease, economic collapse, and social consequences are factors that affect people to have accumulated stress and are at risk of depression. Nowadays, there are many depression screenings such as the assessment of PHQ-2, PHQ-8, CES-D [1 of PHQ-2, 2 of PHQ-8, 3 of CES-D]. that can be used in clinical practice. These assessments are primarily diagnoses, which are reliable and help the assessors evaluate their condition and decide to consult the psychiatrist.

The application of deep learning technology in medical practice is gaining development. Karol Chlastaa et al [4] reported that deep learning Convolutional Neural Network (CNN) was applied for screening depression by using a speech dataset from 2568 audio samples. The results showed that the depression screening accuracy was 77% by ResNet architecture. In addition, Dubagunta S et al [5] examined the modeling of low-pass filtered (LPF) speech signals by implementing a linear signal to predict the residual signal of the sound source. The model learns the sound wave dataset to classify depression and non-depression sounds using a zero-pass filter. This accomplished a sub-model of the linear prediction, residual signal, and zero-frequency filter signal by using sound wave dataset. This led to the development of deep learning systems for modeling sound data. Based on the study of Dubagunta S et al. [5], we hypothesize that data augmentation might improve the detection accuracy of people with depression. The research study of SRI International and faculty by Vikramjit M. et al. [6] have stated that speech of volunteer depression compared to non-depression has explored bio-signatures detection from words. Many feature extracts were investigated, especially the properties of Mel-frequency cepstrum (MFC) such as pitch, energy, and speech rate. The correlation structure features are proposed and showed impressive depression detection accuracy. There are Recurrent Neural Network (RNN) researches for use with the sequencing tasks audio data such as Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).

Therefore, this research aims to construct deep learning models for classifying the depression sound wave audio converted MFC datasets for screening depressed and nondepressed people and including implementing 1D CNN, 2D CNN, LSTM, GRU models.

: 978-1-6654-8903-4/22/$31.00 ©2022 IEEE

II. MATERIAL AND METHOD

A. Dataset preparation

In this study, we used the Daic-woz database [7]. This database contains clinical interviews of the questionnaire PHQ8 screening to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. We have divided the dataset into two groups: depression and non-depression. The number of depression datasets is 105 files (35 persons), and non-depression is 105 files (35 persons), for a total of 210 files. It can be seen that non-depression and depression are the same files. The average length of the sound wave is 5 minutes with a fixed sampling rate of 16 kHz. The dataset was split into 60:20:20 for training, validation, and the unseen test set, respectively. The Daic-woz audio files of the dataset were uncompressed in wave audio format (WAV) (Fig. 1). We have converted these Daic-woz audio files to MFC the Mel-frequency cepstrum (MFC) as the example shown in Fig. 2.

B. Data augmentation

To increase the number of MFC datasets for our experiments, we convert from the previously mentioned WAV files. we performed the augmentation by reducing static noise, stretching, adjusting the speed, and pitch, and shifting [8,9]. The total number of each class was balanced at 546 files.

C. Construction of deep learning model

CNN, LSTM, and GRU model deep learning architectures are popular for the sound wave dataset. The 1D and 2D CNN are the Convolutional Neural Network (CNN) where CNN simulates a human vision of space with autonomous feature extraction. The recurrent neural network (RNN) model can process not only single data points (such as images) but also entire sequences of data (such as speech or video). LSTM has feedback connections. GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. GRU performance on certain tasks of polyphonic music modeling, speech signal modeling, and natural language.

We have chosen the MFC feature process with 1D and 2D CNN which are suitable to be in this experiment. Therefore, we have experimented with the binary class of depression and non-depression.

The difference between WAV and MFC data processing, the depression file of WAV and MFC format was sampled as shown in Fig. 1 and Fig. 2, respectively. The procedure of deep learning models' construction with 4 architectures was shown in Fig. 3. The experiments of computer model construction were implemented by Python language and run on a Kaggle platform using the GPU to increase the efficiency of running the model.

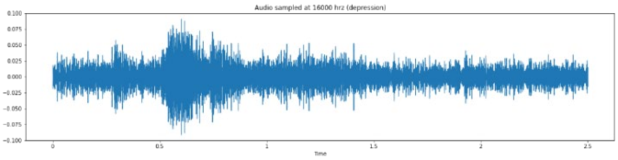

Fig. 1. An example of a recording wave audio format (WAV)

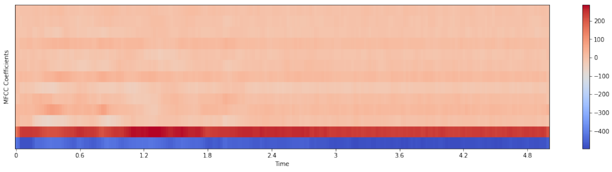

Fig. 2. An example of a converted MFC format

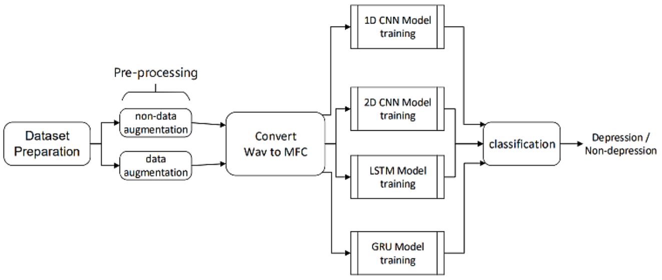

Fig. 3. The construction of the deep learning model

III. EXPERIMENTAL RESULTS

Our classification results are analyzed based on the classification of 4 CNN models including 1D CNN, 2D CNN, LSTM, and GRU. The experiments were performed under the hypothesis of data augmentation and non-data augmentation may affect the model performance. Therefore, the comparable model performance results of non-augmented and augmented datasets were reported.

A. Model performance of non-augmented dataset

The non-augmented datasets were balanced and trained at 200 epochs with 16 batch sizes. The learning curve of training and validation loss of 1D CNN, 2D CNN, LSTM, and GRU were shown in Fig. 4a-d, respectively. The performance results of these constructed models were shown in Table 1.

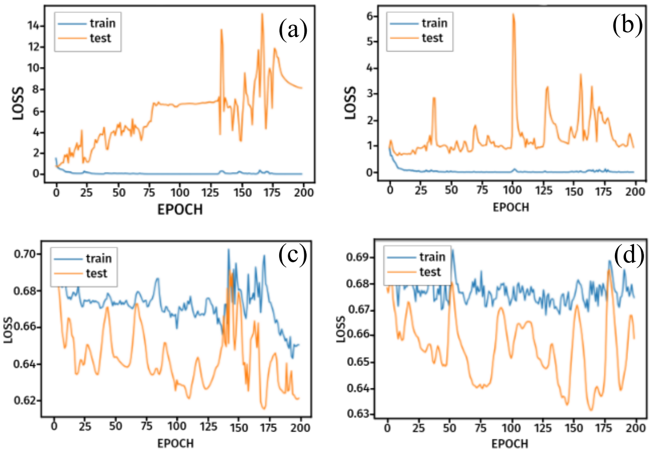

Fig. 4. Training loss and test loss graph of non-augmented dataset of (a) 1D CNN, (b) 2D CNN, (c) LSTM, and (d) GRU

TABLE I. THE MODEL PERFORMANCE OF THE NON-AUGMENTED DATASET

model	Class	precision	recall	f1-score	accuracy
1D CNN	Depression	0.90	0.43	0.58	0.69
	Non-depression	0.62	0.95	0.75	
2D CNN	Depression	0.74	0.67	0.70	**0.71**
	Non-depression	0.70	0.76	0.73	
LSTM	Depression	0.33	0.10	0.15	0.45
	Non-depression	0.47	0.81	0.60	
GRU	Depression	0.33	0.10	0.15	0.45
	Non-depression	0.47	0.81	0.60	

B. Model performance of augmented dataset

The augmented datasets were balanced and trained at 200 epochs with 16 batch sizes. The learning curve of training and validation loss of 1D CNN, 2D CNN, LSTM, and GRU were shown in Fig. 5a-d, respectively. The performance results of these constructed models were shown in Table 2.

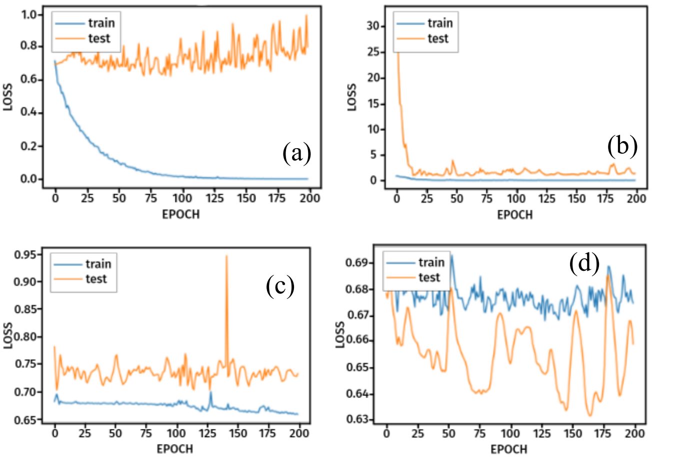

Fig. 5. Training loss and test loss of augmented dataset of (a) 1D CNN, (b) 2D CNN, (c) LSTM, and (d) GRU

TABLE II. THE MODEL PERFORMANCE OF THE AUGMENTED DATASET

model	Class	precision	recall	f1-score	accuracy
1D CNN	Depression	0.91	1.00	0.95	**0.95**
	Non-depression	1.00	0.90	0.95	
2D CNN	Depression	0.59	0.76	0.67	0.62
	Non-depression	0.67	0.48	0.56	
LSTM	Depression	0.43	0.17	0.21	0.48
	Non-depression	0.49	0.81	0.61	
GRU	Depression	0.53	0.48	0.50	0.52
	Non-depression	0.52	0.57	0.55	

IV. DISCUSSION

Based on the learning curve of 1D CNN and 2D CNN non-data augmentation, test loss lines are not steady. There is a noticeable difference in the experimental result when using data augmentation. The non-augmented model data of LSTM and GRU training and validation loss results were shown that their test loss lines are still not steady. Besides, the learning curve of the data augmentation, it is steady, yet the average performance is still not good enough.

In Table 1, The performance of the non-augmented data model shows that 2D CNN has the highest accuracy of 71% and 1D CNN has 69%. The LSTM and GRU had the percentage of accuracy performance not over 50%. It can be concluded that the dataset of non-data augmentation is suitable for 2D CNN models.

In Table 2 of the data augmentation model, we can see that the performance of 1D CNN has the highest accuracy of 95%, and 2D CNN has 62%. Although the performance of the LSTM and GRU models is better than the non-augmented experiment, their average performance is radically less than 1D and 2D. It can be concluded that the 1D CNN models model with data augmentation is the best in all for average performance.

The overall performance criteria include precision, recall, and f1-score, the highest value of data augmentation is 1D CNN model. Their precision and recall are equally valuable for both depression and non-depression.

Here we confirm the performance of the best-trained model by an unseen test set. The confusion matrixes are shown in Fig. 6 and 7 for the non-augmented and augmented datasets, respectively. Without data augmentation, the 2D CNN has the highest accuracy of 71%. Whereas the trained model of the augmented dataset showed higher accuracy of 95% of 1D CNN and corresponds to the confusion matrix of fewer false positives and false negatives.

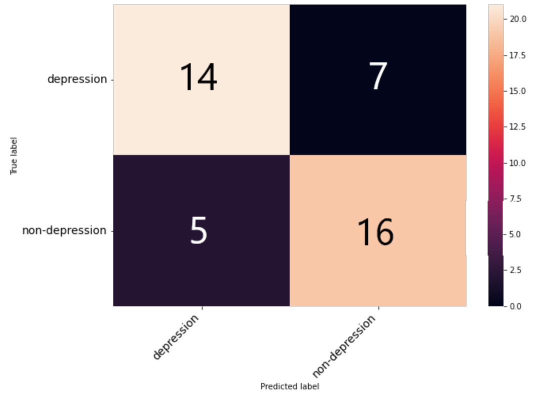

Fig. 6. Confusion Matrix of 2D CNN with the non-augmented dataset

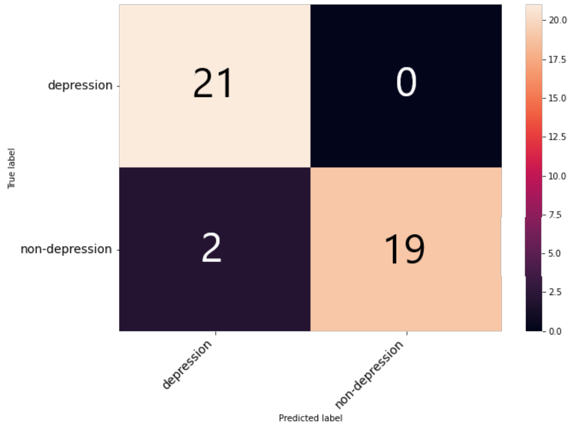

Fig. 7. Confusion Matrix of 1D CNN with the augmented dataset

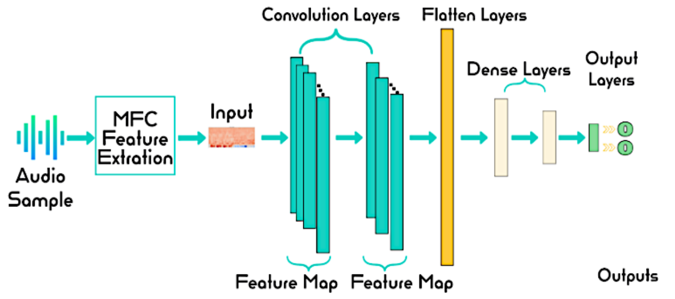

Fig. 8. Model Structure of 1D CNN

Fig. 8 shows the 1D CNN model structure, which achieves the best performance with the dataset more than other architectures. The 1D CNN proceeds the extracted features in 1 direction of kernel 1 dimension convolution process and is mostly used on Time-Series data. On the contrary, the 2D CNN proceeds the extracted features in 2 directions with kernel along 2 dimensions of the convolution process. For the MFC feature extraction data with 1D CNN, the process of 2 dimensional data is progressed 1D CNN by each dimension instead of the kernel along 2 dimensions at the convolutional layers.

TABLE III. SUMMARIZE THE PERFORMANCE OF CNN MODELS FOR THE SOUND WAVE DATASET

No.	Authors	Methodology	Model Architectures (Features Processing)	Accuracy
1	Chlasta, K., et al. 2019 [4]	Automated speech-based screening of depression	CNN Model (Spectrogram)	77%
2	Seneviratne, N., et al. 2021 [10]	Speech-based depression of severity level classification using a multi-stage dilated CNN-LSTM model	CNN Model (MFC)	42.76%
			LSTM (MFC)	45.71%
3	Srimadhur, N., et al. 2020 [11]	An End-to-End model for the detection and assessment of depression levels using speech	CNN Model (Spectrogram)	59.20%
			End-to-end CNN Model (Spectrogram)	61.32%
4	Our Proposed Model	Depression screening system using data-driven audio and video by deep learning model	**1D CNN Model (MFC)**	**95.00%**
			2D CNN Model (MFC)	**75.00%**
			LSTM Model (MFC)	48.00%
			GRU Model (MFC)	52.00%

Furthermore, we reviewed and summarized the efficacy results of our purposed models compared to other published papers as shown in Table 3. Our research compares the accuracy of other studies using the dataset Daic-woz database. Both MFC and spectrogram features have been implemented in related research. Table 3 shows the comparison of model architectures using related work and our methodology with CNN (1D and 2D), and RNN (LSTM and GRU).

V. CONCLUSION AND FUTURE WORKS

We have proposed an audio-based screening for depression by using deep learning. We constructed 4 models of 1D CNN, 2D CNN, LSTM, and GRU architectures trained with the nonaugmented and augmented datasets. The results indicate that the constructed model with the augmented dataset of 1D CNN showed the best performance, which reached 95% prediction accuracy. Therefore, it can be concluded that the sound wave dataset is applicable for constructing the depression recognition model of deep learning. Furthermore, we are now improving the deep learning model using the audio dataset for screening depression, which reports in level based on the result of the PHQ-9 assessment. This model purposes to help people and psychiatrists evaluate depression by using the sound of speech.

REFERENCES

[1] K. Kroenke, L. Robert, B. Janet, "The Patient Health Questionnaire-2: validity of a two-item depression screener," Med. Care, vol. 41, 2003, pp. 1284-1292.

[2] K. Kroenke, W. Strine Tara, L . Robert Spitzer, B. W. Janet Williams, T. Joyce Berry, H. Ali Mokdad, "The PHQ-8 as a measure of current depression in the general population," Journal of affective disorders, val 114, 2009, pp. 163-173.

[3] P. M. Lewinsohn, J. R. Seeley, R. E. Roberts, N. B. Allen, "Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults," Psychol. Aging, vol. 2, 1997, pp. 277-287.

[4] K. Chlasta, W. Krzysztof, K. Izabela, "Automated speech-based screening of depression using deep convolutional neural networks," Procedia. Computer. Science, vol. 164, 2019, pp. 618-628.

[5] S. P. Dubagunta, V. Bogdan, M. Mathew, "Learning voice source related information for depression detection," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, 2019.

[6] M. Vikramjit, T. Andreas, S. Elizabeth, "Noise and reverberation effects on depression detection from speech," 2016.

[7] D. DeVault, K. SimSensei, "A virtual human interviewer for healthcare decision support," Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014.

[8] E. Ma, "Data Augmentation for Audio," unpublished.

[9] Q. HA, "Augmentation methods for audio," unpublished.

[10] N. Seneviratne, C. Espy-Wilson, "Speech based depression severity level classification using a multi-stage dilated cnn-lstm model," arXiv. Preprint. arXiv, 2021, pp. 2104.04195.

[11] N. Srimadhur, S. Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," Procedia. Computer. Science, vol. 171, 2020, pp. 12-21.