Multi-Modal Depression Detection via Deep Learning

Run Peng and Yang Chen and Yang Fei
{roihn, rivachen, yangfei}@umich.edu

Abstract

Depression is a worldwide problem affecting innumerable lives (Organization et al., 2017). Depressed people always find it hard to communicate with people, concentrate on work and may even have the desire to suicide (Kindleberger et al., 1986). Hence the detection and treatment of depression becomes a major issue nowadays. Current techniques of depression detection are mainly based on clinicians' review, or patients' self-reports, which lack objective standard and quantified levels of depression. Therefore, we would like to design a multi-modal depression detection model via deep learning that can smartly detect the level of depression for patients. We take audio features and text-based transcripts as inputs, and generate various types of prediction for different uses. This model is aimed to provide an objective, professional and accurate reference on doctors' diagnosis. Potential ethical issues have been carefully considered, and we believe that our model could make a great contribution to the community of mental health in the near future.

1 Introduction

Major Depressive Disorder (MDD) is one of the major mental health disorders which is increasingly affecting people's lives. This is a world-wide issue, with the modern high-speed lifestyle boosting its spread. Based on the data given by the World Health Organization (WHO), 322 million people over the world are diagnosed with depression (Organization et al., 2017). The number of US adults with MDD increased by 12.9%, from 15.5 to 17.5 million, between 2010 and 2018 (Greenberg et al., 2021). MDD amplifies physical symptoms associated with medical illness, and is also highly correlated with morality (Katon and Sullivan, 1990). In addition to the increasing cases of MDD, the economic burden due to MDD increases in recent years, even when the medical cost for each case has been decreased to relatively low due to the effects of generic competition on antidepressant medications (Greenberg et al., 2021). Take the US as an example. The economic burden on US adults increased from $236 billion in 2010 to $326 billion in 2018 (year 2020 values) (Greenberg et al., 2021).

Therefore, a high-speed, accurate detection of potential MDD patients is quite significant for society. Current depression treatment is limited by low-efficiency assessment methods which still mainly rely on patient-reported or clinical judgements of symptom severity, risking a range of subjective biases and over-consumption of time (Rush and Ryan, 2002). Therefore, we would like to introduce our multi-modal depression detection model, which can make accurate, objective prediction on whether or not the person gets depression. The novelty of our model with respect to traditional diagnosis is that it takes in audio recordings, and text-based transcripts of dialogues between patients and therapists as input, and uses pre-trained models to extract meaningful feature embeddings from the data. As a result, it can provide results of both binary predictions (positive or negative), and quantified PHQ-8 score, which reflects the severity of a patient's level of depression. These prediction results are regarded as different tasks in our learning model, where we apply multi-task learning to extract common knowledge of these tasks to enhance the prediction.

To evaluate our model, we first compared

the performances of models between single-modal inputs and multi-modal inputs. Observations show that text-audio combined inputs lead to a better performance compared with models take in single type of inputs. Next, we tested whether the multi-task structure improves the quality of learned knowledge by singly running the training with binary / quantified outputs. It turns out that single binary prediction has the highest performance, which are quite reasonable because of the relatively low difficulty of task with respect of quantified outputs. After reweighting, we find our multi-task structure has similar performance to model with single binary output.

Based on the high performance we get in the evaluation part, it is convincing to say that our model is greatly capable to support the clinical judgements on depression, and shorten the process of diagnosis and enable early treatments. Also, with pre-designed questionnaires, we can try to implant the self-depression test in mobile apps, and utilize the microphones in mobile phones to enable data collection. We strongly believe that it will play a significant role in future depression treatment, and we hope that such a depression detection technology can be widely applied to hospitals, clinics, schools, etc.

2 Related work

2.1 Audio Emotion Recognition

Audio emotion recognition has been widely studied on various aspects. For traditional investigation on audio features, (Devillers et al., 2010) posted a concept of affective markers as a new classification of audio features, which can be linked to different representation of emotion states; (Han et al., 2014) used both segment-level features such as pitches, MFCC, and utterance-level features to perform emotion recognition.

Also, machine learning related techniques are commonly applied in audio emotion recognition. (Schmidt and Kim, 2011) applied linear regression and deep belief networks to recognize musical emotion. Similarly, (Zhang et al., 2017) used SVM model and deep belief networks to classify six emotion statuses and worked well on cases of both genders.

2.2 Deep learning based architecture

Recent years lots of researchers have explored different kinds of deep learning architecture in depression classifications.

The deep learning architectures can be classified into feed-forward neural networks such as (Ringeval et al., 2015) and (Dham et al., 2017), LSTM-RNN based neural network such as (Rejaibi et al., 2022) and convolution neural network such as (Jain, 2019) and (Othmani et al., 2021).

Researchers also explored ways to combine different learning architectures. For example, Ma and others combine long short-term memory convolutional neural network with convolutional neural network (LSTM-CNN) (Ma et al., 2016). In 2019, Bidirectional long short-term memory convolutional neural network was proposed (Zhang et al., 2021). Researchers also proposed deconvolutional neural networks (DNN) based on idea of convolutional networks (Gupta et al., 2017). DNN can also work with other neural networks. For example, adding convolutional network before DNN to make a deep convolutional neural network-deconvolutional neural network (DCNN-DNN) which makes a good performance (Yang et al., 2017) and adding multiple instance learning makes a deconvolutional neural network multiple instance learning (DNN-MIL) (Salekin et al., 2018). In these approaches DCNN-DNN (Yang et al., 2017) and LSTM-RNN (Rejaibi et al., 2022) perform quite well.

3 Data

3.1 Datasets

The dataset used in this project to assess depression is the DAIC-WOZ depression dataset provided by University of South California Institution for Creative Technologies (Gratch J, 2014) (DeVault and Morency, 1996).

The DAIC-WOZ corpus provides audio recordings of 275 clinical interviews of 275 participants answering the questions of an animated virtual interviewer named Ellie. Each recording is labeled by PHQ-8 binary, PHQ-8 score, PCL-C(PTSD) and PTSD Severity. In this project, PHQ-8 binary and PHQ-8 score are chosen to represent the level of depres-

sion. To be specific, the PHQ-8 binary defines whether the participant is depressed or not and the PHQ-8 score defines the severity level of depression of the participant. All the audio recordings are used.

The repartitions of the participants by their gender, depression, and depression severity level are shown in Fig1, 2, 3.

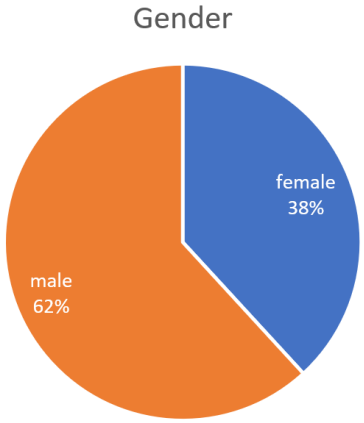

Figure 1: Gender repartitions of the participants within the DAIC-WOZ Corpus

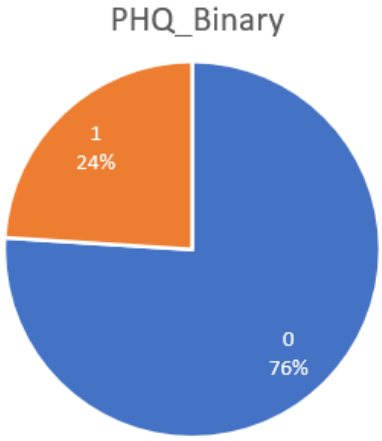

Figure 2: PHQ-8 binary repartitions of the participants within the DAIC-WOZ Corpus

Almost half of the participants are females(38%) and the third of the participants are labeled depressed (Fig.1). The dataset is almost gender-balanced. However, it is class-imbalanced as the number of non-depressed participants is approximately three times higher than the number of depressed participants (Fig.2). The imbalanced problem is solved in Section 4.1. After data preprocessing, 60% of the audio and text (transcripts of the patients) segments are used for training,

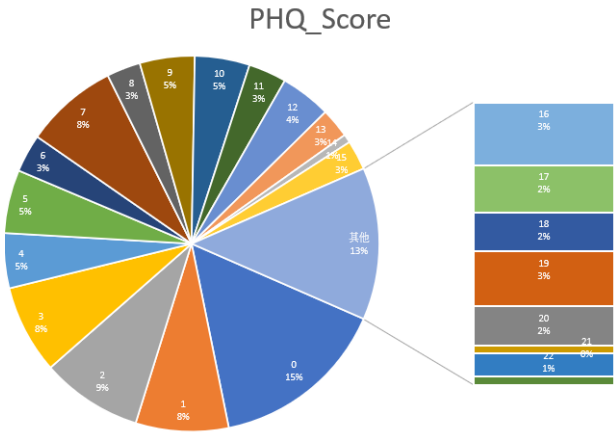

Figure 3: PHQ-8 score repartitions of the participants within the DAIC-WOZ Corpus

20% of them are used for validation and 20% for testing.

3.2 Data Types

3.2.1 audio

The given audio files were processed so as to obtain the voice of the participant only. Hence from the audio file, the voice of the participant was isolated using the speaking times given in transcript file.

3.2.2 text

Transcript files contain speaking times and values of participant. Each row represents the answer for one specific question and all the answers are used in the training and testing process.

4 Methods and Algorithms

4.1 Feature Extraction

In terms of audio feature extraction, we use eGeMAP to generate standardized features used in depression detection (Eyben et al., 2015, 2010). Each utterance can be embedded into a vector with 88 features, including frequency, energy, ectral, cepstral and dynamic information, which provide powerful and condensed features for emotion recognition. Before we ultimately reach eGeMAPs to be our final choice of our model for audio processing, we also tried other techniques, such as MFCC and wav2vec (Schneider et al., 2019). After the evaluation, we find MFCC not powerful enough in distinguishing the emotion of given utterances, and wav2vec requires huge

memory space for calculation, which is beyond our capacity.

In terms of text feature extraction, we use sentence-BERT to extract sentiment information from given transcripts (Reimers and Gurevych, 2019) with . Derived from original version of BERT, sentence-BERT uses siamesed and triplet network structures to derive semantically meaningful sentence embedding (Devlin et al., 2018). It inherits the accuracy of BERT, while boosting the efficiency.

4.2 Data Augmentation & Sampling

Different from Recurrent Neural Network, it is hard for traditional Machine learning to capture temporal relation of data. Therefore, we tried to cluster neighboring n utterances ($n = 10$) together, flattening them as one vector of features for each n utterances.In this way, the model can focus more on the temporal relation of n ordered utterances instead of separately learning from them. After clustering them, we also shuffled the data in batches to ensure that the samples contained in each batch are evenly distributed.

In addition to the lack of temporal relation, the data in DAIC-WOZ are quite imbalanced, which leads to a lower performance of our prediction, and make it hard to bring the model to the ground. As is shown in figure 2, the ratio between positive(1) and negative(0) cases are nearly one-third. Therefore, we chose to randomly sample the data with weights based on the original proportions, to get a more balanced dataset for further study.

4.3 Training Neural Network

4.3.1 Inputs and Outputs of the Model

Three different groups of features are fed to the models: 1.only audio inputs; 2.only text features; 3.concatenation of audio and text features. 1 and 2 are used as two baselines to see if the combination of text and audio features as model inputs can lead to a better model performance.

Depending on the different functionalities of the models, one or more of the following outputs will be predicted: 1. PHQ-8 Binary: 0 for normal persons, 1 for patients. 2. PHQ-8 Score: range from 0-23, represents different level of depression, where 0 represents normal person and 23 represents very severe depression. The scores are divided into 4 groups: no depression: scores 0-5, minor depression: scores 6-10, moderate depression: scores: 11-15 and major depression: score 16-23 to get a better classification performance for the multi-task model.

4.3.2 Data Loader

The batch size for the training models across all the experiments are set to 60, 30, 30 accordingly for training, validation and testing datasets. After evaluating the results obtained with several batch sizes from 10 to 250, the best size range to be used is between 30 and 80; under 30 the model overfits and beyond 80 the model underfits. We shuffle the datasets before both training and testing.

4.3.3 Structure of Neural Network

Three different model structures are used in this paper. The outputs of multi-task classification model are both PHQ-8 binary and PHQ-8 score. The binary and the multi-class model only predict PHQ-8 binary or PHQ-8 score.

Generally, we choose deep learning neural network for implementing the model layers. LSTM or CNN are not used because the input features are 1. high level features after BERT and eGeMAPS; 2. embeddings on sentence group level. And due to the property of our dataset, each sentence group contains answers for a specific question, so there is no relationship between different groups of sentences. Besides, we add three dense layers in the model due to the lengths of our inputs. From our reading of papers, usually 2-5 dense layers are added after the high level depression features. We test on different numbers of layers and find that 4 layers work best in our model.

For learning rate, 1e-5 to 1e-1 are tested and 1e-3 gives the best performance. For weight decay, 1e-7 to 1e-1 are tested and finally 1e-6 gives the best performance.

The detailed model structures for all three model types can be found in Appendix A, B and C.

The multi-task model structure can be visualized in Fig.4. Both of binary outputs and multi-class outputs share the first three hidden layer and have individual output

layers.

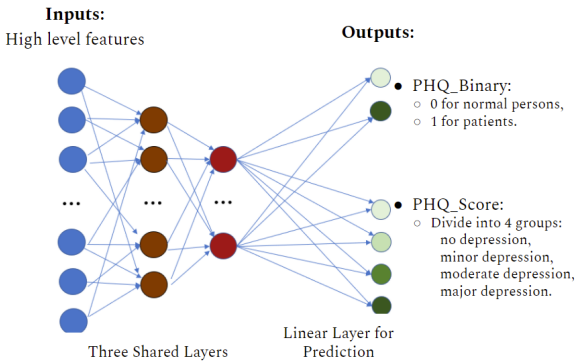

Figure 4: Multi-task Model Structure

Multi-task model has never been used on DAIC-WOZ dataset before. Based on the potential relation of the PHQ-8 binary and PHQ-8 score, in this paper the multi-task model is trained to see if it can better the performance of the prediction. The results are good and are discussed in details in the following section.

5 Results and Discussion

5.1 Results for PHQ-8 Binary Classification

We test PHQ-8 binary classification result based on three kinds of inputs: audio only, text only and audio-text combined input. The following three figures are the loss of training and validation sets.

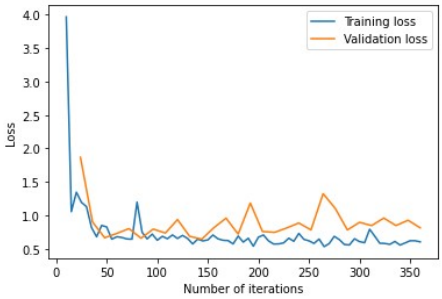

Figure 5: Binary loss figure of audio based input

The validation loss sometimes fluctuates when training loss decreases, indicating that the model doesn't have a good performance when meeting noises.

The confusion matrices of each inputs are in the following figures.

The results of corresponding F1 score and UAR on test sets are in the following table.

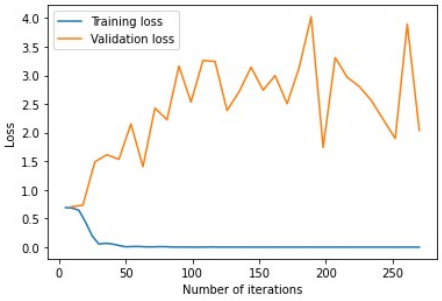

Figure 6: Binary loss figure of text based input

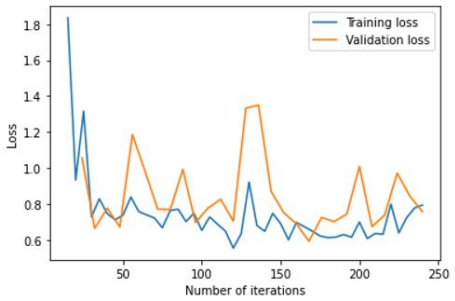

Figure 7: Binary loss figure of audio and text combined input

The confusion matrix shows that our model has a high precision of predicting non-depression but it doesn't do well in predicting depression. The performance of three kinds of input is nearly same. Audio has a better performance with 0.55 UAR and 0.54 F1 score, while text and audio embedding has a worse performance with 0.49 UAR and 0.47 F1 score. The model is better at using single type input to predict binary score than using audio and text combined input.

5.2 Results for PHQ-8 Multi-class Classification

In this section, we use the three same kinds of input in above sections to predict PHQ-8 Multi-class classification.

Directly predicting the whole PHQ-8 24 scores has a very bad performance because high number scores has a very small sample numbers and neighbouring number score has a very slight difference. Therefore, we divide them into four classes as described in 4.2.1.

The following three figures are loss on training and validation set of three kinds of input. Further improvement is needed as the validation loss vibrates when training loss increases.

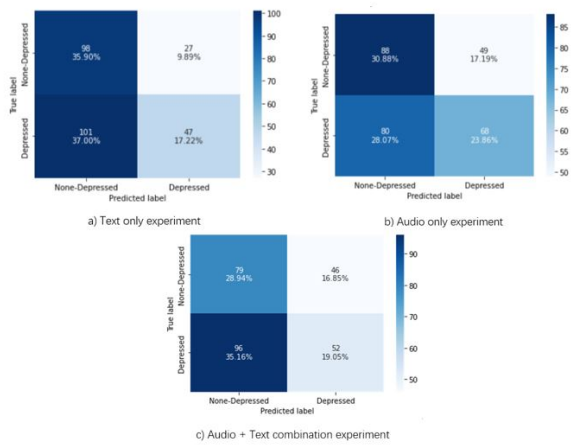

Figure 8: Binary confusion matrix of three kinds of input

Type of inputs	UAR	F1 score
Text	0.55	0.51
Audio	0.55	0.54
Text+Audio	0.49	0.47

Table 1: UAR and F1 score of three kinds of inputs on binary prediction .

The confusion matrices of each inputs are in the following figures.

The confusion matrices show that our model tends to predict minor depression and major depression groups. This shows that difference between none depression and minor depression and difference between moderate depression and major depression are still not clear enough for our model to classify them well. Major depression class which has the most distinguishable features is the easiest to be classified.

The results of three kinds of inputs' F1 score and UAR on test set are shown in the following table.

Type of inputs	UAR	F1 score
Text	0.27	0.27
Audio	0.26	0.21
Text+Audio	0.29	0.34

Table 2: UAR and F1 score of three kinds of inputs on score classes prediction.

For four moderate score classes prediction, all of the input doesn't have a high score. Text and audio combined input has better performance compared with single type input.

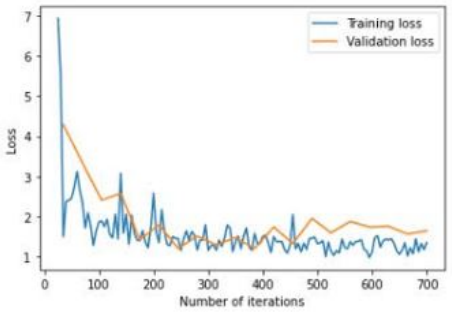

Figure 9: Multi-class loss figure of audio based input

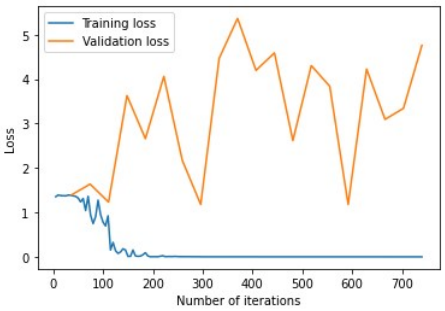

Figure 10: Multi-class loss figure of text based input

5.3 Results for Multi-task Classification

As described in 4.2.3, our model can predicts two kinds of output binary and scores at the same time by sharing hidden layers. Similar to the above sections, we also uses the three kinds of input: audio only, text only and audio-text combined input to test on our models.

The following three figures are the loss in experiment of different types of input. Text based input has a vibrating validation loss, which shows that our model's prediction performance still needs further improvement on text based input.

The results of three kinds of inputs' F1 score are in the following table. Text input has a better F1 score. The combined input and audio based input's performance are nearly the same.

Type of inputs	F1 score
Text	0.45
Audio	0.36
Text+Audio	0.36

Table 3: F1 score of three kinds of inputs on multi-task performance.

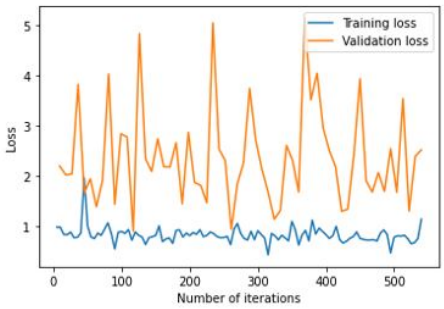

Figure 11: Multi-class loss figure of text and audio combined input

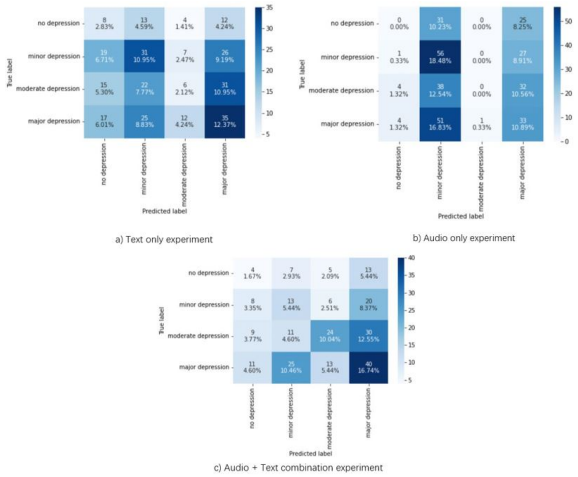

Figure 12: Multi-class confusion matrix of different kinds of input

5.4 Discussion

Based on our experiments, binary classifications have a better performance on any data inputs than multi-class classification. However, after dividing the PHQ-8 scores into four groups, the multi-class prediction results get much better than simply doing the prediction on the 24 different scores. We can develop further on the division criteria of PHQ-8 scores to seek better single model improvements and better collaboration ways with the PHQ-8 binary for multi-task models.

Combination of text and audio features as the input does not always give a better performance in the classification, which means we need to work more on the concatenation methods of the two modalities in the future.

By using multi-task model, the overall F1 score has improved compared with single multi-class classification, which shows the multi-task do help us decide the severity level of depression. This will be a good way to develop in the future when we have better model architectures and way of dealing with features.

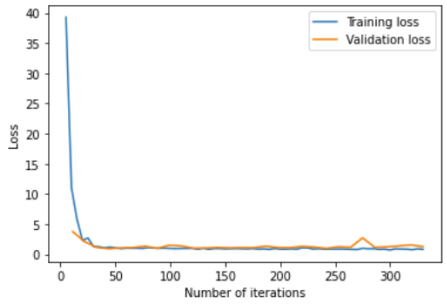

Figure 13: Multi-task loss figure of text based input

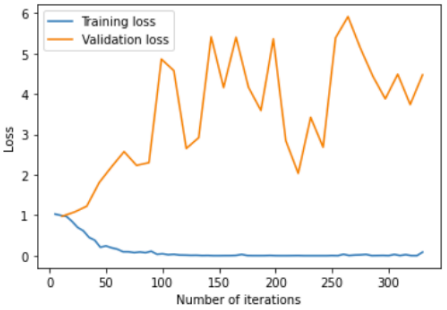

Figure 14: Multi-task loss figure of audio based of input

6 Ethical Considerations

The ethical implications of this technology could be huge and below listed some of major ones and corresponding solutions.

Due to the limited accuracy, the provided prediction of depression could be inaccurate for determining the exact situation of a potential patient. Making decisions largely relies on the model results that may have large potential risks of inaccurate diagnosis. As a result, the predicted results by the model should only be used as an auxiliary reference. Therapists or doctors hold the right to make the final diagnosis. If there happens to be a large deviation between model's detection and therapist' opinion, a double check is required.

The audio and video recording of the patients may be automatically recorded during the progress in order to be tested by the model. Some answers may include personal information of the patient and the leakage of private information would be huge problems for the patients. As a result, only doctors and hospitals can get access to the prediction results

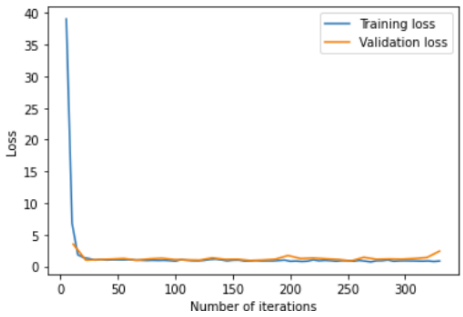

Figure 15: Multi-task loss figure of audio and text combined input

directly. No other institutions should have access to the results unless specially approved. Besides, only assigned patient's id should occur in the recorded datasets, which means no personal information of the patients should be stored. The hospital should also ensure security and confidentiality with the recordings and all the files are not locally stored but updated to the patients' conditional database.

People may act a little differently from normal life when being asked and recorded in front of the camera. So, cases are that their performance change may influence the result of depression diagnosis. As a result, nursing the patients' willingness of being recorded by a machine during the conversation is a must. Patients have the right to stop the recording anytime during the interview. Doctors should make sure that no different questions are asked between recorded and unrecorded interviews.

The text parts may only work on English languages which may cause bias. As a result, more language patterns could be added to the model to make it more universal in the future.

7 Conclusion

In this paper, we proposed a multi-task model which takes in audio and text form of dialogues between patients and therapists, and generate binary and quantified predictions on whether and how much the tester suffers from depression. Based on our evaluation, our model has a high performance on prediction, and we proved that the usage of multi-modal inputs and multi-task structure also have their advantages. As a result, we believe that our model can make a difference in mental health society, and help more patients out of the pain of depression.

References

Artstein R. Benn G. Dey T. Fast E. Gainer A. Georgila K. Gratch J. Hartholt A. Lhommet M. Lucas G. Marsella S. Morbini F. Nazarian A. Scherer S. Stratou G. Suri A. Traum D. Wood R. Xu Y. Rizzo A. DeVault, D. and L.-P. Morency. 1996. FOCUS: the interactive table for product comparison and selection. 13th International Conference on Autonomous Agents and Multiagent Systems.

Laurence Devillers, Laurence Vidrascu, and Omar Layachi. 2010. Automatic detection of emotion from vocal expression. A Blueprint for an Affectively Competent Agent, Cross-Fertilization Between Emotion Psychology, Affective Neuroscience, and Affective Computing, pages 232–244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Shubham Dham, Anirudh Sharma, and Abhinav Dhall. 2017. Depression scale recognition from audio, visual and text analysis. CoRR, abs/1709.05865.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pages 1459–1462.

Lucas GM Stratou G Scherer S Nazarian A Wood R Boberg J DeVault D Marsella S Traum DR Gratch J, Artstein R. 2014. The distress analysis interview corpus of human and computer interviews. InLREC, pages 3123–3128.

Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H Koenigsberg, and Ronald C Kessler. 2021. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). Pharmacoeconomics, 39(6):653–665.

Rahul Gupta, Saurabh Sahu, Carol Y Espy-Wilson, and Shrikanth S Narayanan. 2017. An affect prediction approach through depression severity parameter incorporation in neural networks. In INTERSPEECH, pages 3122–3126.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech 2014.

Royal Jain. 2019. Improving performance and inference on audio classification tasks using capsule networks. CoRR, abs/1902.05069.

Wayne Katon and Mark D Sullivan. 1990. Depression and chronic medical illness. J Clin Psychiatry, 51(Suppl 6):3–11.

Charles Poor Kindleberger et al. 1986. The world in depression, 1929-1939. Univ of California Press.

Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In Proceedings of the 6th international workshop on audio/visual emotion challenge, pages 35–42.

World Health Organization et al. 2017. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. 2021. Towards robust deep neural networks for affect and depression recognition from speech. In International Conference on Pattern Recognition, pages 5–19. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 71:103107.

Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In Proceedings of the 5th international workshop on audio/visual emotion challenge, pages 3–8.

A John Rush and Neal D Ryan. 2002. Current and emerging therapeutics for depression.

Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2(2):1–26.

Erik M Schmidt and Youngmoo E Kim. 2011. Learning emotion-based acoustic features with deep belief networks. In 2011 IEEE workshop on applications of signal processing to audio and acoustics (Waspaa), pages 65–68. IEEE.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal measurement of depression using deep learning models. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17, page 53–59, New York, NY, USA. Association for Computing Machinery.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. Depa: Self-supervised audio embedding for depression detection. In Proceedings of the 29th ACM International Conference on Multimedia, pages 135–143.

Weishan Zhang, Dehai Zhao, Zhi Chai, Laurence T Yang, Xin Liu, Faming Gong, and Su Yang. 2017. Deep learning and svm-based emotion recognition from chinese speech for smart affective services. Software: Practice and Experience, 47(8):1127–1138.

8 Appendix

A Model Structure and Parameters for Binary Classification Model

Model Architecture • Drop Out Layer:
Drop out rate: 0.1

- Fully connected layer 0:
  Input size: depends on the input type,
  Activation Function: ReLU,
  Output size: 2048
- Fully connected layer 1:
  Input size: 2048,
  Activation Function: ReLU,
  Output size: 512
- Fully connected layer 2:
  Input size: 512,
  Activation Function: ReLU,
  Output size:128
- Fully connected layer 3:
  Input size: 128,

Activation Function: ReLU,
Output size: 2

Training Parameters

- Criterion: torch.nn.CrossEntropyLoss
- Optimizer: torch.optim.Adam with 1e-6 weight decay
- Learning rate: 1e-3
- Batch Size: depending on type of dataset

B Model Structure and Parameters for Multi-class Classification Model

Model Architecture

- Drop Out Layer:
  Drop out rate: 0.1
- Fully connected layer 0:
  Input size: depends on the input type,
  Activation Function: ReLU,
  Output size: 2048
- Fully connected layer 1:
  Input size: 2048,
  Activation Function: ReLU,
  Output size: 512
- Fully connected layer 2:
  Input size: 512,
  Activation Function: ReLU,
  Output size:128
- Fully connected layer 3:
  Input size: 128,
  Activation Function: ReLU,
  Output size: 4

Training Parameters

- Criterion: torch.nn.CrossEntropyLoss
- Optimizer: torch.optim.Adam with 1e-6 weight decay
- Learning rate: 1e-3
- Batch Size: depending on type of dataset

C Model Structure and Parameters for Binary Classification Model

Model Architecture

- Drop Out Layer:
  Drop out rate: 0.1
- Shared layer 0:
  Input size: depends on the input type,
  Activation Function: ReLU,
  Output size: 2048
- Shared layer 1:
  Input size: 2048,
  Activation Function: ReLU,
  Output size: 512
- Shared layer 2:
  Input size: 512,
  Activation Function: ReLU,
  Output size:128
- Fully connected layer 3 and 4:
  Input size: 128,
  Activation Function: ReLU,
  Output size: (2, 4)

Training Parameters

- Criterion: using torch.nn.CrossEntropyLoss on both the loss for binary prediction ($loss_b$) and the losses for multi-class prediction($loss_s$). The loss is then calculated by $\alpha_1 * loss_b + \alpha_2 * loss_s$, the model gets the best performance when $\alpha_1 = 0.6$ and $\alpha_1 = 0.4$.
- Optimizer: torch.optim.Adam with 1e-6 weight decay
- Learning rate: 1e-3
- Batch Size: depending on type of dataset