EECS 498-010 Proposal
Multi-Modal Depression Detection via Deep Learning and its Application in Multiple Situations

Executive Summary

Depression is a worldwide problem affecting innumerable lives [3]. Depressed people always find it hard to communicate with people, concentrate on work and may even have the desire to suicide [9]. 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. Recent research has shown its importance of audio features, facial expressions and word choices in emotion analysis. 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. This model is aimed to provide a reliable reference on doctors' diagnosis. In addition, to generalize our model and improve its accessibility, we would like to design a software for self depression test, which also takes in data with multiple modalities, but has higher flexibility and less professionality. Potential ethical issues have been carefully considered, and we hope that our model could make a small contribution to the community of mental health in the future.

High-Level Description

Design intent and the background information

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), 350 million people over the world are diagnosed with depression [3]. The number of US adults with MDD increased by 12.9%, from 15.5 to 17.5 million, between 2010 and 2018 [1]. MDD amplifies physical symptoms associated with medical illness, and is also highly correlated with morality [2]. 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 [1]. Take the US as an example. The economic burden on US adults increased from $US236 billion in 2010 to $US326 billion in 2018 (year 2020 values) [1].

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 [4]. If we can introduce a sensing technology to provide accurate, objective diagnosis, it can not only detect negligible messages (micro facial expressions, habits, tones, etc.) from

potential patients, supporting the clinical judgements on depression, but also 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 cameras and microphones on mobile phones to enable data collection. These are the reasons why we believe that it will play a major role in future depression treatment. And we hope that such sensing technology can be widely applied to hospitals, clinics, schools, etc.

To better detect depression, we come up with the following two possible solutions. Solution I focuses on training a multi-modal depression detection model via deep learning. It takes in audio, text and video as three kinds of input, and generates the depression severity level of patients. It can be mainly used in hospitals or clinics, as an auxiliary reference of doctors and therapists. Solution II is a self-test software for depression detection. Different from solution I, it is no longer required to get a test in specific locations or with professional guidance. We design our software to simulate interviews with users. A virtual assistant will read out the questions, record users' audio answers, and also the behaviors when taking the interviews through cameras. After the collection of data, our software will process the data, extract the features and make a basic judgement on users' depression level. Solution II has its high accessibility to almost every adult, while its professionality and accuracy remains a challenge. Both two solutions have their pros and cons, and also some overlapped parts. We will provide a detailed description on both two solutions in the following part.

Solution I

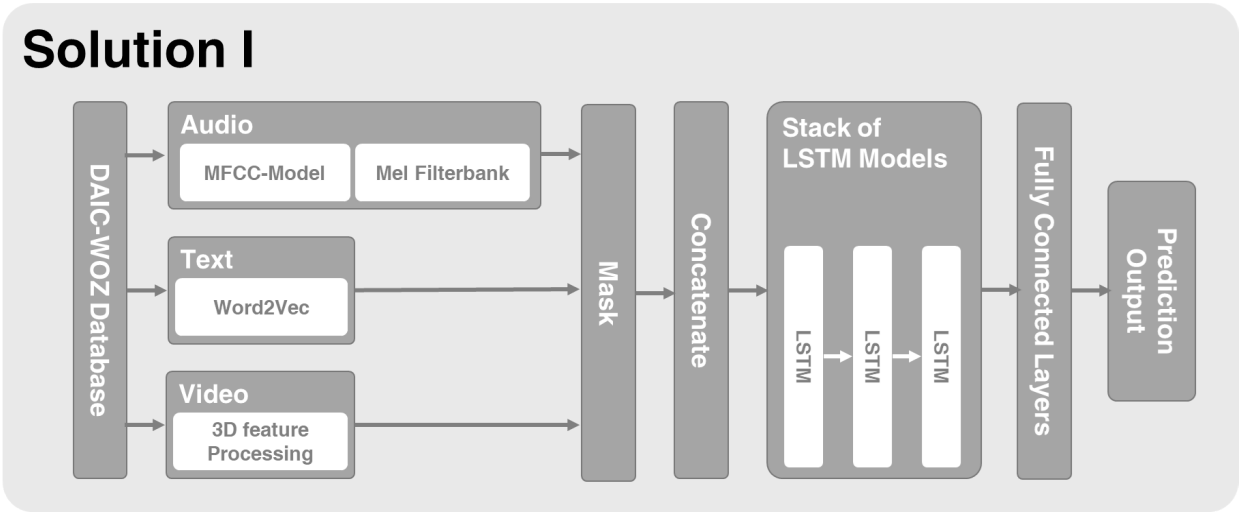

Figure 1. System Diagram of Solution I

In Solution I. We will use the DAIC-WOZ database to construct a deep learning model to classify patients' levels of depression.

The DAIC-WOZ database contains 189 sessions of clinical interviews and extensive questionnaire responses, with the former one including transcripts of interactions, participant audio files, and facial features. All datasets are labeled with different PHQ8 Scores, which is used to describe the depression severity levels. The video and audio features from the dataset will enable us to detect negligible messages

like subtle expressions or speaking styles which are crucial for detecting depression. Text features like words per sentence or ratio of depression words are highly-related features in consideration of depression analysis. As a result, multimodal fusion of audio-visual and text features will be implemented to ensure the completeness and the accuracy of the model. With these data features, we will test on different ML-based classification methods like SVM, GGM or more advanced MFCC-based Recurrent Neural Network [7] on the dataset for benchmarking and performance improvement.

Solution I provides a model that takes various micro but highly-related features to depression into consideration. These features are often hardly captured by human beings. The model should have a high performance on predicting interviewees' depression severity levels, and it is promising to see that hospitals will apply this technique as an auxiliary reference to determine the patients' conditions through conversations. Since the model will give results of different levels of depression, it can also be introduced to detect the potential patients with relatively low levels of depression to take the medical treatment in advance.

Solution II

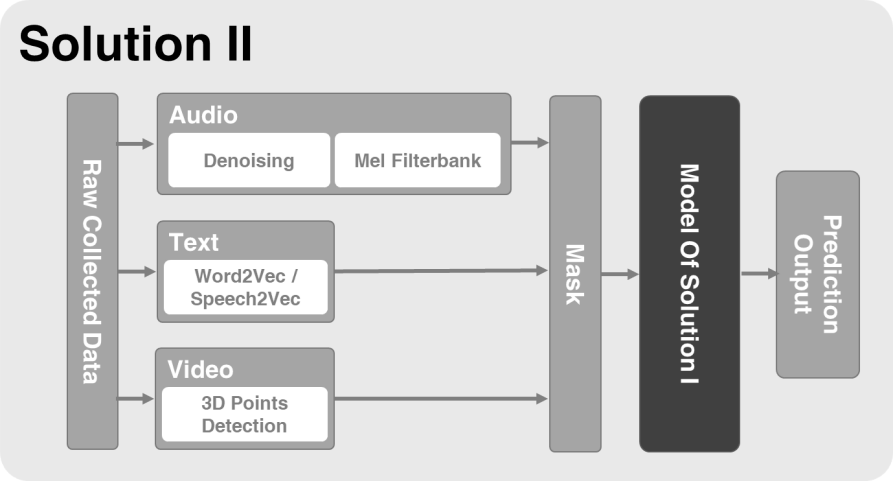

Figure 2. System Diagram of Solution II

Solution II is designed to focus on the individual usage. Based on the DAIC-WOZ database, we will use the questionnaire responses as the bases, combined with one or more features from video, audio or text datasets to implement a self-test app / software. The self-test app is designed to enable the users to test their depression levels with only a mobile phone or a laptop.

The foundation of the app is a depression-level predicted model like we build above. But instead of taking into account all features and having a relatively high accuracy, we make the model more flexible - we offer the users chances to choose different modes, e.g., only questionnaires or doing the video or audio recordings as well. Acquisition of the speech or visual data from the users will be done by using the microphone or camera in the laptop or mobile phone. We want to ensure that our app will be able to process the speech data by denoising the collected speech and using the existing speech-to-text APIs to convert the data. Questionnaires can be read by the machine with the loudspeaker in the eclectic device to

give the users more realistic feelings. All the responses will be collected together for the model to analyze the depression level.

Solution II will give people more flexibility in checking their own mental status. People will be able to use the app with their electronic devices at any time they need. The app will be convenient for the users, cheap for the cost, accurate for results and confidential for the information. Different suggestions will be provided to the users according to the testing severity of depression. We can also collect feedback from users and use them to better the existing model.

Trade-offs

The performance of the model is highly emphasised in Solution I since it is aimed to help hospitals detect the depression of patients. Audio, video and test features all play an important role in analysis depression, so we cannot simply remove any of them to reduce the workload. The problem tends to be the large consumption of time to compute in high dimensional space. To balance the cost and still ensure a rather high accuracy of the model, we will only choose highly related ones w.r.t the three feature fields and use the new machine learning methods with neural networks to ensure a better accuracy compared with the traditional SVM or GMM. We will also work on how to achieve a better fusion of different data modalities to improve the model. In this way, we will try to reduce the time and complexity of the computation while ensuring accuracy.

However, For solution II, models are aimed to be flexible to suit the users' needs and give more user-friendly features. But the complexity of the app will cause us to solve many corresponding problems. Technique problems are related to audio and visual sensing and processing. Ethical problems may mainly relate to privacy. So for the model development part, we choose to build flexible models that can predict results with some of the features. To make the workload reasonable for one semester, we will use existing APIs for converting speech data. We will ensure we can analyze the questionnaire's results as well as audio recordings. Works on other features like visual parts will be decided throughout the semester. Ethical problems will be discussed in detail in the Ethical Consideration part. In general, solution II will provide users with an interactive app that will enable them to test their depression level any time they need without worrying about information leakage.

Implementation and Issues

In Solution I, the first step is to process our raw data from the DAIC-WOZ database. We have input features with three modalities: audio, visual and text. For audio, we need to first select interviewee's audio from the whole speech database. Then we need to extract low-level features. In this part, we choose the MFCC feature not only because it is the most commonly used acoustic feature in speech recognition, especially in clinical consultation but also because it is higher performed compared to other acoustic features [8]. For visual features, we will directly use the extracted features in the dataset such as 3D facial points. For text features, optional features include sentence count, word per sentence and ratio of depression word per sentence.

The following step is to extract higher-level features from lower-level features. For acoustic MFCC features, we need to use recurrent neural networks (RNN) to form the flattened high-level MFCC features. Among a variety of RNN models, we choose Long short-term memory (LSTM) network as it is one of the most common and best performanced RNN networks. LSTM can maintain information in memory for a longer time than other RNN models. Following LSTM layers, we need to concatenate processed high level audio features with facial and text features, and ultimately feed them to fully connected layers to obtain the combined higher level features. In this part, the number of both LSTM layer and fully connected layer are both uncertain, which will be tested during the experiment. Initially, from papers related to our topics, we will choose three LSTM layers for acoustic processing and two fully connected layers for combined multimodal processing [7].

For the output of our neural network, we will use a fully-connected layer to output a vector containing the probability of different PHQ8 scores ranging between zero to one. The predicted result will be the PHQ8 score whose probability is closest to one. This will be the patient's PHQ8 score produced by our model.

In Solution II, the first step is also to preprocess raw data before feeding them into our model. And here are the main issues in the preprocessing part: For audio sources, we need to first denoise the collected speech. By denoising the audio raw data, we can eliminate noise's impact on our model greatly. Denoising the audio raw data can also remove part of background sounds which will protect customer's privacy. For questionnaires, we will apply stop-words reduction and capitalized words modification and bag words transformation on them. For visual features, we need to feed our video data into several layer neural networks to build a network which can help us extract 3D facial points at first. The relative position and velocity change in different time stamps of some important facial points (such as nose lip, peak point of eyebrow, lips points) form our facial features [10].

After raw data is preprocessed, we can input our visual-audio data or text data into the modal very similar in Solution I. For audio or visual data, we will input our MFCC feature from audio data or facial features of multiple timestamps from visual data into LSTM layers and then use dense layers to obtain higher-level output results. For text data, it will directly be input into dense layers to obtain results, as it doesn't need to process time series features which is LSTM's function. The output result is similar to Solution I: an output layer contains the possibilities of different PHQ8 scores. The score whose possibility is closest to one will be the estimated depression PHQ8 score of the tester.

A trade-off point in implementation is the number of LSTM layers and fully connected layers. More layers will extract more features thus improving the accuracy of our model. If we add more layers to our model, our training error will be greatly reduced but it will probably cause an overfit to our data. Moreover, the larger number of neural network layers implies a larger number of parameters, which requires more data. The dataset we obtain may not have adequate data to feed in our neural network. The running time will also increase significantly when adding more layers.

Ethical Considerations

1. Doctors can be euphemistic about the results to the patients but machines always tell the truth. Sometimes it's not proper for the patients to know their conditions. As a result, we need to ensure that only doctors and hospitals can get access to the prediction results. And this machine predicted results will not be shown to the patients or the families at any time during the treatment.
2. The accuracy of the model cannot be 100% but a wrong detection of a severe depression can cause a lot of problems. We need to set restrictions on the usage of our results, guaranteeing that they will 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. It is recommended to have extra observation for both the model and therapist, and make the final decision after deep consideration.
3. The text parts may only work on English languages. More language patterns could be added to the model to make it more universal in the future.
4. The patients have the rights to be informed of any testing with the model and usage of the recordings. To be specific, patients have the right to discontinue recording at any time during the conversation. Patients and families have the right to decide how the collected information will be used. The hospital should ensure security and confidentiality with the recordings, ensure that all the files are not locally stored but update to the patients' conditional database. Any usage of datasets will be approved by the patients and no leakage of the personal information will occur while accessing the data.
5. For the app, users can choose different modes according to their needs. They can decide whether to use the camera or not, whether to do the recording or not. We should not force the users to use any features while using the app.
6. Also, for the app, we should make sure that the users' privacy is not compromised. Users will be informed that the data will be sent to the cloud where our high performance computer will run the analysis and make the prediction with the given data. After prediction, the data will be deleted and no recorded information will be stored. Users will be informed if the possible leakage of their data occurs, which includes but not limited to: his/her personal information, audio and video recording, etc.
7. Since we do not have the right to give the users' information to any other institutions, we should also think about how to deal with cases of severe depression. The app should offer users ways to contact the nearby hospitals and give them suggestions on the importance of treating depression proactively.
8. Children may have quite different perspectives and judgements compared with adults, which may lead to the inaccuracy of the testing results. Hence, we will highly suggest users who are over 16 years of age to use the mobile app while the younger children should talk with their guardians when feeling prone to depression.

Way Forward and Schedule

As we mentioned above, since the time required for testing is unknown at the current stage, we provide two versions of the time schedule for reference. We would like to dynamically follow the first schedule, which can accomplish both solution I & II at the time for presentation; and if some unpredictable issues

happen, we will switch to our backup schedule, which focuses on polishing our solution I in the following two months.

Based on the details we have mentioned in the implementation section, we need to accomplish the following major tasks:

Solution I:

1. Access to Dataset: we would like to use the dataset DAIC-WOZ, while it requires the author's permission for academic use. We are currently approaching the author and asking about whether we can have access to the dataset.
2. Environment Presetting: Codes of Machine Learning may require a special running environment, which may take some time before all the environment setting is done.
3. Base Code Testing: We need to first test the base code, and check whether we can get the same result as is written in the related paper, in order to guarantee that the starting point is correct. Also, we need to measure the time and computational resources required for a single trial of tests. It is necessary for us to make a more reasonable and realistic time schedule for the whole project.
4. Code Improvement & Fine Tuning: After the base code testing, we need to add our own stuff to the existing model, test our hypotheses, and fine tune the model to get a better performance. Among them, there are two challenging or time-consuming tasks: i) Concatenation and Alignment of three modalities (audio, text, video); ii) Comparison Test of models with single or two modalities as input.

Solution II:

5. Data Preprocessing: From here, we need to cope with the raw data, and try our best to standardize the collected data to our previous model; Three predictable challenges are: audio denoising, speech to embedding, and 3D feature Capturing.
6. Fine-tuning model & Get a Good Pre-trained Model: After data preprocessing, we would like to fine-tune our model and get a good set of pre-trained parameters; This is actually quite challenging, since we do not have enough raw data to train with, which means we may reuse the data from our dataset, tune the hyperparameters, and then check whether the model can perform well on our collected data. This part is quite time-consuming.
7. Preparation for Demonstration
8. Report Writing

The detailed Gantt chart for preferred schedule has been shown below:

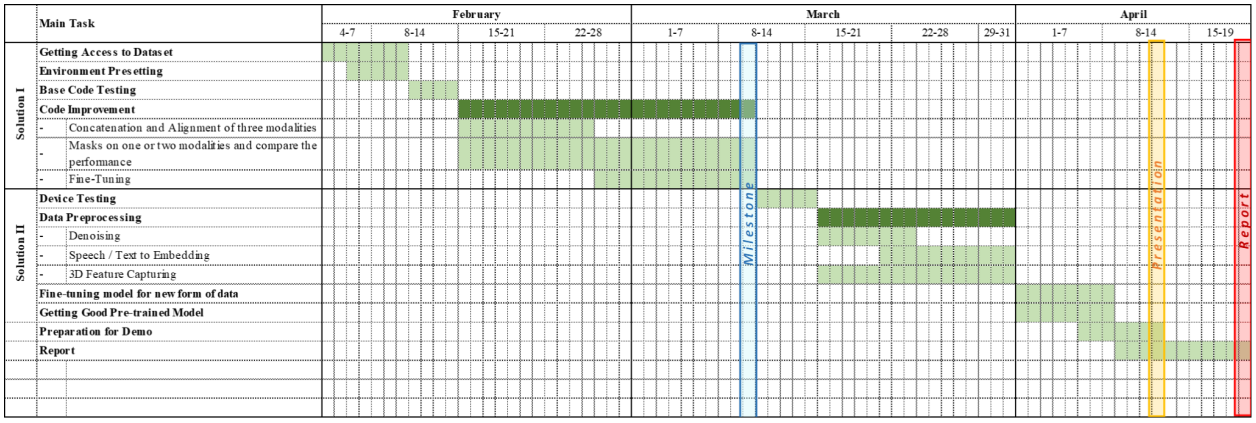

Table 1. Gantt chart for preferred schedule

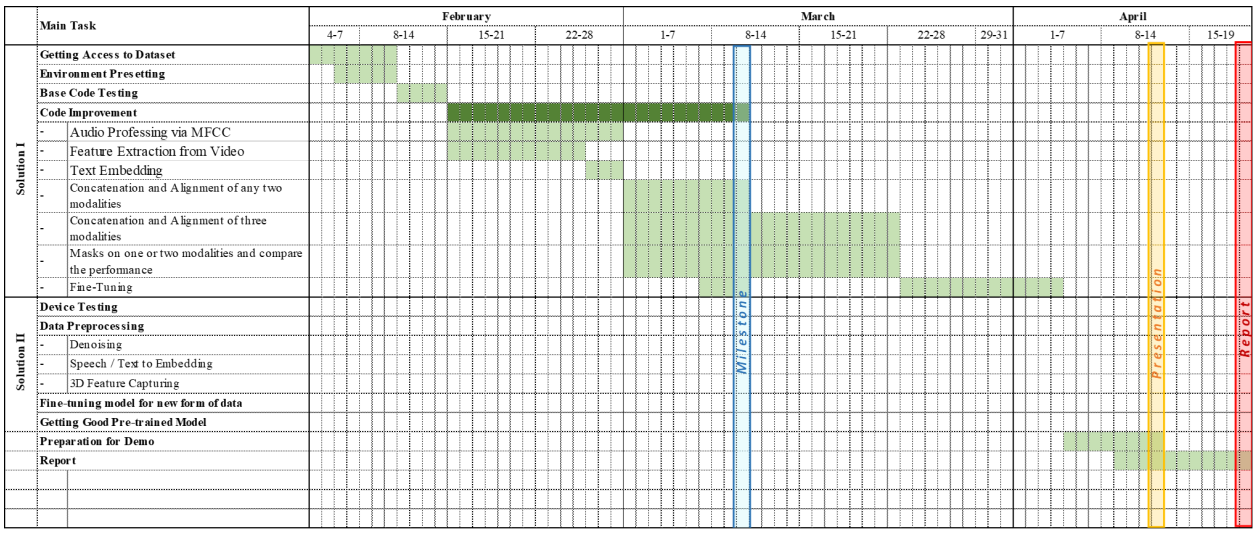

Table 2. Gantt chart for backup schedule

Final Presentation

If we have only completed solution I, we will put emphasis on the benchmarking (comparison among different choices of modalities) and results of evaluations in the final presentation.

If we have finished both solution I and solution II, we will show a demo video of how to do the self test via our software. Demo video could be a recorded version or a live version, depending on the degree of completion.

Tools and Materials

Solution I & Solution II:

1. Mel Filterbank (MFB) features, which are frequently used in speech processing applications, including speech recognition, and emotion recognition.
2. Word2vec model: a pre-trained model from [5] which is used to convert word to a vector of embeddings.
3. MFCC-based LSTM: a RNN model which takes in Mel Fre-quency Cepstral Coefficients(MFCC) as input to learn the classification task.

Solution II Only:
Possible tools for Speech to text / embeddings conversion:

1. Google Cloud Speech-to-Text API + Word2vec model [5]
2. Microsoft Azure Speech to Text + Word2vec model [5]
3. Speech2vec model: a pre-trained model from [6]

The Team

(in alphabetical order)
Run Peng: a senior student majoring in computer science; Responsible for main design, code implementation, and demonstration.
Yang Chen: a junior student majoring in data science; Responsible for code implementation, testing and fine-tuning, and ethics consideration.
Yang Fei: a junior student majoring in computer engineering; Responsible for code implementation and software design.

Reference

[1] P. E. Greenberg et al., ‘The economic burden of adults with major depressive disorder in the United States (2010 and 2018)’, Pharmacoeconomics, 2021[Online].
[2] Katon, Wayne, and Mark D. Sullivan. ”Depression and chronic medical illness.” J Clin Psychiatry 51.Suppl 6 (1990): 3-11.
[3] World Health Organization. Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). (World Health Organization, 2017).
[4] Rush, A. John, and Neal D. Ryan. "Current and emerging therapeutics for depression." 2002.
[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *Proc. Advances Neural Information Processing Systems*, pp. 3111-3119, 2013.
[6] Y.-A. Chung and J. Glass, "Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech", *Proc. Interspeech*, pp. 811-815, 2018.
[7] Rejaibi, Emna, et al. "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech." Biomedical Signal Processing and Control 71 (2022): 103107.
[8] P. Lopez-Otero, L. Dacia-Fernandez and C. Garcia-Mateo, "A study of acoustic features for depression detection," *2nd International Workshop on Biometrics and Forensics*, 2014, pp. 1-6, doi: 10.1109/IWBF.2014.6914245.
[9] Kindleberger, Charles Poor. *The world in depression, 1929-1939*. Univ of California Press, 1986.
[10] S. Dham, A. Sharma and A. Dhall, "Depression scale recognition from audio visual and text analysis", *CoRR*, 2017, [online] Available: http://arxiv.org/abs/1709.05865.