Multimodal Deep Learning for Mental Disorders Prediction from Audio Speech Samples

Habibeh Naderi
Faculty of Computer Science
Dalhousie University
Halifax, Canada
habibeh.naderi@dal.ca

Behrouz H. Soleimani
Faculty of Computer Science
Dalhousie University
Kinaxis Inc., Ottawa, Canada
behrouz.hajisoleimani@dal.ca

Stan Matwin
Faculty of Computer Science
Dalhousie University
Halifax, Canada
stan@cs.dal.ca

Sheri Rempel
Nova Scotia Health Authority
Halifax, Canada
sheri.rempel@dal.ca

Rudolf Uher
Department of Psychiatry
Dalhousie University
Halifax, Canada
uher@dal.ca

Abstract

Key features of mental illnesses are reflected in speech. Our research focuses on designing a multimodal deep learning structure that automatically extracts salient features from recorded speech samples for predicting various mental disorders including depression, bipolar, and schizophrenia. We adopt a variety of pre-trained models to extract embeddings from both audio and text segments. We use several state-of-the-art embedding techniques including XLNet, BERT, FastText, and Doc2VecC for the text representation learning and WaveNet and VGG-ish models for audio encoding. We also leverage huge auxiliary emotion-labeled text and audio corpora to train emotion-specific embeddings and use transfer learning in order to address the problem of insufficient annotated multimodal data available. All these embeddings are then combined into a joint representation in a multimodal fusion layer and finally a recurrent neural network is used to predict the mental disorder. Our results show that mental disorders can be predicted with acceptable accuracy through multimodal analysis of clinical interviews.

1 Introduction

Human brain recognizes linguistic content and emotional intent of an expressed opinion by integrating multiple sources of information. Our communicative perception is not only obtained from verbal analysis of what words have been delivered but also acquired by investigating additional modalities including speech audio and visual cues of how that utterance has been expressed [2]. More importantly, a single source of information (e.g. text-based mental mood understanding) may not be enough to detect and handle ambiguity due to the plurality of meanings. For instance, the emotive content conveyed by the spoken opinion "This was a different experience." may not be clear by itself while considering the tonality, pitch, and intonation of the speaker, it can be taken as a happy or sad narrative. This indicates the textual and audio characteristics of a statement are strongly related and learning how to model these inherent interactions between them can resolve ambiguity to some extent. Previous work in modeling human language often utilizes word embeddings pre-trained on a large textual corpus to represent the meaning of language. However, these methods are not sufficient for modeling highly dynamic human multimodal language. Therefore, to detect the mental state of the speaker, we
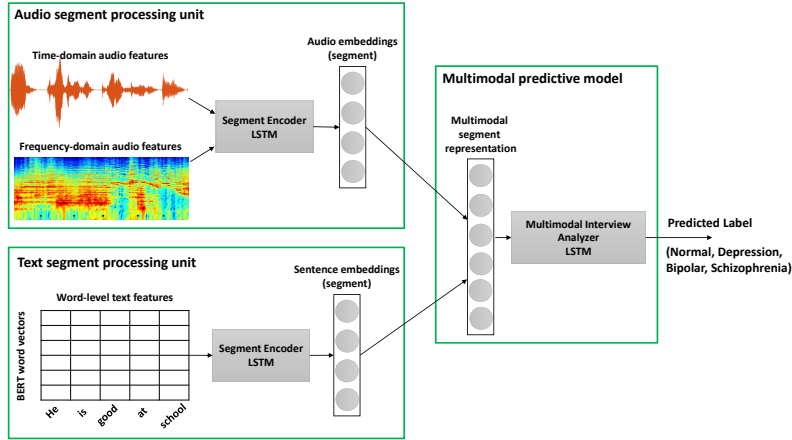
Figure 1: Model architecture.

not only require to consider multiple modalities that are involving in the message conveyance but also need to utilize adequate techniques which can learn complex interactions between those modalities.

Moreover, aspects of speech and language content can inform the diagnosis and outcome prediction in mental disorders [13, 7]. Clinicians use these characteristics in mental state examination by detecting key linguistic elements of their patient's statement in addition to its acoustic cues. However, systematic coding of speech can be laborious and there is lack of agreement about which speech characteristics are most important for diagnostic and prognostic purposes. This motivates us to learn an effective representation of key audio and language characteristics that can identify the presence and severity of mental illnesses. In this paper we introduce a multimodal deep learning structure that automatically extracts salient audio features from audio speech samples (e.g. pitch, energy, voice probability) and linguistic cues extracted from their transcribed texts (e.g. vocabulary richness, cohesiveness, average positive/negative sentiment score) to predict a variety of mental disorders. We use pre-trained WaveNet model [12] and VGG-inspired acoustic model [14] to extract two audio feature encodings. For textual features representation learning, we use pre-trained XLNet [25] and BERT (Bidirectional Encoder Representations from Transformers) [9] language models, in addition to other unsupervised word and document embeddings algorithms to learn text-based features embeddings. Our ultimate text-based and audio-based feature representations obtained from concatenating the learned text and audio embedding vectors. Then, we learn an optimal configuration to combine these two heterogeneous feature sets into a joint representation in a bimodal fusion layer. Next, we train an LSTM with attention mechanism over this multimodal fusion layer to make the final prediction. Figure 1 shows the architecture of our multimodal framework. We demonstrate the validity of this approach using a dataset of recorded speech samples from individuals with mental illness.

2   Related Works

When modeling human language, it is essential to not only consider the literal meaning of the words but also the nonverbal contexts such as vocal patterns and facial expressions in which these words appear. With respect to the modalities interactions learning, many efforts have been done in multimodal sentiment analysis and emotion recognition. Some earlier work introduced acoustic and paralinguistic features to the text-based analysis for the purpose of subjectivity or sentiment analysis [18]. In [21], multimodal cues including visual ones, have been used for the sentiment analysis in product and movie reviews. Their approach directly concatenated modalities in an early fusion representation, without studying the relations between different modalities. [28] has introduced an opinion-level annotated corpus of sentiment and subjectivity analysis in online videos by jointly modeling the spoken words and visual gestures. Most recently, Wang et al. [24] introduced a human language model that learns how to modify word representations based on the fine-grained visual and acoustic patterns that occur during word segments. They modeled the dynamic interactions

Table 1: Statistics of the data

Attribute	Count	Attribute	Count
Total number of subjects	363	Total number of segments	17,565
Average word count in segments	17	Average length of audio segments (seconds)	6.47
Number of objective segments	7,441	Number of subjective segments	10,124
Number of segments with positive sentiment	5,761	Number of segments with negative sentiment	3,417
Number of segments with anger emotion	1,294	Number of segments with fear emotion	807
Number of segments with joy emotion	4,649	Number of segments with sadness emotion	1,150
Number of segments with neutral emotion	9,398	Number of segments with neutral sentiment	8,268
Number of cohesive segments	2,896	Number of ruminated segments	229
Number of overinclusive segments	481	Number of worry segments	1,302
Number of criticism segments	1,750		

between intended meaning of a word and its accompanying nonverbal behaviors by shifting the word representation in the embedding space.

In recent years, automatic mental depressive disorders prediction from speech samples has been extensively studied [6, 1]. It has been shown that verbal interaction reduction and monotonous voice sound are indicative of depression [13]. Moreover, there is a perceptible acoustic change in the pitch, speaking rate, loudness, and articulation of depressed patients before and after treatment [7]. Moore et al. [20] have been explored the emotional content of speech (i.e. vocal affect) and its relationship with the overall mental mood of the patient. While previous works have been successful with respect to accuracy metrics, they have not created new insights on how the fusion is performed in terms of what modalities are related and how modalities engage in an interaction during fusion. Zadeh et al. [26, 28] proposed a Graph Memory Fusion Network(Graph-MFN) model that considers every combination of modalities as vertices inside a graph and calculates the efficacies of the connections between different nodes to learn the best fusion mechanism for modalities in multimodal language.

3 Dataset

The data consists of audio speech samples from 363 subjects participating in the Families Overcoming Risks and Building Opportunities for Well Being (FORBOW) research project. Participants are parents (261 mothers and 102 fathers) in the age range of 28-51 years. In these clinical interviews, parents were asked to talk about their children for five minutes without interruption. These 363 speech samples belong to 222 unique individuals from 180 unique families. Out of these subjects, 149 were diagnosed with Major Depressive Disorder (MDD), 66 with Bipolarity Disorder (BD), 19 with Schizophrenia, and 129 were the control group with no major mood disorders.

We transcribed these audio files using Google Cloud Speech API and after extracting the text, we broke down each sample into multiple segments based on changes in emotion, sentiment, objectivity/subjectivity, etc. which resulted in 17,565 segments. A segment has been coded as subjective if it includes expression of opinion, beliefs, or personal thoughts of the speaker. In contrary, if the segment consists of facts or observations of the speaker, it has been coded as objective. Four basic emotions are considered in this analysis including anger, fear, joy, and sadness. Six multidisciplinary researchers rated each segment for sentiment, objectivity/subjectivity, emotion (anger, fear, joy, sadness, neutral), cohesion, rumination, over-inclusiveness, worry, and criticism. 5,818 segments were rated by two or more researchers and the intraclass correlation for ratings of different researchers was high showing strong agreement in the labeling. In addition to the segment-level labeling, they also rated affect, warmth, overprotection, cohesion, and criticism at the document-level (i.e. for each audio sample). Document-level assessments are provided as nominal ratings between 1 and 5. Table 1 shows the basic statistics of the data and the segment-level labels. Figure 2 illustrates the heatmaps of ratings for segment-level and document-level labels.

4 Proposed Method

To address multilateral dynamic of human language as well as automatic extraction of the most salient speech characteristics, we propose a multimodal deep learning algorithm for automatic clinical speech samples analysis that effectively learns a non-linear combination between textual and acoustic modalities using an attention gating mechanism. In multimodal dynamics, we first build a
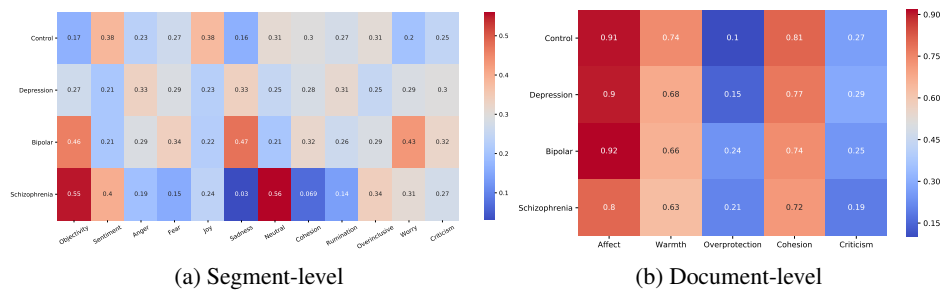
| | (a) Segment-level | (b) Document-level |

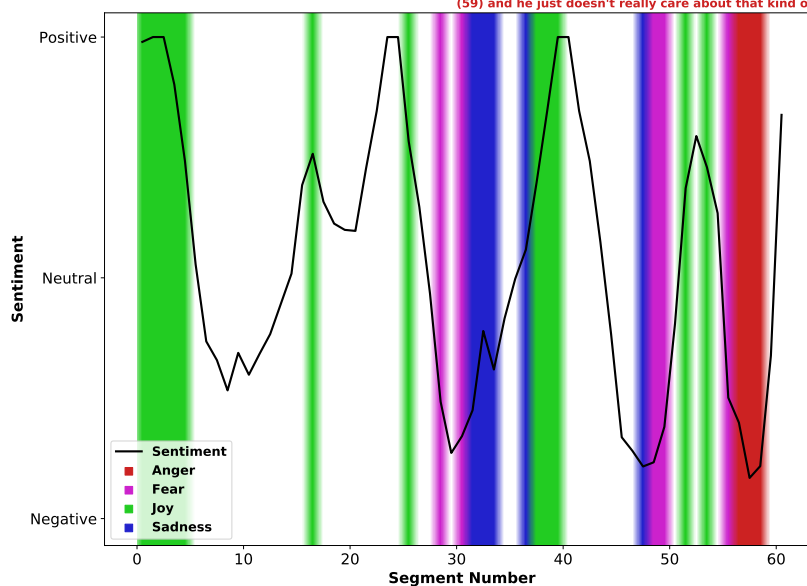Figure 2: Heatmaps of ratings for (a) Segment-level (b) Document-level

Figure 3: Our model prediction for emotional content of every segment in a randomly selected speech sample. The picture shows how the sentiment and emotions changes for each segment during the 5 minute interview. White areas are associated with neutral emotion. This subject has been diagnosed with bipolar disorder.

model for each modality independently with its own structure. We have a sequence of observations and we want to do inference in a sequential supervised learning manner. Then, to learn a joint representation of audio and text, we need to adopt an efficient fusion strategy to map these two sets of heterogeneous features into a common space. We analyze every modality in fine-grained (i.e. segment-level) and coarse-grained (i.e. document-level) and combine the textual and acoustic learned feature representations in two levels. The key insight to our model is that depending on the encoded information in textual and acoustic modalities, the relative importance of their associated learned embeddings may differ in the bimodal feature fusion layer. Here, our unimodal representation learning algorithms for audio and text features extraction are discussed separately.

4.1 Textual Features Representation Learning

Our textual features representation learning module has two major components: 1) segment-level features extraction to learn fine-grained textual embeddings for every segment, and 2) emotion-specific representation of text segment which extracts emotion information contained in every segment. These

4

two textual feature embeddings are then concatenated to create our ultimate segment-level text features representation.

After learning segment-level textual features representation, we feed this sequence of segment embeddings to another recurrent network (i.e. LSTM) with an attention gating mechanism and train it to make the final prediction of mental disorders. Moreover, we consider the learned representation of the last dense layer of this LSTM network as a document-level representation of every transcribed speech sample. The attention vector values demonstrate the relative importances of the segments in a document regarding the mental disorders prediction task. Then, we train different classifiers including Random Forest (RF), Support Vector Machines (SVM), k Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Naive Bayes over this coarse-grained encoding of textual features to predict mental disorders. We refer to this layer as our unimodal text representation layer. The following subsections discuss the details of the above two components of our segment-level textual features representation learning module.

4.1.1 Segment-level Textual Feature Extraction

To extract segment-level textual features, we use two pre-trained language models: 1) BERT language model [9] which is basically a multi-layer bidirectional LSTM networks trained with attention mechanism to learn text-based features embeddings. 2) XLNet [25] which is a generalized autoregressive model that captures longer-term dependency. More specifically, XLNet maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order and so does not suffer from BERT pretrain-finetune discrepancy. After learning BERT (or XLNet) representation of every token in the text segment, we take the average of learned representations to obtain the representation associated with the whole segment. However, since language models provides us with context-dependent word embeddings, we also employ a pre-trained FastText model [3], trained on Wikipedia, to learn another distributed word representations for every token in the text segment. FastText model incorporates subword information and considers character ngrams. Hence, it can learn the compositional representations from subwords to words which allows it to infer representations for words do not exist in the training vocabulary. Similar to BERT and XLNet segment representations, we take average of the learned FastText word embeddings of all the tokens in a segment to achieve FastText segment representation.

Moreover, to make sure our learned segment-level representation contains the most distinctive linguistic content of the clinical interviews - as there is an strong association between some mental disorders and patients' use of words, we apply a pre-trained Document Vector through Corruption (Doc2VecC) model [5] to learn segment-level text features representation of every segment in the transcribed speech sample. Doc2VecC captures the semantic meaning of the document by focusing more on informative or rare words while forcing the embeddings of common and non-discriminative words to be close to zero. We pre-train our Doc2VecC model on a large corpus of 21M tweets data. Then, we concatenate BERT (or XLNet), fastText, and Doc2VecC segment embeddings to obtain the first part of our segment-level text features representation. We use the embeddings dimensionality of d={1024, 100, 100} for BERT (or XLNet), fastText, and Doc2VecC models, respectively.

4.1.2 Emotion-specific Representation of Text Segment

Additionally, to incorporate the emotion information contained in text data, we train an LSTM network for emotion recognition using an auxiliary annotated dataset and learn the emotion-specific representation of every segment using the transfer learning framework [23]. We use SemEval-2018 AIT DIstant Supervision Corpus (DISC) of tweets [19] which includes around 100M English tweet ids associated with tweets that contain emotion-related query terms such as '#angry', 'annoyed', 'panic', 'happy', 'elated', 'surprised', etc. We collected 21M tweets by polling the Twitter API with these tweet ids and fed them into the LSTM network to predict their emotion labels. The output emotion is the label of the class with the highest probability among the four basic emotions of anger, fear, joy, and sadness. Next, we freeze the LSTM network and remove its softmax output layer. Then, we feed our sequence of segment embeddings learned by pre-trained fastText model and consider the learned representation of the last dense layer of the network as an emotion-specific representation of the input text segment.

5

Figure 4: A random sample from subjects with depression. Each line shows a segment and they are colored based on the attention weights learned in our attention-based LSTM model. Darker colors mean the model is paying more attention to those segments for the final recognition (patient's name is replaced with blank for anonymity).

4.2 Audio Features Representation Learning

Our audio features representation learning module shares quite a similar structure with our textual feature extraction one. There are two major components in our audio feature extraction module: 1) segment-level acoustic features extraction to learn audio embeddings for every segment, and 2) emotion-specific representation of audio segment which extracts vocal affect information contained in every segment. These two set of audio feature embeddings are then concatenated to create our ultimate segment-level audio features representation. To obtain the document-level audio features representation, we need to reduce the dimensionality of the extracted time-domain and frequency-domain audio features for each segment. Therefore, we train an LSTM classifier using our 12 segment-level labels (i.e. subjectivity/objectivity, sentiment, emotions, cohesion, rumination, over-inclusiveness, worry, and criticism) to get the audio segment encoding in the lower dimension. Then, similar to our text unimodal representation learning algorithm, we feed this sequence of low-dimensional audio segment encodings to another recurrent network to predict the mental disorders. We consider the learned representation of the last dense layer of this LSTM network as our audio document-level features representation and train different classifiers over it. We refer to this layer as our unimodal audio features representation layer and train the same classifiers have been used in our text unimodal analysis over this layer to predict mental disorders.

4.2.1 Segment-level Audio Feature Extraction

For segment-level audio features representation learning, we first use a pre-trained WaveNet autoencoder model [12] which basically is a neural audio synthesis network. The input audio signal is encoded to the 16 channel embedding by a deep autoregressive dilated convolutions neural network. Then, a similar decoder is trained to invert the encoding process and reconstruct the input audio signal from the learned 16 channel embedding. We feed the sequence of our audio segments to the pre-trained WaveNet model and take the 16 channel encoding as the learned audio segment features representation. Secondly, we employ a pre-trained VGG-inspired acoustic model [14] as another audio feature extractor. This VGG-like network learns a 128-dimensional embedding from Mel spectrogram of the input audio segment. We take the encoding representation obtained from training this VGG-like network over the spectrogram features of every sound frame. We also extract eight

6

time-domain audio features from each frame such as pitch, energy, Normalized Amplitude Quotient (NAQ), peak slope. Regarding the frequency-domain analysis, we extract 272 Mel-Frequency Cepstral Coefficients (MFCC) in addition to their statistics (e.g. mean, standard deviation, range, skewness, and Kurtosis) for each audio segment. The first part of our segment-level audio features representation is then obtained by concatenating the two audio segment embeddings learned by WaveNet and VGG-like models in addition to the traditional audio features that have been extracted from every audio segment.

4.2.2 Emotion-specific Representation of Audio Segment

To incorporate the emotion information contained in the audio segment into our audio feature representation learning, similar to our text modality feature extraction analysis, we use transfer learning. First, we use the COVAREP software [8] to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features, glottal source parameters [10], peak slope parameters and maxima dispersion quotients [16] for audio speech samples. All extracted features are related to emotions and tone of speech. Next, we train an LSTM model on an auxiliary dataset for emotion recognition task. We train our model on CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [28, 27]. CMU-MOSEI contains 23,453 annotated video segments from 1,000 distinct speakers and 250 topics. Each video segment contains manual transcription aligned with audio to phoneme level. Every segment has been annotated for Ekman emotions [11] of {happiness, sadness, anger, fear, disgust, surprise}. However, we only include the audio segments that have been labeled for four basic emotions {happiness, sadness, anger, fear} to match our speech samples emotion annotation. Then, we freeze the model and remove its softmax output layer and feed the COVAREP features associated with each audio segment to this pre-trained model. We use our audio segments' labels to fine-tune the pre-trained model and take the learned representation of the last dense layer of the LSTM network as the emotion-specific COVAREP-based feature representation of the audio segment.

Secondly, we learned emotion-specific features representation for audio segments based on their spectrograms. We extract the spectrogram features of every audio segment and feed it as an input to a Convolutional Neural Network (CNN) plus LSTM model to predict the segment's emotion. By applying 2D-Convolutional layer on spectrogram, we learn the most distinctive spatial and temporal audio features. We use our emotion labels to train this CNN plus LSTM model and take the learned representation of the last dense layer of the network as the emotion-specific spectrogram-based feature representation of the audio segment. Then, we concatenate the two COVAREP-based and spectrogram-based emotion-specific audio segment representations to obtain the emotion-specific audio features representation for every segment.

4.3 Multimodal Fusion Learning

After learning features representation for each modality, we adopt two different feature-level fusion strategies: (1) document-level fusion which combines the two document-level feature representations of audio and text in one multimodal layer as a feature representation of the entire speech sample, and (2) segment-level fusion which concatenates the text and audio representations of each segment and outputs the bimodal segment-level feature representation for every segment.

4.3.1 Document-level Fusion

In document-level fusion, we fuse the two heterogeneous document-level feature sets of text and audio into a joint representation in a bimodal fusion layer. Moreover, we train an LSTM with attention gating mechanism over this multimodal fusion layer of audio-textual learned representation. The attention layer learns to assign different weights to language and audio embeddings depending on the information encoded in the words that are being uttered and acoustic modalities. Eventually, we train a sigmoid output layer on top of this weighted bimodal fusion layer to make the final prediction. Additionally, similar to the unimodal analysis we take the representation of the last hidden layer and train a variety of classifiers to predict the final label. To formulate a segment of speech sample, we have the sequence of uttered words in language modality $L^{(i)} = [l_1^{(i)}, l_2^{(i)}, \ldots, l_{t_{l_i}}^{(i)}]$ accompanying by the sequence of audio frames in acoustic modality $A^{(i)} = [a_1^{(i)}, a_2^{(i)}, \ldots, a_{t_{a_i}}^{(i)}]$ where i denotes the span of the ith segment. To model the temporal sequences of textual and audio information coming

Table 2: Accuracy (%) of mental disorder recognition for our unimodal and multimodal systems over 5-fold cross-validation. The text results correspond to the XLNet language model since XLNet outperformed BERT in our experiments.

	Control			Depression			Bipolar			Schizophrenia		
	Text	Audio	Multi	Text	Audio	Multi	Text	Audio	Multi	Text	Audio	Multi
LSTM	70.7	67.52	67.52	65.62	58.33	54.17	55.56	54.32	55.56	67.39	63.04	63.04
RF	71.97	67.52	77.07	66.67	60.42	71.88	55.56	49.38	**71.6**	69.57	58.7	63.04
SVM	71.97	67.52	70.7	64.58	58.33	54.17	53.09	48.15	60.49	65.22	56.52	58.7
KNN	70.06	54.78	76.43	61.46	57.29	64.58	55.56	54.32	65.43	71.74	71.74	**73.91**
LDA	73.25	68.79	70.7	**72.92**	56.25	56.25	60.49	50.62	65.43	**73.91**	58.7	69.57
NB	71.97	**78.98**	**78.98**	64.58	63.54	60.42	53.09	54.32	62.96	69.57	58.7	71.74
tf-idf+SVM	56.42	-	-	54.74	-	-	57.59	-	-	65.1	-	-
BOW+SVM	55.98	-	-	52.30	-	-	56.47	-	-	62.29	-	-

from each modality and compute the joint embeddings, we use an LSTM networks. LSTMs have been successfully used in modeling temporal data in both natural language processing (NLP) and acoustic signal processing [15]. We apply two LSTMs separately for each modality:

$$h_l^{(i)} = LSTM_l(L^{(i)}), \quad h_a^{(i)} = LSTM_a(A^{(i)}) \tag{1}$$

where $h_l^{(i)}$ and $h_a^{(i)}$ refer to the final states of the language and acoustic LSTMs that we call document-level feature representation (or LSTM embedding) of text and audio modalities. We then combine these two LSTM embeddings using an attention gating mechanism to model the relative importance of every segment in each modality.

$$w_l^{(i)} = \sigma(W_{hl}[h_l^{(i)}] + b_l), \quad w_a^{(i)} = \sigma(W_{ha}[h_a^{(i)}] + b_a) \tag{2}$$

where $w_l^{(i)}$ and $w_a^{(i)}$ are the language and acoustic gates, respectively. W_{hl} and W_{ha} are weight vectors for the language and acoustic gates and b_l and b_a are scalar biases. The sigmoid function $\sigma(x)$ is defined as $\sigma(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$. Then, we calculate the bimodal fusion layer by fusing the language and acoustic embeddings multiplied by their corresponding gates.

$$h_{la}^{(i)} = w_l^{(i)}.(W_l h_l^{(i)}) + w_a^{(i)}.(W_a h_a^{(i)}) + b_{la}^{(i)} \tag{3}$$

where W_l and W_a are weight matrices for the language and acoustic embeddings and b_{la} is the bias.

4.3.2 Segment-level Fusion

In segment-level fusion, we first combine the feature representations of text and audio modalities for each segment and then train one mutual LSTM network over this sequence of multimodal feature embedding.

$$h^{(i)} = LSTM([L^{(i)}; A^{(i)}]) \tag{4}$$

where $[;]$ denotes the operation of vector concatenation and $h^{(i)}$ refers to the final state of the LSTM. Then, we apply an attention gate on top of the LSTM embedding. The attention layer learns to assign greater weights to more discriminative segments and hence improves our prediction accuracy.

$$w^{(i)} = \sigma(W_h[h^{(i)}] + b), \quad h_{la}^{(i)} = w^{(i)}.(W h^{(i)}) + b_{la}^{(i)} \tag{5}$$

where W_h is the weight vector for the attention gate, b is a scalar bias, $w^{(i)}$ is the attention gate, $W h$ is a weight matrix for the bimodal segment embeddings, and b_{la} is the bias vector.

5 Experiments

In this section, we present and analyze the results of our unimodal and multimodal mental disorder recognition systems. We have trained and validated the models using 5-fold cross-validation. Very often in the data we have different recordings from the same parent talking about their different children. Moreover, there are cases where we have recordings from both parents from the same family speaking about the same child. It has been shown that family history is strongly correlated with the development of several mental disorders [17]. Therefore, we take this information into account while

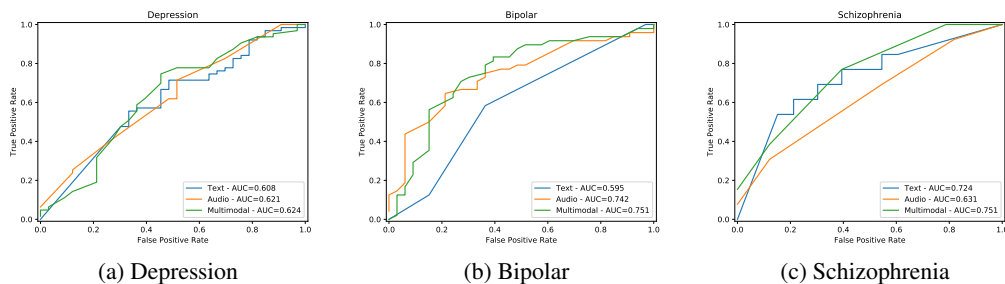| (a) Depression | (b) Bipolar | (c) Schizophrenia |

Figure 5: ROC plots for (a) Depression (b) Bipolar (c) Schizophrenia

splitting the data into different folds. More specifically, we group all the speakers with the same family ID together and use that data either in train or test portion for the folds. This helps us to keep the correlated data points together and makes our training and test sets as independent as possible.

Additionally, our data has imbalance distribution in different categories of mental disorders (Control: 129, Depression: 149, Bipolar: 66, Schizophrenia: 19). To address this problem, we use random oversampling [4] technique and duplicate the randomly selected samples from our two minority classes (i.e. Bipolar and Schizophrenia) and augment them into our data set. Figure 3 illustrates sentiment and mood changes during a five-minute interview for a randomly selected subject with bipolar disorder. The colored vertical bars shows the ground-truth emotion labels in the dataset and the colored text segments above the figure show our model's predicted emotions that match the true emotions. Since there are more than 50 segments in the audio file, we randomly sampled 2 segments from each emotion for the sake of readability of the figure. Figure 4 shows a sample speech from the depression group. Each line represents a segment and the segments are colored based on the attention weights learned in our multimodal attention-based RNN. As we can see from the figure, the segments where the parent talks about the anxiety level of their kids and their communication problems have higher weights showing that the network is paying more attention to those segments.

Table 2 shows the correct classification rate or accuracy of recognition for different mental disorders. The control columns in the table are the accuracies of predicting control group against any other disorder. As we can see from the table, the proposed multimodal architecture has better accuracy than the unimodal systems in most cases. We have achieved an accuracy of 74.35% on average for predicting different mental disorders. As we expected the contextualized word features from the XLNet and BERT language models are more reliable than traditional feature extraction methods such as bag-of-words (BOW). Figure 5 illustrates the Receiver Operating Characteristic (ROC) diagrams of unimodal and multimodal systems for Depression, Bipolar, and Schizophrenia classes. As we can see from the figure, the multimodal architecture has better ROC curve and consequently higher Area Under the Curve (AUC). The AUC score of 0.751 for Schizophrenia which was the most imbalanced class with only 13 positive samples shows the ability of our model in handling imbalanced data.

6   Conclusions & Future Works

Automated classification with multimodal deep learning adds scalability to the use of speech in the prediction of mental health outcomes. In this research, we propose a multimodal deep learning framework for automatic mental disorders prediction. Our results show that mental disorders can be predicted automatically through multimodal analysis of speech samples and language contents extracted from clinical interviews. Using weighted feature concatenation fusion algorithm has achieved the average accuracy of 74.35% (RF trained on learned document representations of two-level LSTMs). The average AUC of 70.5% for RF, over 5-fold cross-validation, indicates that our model could have successfully handled the imbalance dataset. Future steps include investigating offspring's recorded audio samples alongside their parents' speech samples since family history has a great impact on most of the major mental disorders occurrences. Moreover, we would like to improve our mental mood prediction analysis by incorporating clinical narrative summary for every subject.

References

[1] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *PAMI*, 41(2):423–443, 2019.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[4] James C Candy and Gabor C Temes. *Oversampling delta-sigma data converters: theory, design and simulation*. IEEE Press,, 1992.

[5] Minmin Chen. Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*, 2017.

[6] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[7] John K Darby and Harry Hollien. Vocal and speech patterns of depressive patients. *Folia Phoniatrica et Logopaedica*, 29(4):279–291, 1977.

[8] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964. IEEE, 2014.

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

[10] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006, 2012.

[11] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[12] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1068–1077. JMLR. org, 2017.

[13] Judith A Hall, Jinni A Harrigan, and Robert Rosenthal. Nonverbal behavior in clinician patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1995.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[15] Thad Hughes and Keir Mierle. Recurrent neural networks for voice activity detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7378–7382. IEEE, 2013.

[16] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013.

[17] Thomas Munk Laursen, Trine Munk-Olsen, Merete Nordentoft, and Preben Bo Mortensen. Increased mortality among patients admitted with major psychiatric disorders. *The Journal of clinical psychiatry*, 2007.

[18] François Mairesse, Joseph Polifroni, and Giuseppe Di Fabbrizio. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5093–5096. IEEE, 2012.

[19] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, 2018.

[20] Elliot Moore II, Mark A Clements, John W Peifer, and Lydia Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107, 2008.

[21] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM, 2011.

[22] Habibeh Naderi, Behrouz Haji Soleimani, Sheri Rempel, Stan Matwin, and Rudolf Uher. Multimodal Deep Learning for Mental Disorders Prediction from Audio Speech Samples. *arXiv e-prints*, page arXiv:1909.01067, Sep 2019.

[23] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[24] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. 2019.

[25] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[26] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[27] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Association for Computational Linguistics (ACL)*, pages 2236–2246, 2018.