We are using shared late days as indicated on the spreadsheet.

Multimodal Patient Evaluation for Depression and Anxiety

Stanford CS224N {Custom} Project

Ally Nakamura
Department of Computer Science
Stanford University
allynak@stanford.edu

Roshan Swaroop
Department of Computer Science
Stanford University
swaroop1@stanford.edu

Abstract

We introduce a multimodal approach for predicting Generalized Anxiety Disorder-7 (GAD-7) and Patient Health Questionnaire-9 (PHQ-9) scores from patient interviews. Leveraging data from Stanford Medicine's Partnership in AI-Assisted Care (PAC), we employ three models to analyze text, audio, and video modalities: SieBERT for sentiment analysis, ViT FacialEmoRecog for video-based emotion recognition, and Wav2Vec2 for audio-based emotion detection. Principal component analysis (PCA) is utilized for dimensionality reduction to improve model performance in data-constrained, class-imbalanced scenarios. Our experiments evaluated various combinations of modalities, including singular, bimodal (Text-Audio, Text-Video, and Audio-Video), and the fully multimodal approach consisting of text, audio, and video. The highest performance for GAD-7 prediction was achieved with the fully multimodal approach, resulting in a weighted F1 score of 0.64. In contrast, the best result for PHQ-9 prediction was observed using the Text-video combination, yielding a weighted F1 score of 0.70. Our Text-Audio-Video approach outperforms the baseline across all metrics and showcases the potential for multimodal analysis in predicting mental health scores. These results highlight the importance of understanding the nuances of each modality and their potential interactions for mental health assessment.

1 Key Information

Mentor: Elaine Sui (224N TA) **External Collaborators:** (if you have any): Zane Durante, Neha Srivathsa, PAC colleagues **Sharing project:** No

2 Introduction

Clinical depression affects over 280 million people globally, with more than 800,000 suicides per year (Flores et al., 2022). The current diagnosis process, involving psychiatric evaluations by licensed psychiatrists, is resource-intensive, time-consuming, and unscalable, leading to inadequate treatment for over 80% of patients. Therefore, more efficient and accurate diagnosis methods are needed to alleviate the burden on healthcare providers and deliver faster care.

Clinicians consider patients' clinical history and behavior during screening for depression, with the latter being more significant for accurate diagnoses (Krishna and Anju, 2020). Key components of patient interviews include facial expressions, body movements, and speech patterns. People with depression often speak in short phrases and avoid eye contact compared to those without depression (Flores et al., 2022). This study aims to improve prediction performance by combining multiple data modalities. We use PHQ-9 (Patient Health Questionnaire) and GAD-7 (Generalized Anxiety Disorder) scores to evaluate depression (Williams, 2014; Spitzer et al., 2006). Both questionnaires measure the frequency of various symptoms experienced in the last two weeks, with scores used to interpret the severity of depression or anxiety.

Deep learning techniques are promising for understanding the complex nature of anxiety and depression by analyzing patient behavior during screening. We employ SiEBERT, a fine-tuned checkpoint of the RoBERTa-large model, for binary sentiment classification and expand it by incorporating audio and video modalities using a multi-layer perceptron (MLP) to combine outputs of Wav2Vec2-base pretrained for speech emotion recognition (SER) and Google's Vision Transformer (ViT-base) pre-trained on fer2013, a prominent facial expression recognition (FER) dataset. Our experiments show limitations of data scarcity on multimodal learning and marginal performance improvements over solely SiEBERT-based prediction.

3 Related Work

Multimodal deep learning techniques for depression detection have gained popularity in recent years. Traditional components such as RNNs and CNNs perform poorly with limited data, which is common in clinical applications. Recent studies found success with transfer learning, using pre-trained models like VGGish for audio tasks (Hershey et al., 2016) and fine-tuned BERT for text classification (Devlin et al., 2018). Two common ways to integrate modalities are early fusion and late fusion (Barnum et al., 2020; Huang et al., 2020).

3.1 Multimodal Approaches Using Early Fusion

Guohou et al. (2020) proposed a two-layer model that extracts vocal, visual, and text features, feeding them to nine sub-models before aggregating by question category and passing into Support Vector Regressor and Random Forest models for depression prediction. They found that multimodal features improved performance and question-level featurization was more effective than interview-level featurization. In Pampouchidou et al. (2016), they compared early feature fusion and post-decision fusion using high and low-level features from audio, video, and text. In feature-level fusion, modalities were combined into a single feature vector, while decision fusion combined labels produced from individual classifiers through intersection and union operations. The decision fusion model outperformed the feature fusion model, with F1 scores of 0.63 vs. 0.5 for depressed classes and 0.91 vs 0.86 for non-depressed classes.

3.2 Multimodal Approaches Using Late Fusion

AudiBERT, a framework combining BERT with pre-trained audio models like VGGish, SincNet, and Wav2Vec, extends a self-attention mechanism and BiLSTM to text and audio outputs before fusing the embeddings in the final classification layer. It also considers personal features such as gender and age, known to be confounding factors. AudiBERT variants achieved higher F1 scores than baselines, improving scores by 6 to 30% (Toto et al., 2021). Another study used a speech and linguistic approach, employing a pre-trained VGG-15 network and Gated Convolutional Neural Network followed by an LSTM for speech, and BERT followed by a CNN and LSTM for text embeddings. Combining the two embeddings and feeding them into a fully-connected layer led to a 0.283 CCC score increase (Rodrigues Makiuchi et al., 2019).

AudiFace, an extension of AudiBERT, includes temporal facial features as input (landmarks, eye gaze, and action units), generating multivariate feature vectors from video frames. After an LSTM layer, the embeddings across modalities are concatenated for the last classification layer. AudiFace outperformed AudiBERT for 13 out of the 15 datasets, suggesting a multimodal approach with image, text, and audio achieves the highest performance (Flores et al., 2022).

4 Approach

Due to data constraints, this study uses three pre-trained models to evaluate inputs by modality before combining the results through late fusion. The models chosen are described below, as well as the approach for fusing outputs. Thematically, our approach to our late fusion involved corresponding input features to each question of the survey presented in our data section (5.1).

4.1 Text Inputs

For text inputs, a Hugging Face model "SieBERT" was used, which is a fine-tuned checkpoint of RoBERTa-large designed for binary sentiment analysis. It outperforms other models like DistilBERT SST-2 in binary sentiment analysis for English texts (93.2% vs 78.1% average accuracy) Hartmann et al. (2022). SieBERT also demonstrates improved performance on multi-class sentiment datasets with small examples compared to RoBERTa (3x improvement within 1 epoch). To use this model for inference, we took our preprocessed text data for each interview that was further subdivided into each particular question. Given the 512-token limitation for this transformer architecture, it proved to be a semantically viable chunking strategy to fit our model constraint to feed our model text inputs question-by-question. The outputs were then grouped by interview, and the first two scores corresponding to the first few questions were knocked down to the lack of information they contained, as described in 5.1 below. The resulting 9 binary sentiment scores in the inclusive range $[0, 1]$ formed a vector that was used as a low-dimensional feature set to feed into our MLP. We did not invoke any dimensionality reduction techniques on text, given that we had just a single sentiment score feature for each question.

4.2 Video Inputs

For video inputs, we used Google's ViT for Image Classification pretrained on fer2013, a dataset of 30,000 RGB human facial expressions images sized 48×48. The index to emotion mapping is supplied by the dictionary {"angry": 0, "disgust": 1, "fear": 2, "happy": 3, "neutral": 4, "sad": 5, "surprise": 6 } We hypothesized that this model's expressiveness would yield features for anxiety and depression classification. Due to compute constraints, we used ViT-B/16, the smaller ViT architecture with 12 layers in the Transformer encoder. The model achieves nearly 70% accuracy on fer2013, competitive with the benchmark on PapersWithCode (Dosovitskiy et al., 2020). Due to compute constraints, sampling large numbers of frames was prohibitively expensive, so we ultimately decided to sample sparsely. Consistent with our survey question based approach to designing input features, we sampled with respect to each question, extracting three frames per question, or roughly one frame every 10 seconds per interview. We had 33 frames per interview, each with an inference output of a 7-vector encoded by index with emotion scores, producing a 231 input vector into our MLP.

4.3 Audio Inputs

For audio, we opted for Wav2Vec2-base, a lighter-weight automatic speech recognition (ASR) model developed by Facebook AI. Wav2Vec2 is designed to convert raw audio waveforms into text and has shown state-of-the-art performance on a variety of ASR benchmarks. The particular model we ran inference on used the most widely recognized emotional recognition (ER) dataset IEMOCAP, and it follows a conventional evaluation protocol for the dataset where the unbalanced emotion classes are dropped to leave the final four classes with a similar amount of data points and then cross-validation is run on five folds of the standard splits. (Baevski et al., 2020) Using timestamp data, we take an audio sample per each question that was the minimum between 20 seconds and the full length of the question. After running inference, we receive a 4-vector in the same scheme as images, with the emotion scoring occurring in the order happy, neutral, angry, sad. This ultimately produces a 44-vector per interview, and we used the same technique as in the image modality to determine number of principal components to input to our model.

4.4 Late Fusion

Due to our dataset's scarcity, we employed a late fusion approach, leveraging pre-trained models with our data mainly used for strategic fusion of the independent modalities. Text, audio, and video inputs were segmented per question and evaluated separately. The outputs formed question-specific feature vectors as mentioned above, which were fused as input for a single hidden-layer multi-layer perceptron (MLP). Compared to simply ensembling results, the MLP allows the model to achieve more expressiveness by learning non-linear mappings from the fused results across all of the modalities. We also implemented a logistic regressor to illustrate if linear mappings could adequately predict from our feature set. To reduce input dimensionality, we applied principal component analysis (PCA) using Scikit-learn's implementation, which employs Singular Value Decomposition (SVD).

We conducted experiments to determine the optimal number of principal components, balancing data complexity and avoiding the curse of dimensionality.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{1}$$

where $\mathbf{X}$ is the original data matrix (m x n), $\mathbf{U}$ is the left singular vectors matrix (m x r), $\mathbf{S}$ is the diagonal matrix of singular values (r x r), $\mathbf{V}^T$ is the transpose of the right singular vectors matrix (r x n), and r is the rank of the matrix $\mathbf{X}$. To reduce the dimensionality of the original matrix $\mathbf{X}$, we truncate the matrices $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}^T$ by selecting the first k columns of $\mathbf{U}$, the first k singular values of $\mathbf{S}$, and the first k rows of $\mathbf{V}^T$, where k is the desired reduced dimension. We denote the reduced matrices as $\mathbf{U}_k$ (m x k), $\mathbf{S}_k$ (k x k), and $\mathbf{V}_k^T$ (k x n). Then, the reduced matrix $\mathbf{X}_k$ can be obtained by multiplying these reduced matrices:

$$\mathbf{X}_k = \mathbf{U}_k\mathbf{S}_k\mathbf{V}_k^T \tag{2}$$

Here, $\mathbf{X}_k$ is the reduced matrix of size m x n, which is a low-rank approximation of the original matrix $\mathbf{X}$. We evaluated SVD results on each modality with an MLP and logistic regression, as well as all combinations of modalities.

4.5 Baselines

Sentiment analysis on patients' text responses served as a reasonable proxy for assigning GAD-7 and PHQ-9 scores. Binary sentiment analysis on a positive-to-negative scale was chosen as the most performant task with Code (Accessed March 4, 2023b) with Code (Accessed March 4, 2023c) with Code (Accessed March 4, 2023a). SieBERT was selected as the primary baseline due to its performance and ability to process longer sequences. This architecture was chosen because it can handle up to 512 tokens. Since interview transcripts were longer, we analyzed each of the 11 questions individually, averaging sentiment scores and excluding the first two questions due to their procedural nature. SieBERT's question-level results were passed through a logistic regressor, cross-validated, and weighted F1 score was calculated.

5 Experiments

5.1 Data

The data is provided by the Partnership in AI-Assisted Care (PAC), a Stanford's Vision Lab subgroup in collaboration with Stanford Medicine. The raw data, collected during a 2020 behavioral health study, involved patients answering an 11-question survey about their emotional well-being before various appointments, including specialists and primary physicians, both in-person and via Zoom.

1. *What is your name and what is the date?*
2. *What is the purpose of your visit?*
3. *How are you feeling today?*
4. *How were you feeling emotionally the last week?*
5. *How were you feeling physically the last week?*
6. *What experiences and comments have other people made about your condition?*
7. *How have your emotional and physical abilities affected your life in the past week?*
8. *Tell us about a recent good experience and how it made you feel.*
9. *What puts you in a good mood?*
10. *How often do you feel this way lately?*
11. *When was the last time you felt really happy?*

110 interviews were conducted and recorded, with video lengths between 143 and 904 seconds and total frames between 3597 and 40812. Discrepancies existed between Zoom and in-person interviews, with Zoom having 25 frames per second, 640x360 frame size, and .mp4 file type, while in-person interviews had 60 frames per second, 1920x1080 frame size, and .mkv file type. We divided these videos into three data forms: text, image, and sound. Python's moviepy library was used to convert video formats into .wav recordings (sound modality) and sample frames (image modality). Videos were transcribed using Google's tool, with interviewer speech timestamps recorded in a JSON object.

We segmented videos into 11 questions, recording and sampling text, audio, and images only when the patient was speaking. This allowed us to design input features on a question basis. External collaborator Neha extracted the three modalities from video. While the data is novel, it is limited and insufficient to train a deep learning model from scratch. Class imbalances were present: for PHQ-9, about 70% of interviews fell in the 0 bucket, while nearly 60% did for GAD-7. Only 11 patients were 'severe' for GAD-7 and 5 for PHQ-9.

5.2 Evaluation method

For each experiment, cross-validated estimates were calculated for the results. Due to the class imbalances present in the original dataset, we calculated weighted F1 scores and generated confusion matrices to illustrate the model's performance. The formula for a weighted F1 score calculation can be expressed as:

$$WeightedF1 = \sum_{i=1}^{n} w_i \cdot \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{3}$$

where n is the number of classes, w_i is the weight of class i, $Precision_i$ is the precision of class i, and $Recall_i$ is the recall of class i. We also noted raw accuracy, which is simply the fraction of correct predictions over total predictions made. Our presented evaluations have one strong limitation; our total dataset has just 109 examples, so evaluation was typically done on just 11 samples set aside for testing. This introduces a great deal of stochasicity into our metrics, especially given the class imbalances present. We aimed to combat by re-training our models 5 times and presenting the median metrics achieved.

5.3 Experimental details

5.3.1 Multimodal Combinations

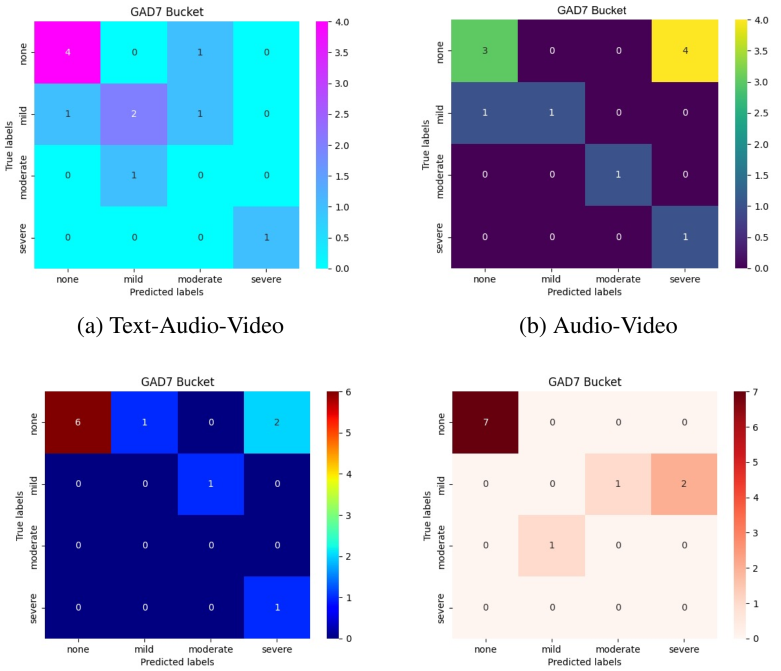
The following combinations were evaluated in this study: 1) singular modalities (SieBERT, the Wav2Vec2 variant, and FacialEmoRecog), 2) paired modalities (Text-Audio, Text-Video, and Audio-Video), and 3) the fully multimodal approach consisting of text, audio, and video (TAV). The pre-trained models were run off the shelf. For multimodal approaches, results from each model were concatenated into larger feature vectors before being processed using PCA to reduce the dimensions. These final results were passed into an MLP or logistic regressor as described below.

5.4 MLP Settings

Both our single hidden layer Pytorch MLP and scikit-learn logistic regressor were trained with a 90% train-validation, 10% test split. The datapoints were sampled at random. We used k-fold cross validation to combat class imbalances and lack of data, where k was fixed to 10 folds. Input size varied between 9 and 231 depending on the experiment, hidden size was fixed at 32 neurons for all experiments, and output size was fixed at 4 to account for our 4 discrete buckets. All MLP training had epochs limited to 100, batch size 16, a learning rate between 0.0007 and 0.0013 (with most experiments ran at 0.001), class weights set to balanced to account for class imbalances, and Adam for our optimizer. Given our small dataset, training took a minute at most per experiment, and each experiment was run 5 times to account for stochasicity in metrics.

5.4.1 Dimension Reduction Using PCA

Due to the high dimensionality of the concatenated results, the results were first processed using principal component analysis (PCA) before being passed into the MLP or logistic regressor. To determine the optimal number of components to use, we plotted curves for logistic regression of number of components vs. performance, whereas for MLP we sampled 5 component configurations at random and then manually adjust component number based on which sample led to the best evaluation scores. The optimal component settings used are shown below:



(a) Text-Audio-Video

(b) Audio-Video

(c) Text-Video

(d) Text-Audio

Figure 1: GAD7 Confusion Matrices

Table 1: Optimal PCA Component Settings

	Baseline	Wav2Vec2	FacialEmoRecog	Text-Audio	Text-Video	Audio-Video	TAV
Components	9	11	25	20	20	20	35

5.5 Results

The F1 scores, PCA component plots, and confusion matrices for each of the multimodal approaches are provided below. For the MLP experiments, the Text-Audio-Video model had the highest average accuracy at $\approx 63\%$, beating the baseline by 40%. The Text-Audio-Video model also had the highest F1 scores for all three metrics (GAD-7, PHQ-9, and average), beating the baseline by $\approx 27\%$.

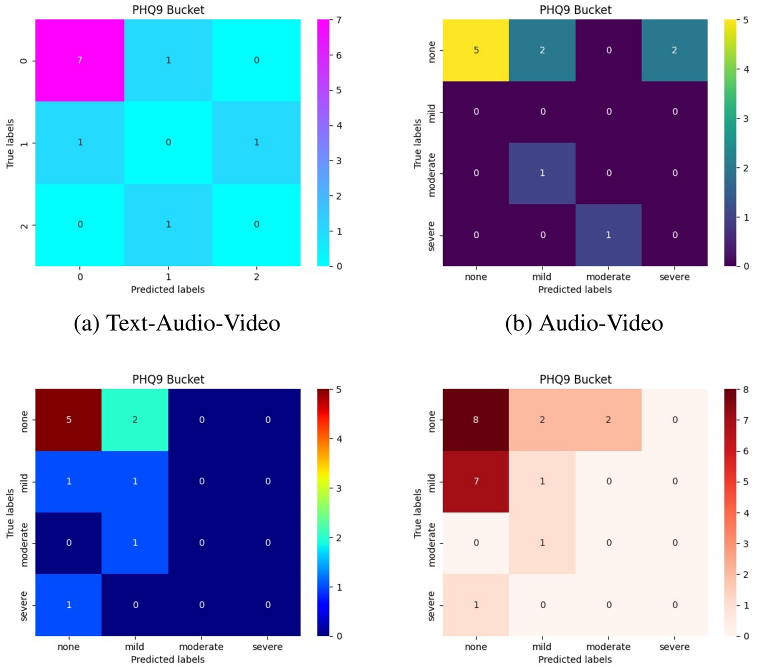
Table 2: Accuracy Using MLP

	Baseline	Wav2Vec2	FacialEmoRecog	Text-Audio	Text-Video	Audio-Video	TAV
GAD-7	0.4545	0.3182	**0.6364**	0.4545	**0.6364**	0.5455	**0.6364**
PHQ-9	0.4545	0.3636	0.4545	0.4545	0.5455	0.4545	**0.6364**
Average	0.4545	0.3409	0.54545	0.4545	0.59095	0.5	**0.6364**

Table 3: F1 Scores Using MLP

	Baseline	Wav2Vec2	FacialEmoRecog	Text-Audio	Text-Video	Audio-Video	TAV
GAD-7	0.5179	0.3285	0.5956	0.5289	**0.7**	0.5895	0.6623
PHQ-9	0.5082	0.4727	0.5333	0.4312	0.5152	0.5844	**0.6364**
Average	0.51305	0.4006	0.56445	0.48005	0.6076	0.58695	**0.64935**

For the log regression experiments, the results show that the Text-Audio-Video model performed best for GAD-7 task with an F1 score of 0.52, beating the baseline by 44%. The Text-Audio model performed best for both the PHQ-9 task and overall tasks, with F1 scores of 0.60 and 0.55, beating the baseline by 5.26% and 17.02%, respectively.



(a) Text-Audio-Video

(b) Audio-Video

(c) Text-Video

(d) Text-Audio

Figure 2: PHQ9 Confusion Matrices

Table 4: F1 Scores Using Log Regression

	Baseline	Wav2Vec2	FacialEmoRecog	Text-Audio	Text-Video	Audio-Video	TAV
GAD-7	0.36	0.34	0.40	0.49	0.47	0.46	**0.52**
PHQ-9	0.57	0.32	0.48	**0.60**	0.56	0.45	0.56
Average	0.47	0.33	0.44	**0.55**	0.52	0.46	0.54

6 Analysis

6.1 Performance Based On Use Case

While it is common for patients to be diagnosed with both depression and anxiety, it is important to treat them as separate entities and recommend model architectures per use case. As shown in Table 2, while the Text-Audio-Video approach (TAV) had the highest average performance and PHQ-7 performance, the Text-Video model (TV) performed the best for predicting GAD-7 scores. Based on these results, we recommend a TAV approach for detecting depression, but a TV approach for detecting the presence of anxiety.

6.2 Which modality is the most significant?

While the TAV model performed best overall, we recognize that preparing text, audio, and video data is expensive and not always feasible. Therefore, we analyze which modalities are most significant ensuring accurate predictions. The results in Table 2 show that out of the unimodal approaches, FacialEmoRecog performs the best across all metrics, suggesting that it is the most informative modality compared to text and audio input. Out of the bimodal approaches, Text-Video has the highest average and GAD-7 performance, while Audio-Video has the highest PHQ-9 performance. These results suggest that video is the most important modality since it is present in both approaches, and it should be prioritized in cases where there are resource constraints and not all modalities are available.

6.3 Does the model overestimate or underestimate GAD-7 and PHQ-9 Scores?

In addition to overall F1 scores, we analyze the confusion matrices to determine whether the models tend to overestimate or underestimate cases of depression and anxiety in the test sets. We define "overestimate" as predicting a higher severity score than the true label, and "underestimate" as predicting a lower score. The count reflects how severe the over or underestimation is. For example, if a model predicts a score of 2 when the true label is 0, we increment the overestimation score by 2.

Table 5: Over and Underestimate Scores

	Text-Audio	Text-Video	Audio-Video	Text-Audio-Video
GAD-7 (Over)	5	8	12	3
GAD-7 (Under)	1	0	1	2
GAD-7 Bias	83% over	100% over	92% over	50% over
PHQ-9 (Over)	6	2	8	2
PHQ-9 (Under)	11	5	2	2
PHQ-9 Bias	65% under	71% under	80% over	equal

Table 5 shows that the TAV model is the least likely to over and underestimate scores for both GAD-7 and PHQ-9, further making it an ideal choice for real application. Across all approaches, all models overestimated for GAD-7 scores, with Text-Video most likely to severely overestimate levels of anxiety. Therefore, future models may benefit from adjustments that reduce this bias towards overestimation. For PHQ-9, both text-based bimodals were likely to underestimate depression scores, while the Audio-Video model was likely to overestimate depression scores. These results suggest that relying solely on text-based inputs can lead to underestimates of depression severity (perhaps because people have more control over the content of their speech), while relying on audio and video data can lead to overestimates.

6.4 Limitations

This study has several limitations. The small, imbalanced dataset makes results susceptible to stochasticity. Although using 10-fold cross-validation and a shallow MLP, the limited final test set remains a challenge. Sampling audio and video frames based on interview timestamps might not capture the subject's actual speech, potentially missing key behavioral timeframes. Aligning features with questions and using multiple samples aimed to mitigate this issue. Another limitation is the unequal representation of modalities in feature concatenation. Audio was reduced to 11 features from 44, close to text's 9, and video's 231-dimension feature space was limited to 25, but parity between modalities wasn't achieved. Running GAD and PHQ trials independently could have improved optimal component selection; running parallel experiments may have influenced both.

7 Conclusion

Experimenting with this dataset illuminated how to model problems in data-constrained, class-imbalanced scenarios, and we achieved better than baseline performance while keeping the dimensionality of our MLP's input features rather small. In the future, we aim to improve model performance as the primary objective. One concrete avenue to explore is supplementing our data with USC's Distress Analysis Interview Corpus (DAIC-WOZ) dataset and/or their Extended DAIC dataset, which at a minimum would increase our data size by 3x. By interleaving these two data sources, we hope to unlock more complex relationships between our data sources. We may gain the ability to input more features from each model into a deeper MLP, and to potentially extract embeddings from the last layer of each corresponding model as an MLP input rather than using the result of each output layer. Because of current data limitations, we lost much of the complexity and granular features that could be achieved if these methods were viable. Given the recent rise of multi-modal large language models such as GPT-4, we also look forward to the near future where similar models will be available offline or in a HIPAA-compliant fashion to service data that is either personal health information (PHI) or personal identifiable information (PII). Such model architectures may greatly outpace current methods given early trials in orthogonal domains, and we hope to implement PHQ-9 scoring with them.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning.

Papers with Code. Accessed March 4, 2023a. Emotion recognition in conversation on meld. `https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld`.

Papers with Code. Accessed March 4, 2023b. Sentiment analysis on SST-2 binary classification. `https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary`.

Papers with Code. Accessed March 4, 2023c. Speech emotion recognition on crema-d. `https://paperswithcode.com/sota/speech-emotion-recognition-on-crema-d`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Ricardo Flores, ML Tlachac, Ermal Toto, and Elke Rundensteiner. 2022. Audiface: Multimodal deep learning for depression screening. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 609–630. PMLR.

Shan Guohou, Zhou Lina, and Zhang Dongsong. 2020. What reveals about depression level? the role of multimodal features at the level of interview questions. *Information Management*, 57(7):103349.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2016. CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3.

Swathy Krishna and J. Anju. 2020. Different approaches in depression analysis : A review. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 407–414.

Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, Fan Yang, and Manolis Tsiknakis. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 27–34, New York, NY, USA. Association for Computing Machinery.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 55–63, New York, NY, USA. Association for Computing Machinery.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Ermal Toto, M. L. Tlachac, and Elke A. Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4145–4154. ACM.

Nerys Williams. 2014. PHQ-9. *Occupational Medicine*, 64(2):139–140.