AudiFace: Multimodal Deep Learning for Depression Screening

Ricardo Flores RFLORES@WPI.EDU
ML Tlachac MLTLACHAC@WPI.EDU
Ermal Toto TOTO@WPI.EDU
Elke Rundensteiner RUNDENST@WPI.EDU
Data Science & Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA

Abstract

Depression is a very common mental health disorder with a devastating social and economic impact. It can be costly and difficult to detect, traditionally requiring a significant number of hours by a trained mental health professional. Recently, machine learning and deep learning models have been trained for depression screening using modalities extracted from videos of clinical interviews conducted by a virtual agent. This complex task is challenging for deep learning models because of the multiple modalities and limited number of participants in the dataset. To address these challenges we propose AudiFace, a multimodal deep learning model that inputs temporal facial features, audio, and transcripts to screen for depression. To incorporate all three modalities, AudiFace combines multiple pre-trained transfer learning models and bidirectional LSTM with self-Attention. When compared with the state-of-the-art models, AudiFace achieves the highest $F1$ scores for thirteen of the fifteen different datasets. AudiFace notably improves the depression screening capabilities of general wellbeing questions. Eye gaze proved to be the most valuable of the temporal facial features, both in the unimodal and multimodal models. Our results can be used to determine the best combination of modalities, temporal facial features, as well as clinical interview questions for future depression screening applications.

1. Introduction

1.1. Background

Depression is a very common mental illness with large social economic impacts. According to (Bloom et al., 2012) the global cost projected of mental health conditions in 2030 is 6.0 trillion US dollars. Traditionally, depression is diagnosed by a psychiatrist or psychologists through clinical interviews. Due to the limited supply of trained clinicians and length of interviews, these interviews can be very costly for healthcare systems. Worse yet, during the global pandemic, the number of people with depression has significantly increased (Czeisler et al., 2020; Hamouche, 2020). On the other hand, with the increasing technology on speech recognition and chat-bots, virtual agents represent an affordable alternative for solving this problem of gathering and analyzing clinical interview data. Therefore, detecting depression using machine and deep learning has become an important research task.

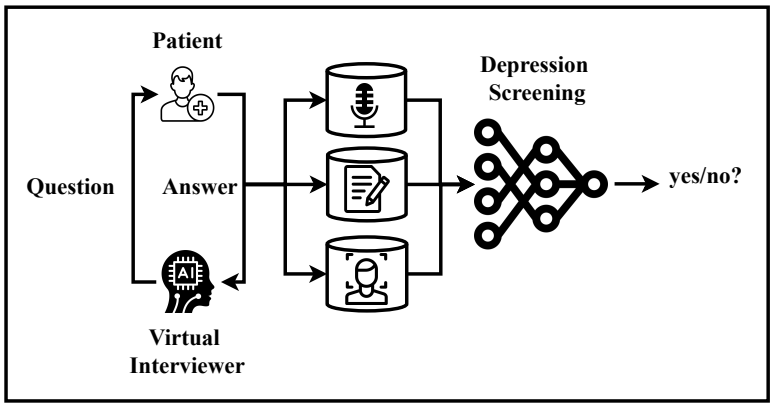

Figure 1: Depression screening through a virtual agent that asks questions to a patient. The clinical interview voice recordings, transcripts, and facial features are then used as input for training a deep learning model to screen for depression.

Prior research has used machine and deep learning for successful diagnostic and prognostic modeling (Dwyer et al., 2018). In particular, depression screening models have used audio (Flores et al., 2021; Di Matteo et al., 2020; McGinnis et al., 2019) and text (Tlachac and Rundensteiner, 2020; Senn et al., 2022) as input. Voice recordings represent the one of the most important inputs, because they allow classification models to leverage both the audio and text (Asgari et al., 2014; Rodrigues Makiuchi et al., 2019; Tlachac et al., 2021; Toto et al., 2021; Tlachac et al., 2022a). However, patient screening depression is essentially a multimodality problem (Schoneveld et al., 2021) since each modality may contain important information for screening models. Leveraging temporal facial features, audio, and transcripts in a multimodal deep learning model to screen for depression is an open problem.

1.2. Motivating Example

As cost of clinical interviews increase due to a shortage of trained clinicians (Bureau of Health Workforce et al., 2021), a virtual agent offers an inexpensive and available alternative (DeVault et al., 2014). The aim of the virtual agent is to ask relevant clinical interview questions to ensure sufficient self-disclosure of the patient about her mental state. Then, the core questions are stored as audio clips, text transcript, and temporal facial features to be uses as input into a deep learning model, which then predicts if the patient has depression, as displayed in Figure 1. However, no research has yet been conducted to determine which is the most relevant question to ask using these three modalities. In the same context, there is no deep learning model that study the effect of each type of facial feature on predict depression. Notably, the facial features relevance may even differ by the specific question. Therefore, this research to identify the relevant question and facial features that is crucial for training a deep learning model to power an effective and efficient virtual screening agent.

1.3. Problem Definition and Challenges

Given a set of temporal facial features, voice recordings, and their respective transcripts with each composed of a patient's answers to a core question as well the patient's depression label, our goal is to determine each participant as depressed or not depressed. Also, we want to know which are the relevant question that deep learning models need to achieve the best patient depression predictions, and the effect of including temporal facial features. Notably, we use both unimodal and multimodal models to screen for depression with facial features.

Training sequential deep learning models on long audio, transcript, and temporal facial features recordings are challenges as these models can suffer from a vanishing problem (Pascanu et al., 2013). Also, long sequences have a high computational cost, especially if leveraging big deep learning architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and recent state-of-the-art transformers models. Further, most audio-visual datasets with depression labels consist of a small sample size (Cummins et al., 2015). This makes it difficult to train deep learning models from scratch. Consequently, deep learning models trained with these datasets tend to report low evaluation metrics (Ma et al., 2016; Huang et al., 2018).

1.4. Our Deep Learning Approach

In this research, we thus seek to improve depression screening capabilities of videos by leveraging multiple modalities. To accomplish this, we introduce AudiFace, a multimodal deep learning that leverages temporal facial features, audio, and transcripts. In particular, our approach to modeling the temporal facial features is unique. Thus, we first compare the depression screening abilities of landmark, eye gaze, action unit, and all temporal facial features in unimodal deep learning models. We then incorporate these different types of facial feature into AudiFace to also take advantage of the predictive abilities of corresponding audio and transcripts. We tackle the aforementioned challenge of small sample size by leveraging transfer learning models pre-trained on large corpuses of either audio or text. Further, to accommodate the longer sequences, we use self-Attention in the models.

To validate our approach, we train the unimodal and multimodal models on modalities extracted from videos of responses to 15 clinical interview questions in the popular Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) (Gratch et al., 2014; DeVault et al., 2014). Thus, we can identify which questions are most useful for depression screening with different facial features types and combinations of modalities. Our research is important in that it informs the development of future depression screening applications.

Generalizable Insights about Machine Learning in the Context of Healthcare

1. A unique and valuable deep learning approach to classify temporal facial features, allowing us to assess the depression screening ability of three types of facial feature.

2. **AudiFace**, a state-of-the-art multimodal deep learning method to classify videos, is applicable for the small datasets common in the healthcare domain.

3. Comparison of the depression screening performance of unimodal and multimodal models that leverage temporal facial features, audio, and transcripts.

4. Identification of which individual questions are most valuable for depression screening.

2. Related Work

Prior research has detected depression with many different modalities and methods. In addition to using individual modalities in classification models, multimodal learning classification using text, audio, or image as input (Ramachandram and Taylor, 2017) has grown popular in the last decade. The different modalities introduce different challenges for machine learning and deep learning, due to the complex nature of data (Baltrušaitis et al., 2018). While multiple studies have attempted to use deep learning for depression screening (Sharifa et al., 2012; Ma et al., 2016; Dibeklioğlu et al., 2015; Huang et al., 2018; Al Hanai et al., 2018; He et al., 2021), the models tend to perform poorly when the datasets have small sample size, due to the challenge of training deep learning models from scratch.

Multimodal deep learning models typically use traditional architectural components (Victor et al., 2019) such as Recurrent Neural Networks (RNNs) for audio and text as well as Convolutional Neural Networks (CNNs) for audio and images. For example, from patient video interviews, He et al. (2021) detects depression severity by using CNNs with audio and videos input while Al Hanai et al. (2018) trained RNNs with text and audio data to detect depression. Most of these studies leverage the OpenFace software (Baltrušaitis et al., 2016), which allows to extract temporal facial features from images or videos. Then, these features can be used for emotional or neurodevelopmental disorder detection. For example, in Cuve et al. (2021) the temporal eye gaze tracking features are used to detect autism.

Transfer learning (West et al., 2007) leverages models pre-trained in a similar task to achieve great classification performance for small target datasets. The most popular pre-trained audio representation model is VGGish (Hershey et al., 2017) used for diverse audio tasks (Xie and Virtanen, 2019; Cerutti et al., 2019; Gemmeke et al., 2017; Brown et al., 2020a). Meanwhile, for text classification the famous Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a notably well-established language representation model. Recent research has notably used such pre-trained audio and text models together to detect depression from a small datasets of voice recordings (Rodrigues Makiuchi et al., 2019; Toto et al., 2021; Flores et al., 2022; Tlachac et al., 2022b). These deep transfer learning models performed better than their traditional deep learning and machine learning counterparts. Notably, we extended the work done by Toto et al. (2021), adding temporal facial features to the voice recordings.

In recent years, there has been many proposed natural language processing models that could be substituted for BERT within multimodal depression screening frameworks. However, related research (Senn et al., 2022) discovered that BERT with fine-tuning layers performed better than more recent variations such as RoBERTa (Liu et al., 2019) or DistilBERT (Sanh et al., 2019) when classifying clinical interview transcripts. Additionally, there are some other new transformer-based deep learning neural networks that have been recently proposed, such as DeBERTa (He et al., 2020), BART (Lewis et al., 2019), GPT3 (Brown et al., 2020b), and T5 (Raffel et al., 2020). Furthermore, Zaheer et al. (2020) proposed a variation of BERT to deal with long sequences. As our research in this paper focuses on temporal facial features, we leverage the text and audio components in our mul-

timodal model framework that have previously demonstrated the most success at screening for depression with clinical interview data (Toto et al., 2021; Senn et al., 2022).

3. Methods

3.1. Deep Learning Methods for Audio Classification

For the voice recordings, we use the popular pre-trained audio architecture VGGish (Hershey et al., 2017) to create audio feature embeddings. VGGish transforms voice clips to log Mel spectrograms that are processed by a multilayer convolutional network to extract embeddings vector of size 128 for each second of voice, forming a 2D array that can be used for depression screening. We also experiment with a variation of the VGGish method, including a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) layer over the embeddings layers of VGGish (Gers et al., 2000). In order to capture longer relationships between audio clips, we add self-Attention (Lin et al., 2017) on top of the LSTM.

3.2. Deep Learning Methods for Text Classification

Transformer models have become well-established in text classification in recent years with pre-trained variations applicable for smaller text. The famous Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) in particular has demonstrated previously unattainable success when classifying smaller text datasets. Thus, for the transcripts, we use BERT to create text feature embeddings. Notably, the pre-trained BERT model can be fine-tuned using an extra linear layer on top of the CLS token, which encodes the semantic representation. Also, BERT uses Attention to capture better relations withing words, however, if the sentences sequences are too long, BERT might not get these relations. Therefore, we also implement a bidirectional LSTM with an extra self-attentions layer on top of BERT embedding to capture the longer relations within the transcripts.

3.3. Deep Learning Methods for Temporal Facial Features Classification

Sequential deep learning models can capture complex semantic information in the context of language modeling. For example, recurrent neural network (RNN), Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (GRU) (Cho et al., 2014) have all been used successfully for text classification with large datasets. The same approach can be apply to temporal facial features, where instead of images or video input, we input multivariate features vectors. These features can be generated through the OpenFace software (Baltrušaitis et al., 2016), which is an open source facial behavior analysis toolkit. Similar to audio and text classification, we can apply a bidirectional LSTM with self-Attention over the temporal hidden states from facial features. The self-Attention layer helps to capture long relationship within facial features, this is specially important for DAIC-WOZ dataset (Gratch et al., 2014), where some temporal facial features have long time steps within clinical interviews. Then, we implement two models to screening for depression, LSTM and LSTM with Attentions, we called the last one LSTM Att.

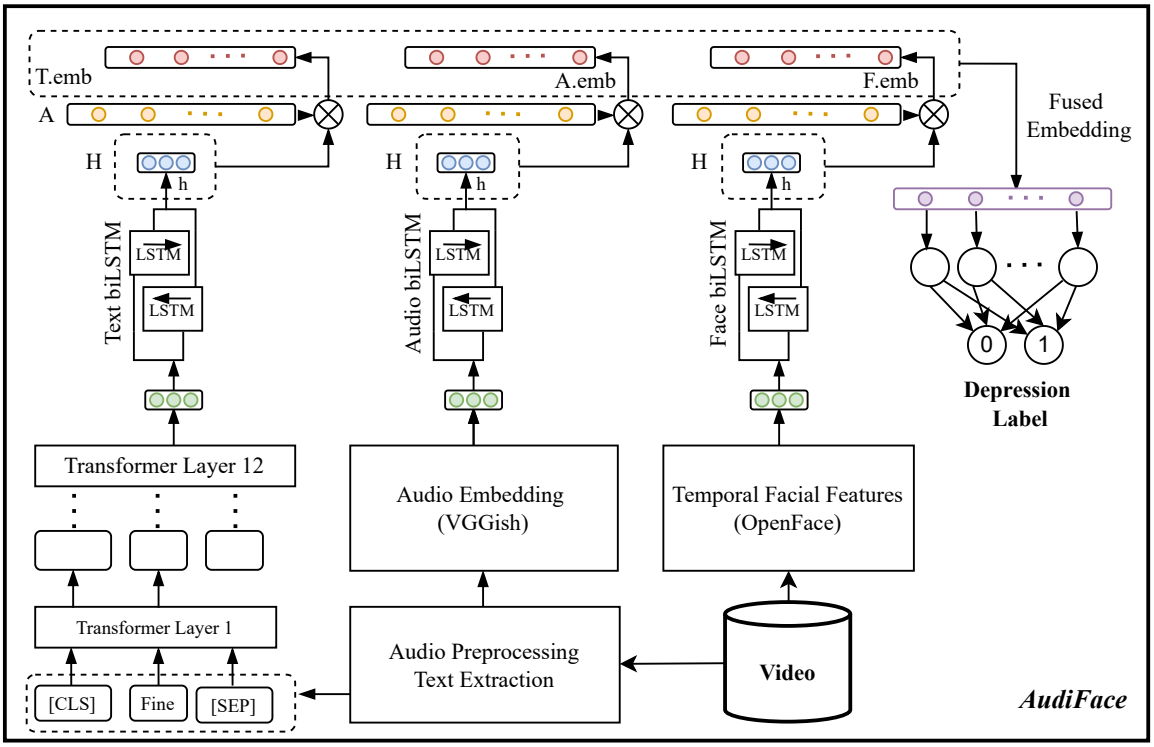

Figure 2: **AudiFace Framework, adapted from Toto et al. (2021).**

3.4. AudiFace Method

As discussed, we use VGGish for audio classification and BERT for transcript classification, as well as LSTM for temporal facial features classification. Therefore, a straightforward solution is combining their embeddings in a single model which we refer to as AudiFace in Figure 2. We leverage self-Attention to each modality after the bidirectional LSTM (biLSTM), then a embedding was created for text (T.emb), audio (A.emb), and facial features (F.emb). Finally, we fuse the embeddings through a linear layer for binary classification.

Recently, Audio Assisted BERT (AudiBERT) (Toto et al., 2021) was proposed for voice classification. Like AudiFace, this deep transfer learning multimodal classification framework also joins the audio and text hidden representations through a linear fusion layer. However, AudiFace also includes the temporal facial features aspect. From DAIC-WOZ dataset, there are three types of facial features: landmark position, eye gaze, and action units. Then, we implement four variations of AudiFace, which are AudiFace-Landmark, AudiFace-Eye Gaze, AudiFace-Action Unit, and AudiFace-All. This last one model, combines all temporal facial features. For the implementation, we use cross entropy loss function, Adam optimizer, $2e^{-5}$ for learning rate, a step size of $2e^{-8}$, and 0.1 as dropout parameter.

4. Datasets

4.1. Dataset Description

The DAIC-WOZ interviews are designed to support the diagnosis of depression (Gratch et al., 2014). This dataset was collected by a virtual agent that interviewed people so the data could be used identify verbal and nonverbal indicators of mental illness (DeVault et al., 2014). The interview was recorded to then extract data from 189 participants, which includes audio, transcript, facial features, and labels with depression screening scores. Audio records ranging between 7 to 33 minutes (with an average of 16 minutes). Facial features, like landmarks, head pose, eye gaze, and facial action units, were extracted using OpenFace software (Baltrušaitis et al., 2016) over the video recording of each sessions. The scores were obtained by administering the first eight questions of the Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2001). The depression screening score is the sum of all PHQ-8 questions, which ranges from 0 to 24. We define the label for each patient as, if a patient screens positive for depression based on the common screening threshold of 10 (Kroenke et al., 2001), in (1).

$$label = \begin{cases} 1, & \text{if score} \geq 10. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

4.2. Methodology for Identifying Questions

We experiment with 15 core questions from the DAIC-WOZ dataset that have at least 90 participant response instances. Alongside each core question, there are a set of follow-up questions to mimic a realistic conversation and increase self-disclosure (DeVault et al., 2014; Bickmore et al., 2005). We follow Toto et al. (2021), and concatenate all follow-up questions.

An important part of the methodology is to separate core and follow-up questions as question wording may vary between interviews with different participants. To this end, we employ a bag of words approach using a noun and verb pair to identify each core question. For example, the question "Is there anything you regret?" can be identified by the pair "anything, regret". We categorize core questions if they contain a "bag of words" set. and as follow-up otherwise. Next, we construct two files, one for audio and other for the facial features, containing each question and participant records. We then use the Natural Language Toolkit (Bird et al., 2009) to implement this bag of words approach.

Table 1 contains each question with their bag of words and the number of participants who answered the question, also we can see the average depressed ratio for each question, which is around 29%, reveling an unbalance datasets. Examples of the real question are in appendix Figure 4.On the other hand, Figure 3 displays the distribution of duration (seconds) and time steps video recorded by each question. Time steps distribution shows shorter answer questions for D5 (diagnosed, depression), D6 (diagnosed, ptsd), and D7 (doing, today), with a median of 2.5, 3.0, and 5.4 seconds respectively. On the other hand, D3 (argued, someone) and D10 (felling, lately) are the longer answered questions. The median time steps for all 15 datasets is 642 which is 21 seconds (0.03 seconds by time step).

Table 1: Questions dataset description with topical bag of words, number of instances, and the ratio of depressed participants, ordered alphabetically.

Dataset	Bag of Words	Instances	Depressed
D1	advice, yourself	102	28.43%
D2	anything, regret	94	29.79%
D3	argued, someone	103	29.13%
D4	controlling, temper	100	30.00%
D5	diagnosed, depression	94	21.28%
D6	diagnosed, ptsd	92	27.17%
D7	doing, today	105	28.57%
D8	dream, job	95	30.53%
D9	easy, sleep	98	27.55%
D10	feeling, lately	92	29.35%
D11	friend, describe	96	26.04%
D12	last, happy	99	28.28%
D13	proud, life	99	28.28%
D14	study, school	95	30.53%
D15	travel, lot	94	27.66%

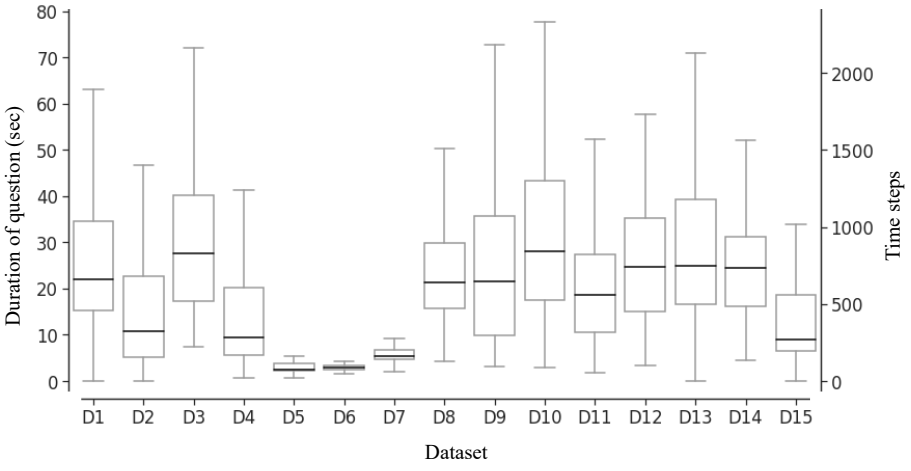

Figure 3: Box plots for the distribution of the duration (left axis) and time steps (right axis) video recording by question. D5 (diagnosed, depression), D6 (diagnosed, ptsd), and D7 (doing, today) have lower median duration and time steps.

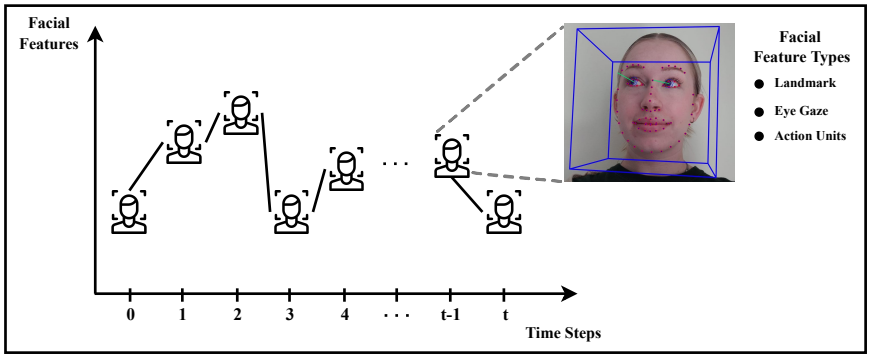

Figure 4: Temporal facial features diagram, where for each participant and time steps we extract the facial landmarks, eye gaze, and action unit.

4.3. Methodology for Temporal Facial Features Questions

One of the main contributions of AudiFace is including the temporal facial features. The features were provided thanks to OpenFace software (Baltrušaitis et al., 2016) that was applied to each patient video recording from DAIC-WOZ depression database (Gratch et al., 2014). In Figure 4, we can see the temporal representation of three types of facial features; landmarks, eye gaze, and action unit. These types of fractures represent a multivariate time series, where in case of landmark has 136 dimensions, eye gaze has 12 dimensions, and action unit has 14 dimensions. We also created an additional datasets with all facial features, carefully standardizing them between -1 and 1. An important part of the methodology is extracting the core question, mention in Table 1, to do so, we use the initial and final time steps from each core question that we got from audio recording, then we extract the facial features from these specific period of time, and order by question and participant.

5. Results

5.1. Classification Evaluation

While accuracy is the most common metric for classification tasks, the $F1$ score is preferred in healthcare because its higher emphasis on the number of true positives. Further, unlike the $F1$ score, accuracy is not the most suitable for unbalanced data. From Table 1, we have an average of 30% of participants who screened positive for depression. Additionally, $F1$ score lets us to easily compare with similar studies like used by Toto et al. (2021). Therefore, we use $F1$ score defined in equation (2), as the metric to evaluate our models. The $F1$ score is notably the harmonic mean of precision and recall. In the Appendix, we also report on the average precision, recall, and AUC for our models.

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{2}$$

where TP refers to the number of true positives, FP refers to the number of false positives, and FN refers to the number of false negatives. For model evaluation, we formed test

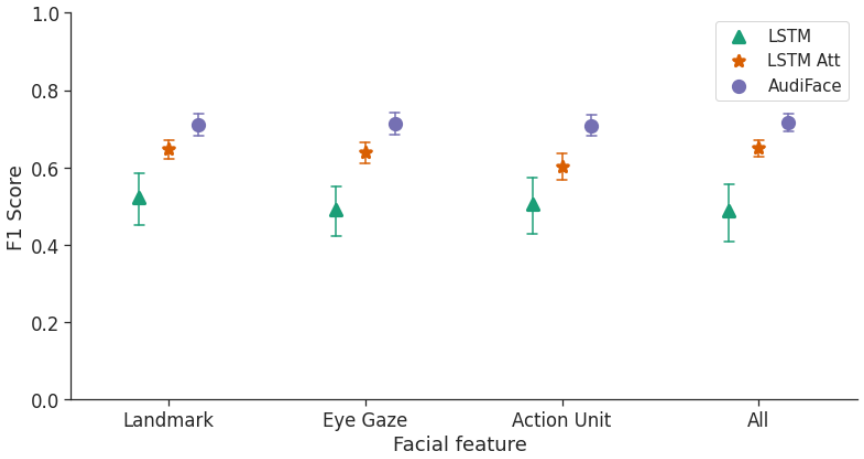

Figure 5: Aggregated results of $F1$ score and facial features, averaging the 15 datasets questions, by the unimodal LSTM and LSTM Attention (LSTM Att) models, and the multimodal AudiFace model. Note, Attention increases the $F1$ score.

sets with 20% of each dataset by selecting a randomly stratified sample. The training set was balanced using upsampling before training the models. Each model configuration was repeated ten times with randomly initialized weights and we reported the average of the 5 highest $F1$ scores. To prevent overfitting, early stopping was employed during training. For temporal facial features, we set to $1,000$ the maximum time steps to input in models.

5.2. Aggregated Results

Figure 5 compares the depression screening capabilities of landmark, eye gaze, action unit, and all temporal facial features. These results are aggregated over all 15 questions. Adding Attention to the LSTM greatly increases the average $F1$ scores and greatly decreases the standard deviation of the unimodal models. Thus, the LSTM Attention models are more robust than the simpler LSTM models, presumably due to their ability to seamlessly handle longer sequences. As such, we recommend using Attention when modeling temporal facial features. The LSTM Attention models using landmark and eye gaze both achieved an average F1 score of 0.64. AudiFace further improved the average $F1$ scores to 0.71.

5.3. Unimodal Models

The unimodal results from Table 2 reveal that different screening modalities are preferred for different individual interview questions. While text-based BERT classifiers obtained the highest average $F1$ score of any question, it was only the best for five of the individual screening questions. Notably, these questions include datasets D5 (diagnosed, depression), D6 (diagnosed, ptsd), and D7 (doing,today), which by far have the smallest average number of time steps, as displayed in Figure 3. We hypothesize this is due to the smaller sequences not containing sufficient information for the models that screen with audio and temporal

Table 2: **Unimodal Models**: average $F1$ scores from BERT, VGGish, (1) LSTM, (2) LSTM Att models by dataset.

	Text	Audio	Temporal Facial Features							
Dataset	BERT	VGGish	Landmarks		Eye Gaze		Action Unit		All	
			(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
D1	0.48	0.71	0.71	0.71	0.24	0.71	0.43	0.60	0.72	**0.74**
D2	0.66	0.69	0.65	0.69	0.64	0.66	**0.74**	0.67	0.64	0.67
D3	0.69	**0.71**	0.64	0.67	0.67	0.68	0.67	0.48	0.68	0.67
D4	0.71	**0.72**	0.69	0.69	0.68	**0.72**	0.67	0.69	0.68	0.69
D5	**0.89**	0.41	0.00	0.51	0.00	0.40	0.00	0.42	0.00	0.47
D6	**0.83**	0.61	0.41	0.60	0.30	0.57	0.10	0.67	0.20	0.61
D7	**0.79**	0.72	0.23	0.70	0.23	0.73	0.00	0.65	0.11	0.71
D8	0.68	**0.72**	0.58	0.69	0.58	0.69	0.64	0.65	0.58	0.65
D9	0.44	**0.64**	0.38	0.44	0.38	0.46	0.40	0.50	0.38	0.57
D10	0.64	0.64	0.70	0.69	0.68	0.67	**0.74**	0.50	0.64	0.73
D11	**0.66**	0.59	0.55	0.64	0.55	0.60	0.55	0.54	0.61	0.59
D12	0.42	0.69	0.57	0.64	0.58	0.66	0.56	**0.75**	0.56	0.67
D13	**0.82**	0.70	0.50	0.75	0.62	0.71	0.74	0.79	0.25	0.71
D14	0.69	0.69	0.59	0.68	0.58	**0.73**	0.69	0.49	0.64	0.66
D15	0.61	**0.64**	**0.64**	**0.64**	**0.64**	0.63	**0.64**	**0.64**	**0.64**	0.61

facial feature. In contrast, the pre-trained BERT classifiers are known to be successful at capturing relevant semantic information in shorter text sequences.

The audio VGGish models and temporal facial feature LSTM models tie in regards to average $F1$ scores for two datasets, namely, D4 (controlling, temper) and D15 (travel, lot). Interestingly, after D7, these datasets contain the smallest average number of time steps. VGGish only achieved the highest average $F1$ scores for another three of the individual screening questions. Despite the lower computational cost of the models, the temporal facial features achieved the highest average $F1$ scores for five of the individual screening questions. Action unit features were most useful for D2 (anything, regret), D10 (feeling, lately) and D12 (last, happy). The remaining two datasets were D1 (advice,yourself) with all features and D14 (study, school) with eye gaze features.

5.4. Multimodal Models

The $F1$ score for the multimodal AudiBERT and AudiFace models are displayed in Table 3 for each of the different question datasets. In general, AudiFace variations achieve higher $F1$ score for 13 of the 15 datasets than AudiBERT. AudiFace with eye gaze features and AudiFace with action unit features screened for depression with a remarkable average $F1$ score of 0.93 for dataset D5 (diagnosed, depression). However, AudiBERT was also very

Table 3: **Multimodal Models**: average ± standard deviation of $F1$ scores from the AudiBERT, AudiFace, with different temporal facial features, models by dataset.

	AudiBERT	AudiFace			
Dataset		Landmarks	Eye Gaze	Action Unit	All
D1	0.65 ± 0.04	0.59 ± 0.02	0.52 ± 0.03	0.59 ± 0.03	$\mathbf{0.70 \pm 0.01}$
D2	$\mathbf{0.78 \pm 0.05}$	0.71 ± 0.01	0.77 ± 0.00	0.76 ± 0.01	0.72 ± 0.00
D3	0.71 ± 0.02	0.72 ± 0.02	$\mathbf{0.74 \pm 0.02}$	0.67 ± 0.01	$\mathbf{0.74 \pm 0.03}$
D4	0.73 ± 0.01	0.69 ± 0.00	0.74 ± 0.02	0.69 ± 0.01	$\mathbf{0.76 \pm 0.02}$
D5	0.92 ± 0.05	$\mathbf{0.93 \pm 0.06}$	$\mathbf{0.93 \pm 0.06}$	$\mathbf{0.93 \pm 0.06}$	0.90 ± 0.01
D6	$\mathbf{0.73 \pm 0.03}$	0.72 ± 0.10	$\mathbf{0.73 \pm 0.00}$	0.69 ± 0.03	0.71 ± 0.04
D7	0.86 ± 0.00	$\mathbf{0.89 \pm 0.02}$	0.86 ± 0.00	0.87 ± 0.02	0.87 ± 0.03
D8	0.69 ± 0.01	0.70 ± 0.02	$\mathbf{0.71 \pm 0.00}$	0.70 ± 0.01	0.69 ± 0.00
D9	0.54 ± 0.04	$\mathbf{0.61 \pm 0.07}$	0.58 ± 0.02	0.58 ± 0.02	$\mathbf{0.61 \pm 0.04}$
D10	0.72 ± 0.03	0.75 ± 0.03	$\mathbf{0.76 \pm 0.00}$	0.71 ± 0.03	0.72 ± 0.01
D11	0.62 ± 0.04	0.65 ± 0.03	0.64 ± 0.06	$\mathbf{0.68 \pm 0.08}$	0.65 ± 0.03
D12	0.63 ± 0.02	0.64 ± 0.05	0.61 ± 0.03	0.65 ± 0.04	$\mathbf{0.66 \pm 0.01}$
D13	0.66 ± 0.08	0.66 ± 0.01	0.67 ± 0.00	$\mathbf{0.71 \pm 0.01}$	0.66 ± 0.03
D14	0.70 ± 0.02	$\mathbf{0.74 \pm 0.05}$	0.72 ± 0.00	0.70 ± 0.03	0.69 ± 0.05
D15	0.69 ± 0.03	0.64 ± 0.00	$\mathbf{0.71 \pm 0.02}$	0.67 ± 0.00	0.64 ± 0.00

successful for this dataset with an average $F1$ score of 0.92. This is an improvement over the unimodal BERT classifier which achieved an average $F1$ score of 0.89.

Of the individual facial features, AudiFace performed best with the eye gaze. This AudiFace variation obtained the highest average $F1$ scores for 6 datasets (with 2 ties). In contrast, AudiFace with landmark features won for 4 datasets (with 2 ties) while AudiFace with action unit features won for 3 datasets (with 1 tie). This is in contrast with the unimodal models where action unit features won more than the other temporal facial feature types. The results of the unimodal models suggest that eye gaze is most complementary to text and audio.

Combining all of the different types of temporal facial features helps improve depression screening for some of the individual questions. This is most noticeable for D1 (advice, yourself) where AudiBERT obtained an average $F1$ score of 0.65, and AudiFace with all temporal facial features increased the average $F1$ score to 0.70. Yet, the unimodal model with all temporal facial features achieved an even higher F1 score, indicating that the audio and text introduced noise. In contrast, multimodal models achieved a higher average $F1$ score than unimodal models for D2 (anything, regret) dataset, with the $F1$ score for AudiBERT 0.01 higher than AudiFace with the temporal eye gaze features. However, for D4 (controlling, temper), AudiFace with all temporal facial features achieves a higher $F1$ score than any of the other unimodal or multimodal models.

Overall, the multimodal models achieved higher average $F1$ scores than the unimodal models for 9 of the 15 datasets. For those 8 of those 9 datasets, AudiFace achieved higher

average F1 scores than AudiBERT. In some cases, the improvement in average $F1$ score from adding temporal facial features is negligible given the standard deviation, like for D5 (diagnosed,depression). In other cases, like D11 (friend,describe), the temporal facial features greatly improved the average $F1$ score. Thus, we must conclude that incorporating temporal facial features is advantageous for depression screening.

5.5. Question Selection

Understandably, dataset D5 (diagnosed, depression) achieved the highest $F1$ score of 0.93 when screening for depression with AudiFace in Table 3. However, in practice, this may not be the best screening question, as it could be perceived as invasive. Therefore, other questions may be preferred, such as D7 (doing, today) which achieved a respectable $F1$ score of 0.89 with AudiFace. This is a more general wellbeing question, though the ability to screen with this question outside of a clinical interview setting still needs to be assessed.

If asking a variety of questions in a single session or over time, other questions that were useful for screening included D2 (anything, regret) with AudiBERT, D10 (feeling, lately) with AudiFace, and D13 (proud, life) with BERT. Interestingly, each of these three questions perform best with different subsets of modalities. This is perhaps due to the length of the responses or the topic of the question. D2 had fewer average time steps while D10 and D13 had higher average time steps. Both D2 and D13 ask about personal history. On the other hand, D10 is another general wellbeing question like D7, so AudiFace performed well when screening for depression with general wellbeing questions.

6. Discussion

Clinical Implications. AudiFace is a promising approach that could decrease the burden of mental illness screening by replacing screening surveys that assess the existence and severity of mental illnesses. AudiFace leverages the text, audio, and temporal facial features extracted from a video. While we used these inputs from responses to individual questions in the DAIC-WOZ datasets, AudiFace can be applied to any short video. Thus, patients seeking healthcare could consent to be recorded during patient intake rather than complete screening surveys. Alternatively, in the future, AudiFace could be incorporated into a mobile video diary application that automatically performs mental illness screening. As such, AudiFace could serve as part of a healthcare ecosystem that involves trained clinicians.

Technical Implications. Most of the existing multimodal models in the healthcare domain leverage only two modalities. Thus, AudiFace is unique in that combines embeddings from three input channels extracted from video input. To accomplish this, we innovated a strategy to use temporal facial features to successfully screen for depression. AudiFace further improved these screening results demonstrating the value of using all three video input channels. In particular, we assessed which of the facial feature types was most useful for depression screening when using an individual interview question in unimodal and multimodal models. Our results are important in that they inform the combination of question and facial feature type to include when developing future video screening applications.

The integration of temporal facial features has the potential to make AudiFace more robust than models leveraging only two video input channels. Unlike the audio and tran-

scripts, the predictiveness of the temporal facial features would not compromised by poor audio quality, heavy accents, or different dialects. Further, the stigmatization of depression (Barney et al., 2006) can lead to conscious and unconscious attempts to influence screening results. In this case, temporal facial features would likely be more useful than the other modalities given that it is difficult to intentionally alter eye gaze, landmark, and activation unit features. In AudiFace, the temporal facial feature modeling component is less computationally expensive than the audio and transcript modeling components. Thus, there is not a trade-off between robustness and computational cost.

Privacy Concern. If the multimodal AudiFace is adopted as part of the healthcare ecosystem, the recorded video interviews would receive the same protections as any other medical data. There is unfortunately always the concern of bad agents obtaining access to protected data, but these interviews would likely be less of a concern than other stored medical data. In fact, after the modeling occurs, the recording could be deleted, thus reducing the privacy risk. Our work focuses on the incorporation of temporal facial features, which are not stored in their raw form and therefore do not pose the privacy risk associated with the audio or transcripts of the interviews. As such, it would be possible to extract the temporal facial features for storage during the interview without recording the video. While the models leveraging only the facial features did not perform as well as AudiFace, these unimodal models are still an option that would preserve privacy, especially if modeling is performed outside of a healthcare setting.

Limitations and Future Work. As is common for healthcare datasets, this research was limited by the number of participants in the existing dataset. Further, the unequal distribution of depressed participants (Table 1) added an additional modeling challenge. As such, we are unable to advise regarding the generalizability of the results, which would need to be assessed before deployment to ensure that the model is not biased by the participant population. Though, it is worth noting, a strength of AudiFace is it's applicability to datasets with limited number of participants. Additionally, AudiFace is applicable for other video datasets within and outside of the healthcare domain.

This research is also unfortunately limited by the lack of datasets containing videos with depression labels. We used the DAIC-WOZ dataset which contained facial features derived from the videos instead of raw videos to protect participant privacy. While we innovated a unique strategy to utilize these facial features in a temporal manner, we were limited to using the set of features provided. As such, we were unable to extract additional features or feature embeddings from raw videos. Further, we were not able to establish an end-to-end multimodal depression screening pipeline.

Thus, our future work involves collecting a datasets of mobile video recordings with mental illness labels to further research in this domain. Such datasets would permit us to determine the screening capabilities of short videos recorded outside of the clinical interview setting as well as assess the generalizability of the results. Furthermore, raw videos recording would allow us to extract additional temporal facial features as well as feature embeddings not present within publicly available datasets in the healthcare domain. Lastly, our future research involves incorporating more recent transformers within our multimodal framework.

Acknowledgments

This work was supported by Fulbright Foreign Student Program, National Agency for Research and Development Chile, US Department of Ed. P200A180088: GAANN Fellowship, and AFRI Grant 1023720. Results were obtained using a computing cluster acquired with NSF MRI grant DMS-1337943 to WPI. We thank Avantika Shrestha and the DAISY research lab at WPI for advice and support.

References

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.

Meysam Asgari, Izhak Shafran, and Lisa B Sheeber. Inferring clinical depression from speech and spoken utterances. In *2014 IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–5. IEEE, 2014.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54, 2006.

Timothy Bickmore, Amanda Gruber, and Rosalind Picard. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30, 2005.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

David E Bloom, Elizabeth Cafiero, Eva Jané-Llopis, Shafika Abrahams-Gessel, Lakshmi Reddy Bloom, Sana Fathima, Andrea B Feigl, Tom Gaziano, Ali Hamandi, Mona Mowafi, et al. The global economic burden of noncommunicable diseases. Technical report, Program on the Global Demography of Aging, 2012.

Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3474–3484, 2020a.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020b.

Bureau of Health Workforce, Health Resources and Services Administration, and U.S. Department of Health & Human Services. Designated health professional shortage areas statistics: Designated hpsa quarterly summary, 2021.

Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, and Elisabetta Farella. Neural network distillation on iot platforms for sound event detection. In *Interspeech*, pages 3609–3613, 2019.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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.

Hélio Clemente Cuve, Santiago Castiello, Brook Shiferaw, Eri Ichijo, Caroline Catmur, and Geoffrey Bird. Alexithymia explains atypical spatiotemporal dynamics of eye gaze in autism. *Cognition*, 212:104710, 2021.

Mark É Czeisler, Rashon I Lane, Emiko Petrosky, Joshua F Wiley, Aleta Christensen, Rashid Njai, Matthew D Weaver, Rebecca Robbins, Elise R Facer-Childs, Laura K Barger, et al. Mental health, substance use, and suicidal ideation during the covid-19 pandemic—united states, june 24–30, 2020. *Morbidity and Mortality Weekly Report*, 69(32), 2020.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1061–1068, 2014.

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.

Daniel Di Matteo, Kathryn Fotinos, Sachinthya Lokuge, Julia Yu, Tia Sternat, Martin A Katzman, Jonathan Rose, et al. The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study. *JMIR Formative Research*, 4(8):e18751, 2020.

Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM, 2015.

Dominic B Dwyer, Peter Falkai, and Nikolaos Koutsouleris. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14:91–118, 2018.

Ricardo Flores, ML Tlachac, Ermal Toto, and Elke A Rundensteiner. Depression screening using deep learning on follow-up questions in clinical interviews. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 595–600. IEEE, 2021.

Ricardo Flores, ML Tlachac, Ermal Toto, and Elke Rundensteiner. Transfer learning for depression screening from follow-up clinical interview questions. *Deep Learning Applications*, 4, 2022. In Press.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE, 2017.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Language Resources and Evaluation*, pages 3123–3128. CiteSeer, 2014.

Salima Hamouche. Covid-19 and employees' mental health: stressors, moderators and agenda for organizational actions. *Emerald Open Research*, 2, 2020.

Lang He, Jonathan Cheung-Wai Chan, and Zhongmin Wang. Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, 422:165–175, 2021.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135. IEEE, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.

Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su, and Yu-Ting Kuo. Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. *IEEE Transactions on Affective Computing*, 2018.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016.

Ellen W McGinnis, Steven P Anderau, Jessica Hruschak, Reed D Gurchiek, Nestor L Lopez-Duran, Kate Fitzgerald, Katherine L Rosenblum, Maria Muzik, and Ryan S McGinnis. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE journal of biomedical and health informatics*, 23 (6):2294–2301, 2019.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR, 2013.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *AVEC*, pages 55–63, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021.

Saskia Senn, ML Tlachac, Ricardo Flores, and Elke Rundensteiner. Ensembles of bert for depression classification. In *44th International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022. In press.

M Sharifa, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, Gordon Parker, et al. From joyous to clinically depressed: Mood detection using spontaneous speech. In *Twenty-Fifth International FLAIRS Conference*, 2012.

ML Tlachac and Elke Rundensteiner. Screening for depression with retrospectively harvested private versus public text. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3326–32, 2020.

ML Tlachac, Ermal Toto, Joshua Lovering, Rimsha Kayastha, Nina Taurich, and Elke Rundensteiner. Emu: Early mental health uncoveringframework and dataset. In *20th International Conference of Machine Learning Applications (ICMLA) Special Session Machine Learning in Health*, pages 1311–1318, 2021.

ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering, Ermal Toto, and Elke Rundensteiner. StudentSADD: Mobile depression and suicidal ideation screening of college students during the coronavirus pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–32, 2022a.

ML Tlachac, Ricardo Flores, Ermal Toto, and Elke Rundensteiner. Early mental health uncovering with short scripted and unscripted voice recordings. *Deep Learning Applications*, 4, 2022b. In Press.

Ermal Toto, ML Tlachac, and Elke Rundensteiner. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *30th ACM International Conference on Information and Knowledge Management (CIKM) Applied Research Track*, pages 4145–4154, 2021.

Ezekiel Victor, Zahra M Aghajan, Amy R Sewart, and Ray Christian. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological assessment*, 31(8):1019, 2019.

Jeremy West, Dan Ventura, and Sean Warnick. Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young University, College of Physical and Mathematical Sciences*, 1(08), 2007.

Huang Xie and Tuomas Virtanen. Zero-shot audio classification based on class label embeddings. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 264–267. IEEE, 2019.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Appendix

Table 4: Questions dataset description, ordered alphabetically.

Dataset	Question Example
D1	What advice would you give to yourself ten or twenty years ago?
D2	Is there anything you regret?
D3	When was the last time you argued with someone and what was it about?
D4	How are you at controlling your temper?
D5	Have you been diagnosed with depression?
D6	Have you ever been diagnosed with PTSD?
D7	How are you doing today?
D8	What's your dream job?
D9	How easy is it for you to get a good night's sleep?
D10	How have you been feeling lately?
D11	How would your best friend describe you?
D12	When was the last time you felt really happy?
D13	What are you most proud of in your life?
D14	What'd you study at school?
D15	Do you travel a lot?

Table 5: **Landmarks**: Average precision, recall, and AUC (also known as balanced accuracy for binary classification) by dataset and models.

	LSTM			LSTM Att			AudiFace		
Dataset	precision	recall	AUC	precision	recall	AUC	precision	recall	AUC
D1	0.59	0.89	0.60	0.63	0.81	0.64	0.47	0.79	0.41
D2	0.53	0.84	0.51	0.53	1.00	0.50	0.59	0.89	0.60
D3	0.62	0.66	0.63	0.61	0.74	0.62	0.66	0.79	0.68
D4	0.53	0.99	0.50	0.54	0.96	0.53	0.55	0.93	0.54
D5	0.00	0.00	0.50	0.38	0.78	0.55	1.00	0.87	0.93
D6	0.29	0.70	0.50	0.44	0.94	0.50	0.87	0.61	0.77
D7	0.18	0.32	0.52	0.57	0.91	0.59	0.87	0.91	0.88
D8	0.50	0.69	0.46	0.59	0.83	0.55	0.55	0.96	0.54
D9	0.43	0.34	0.47	0.43	0.45	0.47	0.52	0.74	0.57
D10	0.62	0.80	0.65	0.61	0.79	0.64	0.73	0.77	0.73
D11	0.40	0.88	0.48	0.60	0.69	0.63	0.49	0.97	0.63
D12	0.54	0.60	0.50	0.60	0.69	0.58	0.55	0.77	0.59
D13	0.20	0.60	0.57	0.60	1.00	0.69	0.50	0.97	0.55
D14	0.51	0.70	0.47	0.54	0.92	0.51	0.65	0.86	0.65
D15	0.47	1.00	0.50	0.47	1.00	0.50	0.47	1.00	0.50

Table 6: **Eye Gaze**: Average precision, recall, and AUC (also known as balanced accuracy for binary classification) by dataset and models.

	LSTM			LSTM Att			AudiFace		
Dataset	precision	recall	AUC	precision	recall	AUC	precision	recall	AUC
D1	0.20	0.30	0.53	0.76	0.67	0.64	0.43	0.66	0.35
D2	0.53	0.81	0.51	0.53	0.87	0.51	0.79	0.80	0.75
D3	0.64	0.70	0.65	0.55	0.89	0.55	0.61	0.93	0.67
D4	0.53	0.95	0.51	0.59	0.92	0.60	0.64	0.88	0.67
D5	0.00	0.00	0.50	0.46	0.35	0.52	1.00	0.87	0.93
D6	0.22	0.47	0.50	0.45	0.78	0.50	1.00	0.57	0.79
D7	0.18	0.32	0.52	0.62	0.89	0.66	0.82	0.90	0.85
D8	0.50	0.69	0.46	0.62	0.78	0.59	0.56	1.00	0.56
D9	0.43	0.34	0.47	0.45	0.47	0.47	0.52	0.67	0.55
D10	0.60	0.78	0.63	0.53	0.91	0.54	0.73	0.80	0.75
D11	0.40	0.88	0.48	0.62	0.58	0.64	0.53	0.90	0.62
D12	0.44	0.85	0.45	0.53	0.87	0.56	0.50	0.78	0.54
D13	0.30	0.50	0.60	0.59	0.89	0.61	0.67	0.70	0.67
D14	0.50	0.69	0.46	0.63	0.87	0.63	0.60	0.90	0.62
D15	0.47	1.00	0.50	0.49	0.88	0.53	0.63	0.85	0.68

Table 7: **Action Unit**: Average precision, recall, and AUC (also known as balanced accuracy for binary classification) by dataset and models.

	LSTM			LSTM Att			AudiFace		
Dataset	precision	recall	AUC	precision	recall	AUC	precision	recall	AUC
D1	0.77	0.30	0.54	0.63	0.57	0.60	0.48	0.76	0.43
D2	0.59	1.00	0.61	0.62	0.73	0.60	0.64	0.93	0.67
D3	0.64	0.70	0.65	0.52	0.45	0.52	0.51	1.00	0.52
D4	0.53	0.91	0.50	0.56	0.90	0.55	0.53	1.00	0.52
D5	0.00	0.00	0.50	0.31	0.65	0.54	1.00	0.87	0.93
D6	0.07	0.18	0.50	0.59	0.78	0.65	0.87	0.57	0.75
D7	0.00	0.00	0.50	0.61	0.70	0.60	0.85	0.90	0.87
D8	0.54	0.79	0.52	0.63	0.67	0.61	0.55	0.97	0.54
D9	0.52	0.33	0.52	0.45	0.56	0.48	0.50	0.70	0.54
D10	0.82	0.67	0.76	0.44	0.58	0.42	0.63	0.83	0.65
D11	0.40	0.88	0.48	0.42	0.76	0.51	0.67	0.76	0.71
D12	0.42	0.84	0.45	0.68	0.84	0.69	0.52	0.85	0.58
D13	0.58	1.00	0.67	0.68	0.94	0.74	0.59	0.89	0.66
D14	0.56	0.90	0.56	0.53	0.46	0.50	0.59	0.87	0.60
D15	0.47	1.00	0.50	0.53	0.81	0.57	0.50	1.00	0.55

Table 8: **All**: Average precision, recall, and AUC (also known as balanced accuracy for binary classification) by dataset and models.

	LSTM			LSTM Att			AudiFace		
Dataset	precision	recall	AUC	precision	recall	AUC	precision	recall	AUC
D1	0.59	0.92	0.61	0.65	0.86	0.66	0.55	0.97	0.55
D2	0.53	0.81	0.51	0.60	0.76	0.53	0.57	0.97	0.58
D3	0.64	0.73	0.65	0.62	0.73	0.62	0.61	0.97	0.67
D4	0.53	0.95	0.50	0.61	0.79	0.62	0.65	0.91	0.69
D5	0.00	0.00	0.50	0.38	0.62	0.56	0.94	0.87	0.92
D6	0.29	0.15	0.50	0.48	0.84	0.55	0.73	0.71	0.75
D7	0.25	0.07	0.51	0.70	0.72	0.68	0.87	0.87	0.87
D8	0.50	0.69	0.46	0.64	0.66	0.57	0.53	1.00	0.50
D9	0.43	0.34	0.47	0.48	0.70	0.50	0.53	0.74	0.57
D10	0.62	0.66	0.65	0.62	0.89	0.66	0.61	0.90	0.65
D11	0.44	1.00	0.55	0.65	0.54	0.63	0.49	0.95	0.63
D12	0.53	0.59	0.50	0.54	0.88	0.58	0.71	0.70	0.65
D13	0.60	0.16	0.70	0.61	0.85	0.69	0.49	1.00	0.53
D14	0.53	0.81	0.51	0.60	0.73	0.57	0.54	0.97	0.52
D15	0.47	1.00	0.50	0.49	0.81	0.51	0.47	1.00	0.50