Multi-Task Learning to Screen for Major Depressive Disorder and Post-Traumatic Stress Disorder

by

Avantika Shrestha

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Data Science

March 2024

APPROVED:

_____

Professor Elke A. Rundensteiner, Thesis Advisor

_____

Professor Xiaozhong Liu, Thesis Reader

ABSTRACT

Prevalent mental health conditions, major depressive disorder (MDD), and post-traumatic stress disorder (PTSD) have severe physical and social impacts. Detection is difficult and expensive, requiring substantial time from trained mental professionals. To counteract this issue, recent research has explored the diagnostic potential of deep learning models trained on modalities extracted from virtual agent-conducted clinical interview videos. However, deep learning models are challenging to train due to long sequences and the small number of participants that are common in the mental health community. In this thesis, a solution to combat these challenges is developed by leveraging a multi-task learning framework that uses temporal facial features as input to screen for MDD and PTSD. The multi-task framework is based on a bidirectional GRU model with self-attention. This thesis evaluates the multi-task model on temporal facial features extracted from the responses to 15 clinical interview questions conducted by a virtual agent. The results suggest that multi-task learning increases the generalization performance compared to single-task learning. For MDD screening, multi-task learning improved the balanced accuracy over single-task learning for 11 of the 15 datasets. The multi-learning model increased the MDD screening ability by 25 percent to a balanced accuracy of 0.87 in some scenarios. This work provides valuable findings for the future of mental screening applications leveraging temporal facial features. [1]

[1]This thesis is adapted from my paper in collaboration with Dr. Ricardo Flores from Sanofi, *A. Shrestha*, R. Flores*, M. Tlachac and E. A. Rundensteiner, "Multi-Task Learning Using Facial Features for Mental Health Screening," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 4881-4890, doi: 10.1109/BigData59044.2023.10386191*, where Dr. Flores and I shared equal contributions towards the paper and has been featured in his dissertation as well.

CONTENTS

1 INTRODUCTION

1.1 BACKGROUND

Major depressive disorder (MDD) and post-traumatic stress disorder (PTSD) are mental illnesses experienced by millions of individuals annually in the U.S. (1; 2). Individuals with MDD and/or PTSD have a reduced quality of life (3; 4), as the mental illnesses have negative impacts on physical, social, and financial health (5; 6). MDD and PTSD are both associated with suicidal behavior (6), particularly when the mental illnesses are undertreated (7). While prevalence is slightly higher among U.S. veterans (8), PTSD is a disability that can impact anyone who experiences trauma (2; 6).

MDD and PTSD have high comorbidity, with approximately 50% of people who screen positive for PTSD also screening positive for MDD (9). Having both mental illnesses notably increases the likelihood of developing physical health disorders, with simultaneous MDD and PTSD increasing the likelihood of disability by three in veterans (10). Both mental illnesses can go unnoticed among the general population due to the lack of awareness regarding symptoms (11) and mental health stigma (12). However, early treatment can greatly reduce the symptoms of MDD and PTSD, and even result in depression remission (13) and PTSD recovery (2).

Clinicians play a vital role in diagnosing MDD and PTSD, as these mental illnesses can present themselves with different symptom profiles for different people (14). Further, the symptoms can be easily attributed to other health conditions (15), as patients often may not recognize and communicate more subtle or stigmatized symptoms (11; 16). As a result of the COVID-19 pandemic, the prevalence of MDD and PTSD has increased (17; 18). This has caused further strain on the already limited mental health resources (19; 20). The shortage of clinicians has a devastating effect on community health, especially following trauma exposure from catastrophic events (21) like the COVID-19 pandemic.

1.2 MOTIVATING EXAMPLE

Traditional methods of diagnosing MDD and PTSD usually involve the administration of screening questionnaires (22; 23) and lengthy clinical interviews. Given the comorbidity between MDD and PTSD, clinicians also need to be able to distinguish between them, as they require differing treatment plans (24; 25). This requirement for differential diagnosis can result in even lengthier clinical interviews. With the limited availability of mental health professionals (19; 20), conducting lengthy clinical interviews can be costly and have long wait times, resulting in delays in care that often have disastrous impacts on patients' health (26; 13). Differential diagnoses are one of the psychiatric tasks that could most benefit from the application of machine learning (27).

Virtual agent-conducted clinical interviews can be used to expedite mental health screening. When integrated with mental health screening models, they present an inexpensive and efficient alternative to the traditional interview process, which allows for the interview to be conducted asynchronously of clinician availability. The goal of the virtual agent in this setting is to encourage the patients to disclose sufficient information for the screening models by asking a series of questions to simulate conversation (28). Data collected during the clinical interview can be used by screening models and provided to the clinician for review.

Facial features represent a particularly promising mental illness screening modality as they maintain patient privacy, are language agnostic, and do not require computationally prohibitive models. Further, recent research has found that temporal facial features (29) were much more robust at depression screening than audio or transcripts across multiple video datasets (30). As such, the patient's facial features can be recorded during the clinical interview and subsequently used by deep learning models to screen for mental illnesses.

1.3 Problem Definition and Challenges

Given a set of temporal facial features extracted while patients answer questions and the corresponding MDD and PTSD screening labels, the goal of this thesis is to simultaneously predict whether each participant has MDD and/or PTSD. Additionally, this thesis aims to identify which clinical interview questions are most useful to ask, as evaluated by the use of the resulting temporal facial features in the MDD and PTSD screening models.

Training a deep learning model on temporal facial features for mental illness screening poses certain challenges. The temporal facial features must span enough time to hold predictive signal, but deep learning models can suffer from the vanishing problem (31) when the input sequences are too long. In particular, long sequences have a high computational cost for training, especially if leveraging large architectures, such as transformer-based models (32; 33; 34). Further, most visual-mental health datasets are composed of data from a relatively limited number of participants (35). Consequently, deep learning models trained to screen for depression on such datasets tend to report low evaluation metrics (36; 37).

1.4 The Deep Learning Approach

In this thesis, the aim is to improve the MDD and PTSD screening capabilities of facial features extracted from clinical interviews conducted by a virtual agent. To achieve this, a multi-task learning approach leverages temporal facial features as displayed in Fig. 1. Notably, by incorporating three different types of facial features, landmark, eye gaze, and action unit. The aforementioned challenge

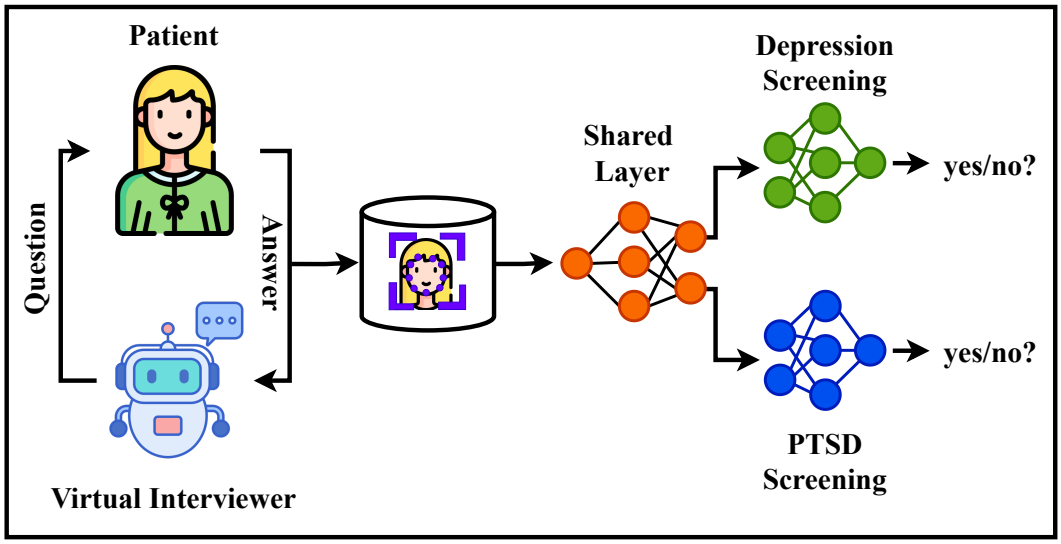

Figure 1: **Multi-task learning**. A virtual interviewer asks questions to a patient. The clinical interview facial features are leveraged for a multi-task learning model with a shared layer and two final layers for major depressive disorder (MDD) and post-traumatic stress disorder (PTSD) screening.

of long sequences is addressed through the use of self-attention (32) in the models (38). To address the small sample size, the thesis incorporates multi-task regularization techniques (39). As the screening model is designed to use facial features rather than raw videos, this approach also preserves patient privacy.

To corroborate the approach, the single and multi-task models are trained on the facial features in the popular Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset. In addition to the MDD screening labels in DAIC-WOZ, this thesis also leverages the PTSD screening labels for the clinical interviews from the Extended Distress Analysis Interview Corpus (E-DAIC) (40). The clinical interviews are divided into sub-interviews based on a set of core questions, to determine which questions are most useful for MDD and PTSD screening. This thesis is significant in that it provides knowledge into how to leverage temporal facial features and multi-task learning in general and applied to temporal facial features in particular to improve the development of the next generation of mental health screening applications. The developed approach is general, and could also be evaluated on other modalities like audio or textual responses. This thesis contributes:

1. A multi-task learning approach to classify temporal facial features, allowing simultaneous screening for major depressive disorder (MDD) and post-traumatic stress disorder (PTSD).
2. Comparison of the ability of single and multi-task models to screen for MDD and PTSD.
3. Assessment of which specific clinical interview questions are most useful for MDD and PTSD screening.

2 Related Work

2.1 Single-Task Learning for Mental Health Assessment

There is a substantial amount of work applying traditional single-task learning models to mental illness screening with digital data ranging from location (41) to videos (42). Much of the depression screening research on video modalities has been conducted with the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset (43), given that it featured in the Audio/Visual Emotion Challenge and Workshop (AVEC) (44). Single-task learning models have been used for depression screening with audio (45; 46; 47; 48; 36; 49), transcripts (50; 51; 52; 53), facial features (29), and multiple modalities (47; 54; 30) extracted from the DAIC-WOZ clinical interviews.

The most successful of the depression screening models (47) were computationally expensive, language-dependent, and posed a privacy risk as they require raw data. In contrast, facial features extracted from videos do not share these limitations. Facial features that were extracted in this manner have been leveraged for health applications such as depression screening (55; 56; 29), suicide risk assessment (57), and autism identification (58).

2.2 Multi-Task Learning for Mental Health Assessment

The concept of multi-task learning was first introduced by the observation of the knowledge transfer that occurs naturally in human communication and education (59). The same can be said for machine learning models, as the knowledge transfer between related tasks often allows the model to robustly learn data which allows better generalization performance over the data. Multi-task learning allows eavesdropping of information between tasks to increase performance and generalization (60). Health researchers have used this transfer of knowledge among tasks to explore research in tumor segmentation and response prediction (61), construction of personalized brain-computer interfaces (62), Alzheimer's disease screening (63; 64; 65), and survival analysis (66).

Multi-task learning has also been used with social media text (67; 68) and sensor data from mobile phones and wearable devices (69) to improve mental health screening. Prior mental illness screening research (67) found that multi-task learning can greatly improve upon the performance of their single-task counterparts. Using multi-task learning for mental health screening mirrors the clinical decision-making process as multiple assessments are performed simultaneously by mental health professionals during clinical interviews (70). Overall, multi-task learning can be used to help diagnostic models implicitly capture comorbidities in their hidden layers. As such, multi-task learning is particularly beneficial for small datasets (39), which are common in healthcare research (35).

2.3 Facial Features for Mental Health Assessment

There are multiple types of facial features. Eye gaze is the most common in digital health research, having been used for depression screening (55; 29), suicide risk assessment (57), and autism identification (58). Head pose (56), landmark (29), and action unit (29) facial features have also been used for depression screening. Additionally, the facial features have been used to train multi-modal models for depression screening (71; 72; 73; 30). Only a limited number of these studies model the facial features temporally (58; 29; 30).

In this thesis, three types of facial features are leveraged to train a deep learning sequential model that can simultaneously output MDD and PTSD screening predictions. The strong comorbidity between the two mental illnesses (9) is favorable for multi-task learning as there can be information transfer between the two classification tasks (59). Further, the single-task models provide unique insights into the ability of temporal facial features to screen for PTSD.

3 DATA

3.1 DISTRESS ANALYSIS INTERVIEW CORPUS

This thesis trains models leveraging the Extended Distress Analysis Interview Corpus (E-DAIC) (43; 40), which encompasses the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset (43; 28). Featured in the 2016 AVEC workshop (44), the DAIC-WOZ dataset was intended to support the digital screening and diagnosis of major depressive disorder (MDD) by providing data in which to identify verbal and nonverbal indicators of mental illness (43; 28). The DAIC-WOZ dataset consists of data from clinical interviews with 189 unique participants, which were conducted by a virtual agent named Ellie (28). For the WOZ subset, Ellie was animated by a researcher (known as a wizard) in a different room, a fact unknown to the participants being interviewed(43; 28).

The clinical interview data that was made available as part of the DAIC-WOZ dataset include audio recordings, transcripts, and facial features. The three types of facial features, extracted from clinical interview videos by the OpenFace Software (74), include landmark, eye gaze, and action unit. The interview data was labeled with MDD screening scores, permitting supervised modeling of the data. The interviews ranged from 7 to 33 minutes (with an average of 16 minutes), demonstrating great variability in the quantity of data depending on the participant being interviewed.

E-DAIC similarly contains data from clinical interview recordings (40; 43), including the DAIC-WOZ clinical interview data. Notably, E-DAIC was designed to support the diagnosis of multiple mental illnesses and is therefore more broad than DAIC-WOZ in that regard. In addition to MDD labels, E-DAIC contains PTSD labels for the clinical interviews. Further, E-DAIC contains clinical interviews with more participants than DAIC-WOZ. However, these additional interviews were conducted with different setups for Ellie such as having Ellie act completely autonomously (43; 28). The E-DAIC train set (of 163 participants) and development (of 56 participants) set are intermixed with clinical interviews conducted by Ellie under the WOZ and autonomous settings, though the test set (of 56 participants) is comprised entirely of autonomously conducted interviews.

This thesis specifically leverages the 189 clinical interviews that comprise the DAIC-WOZ dataset. In addition to the MDD screening labels made available in the DAIC-WOZ dataset, the thesis also makes use of the PTSD screening labels for the WOZ interviews that were made available only as part of the E-DAIC release. Due to the differences in the collection procedures, the 86 clinical interviews that are only available as part of the E-DAIC release are not utilized. In other words, the thesis assesses the ability to screen for mental illnesses with the interviews conducted by Ellie under the WOZ setting rather than under the fully autonomous setting.

3.2 Mental Illness Screening Scores

The major depressive disorder (MDD) screening scores in the DAIC-WOZ dataset were acquired through the administration of the Patient Health Questionnaire (PHQ-8) (22) and the post-traumatic stress disorder (PTSD) labels in the E-DAIC dataset were acquired through the administration of the PTSD Checklist-Civilian version (PCL-C) scores (23). Both of these screening instruments consist of a series of questions that correspond to key symptoms of the mental illness, the scores of which are totaled for interpretation.

The PHQ-8 (22) consists of 8 questions that ask about symptom frequency in the past two weeks on a scale that ranges from '0: Not at all' to '3: Nearly every day'. Thus, the PHQ-8 score ranges from 0 to 24. Scores of 5, 10, and 15 are the cutoffs used to screen for mild depression, moderate depression, and moderately severe depression. For this thesis, the cutoff of 10 for screening positive for depression is adopted as it is commonly used as the threshold for MDD screening.

The PCL-C (23) is a longer questionnaire than the PHQ-8, consisting of 17 questions. The PCL-C questions ask about how often a symptom has been a problem in the last month on a scale that ranges from '1: Not at all' to '5: Extremely', resulting in a PCL-C score that ranges from 17 to 85. Based on prior research (75), the cutoff of 44 is utilized, where scores of and above this threshold were labeled as positive for PTSD.

3.3 Question Datasets

During the DAIC-WOZ clinical interviews (43; 28), each participant was asked a series of core and follow-up questions. The follow-up questions were asked to mimic a realistic conversation and increase self-disclosure (28; 76). As such, the clinical interviews can be divided into subsets based on the core questions, with the follow-up questions ensuring the responses contain sufficient data for modeling purposes. Thus, the responses to each core question as separate datasets, following the proposed data set construction first introduced by Toto et al. (47).

To form question datasets, the interview transcripts are first parsed to identify core questions based on their bag-of-words representation. Then the data is separated based on these core questions and this transformation is applied to the facial features using the transcript timestamps. Notably, a response to a core question extends until the next core question is asked. As such, the responses to the follow-up questions are concatenated with the response to the preceding core question.

Each DAIC-WOZ participant is asked a different series of core questions based on their responses to prior core questions (28; 43). There were 15 core questions that were answered by at least 90 of the DAIC-WOZ participants. As such, individual question datasets are created for these 15 questions,

Table 1: Question dataset descriptions with an example of question, and number of instances

Dataset	Question Example	Instances
D1	What advice would you give to yourself ten years ago?	102
D2	Is there anything you regret?	94
D3	When was the last time you argued with someone?	103
D4	How are you at controlling your temper?	100
D5	Have you been diagnosed with depression?	94
D6	Have you ever been diagnosed with PTSD?	92
D7	How are you doing today?	105
D8	What's your dream job?	95
D9	How easy is it for you to get a good night's sleep?	98
D10	How have you been feeling lately?	92
D11	How would your best friend describe you?	96
D12	When was the last time you felt really happy?	99
D13	What are you most proud of in your life?	99
D14	What'd you study at school?	95
D15	Do you travel a lot?	94

which are summarized in Table 1. The table includes the dataset name, an example of the question wording, and the number of participants who answered the question. As the largest of these individual question datasets contains 105 participants, the datasets are all relatively similar in size.

Table 2: Question datasets with corresponding the ratio of participants who screened positive for MDD and PTSD, and correlation (Corr.) of the PHQ-8 and PCL-C scores.

Dataset	MDD Ratio	PTSD Ratio	Corr.
D1	28.4%	31.4%	0.83
D2	29.8%	30.8%	0.86
D3	29.1%	32.0%	0.83
D4	30.0%	33.0%	0.84
D5	21.3%	25.5%	0.78
D6	27.2%	29.3%	0.86
D7	28.6%	31.4%	0.83
D8	30.5%	31.6%	0.83
D9	27.6%	28.6%	0.84
D10	29.4%	32.6%	0.82
D11	26.0%	31.3%	0.84
D12	28.3%	31.3%	0.83
D13	28.3%	31.0%	0.83
D14	30.5%	31.6%	0.85
D15	27.7%	29.8%	0.84

Table 2 includes the average MDD and PTSD ratio for each question as well as the correlation between the PHQ-8 and PCL-C scores. Around a third of the participants screened positive for MDD and PTSD. In other words, the question datasets are somewhat unbalanced for both the MDD and the PTSD screening tasks. The average Pearson correlation coefficient between the PHQ-8 and PCL-C scores is 0.83. This confirms that there is a high positive correlation between the two main tasks, which was expected given the literature that states PTSD is comorbid with MDD (9).

According to literature, multi-task learning is known to perform well when the tasks are strongly correlated (59). This means that optimizing for such correlated tasks jointly tends to improve the

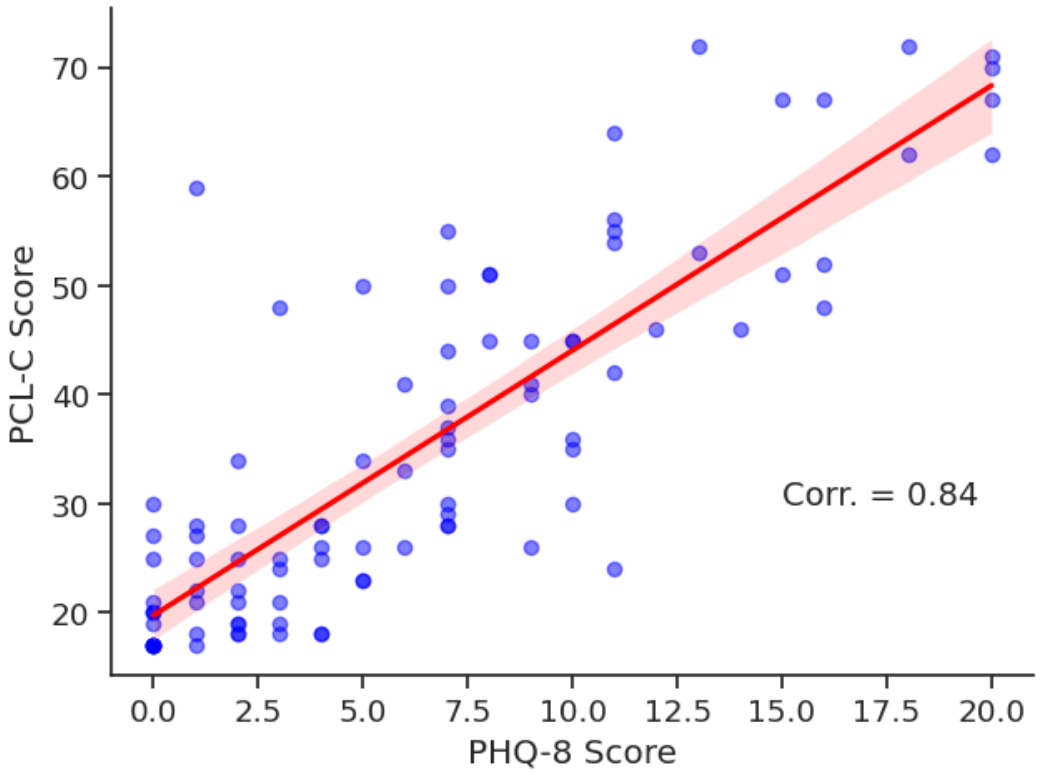

Figure 2: An example of the correlation between the scores of the two main tasks: PHQ-8 and PCL-C scores for the D11 dataset, with a high correlation coefficient of 0.84.

performances of each of the individual tasks. As an example, the correlation of PHQ8 and PCL-C scores for the $D11$ dataset is explored and displays the best-fit line depicted in red in Fig. 2, highlighting the high positive correlation of mental illness screening scores.

3.4 Temporal Facial Features

The DAIC-WOZ dataset contains processed facial features instead of original images to protect participant privacy. The facial features were generated by applying the OpenFace software (74) to the video recordings of the clinical interviews for each participant. The 162 facial features constitute three types of facial features: landmark, eye gaze, and action unit. The features are normalized between -1 and 1.

The temporal nature of these facial features is represented in Fig. 3. In this thesis, the temporal facial features are modeled as a multivariate time series. The landmark time series has 136 dimensions, the eye gaze time series has 12 dimensions, and the action unit time series has 14 dimensions. The multivariate time series of the three facial feature types are concatenated. Further, the sub-sequence of facial features is set for one second and then the facial feature information contained within that second is aggregated, as depicted in Fig. 3.

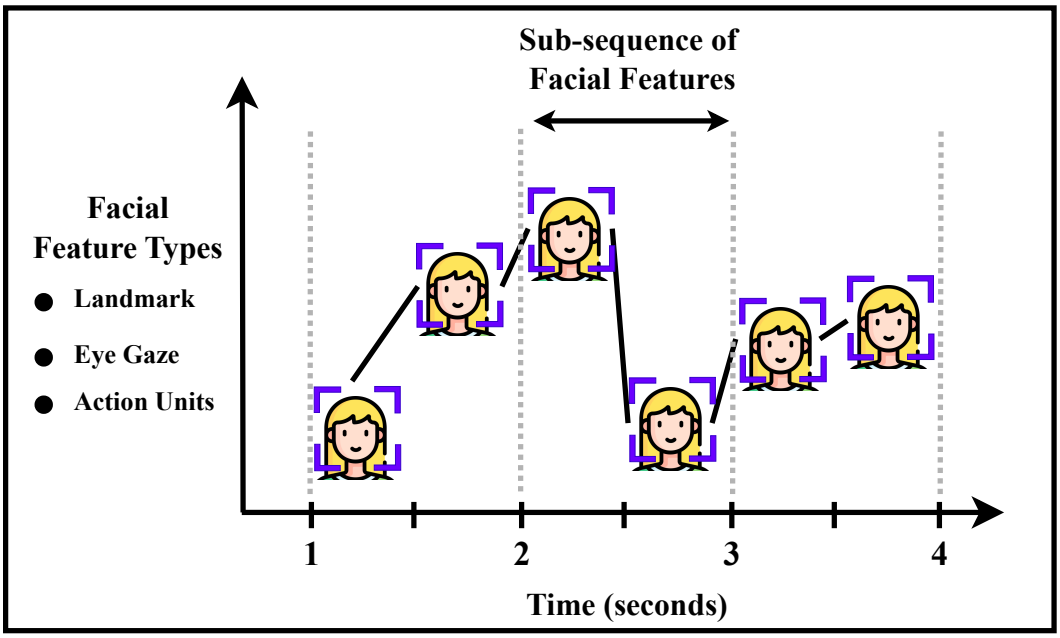

Figure 3: Landmark, eye gaze, and facial action unit are types of temporal facial features represented as multi-variate time series of dimensions 136, 12, and 14 respectively. The information contained within a second is aggregated.

The screening potential of the facial features in the DAIC-WOZ dataset is understudied, with a limited number of previous studies (29; 30) considering their temporal aspects for modeling purposes. The DAIC-WOZ audio and transcripts are decidedly more popular (45; 46; 47; 48; 36; 49; 50; 51; 52; 53; 77). Notably, Lau et al. (77) leveraged both speech and transcript data from the DAIC-WOZ corpus to improve depression assessment of a multi-task multi-modal framework. This thesis in comparison leverages facial features as a singular modality to conduct both MDD and PTSD screening. Despite their privacy advantages, no such prior multi-task research has worked with temporal facial features.

4 METHODOLOGY

4.1 SINGLE-TASK FRAMEWORK

The three types of temporal facial features– landmark, eye gaze, and action unit — are concatenated into a vector f_i of size p containing all facial feature types; i.e., f_i = ($f_{i,1}, f_{i,2}, \cdots f_{i,p}$). Each patient is modeled by a time series of n sub-clips of facial features $S = (f_1, f_2, ..., f_n)$. Consequently, S corresponds to a matrix of size $n \times p$ such as:

$$S = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,p} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,p} \end{pmatrix}. \tag{1}$$

As the time series data is multivariate, the thesis leverages a GRU model for MDD and PTSD screening. GRU is known to be useful for depression screening (78; 79; 80). Further, it is more computationally efficient (81) than its predecessor, long short-term memory (LSTM) (82). This is beneficial in clinical settings where computational resources may be scarce.

The relationship between facial features is learned through a bidirectional gated recurrent unit (biGRU) (83) network with input size p. The GRU can be described as

$$z_t = \sigma(W_z f_t + U_z h_{t-1} + b_z), \tag{2}$$

$$r_t = \sigma(W_r f_t + U_r h_{t-1} + b_r), \tag{3}$$

$$\hat{h}_t = \phi(W_h f_t + U_h(r_t \odot h_{t-1}) + b_h), \tag{4}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t. \tag{5}$$

where z_t, r_t, $\hat{h}_t$, and h_t are the update gate, reset gate, candidate activation gate, and hidden output vectors, respectively. Further, W, U, and b are parameters to estimate, and σ and ϕ are the Sigmoid and Tanh activation functions, respectively.

At each time step t, the hidden state (h_t) is processed by the biGRU. This hidden state changes based on the direction of information. For example, features can go from left to right (forward direction) or from right to left (backward direction). Thus, the forward hidden state ($\overrightarrow{h}_t$) and the backward hidden state ($\overleftarrow{h}_t$) set as $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$ to correspond to the concatenation of the forward and backward hidden states. H is the set of all hidden states of the biGRU, such that $H = (h_1, ..., h_n)$.

Like the prior studies (47; 30), this thesis applies a self-attention mechanism (32) to the single-task GRU models. This helps the model prioritize meaningful embeddings, along with preventing vanishing problems (38) from long sequences. The vanishing problem results when the gradient of a neural network is too small, and the learning process stops, as a result (84). Further, an addition of a self-attention mechanism A is done to the previous deep learning architectures such that,

$$A = softmax(W_{s2}tanh(W_{s1}(H)^T)), \tag{6}$$

where H represents all hidden states from the GRU. The weight matrix W_{s1} and W_{s2} are both learned during training. The attention A is multiplied with the hidden state H to generate the embeddings e, such as:

$$e = A \times H, \tag{7}$$

$$\hat{y} = FCL(e). \tag{8}$$

where the embedding e is used as input for a fully connected layer (FCL) for the binary prediction label ($\hat{y}$). For comparison, the single-task framework described above for each of the MDD and PTSD screening tasks is trained separately.

4.2 Multi-Task Framework

The multi-task learning framework consists of two primary tasks, MDD and PTSD screening. In addition to the aforementioned tasks, there is also the addition of an auxiliary reconstruction task. The model design is depicted in Fig. 4. In the framework, the input facial features are passed through a shared bidirectional GRU module and then through attention layers. After which, the outputted shared embedding from the shared component is then passed through the task-specific layers. For the primary classification tasks, the model simply passes the shared embeddings through additional fully connected layers.

To model the reconstruction task, the encoder compresses the shared embeddings to a latent variable (85). This is then passed to a decoder that reconstructs the embedding back into the reconstructed input. This is supervised by the loss function minimizing the difference between the feature input and reconstructed feature output of this encoder-decoder model. The intuition behind adding the reconstruction task as an auxiliary task is that providing the multi-task model with a simple yet unrelated task would help the multi-task model learn more information about the features (86; 87), and thus could potentially avoid common multi-task learning pitfalls such as negative transfer (60).

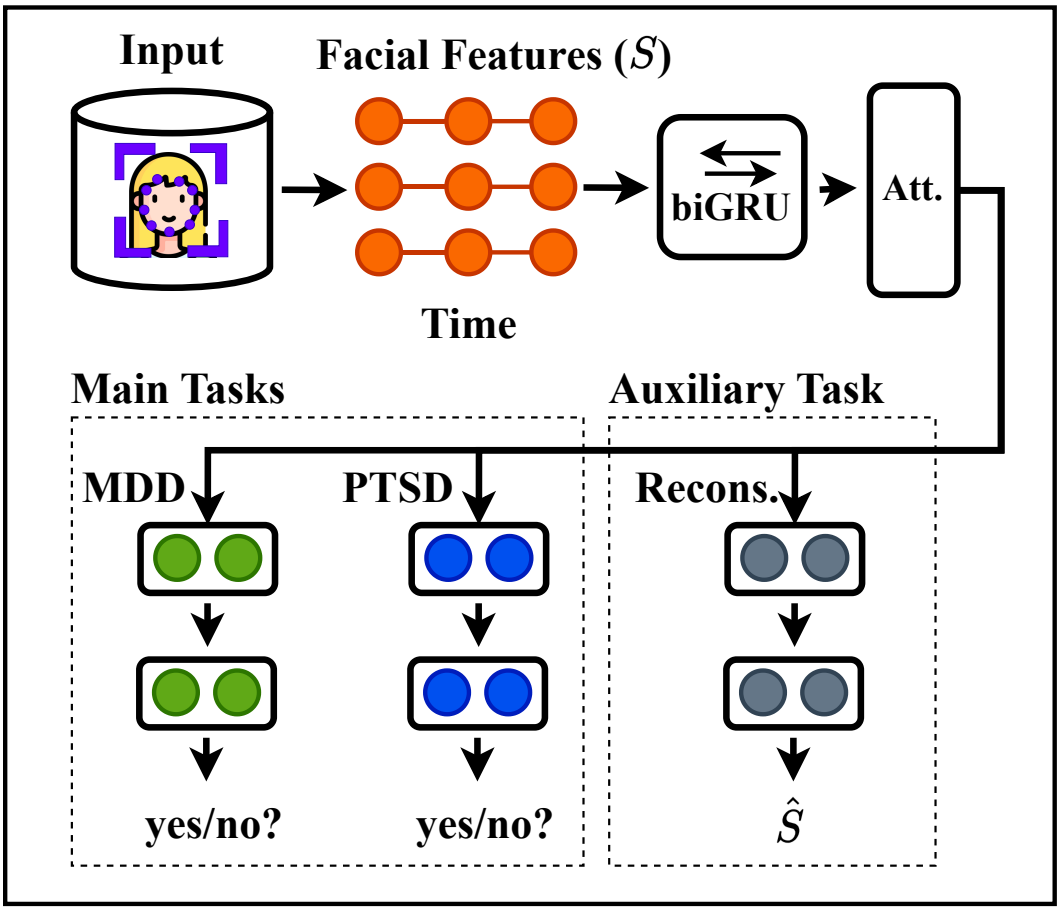

Figure 4: The multi-task learning framework, where the temporal facial features (S) are fed into a bidirectional GRU (biGRU) with self-attentions (Att) to train for two main tasks, MDD screening and PTSD screening, and an auxiliary task (reconstruction, $\hat{S}$).

As Table 1 and Fig. 2 depict, the two primary tasks are highly correlated. Similar to the single-task framework, all the facial features in the matrix S of size $n \times p$ are leveraged for the multi-task framework. S is the input for the bidirectional GRU (biGRU) which outputs the hidden states H^{mt}. In this context, H^{mt} is the shared layer, which learns the common knowledge of all tasks. Same as before, there is an addition self-attention to H^{mt} to create an embedding such as,

$$A^{mt} = softmax(W_{s2}^{mt} tanh(W_{s1}^{mt}(H^{mt})^T)), \tag{9}$$

$$e^{mt} = A^{mt} \times H^{mt}, \tag{10}$$

where W_{s1}^{mt} and W_{s2}^{mt} are weights to estimate. Finally, the shared embeddings e^{mt} are utilized as input for the task-specific layers, using a fully connected layer (FCL) for each task:

$$\hat{y}_{mdd} = FCL_{mdd}(e^{mt}), \tag{11}$$

$$\hat{y}_{ptsd} = FCL_{ptsd}(e^{mt}). \tag{12}$$

For training, both the losses from MDD and PTSD prediction labels are leveraged such that,

$$Loss_{mdd} = CE(y_{mdd}, \hat{y}_{mdd}), \tag{13}$$

$$Loss_{ptsd} = CE(y_{ptsd}, \hat{y}_{ptsd}). \tag{14}$$

Additionally, the model also utilize the loss of the auxiliary reconstruction task consuming the same features:

$$Loss_{auxiliary} = CE(S, \hat{S}). \tag{15}$$

Here, CE is the cross-entropy loss commonly used for classification purposes. For backpropagation, the summation of the three losses is calculated and leveraged. Doing so allows the model to equally prioritize each task. Thus, the model not only learns the data concerning the MDD and PTSD labels but also learns a better representation of the feature space (87).

Table 3: Balanced accuracy of single-task (ST) and multi-task (MT) models for MDD and PTSD screening, with task Gain (MT-ST). The higher value between MT and ST as well as positive gain are in bold.

	MDD			PTSD		
Dataset	**Single-task**	**Multi-task**	**Gain**	**Single-task**	**Multi-task**	**Gain**
D1	0.73 ± 0.01	$\mathbf{0.77 \pm 0.03}$	**+0.04**	0.71 ± 0.03	$\mathbf{0.72 \pm 0.03}$	**+0.01**
D2	0.74 ± 0.03	$\mathbf{0.76 \pm 0.03}$	**+0.02**	0.72 ± 0.06	$\mathbf{0.74 \pm 0.03}$	**+0.02**
D3	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.72 \pm 0.03}$	0.00	$\mathbf{0.75 \pm 0.00}$	0.73 ± 0.03	-0.02
D4	0.65 ± 0.00	$\mathbf{0.68 \pm 0.03}$	**+0.03**	0.72 ± 0.03	$\mathbf{0.75 \pm 0.00}$	**+0.03**
D5	0.73 ± 0.03	$\mathbf{0.87 \pm 0.03}$	**+0.14**	$\mathbf{0.75 \pm 0.05}$	0.68 ± 0.03	-0.07
D6	$\mathbf{0.75 \pm 0.02}$	0.72 ± 0.01	-0.03	$\mathbf{0.73 \pm 0.05}$	0.69 ± 0.03	-0.04
D7	0.65 ± 0.05	$\mathbf{0.72 \pm 0.03}$	**+0.07**	$\mathbf{0.65 \pm 0.00}$	0.57 ± 0.08	-0.08
D8	0.74 ± 0.01	$\mathbf{0.75 \pm 0.00}$	**+0.01**	0.67 ± 0.05	$\mathbf{0.71 \pm 0.04}$	**+0.04**
D9	$\mathbf{0.59 \pm 0.00}$	0.57 ± 0.00	-0.02	$\mathbf{0.58 \pm 0.03}$	$\mathbf{0.58 \pm 0.06}$	0.00
D10	$\mathbf{0.76 \pm 0.03}$	0.71 ± 0.03	-0.05	$\mathbf{0.66 \pm 0.04}$	0.62 ± 0.05	-0.04
D11	0.69 ± 0.12	$\mathbf{0.73 \pm 0.00}$	**+0.04**	$\mathbf{0.78 \pm 0.04}$	$\mathbf{0.78 \pm 0.03}$	0.00
D12	0.64 ± 0.04	$\mathbf{0.76 \pm 0.06}$	**+0.12**	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.78 \pm 0.06}$	0.00
D13	0.74 ± 0.01	$\mathbf{0.82 \pm 0.03}$	**+0.08**	0.79 ± 0.01	$\mathbf{0.81 \pm 0.03}$	**+0.02**
D14	0.69 ± 0.01	$\mathbf{0.74 \pm 0.01}$	**+0.05**	0.62 ± 0.02	$\mathbf{0.69 \pm 0.05}$	**+0.07**
D15	0.64 ± 0.03	$\mathbf{0.68 \pm 0.00}$	**+0.04**	0.63 ± 0.01	$\mathbf{0.68 \pm 0.00}$	**+0.05**
Avg.	0.69	**0.73**	**+0.04**	**0.70**	**0.70**	0.00
Max.	0.76	**0.87**	**+0.11**	0.79	**0.81**	**+0.02**

5 Experimental Results

5.1 Classification Evaluation

Accuracy, a common metric to evaluate classification tasks, is not suitable for unbalanced data. Therefore, the model is evaluated using balanced accuracy (BA) and $F1$ defined in Equations 16 and 17 respectively. Both utilize the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. The average of sensitivity and specificity, BA is well suited to evaluate diagnostic models. Meanwhile, the high emphasis that $F1$ places on true positive predictions makes it appropriate for evaluating unbalanced data commonly found in healthcare.

$$BA = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \tag{16}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{17}$$

For model evaluation, the formed test sets contain 20% of each dataset by selecting a random stratified sample. As this is a multi-task problem, the training data is upsampled until the upsampled training set is approximately balanced for both the MDD and the PTSD labels. Each model was repeated ten times with randomly initialized weights. The results reported are the average of the five highest values. To prevent overfitting, early stopping, and dropout layers were employed during training.

Table 4: F1-Score of single-task (ST) and multi-task (MT) models for MDD and PTSD screening, with task Gain (MT-ST). The higher value between MT and ST as well as positive gain are in bold.

	MDD			PTSD		
Dataset	**Single-task**	**Multi-task**	**Gain**	**Single-task**	**Multi-task**	**Gain**
D1	0.70 ± 0.00	$\mathbf{0.75 \pm 0.02}$	**+0.05**	$\mathbf{0.71 \pm 0.02}$	0.70 ± 0.04	-0.01
D2	0.70 ± 0.06	$\mathbf{0.74 \pm 0.01}$	**+0.04**	$\mathbf{0.70 \pm 0.01}$	0.69 ± 0.02	-0.01
D3	0.70 ± 0.04	$\mathbf{0.74 \pm 0.01}$	**+0.04**	$\mathbf{0.76 \pm 0.00}$	0.74 ± 0.02	-0.02
D4	0.69 ± 0.03	$\mathbf{0.71 \pm 0.02}$	**+0.02**	$\mathbf{0.74 \pm 0.02}$	0.72 ± 0.02	-0.02
D5	0.65 ± 0.03	$\mathbf{0.81 \pm 0.02}$	**+0.16**	$\mathbf{0.66 \pm 0.06}$	0.59 ± 0.02	-0.07
D6	$\mathbf{0.72 \pm 0.05}$	0.69 ± 0.04	-0.03	$\mathbf{0.69 \pm 0.07}$	0.66 ± 0.07	-0.03
D7	0.70 ± 0.02	$\mathbf{0.74 \pm 0.02}$	**+0.04**	$\mathbf{0.67 \pm 0.01}$	$\mathbf{0.67 \pm 0.00}$	0.00
D8	$\mathbf{0.76 \pm 0.00}$	0.75 ± 0.03	-0.01	$\mathbf{0.73 \pm 0.03}$	0.68 ± 0.04	-0.05
D9	$\mathbf{0.61 \pm 0.02}$	0.60 ± 0.04	-0.01	0.54 ± 0.02	$\mathbf{0.58 \pm 0.10}$	**+0.03**
D10	$\mathbf{0.75 \pm 0.04}$	0.72 ± 0.02	-0.03	0.60 ± 0.04	$\mathbf{0.64 \pm 0.02}$	**+0.04**
D11	$\mathbf{0.68 \pm 0.11}$	$\mathbf{0.68 \pm 0.02}$	0.00	$\mathbf{0.77 \pm 0.04}$	$\mathbf{0.77 \pm 0.01}$	0.00
D12	0.67 ± 0.00	$\mathbf{0.76 \pm 0.04}$	**+0.09**	0.75 ± 0.02	$\mathbf{0.76 \pm 0.02}$	**+0.01**
D13	0.76 ± 0.02	$\mathbf{0.83 \pm 0.02}$	**+0.07**	0.78 ± 0.03	$\mathbf{0.81 \pm 0.03}$	**+0.03**
D14	0.71 ± 0.02	$\mathbf{0.79 \pm 0.01}$	**+0.08**	0.66 ± 0.05	$\mathbf{0.72 \pm 0.03}$	**+0.06**
D15	0.60 ± 0.02	$\mathbf{0.61 \pm 0.02}$	**+0.01**	$\mathbf{0.65 \pm 0.03}$	0.64 ± 0.02	-0.01
Avg.	0.69	**0.73**	**+0.04**	**0.69**	**0.69**	0.00
Max.	0.76	**0.83**	**+0.07**	0.78	**0.81**	**+0.03**

5.2 SINGLE-TASK VS MULTI-TASK

The average BA and average F1 scores from the experiments are displayed in Tables 3 and 3, respectively. For MDD screening, the highest average balanced accuracy of 0.87 (on $D5$) and the highest average F1 of 0.83 (on $D13$) was achieved by the developed multi-task framework. Likewise, the multi-task framework also achieved the highest average balanced accuracy and F1 of 0.81 (on $D13$) for PTSD screening. When compared to single-task learning, multi-task learning improved the maximum balanced accuracy by 0.11 for MDD screening and 0.02 for PTSD screening as well as the maximum F1 score by 0.07 for MDD screening and 0.03 for PTSD screening.

On average, across the 15 datasets, the multi-task framework outperforms its single-task counterpart for MDD screening, with an average balanced accuracy and F1 of 0.73. In comparison, the single-task framework only achieves an average balanced accuracy and F1 of 0.69 across all datasets. Thus, multi-task learning has a relative gain of 0.04 over single-task learning. Additionally, the multi-task framework achieves a higher balanced accuracy score for 11 out of 15 datasets and a higher F1 score for 10 out of 15 datasets. The largest gains in balanced accuracy were for $D5$ ($+0.14$), $D12$ ($+0.12$), $D13$ ($+0.08$), and $D7$ ($+0.07$) and the greatest gains in F1 were in $D5$ ($+0.16$), $D12$ ($+0.09$), $D13$ ($+0.08$), and $D14$ ($+0.08$).

As for the PTSD screening task, across the datasets, both single-task and multi-task models achieve an average balanced accuracy of 0.70. The single-task models have balanced accuracy scores ranging from 0.58 to 0.79 whereas the multi-task models have balanced accuracy scores ranging from 0.57 to 0.81. Likewise, the single-task models have F1 scores ranging from 0.54 to 0.77 whereas the

multi-task models have F1 scores ranging from 0.58 to 0.81. Thus, the multi-task models have slightly higher minimum and maximum scores across the datasets than the single-task models. Overall, for PTSD screening, the multi-task framework achieves a higher BA score for 7 out of 15 datasets and a higher F1 score for 5 out of 15 datasets. The largest gain in both balanced accuracy and F1 was achieved for $D14$.

The results overall demonstrate that multi-task learning improved MDD screening. While PTSD screening did not observe the same benefit as MDD from multi-task learning, it still maintained equal performance in comparison to its single-task counterpart while aiding the ability to screen for MDD. By providing the model with multiple tasks, the model is made to learn a more robust representation of the hidden embeddings. It is thus able to better perform over the dataset for one of the primary tasks, in this case, MDD screening.

5.3 Question Analysis

Understanding the screening value of each clinical interview question allows for the prioritization of certain questions to be deployed during shorter screening sessions. For MDD screening with temporal facial features, both $D5$ and $D13$ have average BA and F1 scores over 0.80 obtained with multi-task learning. Notably, when screening for MDD with D5, the multi-task framework improved the balanced accuracy by 19% (0.14) and the F1 by 25% (0.16) over the single-task framework. Meanwhile, for single-task learning with temporal facial features, $D10$ was the only dataset with both average BS and F1 scores of at least 0.75 for MDD screening.

For PTSD screening with temporal facial features, both single-task and multi-task learning models achieved the highest average BA and F1 scores on $D13$. The multi-task framework with $D13$ was the only PTSD screening model to obtain a BA and F1 over 0.80. $D11$ and $D12$ were the other two most useful questions for MDD screening, with both the single and multi-task models obtaining BA scores of 0.78 and F1 scores of at least 0.75.

From the results, the recommendation drawn is to include the question represented as $D13$ in multi-task models that simultaneously perform MDD and PTSD screening with temporal facial features. The other datasets that performed above average for both screening tasks in the multi-task framework are $D1$, $D2$, $D8$, $D12$, and $D14$.

6 Discussion

6.1 Benefits of Facial Features for Mental Health Assessment

There are ethical concerns regarding screening for mental illnesses with data that can be collected by bad agents outside of clinical settings. It is generally considered that the benefits of screening outweigh the risk of misuse. Out of the modalities extracted from videos, temporal facial features protect privacy more than alternatives such as voice recordings and transcript content. The facial features could be extracted from a video stream or recording. Either way, the video does not need to be retained. As the facial features extracted from the OpenFace software (74), present a non-identifiable screening modality that retains signal for screening applications.

Both the single-task and multi-task models evaluated in this thesis uses temporal facial features as input for MDD and PTSD screening, thus making them language-agnostic. The usage of transcripts for mental illness research (54; 47; 51; 52; 88; 89; 90) is quite common as this modality allows the usage of large pre-trained models for optimal performance. However, these models are often trained on large corpora that depend on specific languages, making them not generalize well across different languages. Thus, they would require that a separate model be trained for every language in a dataset, increasing computational costs. Since the models focus on non-verbal information, these models can be deployed and adapted in multilingual settings.

6.2 Benefits of GRU for Mental Health Assessment

As previously mentioned, the GRU models implemented in this research do not have high computation costs (81). While they function similarly, GRU models are more efficient than LSTM models (81). Unlike the majority of large language models, GRU models also do not require a graphics processing unit (GPU) for training and therefore would be easier to implement in mobile applications and/or clinical settings. As such, models that screen with temporal facial features would be more accessible for economically disadvantaged groups who are more likely to be affected by mental illnesses (91).

The GRU models do not sacrifice MDD screening ability for computational efficiency. Previously, an LSTM with self-attention achieved the highest balanced accuracy of 0.69 on D13 (proud, life) (30), which is notably lower than both the single-task and multi-task GRU models, which achieved BA scores of 0.74 and 0.82, respectively. Likewise, another dataset with a question similar to D11 (friend, describe) achieved BA scores of 0.63 and 0.53 with deep transfer audio and text models, respectively. In comparison, the models were able to achieve higher BA scores of 0.69 and 0.73 using temporal features in the single-task and multi-task GRU models, respectively.

6.3 Limitations and Future Work

In this thesis, a multi-task learning approach was designed to improve mental illness screening. To do so, a sequential GRU model was implemented with self-attention, which is efficient and works well for MDD screening. The simplicity of the self-attention layer could be considered a limitation. Further, multi-task learning was shown to be more beneficial for MDD screening than PTSD screening. A possible explanation is the simple regularization loss function, where the design of the loss optimization simply sums the three loss values to tally a total loss for backpropagation. A more advanced approach would be to assign learnable weights to each loss, allowing the model to prioritize information gained for backpropagation. Additionally, the benchmark dataset leveraged in this thesis has a relatively small number of participants and lacks demographic attributes for all participants – thus this thesis cannot provide any insights into the effectiveness of the model for particular demographics.

Future work on multi-task learning involves exploring new modeling strategies and datasets. As mentioned, the datasets leveraged in this research were relatively small. Leveraging larger datasets such as datasets that are publicly available on social media platforms such as Facebook (92) is something that can be done to further this work. Additionally, the relation between performance and dataset size was not explored. This work can be further expanded by involving an evaluation of performance over the size of data input to the model. The subsequent work based on this research can also incorporate time series transformers, multi-head attention (32), and other loss regularization techniques. This research can also be extended to a variety of other digital modalities, such as transcripts and audio recordings, and diagnostic screening tasks, such as suicidal ideation or anxiety.

7 CONCLUSION

This thesis provides an assessment of the performance of a multi-task learning architecture that leverages three types of temporal facial features to screen for MDD and PTSD across 15 datasets. Based on balanced accuracy results, multi-task learning outperforms single-task learning on 11 datasets for MDD screening and 7 datasets for PTSD screening. The highest average balanced accuracy for MDD screening was obtained by the multi-task model for the D5 dataset with a score of 0.87, an increase of 25% over the single-task model. The results indicate that temporal facial features and multi-task learning are a promising combination for future mental health screening applications.

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor Elke Rundensteiner and my reader, Professor Liu for their guidance and support in this work. I would also like to thank Dr. Ricardo Flores and Dr. ML Tlachac for their support and collaboration. This thesis was based on a paper (93) where Dr. Flores and I shared equal contributions and has been featured in his dissertation as well. Lastly, I thank EMUTIVO lab and DAISY lab for their continued support. The experiments of this work were conducted using a high-performance computing system at WPI.

REFERENCES

[1] Substance Abuse and Mental Health Services Administration, "Key sub- stance use and mental health indicators in the United States: Results from the 2020 national survey on drug use and health," Tech. Rep., 2021.

[2] U.S. Department of Veteran Affairs, "How common is ptsd in adults?" 2018, accessed Sept 14, 2023. [Online]. Available: https://www.ptsd.va.gov/understand/common

[3] H. Katschnig, "Quality of life in mental disorders: challenges for rese- arch and clinical practice," *World psychiatry*, vol. 5, no. 3, p. 139, 2006.

[4] M. V. Mendlowicz and M. B. Stein, "Quality of life in individuals with anxiety disorders," *Am. J. Psychiatry*, vol. 157, no. 5, pp. 669–82, 2000.

[5] H.-J. Kang *et al.*, "Comorbidity of depression with physical disorders: research and clinical implications," *Chonnam medical journal*, vol. 51, no. 1, pp. 8–18, 2015.

[6] J. Sareen *et al.*, "Physical and mental comorbidity, disability, and suicidal behavior associated with posttraumatic stress disorder in a large community sample," *Psychosomatic medicine*, vol. 69, no. 3, 2007.

[7] E. T. Isometsä, "Psychological autopsy studies–a review," *European psychiatry*, vol. 16, no. 7, pp. 379–385, 2001.

[8] U.S. Department of Veteran Affairs, "How common is ptsd in veterans?" 2018, accessed Sept 14, 2023. [Online]. Available: https://www.ptsd.va.gov/understand/common/common_veterans.asp

[9] J. D. Flory and R. Yehuda, "Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and treatment considerations," *Dialogues in clinical neuroscience*, 2022.

[10] B. Nichter *et al.*, "Physical health burden of ptsd, depression, and their comorbidity in the us veteran population: Morbidity, functioning, and disability," *Journal of psychosomatic research*, vol. 124, 2019.

[11] R. Epstein *et al.*, ""i didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress," *Journal of General Internal Medicine*, vol. 25, no. 9, pp. 954–61, 2010.

[12] M. Tlachac *et al.*, "Impact assessment of stereotype threat on mobile depression screening using bayesian estimation," *Healthcare Analytics*, vol. 2, p. 100088, 2022.

[13] A. Halfin, "Depression: the benefits of early and appropriate treatment," *American Journal of Managed Care*, vol. 13, no. 4, p. S92, 2007.

[14] E. P. Balogh, B. T. Miller, and J. R. Ball, *Improving Diagnosis in Health Care*. National Academies Press, 2015.

[15] M. Savoy and D. O'Gurek, "Screening your adult patients for depression," *Family practice management*, vol. 23, no. 2, pp. 16–20, 2016.

[16] K. Demyttenaere *et al.*, "Comorbid painful physical symptoms and depression: prevalence, work loss, and help seeking," *Journal of affective disorders*, vol. 92, no. 2-3, pp. 185–193, 2006.

[17] M. É. Czeisler *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*, vol. 69, no. 32, 2020.

[18] F. Chamaa *et al.*, "Ptsd in the covid-19 era," *Current neuropharmacology*, vol. 19, no. 12, p. 2164, 2021.

[19] National Council for Behavioral Health, "The psychiatric shortage: causes and solutions," Tech. Rep., 2017.

[20] C. H. A. Andrilla *et al.*, "Geographic variation in the supply of selected behavioral health providers," *American journal of preventive medicine*, vol. 54, no. 6, pp. S199–S207, 2018.

[21] K. M. Magruder, K. A. McLaughlin, and D. L. Elmore Borbon, "Trauma is a public health issue," *European journal of psychotraumatology*, vol. 8, no. 1, p. 1375338, 2017.

[22] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[23] C. A. Blevins *et al.*, "The posttraumatic stress disorder checklist for dsm-5 (pcl-5): Development and initial psychometric evaluation," *Journal of traumatic stress*, vol. 28, no. 6, pp. 489–498, 2015.

[24] A. J. Gelenberg *et al.*, "American psychiatric association practice guidelines for the treatment of patients with major depressive disorder," *American Journal of Psychiatry*, vol. 167, pp. 9–118, 2010.

[25] C. Frueh *et al.*, *Assessment and treatment planning for PTSD*. John Wiley & Sons, 2012.

[26] A. Reichert and R. Jacobs, "The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in england," *Health Economics*, vol. 27, no. 11, pp. 1772–1787, 2018.

[27] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual Review of Clinical Psychology*, vol. 14, pp. 91–118, 2018.

[28] D. DeVault *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Int. Conf. on Autonomous Agents and Multi-Agent Systems*, 2014, pp. 1061–68.

[29] R. Flores *et al.*, "Temporal facial features for depression screening," in *ACM UbiComp*, 2022.

[30] ——, "Audiface: Multimodal deep learning for depression screening," in *Machine Learning for Healthcare Conference*. PMLR, 2022.

[31] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1310–1318.

[32] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[35] N. Cummins *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, 2015.

[36] X. Ma *et al.*, "Depaudionet: An efficient deep model for audio based depression classification," in *6th AVEC*, 2016.

[37] K.-Y. Huang *et al.*, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing*, 2018.

[38] Z. Lin *et al.*, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[39] J. Li *et al.*, "Empirical evaluation of multi-task learning in deep neural networks for natural language processing," *Neural Computing and Applications*, vol. 33, no. 9, pp. 4417–4428, 2021.

[40] F. Ringeval *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *9th AVEC*, 2019, pp. 3–12.

[41] J. Shin and S. M. Bae, "A systematic review of location data for depression prediction," *International Journal of Environmental Research and Public Health*, vol. 20, no. 11, p. 5984, 2023.

[42] L. He *et al.*, "Deep learning for depression recognition with audiovisual cues: A review," *Information Fusion*, vol. 80, pp. 56–86, 2022.

[43] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews." in *Language Resources and Evaluation*. CiteSeer, 2014, pp. 3123–3128.

[44] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *6th AVEC*, 2016, pp. 3–10.

[45] E. Toto *et al.*, "Audio-based depression screening using sliding window sub-clip pooling," in *19th IEEE ICMLA*, 2020, pp. 791–6.

[46] M. L. Tlachac *et al.*, "Topological data analysis to engineer features from audio signals for depression detection," in *19th IEEE ICMLA*, 2020.

[47] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *30th ACM CIKM*, 2021, pp. 4145–4154.

[48] R. Flores *et al.*, "Depression screening using deep learning on follow-up questions in clinical interviews," in *20th IEEE ICMLA*, 2021.

[49] S. Sardari *et al.*, "Audio based depression detection using convolutional autoencoder," *Expert Systems with Applications*, vol. 189, 2022.

[50] A. Mallol-Ragolta *et al.*, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," *Proc. Interspeech 2019*, pp. 221–225, 2019.

[51] S. Senn *et al.*, "Ensembles of bert for depression classification," in *44th IEEE EMBC*, 2022, pp. 4691–4694.

[52] R. Flores *et al.*, "Transfer learning for depression screening from follow-up clinical interview questions," *Deep Learning Applications*, vol. 4, pp. 53–78, 2022.

[53] M. L. Tlachac *et al.*, "Text generation to aid depression detection: A comparative study of conditional sequence generative adversarial networks," in *IEEE BigData*. IEEE, 2022, pp. 2804–2813.

[54] M. Rodrigues Makiuchi *et al.*, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *AVEC*, 2019, pp. 55–63.

[55] S. Alghowinem *et al.*, "Eye movement analysis for depression detection," in *IEEE Int. Conf. on Image Processing*, 2013, pp. 4220–4.

[56] ——, "Head pose and movement analysis as an indicator of depression," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 283–288.

[57] S. Liu *et al.*, "Explainable ai for suicide risk assessment using eye activities and head gestures," in *International Conference on Human-Computer Interaction*. Springer, 2022, pp. 161–178.

[58] H. C. Cuve *et al.*, "Alexithymia explains atypical spatiotemporal dynamics of eye gaze in autism," *Cognition*, vol. 212, 2021.

[59] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, 1997.

[60] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[61] C. Jin *et al.*, "Predicting treatment response from longitudinal images using multi-task deep learning," *Nature communications*, vol. 12, no. 1, p. 1851, 2021.

[62] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *13th international conference on artificial intelligence and statistics*. JMLR, 2010, pp. 17–24.

[63] J. Zhou *et al.*, "A multi-task learning formulation for predicting disease progression," in *17th ACM SIGKDD*, 2011, pp. 814–822.

[64] J. Wan *et al.*, "Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[65] H. Wang *et al.*, "High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression predic- tion," *Advances in neural information processing systems*, vol. 25, 2012.

[66] Y. Li *et al.*, "A multi-task learning formulation for survival analysis," in *22nd ACM SIGKDD*, 2016, pp. 1715–24.

[67] A. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," *arXiv preprint arXiv:1712.03538*, 2017.

[68] U. Lokala *et al.*, "A computational approach to understand mental health from reddit: knowledge-aware multitask learning framework," in *Int. AAAI Conf. on Web and Social Media*, vol. 16, 2022, pp. 640–650.

[69] J. Lu *et al.*, "Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning," *ACM IMWUT*, vol. 2, no. 1, 2018.

[70] A. Laxmisan *et al.*, "The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care," *Int. j. of medical informatics*, vol. 76, no. 11-12, pp. 801–811, 2007.

[71] A. Pampouchidou *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *6th AVEC*, 2016.

[72] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," in *6th AVEC*, 2016, pp. 11–18.

[73] L. Yang *et al.*, "Decision tree based depression classification from audio video and language information," in *6th AVEC*, 2016, pp. 89–96.

[74] T. Baltrušaitis, P. Robinson, and L. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE WACV*, 2016, pp. 1–10.

[75] E. B. Blanchard *et al.*, "Psychometric properties of the ptsd checklist (pcl)," *Behaviour research and therapy*, vol. 34, no. 8, pp. 669–73, 1996.

[76] T. Bickmore, A. Gruber, and R. Picard, "Establishing the computer–patient working alliance in automated health behavior change interventions," *Patient Education & Counseling*, vol. 59, no. 1, pp. 21–30, 2005.

[77] C. Lau, W.-Y. Chan, and X. Zhu, "Improving depression assessment with multi-task learning from speech and text information," in *55th Asilomar Conf. on Signals, Systems, and Computers*. IEEE, 2021, pp. 449–53.

[78] M. Tlachac *et al.*, "Deprest-cat: Retrospective smartphone call and text logs collected during the covid-19 pandemic to screen for mental illnesses," *ACM IMWUT*, vol. 6, no. 2, pp. 1–32, 2022.

[79] Y. Cao *et al.*, "Depression prediction based on biattention-gru," *J. of Ambient Intelligence and Humanized Computing*, vol. 13, no. 11, 2022.

[80] S. Teng *et al.*, "A transformer-based multimodal network for audiovisual depression prediction," in *11th IEEE GCCE*, 2022, pp. 761–764.

[81] R. Cahuantzi, X. Chen, and S. Güttel, "A comparison of lstm and gru networks for learning symbolic sequences," in *Science and Information Conference*. Springer, 2023, pp. 771–785.

[82] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Advances in Neural Information Processing Systems*, 1997.

[83] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *2014 ACL EMNLP*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[84] S. Basodi *et al.*, "Gradient amplification: An efficient way to train deep neural networks," *Big Data Mining and Analytics*, vol. 3, no. 3, 2020.

[85] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, 1991.

[86] P. Zhai *et al.*, "Multi-task learning for lung nodule classification on chest ct," *IEEE access*, vol. 8, pp. 180 317–180 327, 2020.

[87] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, 2018.

[88] M. L. Tlachac *et al.*, "StudentSADD: Mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *ACM IMWUT*, vol. 6, no. 2, pp. 1–32, 2022.

[89] M. Tlachac *et al.*, "Early mental health uncovering with short scripted and unscripted voice recordings," in *Deep Learning Applications, Volume 4*. Springer, 2022, pp. 79–110.

[90] A. Shrestha *et al.*, "Bert variants for depression screening with typed and transcribed responses," in *ACM UbiComp*, 2023, p. 211–215.

[91] N. K. Grote *et al.*, "Engaging women who are depressed and economically disadvantaged in mental health treatment," *Social work*, vol. 52, no. 4, pp. 295–308, 2007.

[92] Y. Wang *et al.*, "Adaptive multi-task positive-unlabeled learning for joint prediction of multiple chronic diseases using online shopping behaviors," *Expert Systems with Applications*, vol. 191, p. 116232, 2022.

[93] A. Shrestha *et al.*, "Multi-task learning using facial features for mental health screening," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 4881–4890.