Temporal Facial Features for Depression Screening

Ricardo Flores
rflores@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

ML Tlachac
mltlachac@wpi.edu
Bryant University
Smithfield, RI, USA

Avantika Shrestha
ashrestha4@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Elke A. Rundensteiner
rundenst@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

ABSTRACT

Depression is a common and debilitating mental illness. Given the shortage of mental health professionals, there are delays in depression detection. Interviews conducted by virtual agents could expedite depression screenings. While the interview audio and transcript have received more attention, facial features offer an attractive privacy-preserving screening modality. Thus, we conduct a comprehensive comparative evaluation of the effectiveness of temporal facial features to screen for depression. We extract time series of eye gaze, landmark, and action unit features from video responses to 15 clinical interview questions. We input them into CNN, LSTM, and recurrent convolutional neural network (RCNN) models. An extra attention layer proved critical for CNN and LSTM performance. For a general wellbeing question, eye gaze features screened for depression with an F1 of 0.81. Our study informs the use of temporal facial features in future digital mental illness screening technologies.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Supervised learning by classification**; • **Mathematics of computing** → **Time series analysis**; • **Applied computing** → *Psychology*; Health informatics.

KEYWORDS

Digital health; Time Series Classification; Convolutional Neural Network; Recurrent Neural Network; Recurrent Convolutional Neural Network

ACM Reference Format:
Ricardo Flores, ML Tlachac, Avantika Shrestha, and Elke A. Rundensteiner. 2022. Temporal Facial Features for Depression Screening. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544793.3563424

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563424

1 INTRODUCTION

Depression is a common mental illness. The third leading cause of global disability [33], it has devastating social economic impacts. The increased rate of depression during the COVID-19 pandemic [5] exacerbated the shortage of mental health professionals [19, 34]. While there are screening surveys [12], diagnoses still require that mental health professionals conduct lengthy clinical interviews with patients. Given the shortage of such professionals, these interviews can be cost prohibitive and have long wait times. Unfortunately, delays in care can result in detrimental impacts on patient health and wellbeing [21].

Facial features from interviews can be useful in diagnostic applications, such as detecting autism with eye gaze [4]. Researchers have also used eye gaze activities and head pose to identify depression and suicidal ideation [1, 2, 16]. Eye gaze, landmark, and action unit facial features are also part of the popular Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) dataset [10] which contains video recordings of clinical interviews conducted by a virtual agent [6]. To probe the ability to expedite depression detection, these clinical interviews were featured in the 2016 Audio/Visual Emotion Challenge and Workshop (AVEC) [30]. The majority of the research on this dataset only uses the interview audio and transcripts in machine learning models that screen for depression [8, 9, 17, 22, 24, 27–29]. However, some research also leverages the DAIC-WOZ facial features in multimodal models [20, 32, 38].

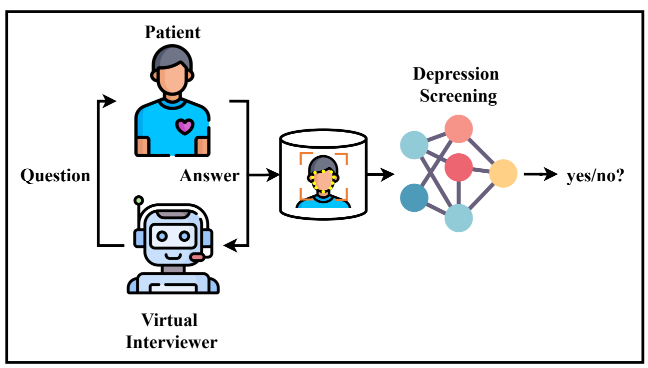

Figure 1: Depression screening conducted by a virtual agent asking a patient relevant questions. The captured temporal facial features (eye gaze, landmark, and action unit) are used to train a set of sequential deep learning model.

The prior studies [1, 2, 16, 20, 32, 38] that have used facial features require additional data transformations and leverage traditional machine learning models, such as support vector classifiers and decision trees. While the temporal aspect of eye gaze features have been used in the detection of autism with traditional machine learning models [4], the temporal aspect of facial features has not been explored for depression screening. Further, in the aforementioned related works, deep learning has not been applied to model the facial features.

To remedy this gap in the related literature, we conduct a comprehensive evaluation of the ability of the temporal facial features to screen for depression, like in Fig. 1. We elect to use deep learning models so that no information about the temporal facial features is lost from data transformations; prior research [1, 2, 4, 16, 20, 32, 38] leveraged feature engineering to transform the data for use in their machine learning models. As deep learning models can struggle to capture the relevant information in long sequences, we address this challenge by assessing the usefulness of adding a self attention layer [15] to our models.

Our novel modeling approach specifically involves constructing multivariate time series of eye gaze, landmark, and action unit features. For this research, we use the facial features extracted from the video recordings of responses to 15 clinical interview questions in the DAIC-WOZ dataset. We then use the constructed multivariate time series in deep learning models with and without an architectural attention layer. The models include convolutional neural network (CNN) which are known for their ability to classify faces as well as long short-term memory (LSTM) which are known for their ability to classify sequences. As the temporal facial features would benefit from both abilities, we also experiment with recurrent convolutional neural network (RCNN) models. Therefore, we compare the depression screening ability of:

(1) Fifteen different clinical interview questions,
(2) Three different types of temporal facial features,
(3) Three architectures of deep learning models, and
(4) Models with and without an extra layer of attention.

2 DATA AND METHODOLOGY

2.1 Datasets of Video Recordings

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) corpus of clinical interviews was collected by a virtual agent so the data could be used to identify verbal and nonverbal indicators of depression [6, 10]. The corpus thus contains audio, transcripts, facial features, and depression screening score labels from 189 participants. The interviews ranged between 7 to 33 minutes (with an average of 16 minutes). Facial features – landmarks, eye gaze, and facial action units – were extracted using OpenFace software [3] from each frame of the video recordings. The depression screening scores were obtained by administering the first eight questions of the Patient Health Questionnaire (PHQ-8) [12]. The sum of the questions, which range from 0 to 24 are used to screen for depression with the common threshold of 10 [12].

Each interview contains a subset of topical core questions with follow-up questions. We treat the responses to each core question as a separate dataset, as described in Toto et al. [28]. Thus, we parse the clinical interviews by topical core questions such that

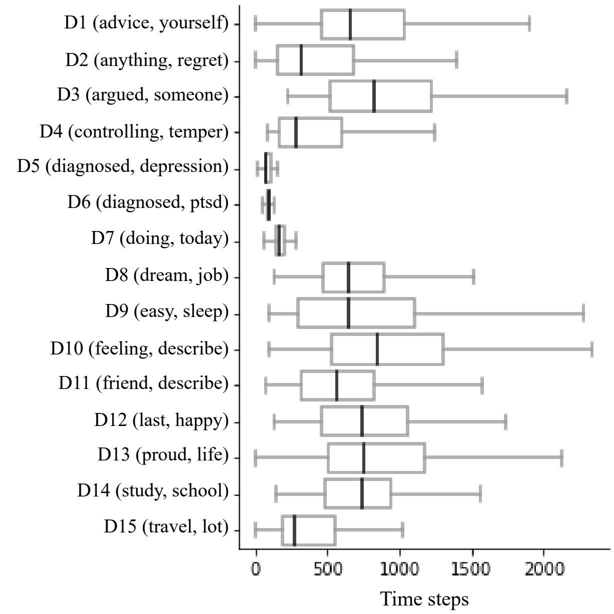

Figure 2: Distribution of the number of time steps in the video recordings for each dataset. The topic of each core question is represented with two keywords.

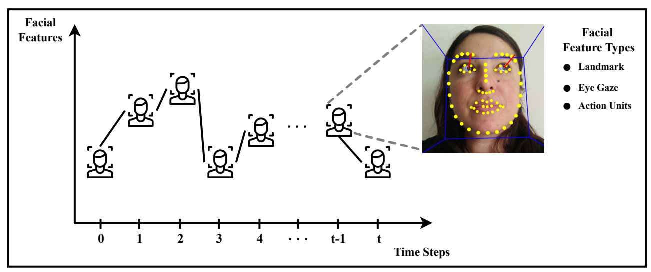

Figure 3: Temporal facial features diagram. For each combination of participant and time step, we extract facial landmark, eye gaze, and action unit features.

each dataset contains all participant responses until the next core question. There are 15 core questions to which at least 92 of the participants responded [9]. In this research, we use the facial features from the responses to these 15 core questions, further referred to as datasets D1 to D15. Between 21.3% (D5) and 30.5% (D8, D14) of participants screened positive for depression. As core question wording varied, we represent the topic of the core questions in Fig. 2. The amount of time steps in the videos vary by dataset, which is also displayed in Fig. 2.

2.2 Methodology for Temporal Facial Features

A main contribution of this work is our approach to using the temporal facial features. Provided by the OpenFace software [3], the facial features types encompass landmark, eye gaze, and action unit. Extracted from each video frame as depicted in Figure 3, these features represent a multivariate time series. There are 136, 12, and

14 dimensions respectively for the landmark, eye gaze, and action unit feature types.

Once we extract the facial features for each participant, then we create a separate data set for each of the 15 core question listed in Fig. 2, to do so, we use time steps from audio recording where we can get the initial and final time steps by each core question. Finally, we extract the facial features from these specific period of time, and order by question and participant. We use up to 1000 time steps as classifier input.

2.3 Deep Learning Classifiers

In deep learning, it is standard to classify images using CNNs [14, 18]. In particular, this architecture is useful for modeling facial landmark features [35]. As the data is temporal, it is also appropriate for recurrent networks such as LSTMs [11]. These two architectural approaches have been combined to form RCNNs [13], designed to improve classification ability by overcoming limitations of the individual architectures.

Attention is known to improve the performance of neural networks [15], including for other modalities in the domain of depression screening [8, 9, 24–26, 28]. As such, we assess the impact of adding a layer of attention to the CNN, LSTM, and RCNN models. For the RCNN model, the attention layer is added to the LSTM prior to the convolutional component. We anticipate that the self-attention layer will capture the relevant relationships in the longer sequences of temporal facial features and therefore improve depression screening results.

For the implementation of the LSTM, RNN, and RCNN models we use the following standard hyperparameters for time series classification: learning rate equal to 0.001, a hidden dimension size of 32, dropout of 0.2, and Adam optimizer. Since each batch represent the facial features of a patient at one specific time, we set batch size equal to 1.

2.4 Classification Evaluation

To evaluate the classifiers, we form a stratified test set with 20% of participants for each dataset. We then upsample the training set to balance the class labels. Recall, between 21.3% and 30.5% of the participants in each dataset screened positive for depression so the test sets remain unbalanced. Unlike accuracy, the F1 score is suitable for assessment with unbalanced test sets. Thus, we use the F1 score to evaluate our classifiers, which is defined as

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

which is calculated using the number of true positive predictions TP, false positive predictions TN, and false negative predictions FN. F1 score is the harmonic mean of precision and recall. We repeat each model 10 times for robustness, reporting on the average and standard deviation of the best 5 models.

2.5 Computational Resources and Updates

These models were run on an internal computing cluster at Worcester Polytechnic Institute (WPI) with CPU resources. We will post research updates on our project website: *emutivo.wpi.edu*.

3 EXPERIMENTAL EVALUATION

3.1 Aggregated Results Across All Datasets

The results aggregated over all 15 datasets are displayed in Fig. 4. The highest average F1 score of 0.69 is achieved by CNN Attention using eye gaze features. However, the LSTM Attention model performed almost as well on this facial feature type, making both viable modeling choices when screen for depression with time series of eye gaze features.

Impact of feature type. While the eye gaze features proved most predictive of depression, the other features performed almost as well. Either CNN Attention or LSTM Attention were good model choices for eye gaze and action unit features. While we expected CNN models to perform best on landmark features, LSTM Attention surprisingly performed slightly better than any of the other models. Without an attention layer, LSTM performed the worst for both the eye gaze and landmark feature types. The overall worst performing model is CNN without attention on action unit features. CNN may have performed better if we used raw images instead of extracted facial features.

Impact of attention. For all three types of features, the best models were those with attention layers. Attention proved useful for obtaining good depression screening results. Attention had the least effect on the RCNN models. For the eye gaze and landmark features, attention had the most impact on the LSTM models. Further, for all feature types, attention also notably reduced the standard deviation of the LSTM models. Attention had the largest impact on the CNN models for action unit features; CNN is the worst performing model while CNN Attention is the best model. Interestingly, attention had no impact on the CNN models for landmark features. Though, attention is overall helpful for the CNN and LSTM models. Notably, attention layer helps to reduce the standard deviation of almost all models, this is because of attention property, which allows to focus on the relevant sequences or spacial features for depression screening, rather than considering the whole information.

Impact of model type. As noted, either CNN Attention and LSTM Attention were the best models for each of the feature types.

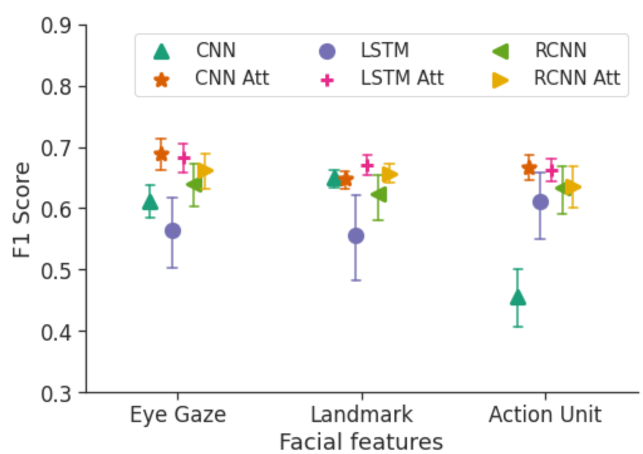

Figure 4: Aggregated results by type of facial features, averaging the 15 datasets.

Table 1: Eye Gaze: Average ± standard deviation of the $F1$ scores.

	CNN		LSTM		RCNN	
Dataset	Base	Attention	Base	Attention	Base	Attention
D1	0.66 ± 0.01	**0.78 ± 0.04**	0.71 ± 0.00	0.72 ± 0.02	0.73 ± 0.01	0.77 ± 0.00
D2	0.63 ± 0.03	0.70 ± 0.01	0.64 ± 0.00	**0.71 ± 0.03**	0.69 ± 0.02	0.70 ± 0.01
D3	0.68 ± 0.04	**0.71 ± 0.07**	0.67 ± 0.00	**0.71 ± 0.04**	0.67 ± 0.00	0.68 ± 0.02
D4	0.66 ± 0.03	0.76 ± 0.03	0.70 ± 0.01	0.76 ± 0.02	0.78 ± 0.02	**0.79 ± 0.03**
D5	**0.61 ± 0.03**	0.49 ± 0.04	0.00 ± 0.00	0.48 ± 0.04	0.42 ± 0.15	0.54 ± 0.07
D6	0.58 ± 0.02	**0.63 ± 0.03**	0.61 ± 0.00	0.60 ± 0.02	0.61 ± 0.00	0.61 ± 0.00
D7	0.73 ± 0.02	**0.81 ± 0.04**	0.69 ± 0.00	0.74 ± 0.07	0.71 ± 0.02	0.70 ± 0.01
D8	0.60 ± 0.01	**0.72 ± 0.02**	0.58 ± 0.00	0.71 ± 0.03	0.58 ± 0.00	**0.72 ± 0.02**
D9	0.58 ± 0.01	0.56 ± 0.09	0.38 ± 0.00	**0.61 ± 0.10**	0.38 ± 0.00	0.38 ± 0.01
D10	0.61 ± 0.03	0.68 ± 0.02	0.70 ± 0.00	0.70 ± 0.01	**0.72 ± 0.00**	0.70 ± 0.01
D11	0.54 ± 0.06	0.62 ± 0.02	0.56 ± 0.02	**0.65 ± 0.02**	0.60 ± 0.06	0.58 ± 0.04
D12	0.39 ± 0.09	0.69 ± 0.02	0.51 ± 0.05	0.68 ± 0.04	0.67 ± 0.02	**0.71 ± 0.02**
D13	0.68 ± 0.02	**0.77 ± 0.02**	0.50 ± 0.43	0.76 ± 0.02	0.76 ± 0.06	0.70 ± 0.06
D14	0.69 ± 0.08	**0.75 ± 0.03**	0.58 ± 0.00	**0.75 ± 0.04**	0.66 ± 0.03	0.72 ± 0.02
D15	0.52 ± 0.02	**0.66 ± 0.01**	0.64 ± 0.00	0.64 ± 0.00	0.64 ± 0.00	0.64 ± 0.00

However, without attention, RCNN is the best performing model for eye gaze and action unit feature types. RCNN is less reliant on feature type than CNN or LSTM models. This suggests that either the addition of an attention layer or a more advanced model is required to obtain better and more robust screening results.

3.2 Individual Dataset Results

For each of the 15 individual datasets, the results for the eye gaze, landmark, and action unit features are in Tables 1, 2, 3, respectively. The highest average F1 score 0.81 is achieved with a CNN Attention model on eye gaze features for D7. Likewise, a CNN Attention model on action unit features for D6 achieved an average F1 score of 0.80. We thus surmise that different temporal facial feature types are more effective at depression screening for different datasets.

Eye gaze. The best datasets for eye gaze features are D7 (doing, today) with CNN Attention, D4 (controlling, temper) with RCNN Attention, D1 (advice, yourself) with CNN Attention, and D13 (proud, life) with CNN Attention. All of these models achieved an average F1 score of at least 0.77. For the eye gaze features, CNN Attention was the best model for six datasets while LSTM Attention was the best model for five of the datasets. With the exception of D5 (diagnosed, PTSD) and D9 (easy, sleep), CNN Attention achieved a score of at least 0.62 and LSTM attention achieved a score of at least 0.64.

Landmark. LSTM Attention achieved the highest average F1 score for every dataset with landmark features. The best dataset is D13 (proud, life) with an average F1 score of 0.75. Multiple models with eye gaze features achieved slightly higher average F1 scores for D13. In fact, landmark features did not achieve the highest average F1 score for any dataset.

Action unit. The best datasets for action unit features are D6 (diagnosed, PTSD) with CNN Attention and D8 (dream, job) with RCNN, and D13 (proud, life) with RCNN Attention. These models achieved average F1 scores between 0.77 and 0.80. Three more questions achieved average F1 scores of 0.75: D2 with RCNN, D4 with CNN, and D10 with LSTM. As four of the six best models did not involve attention, this layer seems less useful for action unit features. CNN Attention and LSTM both achieved the highest scores for four datasets.

4 DISCUSSION AND FUTURE WORK

Ethics. The OpenFace software [3] extracts temporal facial features from a video stream or recording; either way, the video does not need to be retained. Thus, these temporal facial features have the benefit over the other interview modalities because they preserve patient privacy while retaining all positional details of the face. The action unit features can even capture emotion [37]. In summary, no identifiable information need be stored to screen with temporal facial features.

Limitations. While we included the maximum number of participants in each dataset, the participants in each dataset did differ as a result. Further, the number of participants by dataset unfortunately remained a limitation. Answered by 105 participants, D1 was the most populous dataset and among the most predictive for eye gaze. This indicates that more participants may improve depression screening results. Like many diagnostic datasets [7], the datasets suffered from class imbalance. The highest average F1 score for D5, the least balanced dataset, was 0.61. This was lower than the highest F1 score for any of the other datasets, suggesting class imbalance may negatively impact results.

Challenges. We acknowledge there are more advanced sequential deep learning models. For example, Wen et al. [31] summarizes the application of transformer in time series, in particular pre-trained transformer for multivariate time series classification [36, 39, 40]. However, we argue that pre-trained transformer models have high computational cost, requiring expensive GPUs for training huge models. This computational cost discourages the implementation of such models within depression screening applications. Furthermore, transformer-based models have problems dealing with long sequences, which is why some researchers still leverage traditional sequential models like LSTM for long time series modeling [23]. Thus, for our comparative study, we elected to

Table 2: Landmark: Average ± standard deviation of the $F1$ scores.

	CNN		LSTM		RCNN	
Dataset	Base	Attention	Base	Attention	Base	Attention
D1	0.69 ± 0.00	0.69 ± 0.00	0.71 ± 0.00	**0.72 ± 0.02**	0.69 ± 0.00	0.71 ± 0.00
D2	**0.69 ± 0.00**	**0.69 ± 0.00**	0.64 ± 0.00	**0.69 ± 0.00**	0.66 ± 0.03	**0.69 ± 0.01**
D3	**0.67 ± 0.00**	**0.67 ± 0.00**	**0.67 ± 0.00**	**0.67 ± 0.00**	**0.67 ± 0.00**	**0.67 ± 0.00**
D4	0.69 ± 0.00	0.69 ± 0.00	0.68 ± 0.01	**0.72 ± 0.02**	0.69 ± 0.00	0.69 ± 0.00
D5	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00	**0.53 ± 0.00**	0.50 ± 0.00	**0.53 ± 0.00**
D6	**0.61 ± 0.00**	**0.61 ± 0.00**	0.00 ± 0.00	**0.61 ± 0.00**	**0.61 ± 0.00**	**0.61 ± 0.00**
D7	0.67 ± 0.00	0.67 ± 0.00	0.69 ± 0.00	**0.71 ± 0.02**	0.67 ± 0.00	0.69 ± 0.00
D8	**0.69 ± 0.00**	**0.69 ± 0.00**	0.58 ± 0.00	**0.69 ± 0.00**	0.65 ± 0.06	0.62 ± 0.06
D9	**0.64 ± 0.00**	**0.64 ± 0.00**	0.38 ± 0.00	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**
D10	0.67 ± 0.00	0.67 ± 0.00	**0.70 ± 0.00**	**0.70 ± 0.00**	0.67 ± 0.00	**0.70 ± 0.00**
D11	0.59 ± 0.02	0.58 ± 0.00	0.56 ± 0.02	**0.62 ± 0.04**	0.29 ± 0.29	0.58 ± 0.00
D12	0.66 ± 0.03	0.64 ± 0.00	0.64 ± 0.00	**0.67 ± 0.02**	0.64 ± 0.00	0.64 ± 0.00
D13	0.64 ± 0.00	0.64 ± 0.00	**0.75 ± 0.00**	**0.75 ± 0.00**	0.64 ± 0.00	**0.75 ± 0.00**
D14	0.69 ± 0.00	0.69 ± 0.00	0.69 ± 0.00	**0.71 ± 0.03**	0.69 ± 0.00	0.69 ± 0.00
D15	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**	**0.64 ± 0.00**

Table 3: Action Unit: Average ± standard deviation of the $F1$ scores.

	CNN		LSTM		RCNN	
Dataset	Base	Attention	Base	Attention	Base	Attention
D1	0.31 ± 0.00	0.65 ± 0.01	**0.70 ± 0.01**	0.68 ± 0.01	**0.70 ± 0.06**	0.67 ± 0.02
D2	0.51 ± 0.04	0.70 ± 0.01	0.74 ± 0.00	0.69 ± 0.00	**0.75 ± 0.08**	0.72 ± 0.02
D3	0.14 ± 0.13	0.63 ± 0.04	0.67 ± 0.00	0.67 ± 0.00	**0.72 ± 0.02**	0.69 ± 0.02
D4	**0.75 ± 0.02**	0.72 ± 0.02	0.69 ± 0.00	0.72 ± 0.01	0.74 ± 0.00	0.70 ± 0.01
D5	0.38 ± 0.02	**0.54 ± 0.02**	0.00 ± 0.00	0.53 ± 0.00	0.27 ± 0.02	0.26 ± 0.02
D6	0.63 ± 0.03	**0.80 ± 0.02**	0.61 ± 0.00	0.67 ± 0.04	0.68 ± 0.02	0.64 ± 0.03
D7	0.64 ± 0.02	**0.73 ± 0.04**	0.68 ± 0.01	0.69 ± 0.02	0.63 ± 0.06	0.64 ± 0.03
D8	0.51 ± 0.06	0.66 ± 0.04	0.64 ± 0.00	0.72 ± 0.02	**0.77 ± 0.07**	0.74 ± 0.05
D9	0.33 ± 0.02	0.62 ± 0.03	0.43 ± 0.03	**0.66 ± 0.03**	0.52 ± 0.02	0.63 ± 0.03
D10	0.35 ± 0.00	0.67 ± 0.04	**0.75 ± 0.02**	0.68 ± 0.05	0.67 ± 0.01	0.64 ± 0.02
D11	0.40 ± 0.00	**0.63 ± 0.03**	0.55 ± 0.00	0.60 ± 0.01	0.53 ± 0.05	0.58 ± 0.00
D12	0.56 ± 0.10	0.70 ± 0.03	0.62 ± 0.05	**0.72 ± 0.01**	0.69 ± 0.05	0.67 ± 0.04
D13	0.49 ± 0.11	0.71 ± 0.04	0.75 ± 0.00	0.72 ± 0.03	0.76 ± 0.06	**0.77 ± 0.05**
D14	0.35 ± 0.07	0.58 ± 0.05	**0.69 ± 0.00**	0.55 ± 0.07	0.59 ± 0.06	0.55 ± 0.10
D15	0.48 ± 0.04	**0.64 ± 0.03**	**0.64 ± 0.00**	**0.64 ± 0.03**	0.48 ± 0.02	0.63 ± 0.03

use these traditional sequential models since our data consisted of temporal facial features extracted from long video recordings.

Future work. Our findings inform future research in regards to the combination of modeling strategy, temporal facial feature type, and question. This can be used in the development of unimodal or multimodal screening models. For example, future work could combine the time series of temporal facial features with features extracted from the corresponding audio and transcripts. Further, a larger dataset could be collected to determine the generalizability of our results as well as whether more participants improves ability of the deep learning models to screen for depression with temporal facial features. Given the ubiquity of cameras, these future datasets could be collected in a variety of settings.

5 CONCLUSION

Our research provides the first comprehensive assessment of the usefulness of temporal facial features to screen for depression. We experiment with 3 different facial feature types, 6 deep learning architectures, and 15 datasets. Overall, attention proved helpful in improving depression screening capabilities of the CNN and LSTM models, which yielded the highest average F1 scores aggregated across all datasets. For each of the individual datasets, either eye gaze or action unit features produced the highest F1 scores. In summary, our results promise to help future research wisely leverage temporal facial features to screen for mental illnesses.

ACKNOWLEDGMENT

This work was supported by Fulbright Foreign Student Program, National Agency for Research and Development (ANID)/Scholarship

Program/DOCTORADO BECAS CHILE/2015-56150007, Chile. U.S. Department of Education 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 Ermal Toto, Thomas Hartvigsen, and the Data-driven Intelligent Systems (DAISY) research lab at WPI for advice and support.

REFERENCES

[1] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*. IEEE, 4220–4224.

[2] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 283–288.

[3] Tadas Baltrušaitis, Peter Robinson, and L Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *IEEE WACV*. 1–10.

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

[5] Mark É Czeisler, Rashon I Lane, Emiko Petrosky, Joshua F Wiley, Aleta Christensen, Rashid Njai, Matthew D Weaver, Rebecca Robbins, Elise R Facer-Childs et al. 2020. 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).

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proc. of 2014 Int. Conf. on Autonomous Agents and Multi-Agent Systems*. 1061–68.

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

[8] Ricardo Flores, ML Tlachac, Ermal Toto, and Elke Rundensteiner. 2022. Transfer Learning for Depression Screening from Follow-up Clinical Interview Questions. *Deep Learning Applications* 4 (2022). In Press.

[9] Ricardo Flores, Ml Tlachac, Ermal Toto, and Elke A Rundensteiner. 2021. Depression Screening Using Deep Learning on Follow-up Questions in Clinical Interviews. In *20th IEEE ICMLA*. 595–600.

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

[11] Sepp Hochreiter, and Jürgen Schmidhuber. 1997. LSTM can solve hard long time lag problems. In *Advances in Neural Information Processing Systems*.

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

[13] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *29th AAAI Conference on Artificial Intelligence*.

[14] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

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

[16] Siyu Liu, Catherine Lu, Sharifa Alghowinem, Lea Gotoh, Cynthia Breazeal, and Hae Won Park. 2022. Explainable AI for Suicide Risk Assessment Using Eye Activities and Head Gestures. In *International Conference on Human-Computer Interaction*. Springer, 161–178.

[17] Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W Schuller. 2019. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. *Proc. Interspeech 2019* (2019), 221–225.

[18] Alessio Micheli. 2009. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks* 20, 3 (2009), 498–511.

[19] National Council for Behavioral Health. 2017. *The psychiatric shortage: causes and solutions*. Technical Report.

[20] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos et al. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 27–34.

[21] Anika Reichert, and Rowena Jacobs. 2018. The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England. *Health Economics* 27, 11 (2018), 1772–1787.

[22] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. In *AVEC*. 55–63.

[23] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36, 3 (2020), 1181–1191.

[24] Saskia Senn, M L Tlachac, Ricardo Flores, and Elke Rundensteiner. 2022. Ensembles of BERT for Depression Classification. In *44th EMBC*. In press.

[25] M L Tlachac, Ricardo Flores, Miranda Reisch, Katie Houskeeper, and Elke Rundensteiner. 2022. DepreST-CAT: Retrospective Smartphone Call and Text Logs Collected During the COVID-19 Pandemic to Screen for Mental Illnesses. *ACM IMWUT* 6, 2 (2022).

[26] M L Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering et al. 2022. StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic. *ACM IMWUT* 6, 2 (2022).

[27] M. L. Tlachac, Adam Sargent, Ermal Toto, Randy Paffenroth, and Elke Rundensteiner. 2020. Topological Data Analysis to Engineer Features from Audio Signals for Depression Detection. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 302–307.

[28] Ermal Toto, ML Tlachac, and Elke A Rundensteiner. 2021. AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening. In *30th ACM CIKM*. 4145–4154.

[29] Ermal Toto, ML Tlachac, Francis Lee Stevens, and Elke A Rundensteiner. 2020. Audio-based Depression Screening using Sliding Window Sub-clip Pooling. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 791–796.

[30] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie et al. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *6th AVEC*. 3–10.

[31] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125* (2022).

[32] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 11–18.

[33] World Health Organization. 2017. *Depression and other common mental disorders: global health estimates*. Technical Report.

[34] World Health Organization. 2022. World mental health report: transforming mental health for all. Geneva: World Health Organization, 1–296. ISBN 978-92-4-004933-8.

[35] Yue Wu, Tal Hassner, Kanggeon Kim, Gérard Medioni, and Prem Natarajan. 2018. Facial Landmark Detection with Tweaked Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2018), 3067–3074. https://doi.org/10.1109/TPAMI.2017.2787130

[36] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. 2021. Voice2series: Reprogramming acoustic models for time series classification. In *International Conference on Machine Learning*. PMLR, 11808–11819.

[37] Jiannan Yang, Fan Zhang, Bike Chen, and Samee U Khan. 2019. Facial expression recognition based on facial action unit. In *2019 10th IGSC*. IEEE, 1–6.

[38] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 89–96.

[39] Yuan Yuan, and Lei Lin. 2020. Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), 474–487.

[40] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2114–2124.