Research Article ISSN Online: 2783-4859 Computational Mathematics and Computer Modeling with Applications

Received 15 December 2023
Accepted 12 April 2024

DOI: 10.48308/CMCMA.2.1.45
AMS Subject Classification: 68T07; 68T10; 68T45

Automated Depression Recognition Using Multimodal Machine Learning: A Study on the DAIC-WOZ Dataset

Alireza Afzal Aghaei[a] **and Nadia Khodaei**[b]

This paper addresses the escalating global mental health crisis, particularly accentuated by the COVID-19 pandemic, by proposing a robust solution for the automated detection of depression. Leveraging the DAIC-WOZ dataset, a collection of clinical interviews and survey evaluations from over a hundred individuals, the study employs machine learning algorithms to automate and enhance depression recognition. The performance of the proposed models is rigorously evaluated using key metrics, including root mean square error (RMSE) and mean absolute error (MAE). A significant innovation is introduced with the incorporation of a novel attention fusion network, allowing the integration of features extracted from diverse modalities such as video, text, and audio. The study places a distinctive emphasis on intramodality connection, elucidating the intricate interactions among features within and across modalities. Structured into two pivotal sections, the first reviews existing approaches to automatic depression recognition, exploring associated areas and commonly employed modalities. The second section focuses on methodologies related to visual and audio modalities, laying the foundation for the proposed algorithm. The research strives to contribute valuable insights to the field, offering an effective approach to depression recognition through the integration of multi-modal machine learning techniques. The potential ramifications extend to more accurate mental health assessments and the development of targeted intervention strategies. This study emerges as a timely and crucial endeavor to address the pressing challenges posed by the global mental health crisis.

Keywords: Depression detection; Deep learning; Machine learning; Computer vision; Signal processing.

1. Introduction

As long as Oscillations in mood are not critical or intervene in that individuals routine and common living duties, it would not be harmful to peoples emotional lives; oppositely, it provoke psychiatric trouble such as major depressive dysfunction. A prominent mental health disorder that may continue for weeks, months, or years, fluctuates in austerity, and is connected with anxiety and weakness called Major Depressive Disorder (MDD) which plays a vital role in impairing an individuals capability to perform in everyday life. World Health Organization (WHO) has predicted that, from 2021, depression will be the fourth most substantial reason for the inability universal. Moreover, the WHO estimated that 350 million people worldwide are suffering from depression [20]. Depression not only doubles the risk of death but also leads to enormous economic losses [13]. As reported by the World Mental Health Survey Consortium, the need for mental disorder treatment is a major difficulty in both, developed and less-developed countries. Also, it reveals, the association of respondents who obtained treatment for emotional or material use problems is much larger in advanced than in less-developed countries [35], thus coping with depression has a positive effect in many situations [17]. However, it is common to misdiagnose patients with depression. Obviously, still, the dominant mental disorder treatment is clinical depression, which expresses itself in several forms. Self-medication is the only means of analysis, so it is often hard to diagnose causing misjudgment in mental inclinations. As claimed by the WHO Global Burden of Disease News, the obstacles to the practical depression analysis constituted a source shortage and prepared providers of health supervision. Besides, a Psychological clinicians evaluation varies depending on the expertise and the diagnostic techniques used in surveys such

[a] *Independent Researcher.*
[b] *Department of Computer Sciences, Faculty of Mathematical Sciences, Kharazmi University, Tehran, Iran.*
[*] *Correspondence to: A.A. Aghaei. Email: alirezaafzalaghaei@gmail.com*

as PHQ (Physical Health Questionnaire Depression Scale), the HDRS (Hamilton Depression Rating Scale), or the BDI (Beck Depression Inventory), etc. A patients answer is required in all of these applications used in screening which is usually not very reliable due to various personal matters of a person. Depression has no dedicated workroom experiments. As a result, there would be no systematic approach to recognizing depressed mood. We consider that new advancements in powerful sensing technology will probably promote an accurate evaluation. However, computerized recognition of status has been an effective examination field in the preceding decade. Techniques for depression disorder discovery are in their cradle yet. By promoting an accurate multimodal sensing method, we strive to help clinicians during the analysis and monitoring of clinical depression, which in the future, may become a truly valuable device for monitoring depression online for easing conversation between doctor and patient in the meaning of e-health foundation. patient clinical depression estimation relies massively on two domains: the clinical records (i.e., archives of displaying traits, previous chapters, relationship archives, etc.) and the mental element analysis (vision, speech, mobility, published mood, etc.). Then, in specific, we examined the analysis of audio-visual data collected through a clinical conversation with people meeting standards for depression for characteristics that would usually be evaluated for the standard mental state test. Notwithstanding, this examination of behavioral analysis is just for strengthening mental state examination. So, there is no aim for substitution. In this paper, a unique framework is introduced that invokes attention mechanisms at several layers to recognize and focus on significant characteristics from three types of models to a divine level of two datasets such as the Train set and Development set. The network utilizes lower-level and mid-level features from both audio and video proving that attention at different levels gives us the ratio of the greatness of each modality, running to more solid outcomes. We present several practices on each feature that merged several modalities and conclude that the all-feature fusion is moperior to a single-feature network in the case of RMSE (2.6) and MSE (1.91).

2. Related Works

In this section, we shortly present a review of the application region in contemporary times, which is an automatic depression analysis. A multi-featured system that combines complex channels and nods is required to produce more reliable recognition than unimodal methods. However, only less effective sensing methods use multimodal data where different features are combined, such as body mobility, facial appearance, and speech metrics, as reviewed in [30], [10]. Furthermore, the AVEC depression challenges have drawn enthusiasm for evaluating practices to predict depression intensity [39]. Yet, comparatively rare methods applied a multimodal approach as examined below. In general, deep neural network-based depression diagnostic methods target the DAIC-WOZ dataset, however, many of them target only a few applications in speech, computer vision, and natural language processing:

2.1. Deep Automatic Depression Detection

Al Hanai et al. [1] exploit the potential of LSTM (long-short-term memory) networks to model the communication process during the clinical interview. The proposed automated model learns audio and text features from sequences of questions and answers, making explicit content modeling techniques unnecessary. Yang et al. [41] for extracting audio and visual features from the clinical interviews, uses deep convolutional neural networks (DCNN). The extracted features are then fed into another multilayer perception network (MLP) to predict the PHQ-8 score. Tzirakiset al. [37] propose a deep residual network (ResNet) for extracting visual features while extracting audio features with a DCNN. To track the context of the information being processed, both sets of features are then concatenated and fed into an LSTM. A handful of works attempt to grasp more faithfully the clinician-patient communication interactions by combining video, audio, and text features. Subheadings. Williamson et al. [40] compute analytical features for speech recognition and process transcripts using word embedding and feature extraction. A Gaussian staircase model is built from these features alongside the visual features provided by the dataset. Haque et al. [14] follow a similar approach to the work of Williamson et al. [40] but compute features on a sentence level. Processed features are fed to a temporal convolutional network [4, 7], which is shown to outperform an LSTM. Both binary classification and regression tasks are investigated, but no information is provided on the classifier and regressor adopted.

2.1.1. Depression Detection by Natural Language Processing Dang et al. [8] apply verbal properties, secondary speech performance, and word effect. Points to prophesy depression sharpness and emotional descriptions inside the DAIC dataset [12]. Semantic property features include the total number of terms, single-word lexical proficiency, and pronouns during communication. The auxiliary oral response includes identical activities, such as crying or laughing, and other big information, such as word repetition and average phrase length. Word attributes are determined by a set of defined corpora that distribute n-grams conversely evaluations linking to their effective connotative. For example, Choose the type of emotion (anger, hatred, joy, etc.) to a term or number of words from 0 to 10 for emotional characteristics such as arousal, evaluation, dominance, and well-being. These three feature examples are used to predict the degree of depression.

2.1.2. Depression Detection by Speech Processing Ozdas et al. [24] provides a method for determining the likelihood of suicide based on the initial frequency of a person's voice. Alghowinem et al. [3] Performance evaluation of a wide selection of oral features, selected using the Open Smile Toolbox [11] to identify depression. Dibeklioglu et al. [9] In their research on psychomotor delays caused by depressed mood practice acoustic criteria to diagnose depression. Syed et al. [34]Use low-level

descriptors to disrupt thematic text. By indexing depressed and non-depressed members with a depressive dataset, they promote a model for predicting the severity of depression based on the level of disturbance.

2.1.3. Depression Detection by Computer Vision ang et al. [42] present an innovative facial descriptor, an HDR (Histogram of Displacement Range), which represents the number of changes in facial landmarks. The histogram calculates the number of occurrences of a displacement within a certain range of mobility. Where Syed et al. described the amount of change of specific facial features to include psychomotor hindrance, Yang et al. describe the fraction of times the face is distorted, therefore to speak, by landmarks moving a particular quantity. While Joshi et al. practice a categorization method, Dibeklioglu et al. [9] examine universal features describing facial transition dynamics for depression discovery. This approach requires providing mathematical derivations from change features such as speed, dispatch, and facial displacement over some time and then forming their influence. Alghowinem et al. [2] apply eye gaze/activity to achieve the binary classification of depression in cross-cultural datasets. They derive iris and eyelid movements for extracting blink rate, duration of closed eyes, and analytical functionals (i.e, simple derivations) of the amount of activity. Nevertheless, activity is not the only symbol, Scherer et al. [29] examined the span [-60,60] degrees for the average eye gaze vertical orientation (among other features).

2.2. Multi-Modal Hybrid Depression Detection

Song et al. [31] combine sample-level audio features with micro-expressions from facial and bodily modalities. They compare three methods: old merging, which concatenates audio features with features from each visual frame, old fusion, which uses a CCA [15] kernel, and modern merging, which equalizes per-frame forecasts from the visual modalities over the specimen and then merges them with the audio prediction. Dibeklioglu et al. [9] Using feature concatenation, igniter face, head, and oral modalities. They expand on this by using the Min-Redundancy Max-Relevance method to create feature selection on the concatenated vector rather than the source vectors [25]. Huang et al. [18] Train LSTM (long-short-term memory) models on the face, voice, and text modalities, then use SVR (Support Vector Regression) to predict the ultimate regression rates using decision-level fusion. Unlike many other deep learning approaches, this article uses decision-level fusion instead of letting deep learning models find types over feature models. Sharifa Alghowinem [2] provides composite modality fusion, which combines feature vectors from several modalities and works a majority of votes on a single modality on classification prediction. Three classifiers are used in the vote fusion: two mono-modal classifiers and one feature fusion classifier. Huang et al. [18] Train LSTM (long-short-term memory) models on the face, voice, and text modalities, then use SVR (Support Vector Regression) to build a decision-level fusion to assess the regression rates.

3. Methodology

In this section, we present the proposed structure for our research, utilizing the DAIC-WOZ database as depicted in Figure 1. Our approach involves extracting features from both the visual and audio modalities. Specifically, we extract four features from the visual modality and two features from the audio modality. Each of these features undergoes processing via Long Short-Term Memory (LSTM) networks for concatenation, facilitating the integration of temporal information.

Subsequently, the network is constructed with an attention layer for each modality. This attention mechanism enables the network to focus on the most salient characteristics within each modality, thus generating meaningful features for further analysis.

3.1. Database Information

The results of our research have been tested on the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database, which contains 189 clinical interviews and PHQ-8 questionnaire answers of depressed and patients who have no symptoms of depression. The interviews were conducted by a computer-animated avatar called Ellie, which was remotely controlled by a clinician. For each patient, the audio recording and transcripts of the interview are provided, along with 3D facial scans extracted from the video recordings. The dataset contains two main sets, Training and Development. The first one contains 107 samples, 29 out of them being depressed participants. The Development set is composed of 35 samples, 12 out of them being depressed subjects. As shown in Figures 2 and 3, the gender distribution of samples is well-balanced, preventing gender bias in deep learning algorithms. However, the number of non-depressed cases is over four times that of depressed patients, suggesting a bias in favor of non-depressed classification.

3.2. AUDIO modality

Two audio features are extracted, capturing relevant audio cues and patterns. LSTM Concatenation Similar to the visual modality, the audio features are concatenated using LSTM networks to capture temporal dynamics within the audio data.An attention layer specific to the audio modality is integrated into the model, allowing it to focus on crucial auditory features and ignore irrelevant noise or background sounds. Attention Mechanism: An attention layer specific to the audio modality is integrated into the model, allowing it to focus on crucial auditory features and ignore irrelevant noise or background sounds.

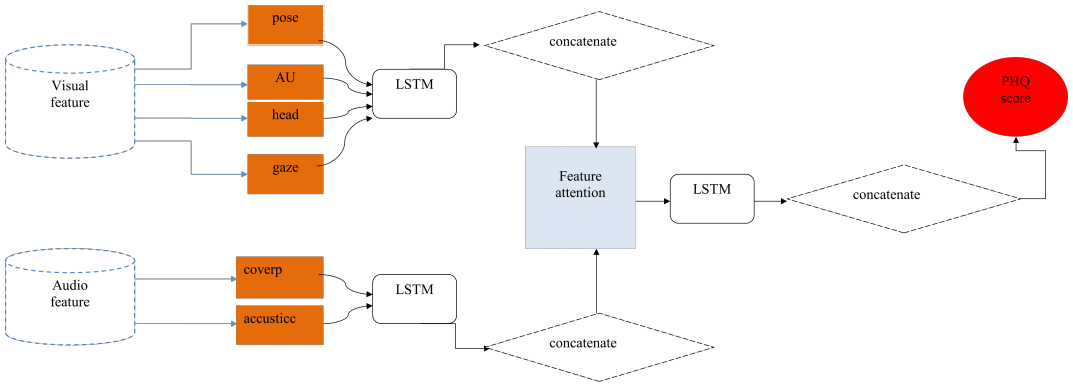
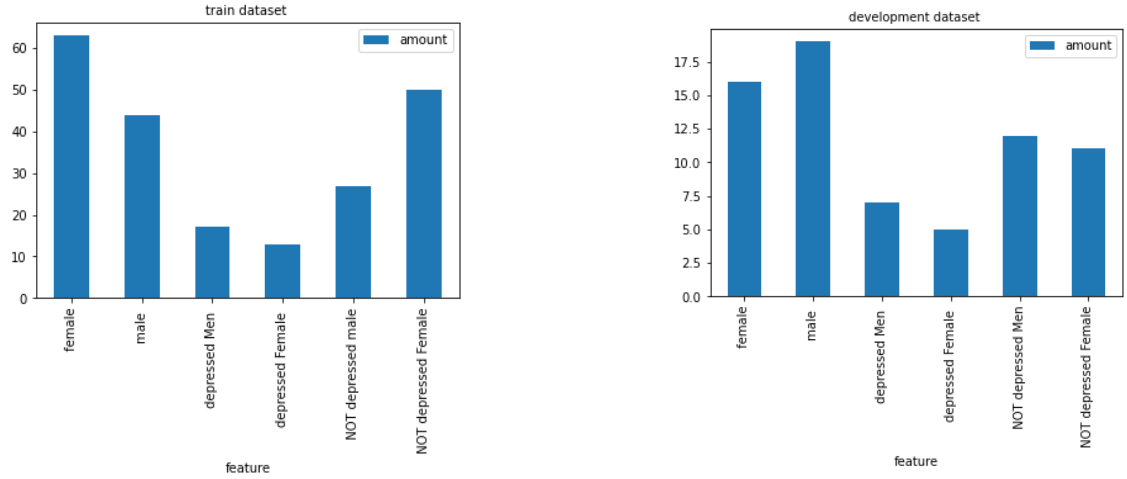


Figure 1. Overall system structure



(a) DAICWOZ distribution of train set

(b) DAICWOZ distribution of Development set

Figure 2. distribution plots of data

3.2.1. AUDIO features consideration The studies indicate response time, speech rate, and interaction involvement rate were longer, and also higher in control subjects in depressed individuals. Therefore, using audio features can be effective for our purpose. Each participant has 189 audio recordings ranging from 733 minutes (avg. 16 minutes) that are linked to his or her PHQ-8 score [18]. The database's PHQ-8 results are split between depressed and non-depressed people. We go over the process in further depth here.

3.2.2. Feature Extracted DCC at AVEC-2017 [27] establishes a baseline using traditional machine-learning approaches namely, an SVM-based (Support Vector Machine) classifier using speech characteristics obtained from COVAREP, including main frequency, formants, energy, normalized amplitude effect, and Mel-Frequency Cepstral Coefficients (MFCC) [22]. Time-series data can be found in the COVAREP and formant feature records. Each row of this data contains 74 and 5 real integers that were captured at a frequency of 100Hz. 13 features of the participants speech or the virtual interviewers sound are recorded in the dataset namely, Peak Slope, VUV, QOQ, F0, NAQ H1H2, PSP, MDQ, Rd-conf, Rd, MCEP, HMPDM. One of the features in both records is a flag labeled VUV (Voiced/Unvoiced), which indicates whether that section is sounded or not. According to the DAIC-WOZ depression dataset manual, rows with VUV flag grades of 0 should not be used. Nonetheless, analyzing log-spectrograms or Mel-scale spectrograms as a characteristic of input was necessary when dealing with deep-learning algorithms in speech-based tasks[26] [19]. We used these two types of auditory features in this experiment, and the findings were comparable to those published in [19] for the depression categorization function.

3.2.3. Formant Tracks In [6, 28], the formant tracker used in this technique is described in detail. Sections of vocal tract noises over time include knowledge regarding speech dynamics associated with articulatory features of depressed speech. A formant tracking algorithm based on Kalman filtering was used to achieve stable estimations of the first three loud beats over an interval [21]. The topic speech was retrieved at every 10 MS from the audio signal, which had not been altered other than being segmented based on the transcripts. It creates an auditory activity tracker in the tracking algorithm, allowing a Kalman to move more horizontally than non-spoken regions. The measures of the third Formant, which increased to the threshold of 4.5 kHz,

Table 1. Outcome of Development set for Audio

Partition	Features	Method	RMSE	MAE
Our approach	COVERP	LSTM	6.59	5.7
		RF	6.79	5.71
Sun [32]	FORMANT	LSTM	6.65	5.65
		RF	6.95	5.75
Baseline [41]	AUDIO FUSION	Concatenate LSTM	6.45	5.64
		Cascade RF	5.5	4.31
		RF	6.74	5.36

were shortened. The speaker's changing segments were then applied for feature processing throughout a one-second period. The Formant similarity structure characteristics for each of these pieces were computed as follows. Using time-delay embedding, a channel-delay similarity matrix was constructed from the formant recordings. The correlation matrix is based on three formant sounds and 15-time pauses per channel, with 3-frame (30 MS) delay spacing, and has a dimensionality of 45 x 45. The 45-dimensional rank-ordered Eigenspectral, which characterizes the within-channel and cross-channel distributional characteristics of the multivariate formant time set, was calculated using this matrix. These models of articulatory coordination have previously been used to measure depressive austerity. Dimensionality loss of the 45-dimensional correspondence construction feature vectors is performed as follows. Each featured part was z-scored across the Training set so that features at each Eigenvalue index were afforded equal weight, and then the principal component analysis (PCA) was used to generate the 4D feature vector. The z-scoring and PCA changes accrued in the Training set were then applied to the test set point vectors.

3.2.4. Delta MFCCs A normal set of 16 MFCCs was created from segmented but unprocessed audio files using Open Smile to begin vocal tract spectral size information [11]. The dynamic speeds of the MFCCs were calculated using delta MFCCs (dMFCCs). The delta coefficients were calculated using a 2-second pause parameter (regression across two frames earlier and later a given stage). A channel pause relationship matrix was calculated from the dMFCCs using time-delay embedding, with dimensionality (240 x 240), based on 16 dMFCC channels and 15 delays per channel with a latency interval of 1 frame, based on 16 dMFCC channels and 15 delays per channel with a latency interval of 1 frame (10 milliseconds). The 240-dimensional rank-ordered eigenspectrum, which describes the within-channel and cross-channel distributional characteristics of the multivariate dMFCC time measure, was calculated using this matrix.

3.2.5. Lower Acoustic Although often ignored in speech processing, the lower vocal tract (VT) plays a major role in speech processing. [16, 33, 36]. The lower VT holes, which are located between the glottis and the pharynx, ameliorate difficulties with the speech production system, resulting in an area of strong source-filter coupling [36]. Importantly, the lower VT plays an important role in the development of an interviewer's formants or pattern of composition, which is rarely referred to as a "resonant" voice. In this sort of composition method, the person uses a combination of methods to calculate his or her voice, including narrowing the epilarynx cavity and increasing reduction at the epilarynx and piriform apertures. This manifests itself in a spectral enhancement of the discrete lower VT vibration example [16], with the epilarynx noise (usually about 3kHz) widening and the piriform null intensifying (around 5kHz). The opposite trend has been observed in people with voice problems (e.g. hoarseness) [23].

3.2.6. Loudness Acoustic A peak-to-RMS size was assessed toward a segmental level, displaying a local loudness metric compared to waveform shape across a few tone intervals, since a whole symbol of loudness was linked to waveform patterns (with a regulation examination pane of 30ms). The global standard deviation of local (mean, sd, range) peak-to-RMS statistics were used as classification features to gather evidence that may be linked to prosodic variance over the session. For auditory frames in 2-second intervals slide with 50 percent extension, the local mean, standard deviation, and range (variation between top and bottom 5 percent rates) statistics were computed. Peak-to-RMS levels increased with time for dismal speech, potentially indicating a combination of modal and nonmodal phonation regions. While variations in the Peak-to-RMS statistic were statistically significant when linked to the PHQ score, a similar loudness characteristic was not, despite the fact that it suggested a trend toward overall more delicate speaking levels for depressed people.

3.2.7. RMSE and MAE for Audio Modality Table 1 shows the results of Trained models on each feature from each modality on the Development set, as well as the performance of audio features on the Development set based on PHQ-8 scores (the dimension of the column vectors is given between brackets). The paper [32] uses Cascade Random Forest decision-level fusion approach and [38] unimodal random forest regressor. In comparison with [32], Each individual feature network outperformed in terms of RMSE for audio, however, the Cascade Random Forest method can give a more accurate performance in the case of audio merging, as indicated in bold.

Table 2. Outcome of Development set for video modality

Partition	Features	Method	RMSE	MAE
Our approach	Head pose	LSTM	6.44	5.23
	Eye gaze	LSTM	6.56	5.45
Sun [32]	Facial landmark	LSTM	6.24	5.30
	Action unit	LSTM	6.52	5.05
		RF	6.95	5.75
Baseline [41]	Visual	Concatenate LSTM	6.58	4.28
		Cascade RF	5.9	4.6
		RF	7.13	5.88

3.3. Video Modality

Utilizing the DAIC-WOZ dataset, which comprises recordings of interviews between patients and virtual agents, we extract four visual features. These features include Head pose, Eye gaze, facial landmark and action unit, all of which convey important non-verbal cues during interactions. Each of these visual features undergoes processing via Long Short-Term Memory (LSTM) networks. LSTM networks are particularly adept at capturing temporal dependencies in sequential data, making them well-suited for analyzing the temporal dynamics of visual cues during the conversation. An attention layer is introduced into the model, allowing it to focus on the most relevant visual cues. This attention mechanism helps prioritize significant visual features, such as sudden changes in facial expressions or prolonged eye contact.

3.3.1. Video Features Consideration Darwin proved that facial emotions are universal, i.e., regardless of ethnicity or religion, the greatest emotions are exhibited in the same manner on the human face [21]. The eyes (blinking rate, pupil size fluctuation, gaze distribution), the lips (lip deformations, mouth movement), the cheeks, and the head as a whole are the main indicators of depression on the individual face (head movements, head speed). Extra facial signs connected with depression include a tight face, pale skin, and twitching eyelids. [28]. In the following sections, we discuss how the publicly accessible baseline feature sets for the video data are generated. Participants can use these features set exclusively or in addition to their features. For ethical purposes, the raw video is unavailable

3.3.2. Features Extracted the DAICWOZ dataset provides video for the screening interviews under the conditional local neural fields (CLNF) representation [12]. These embeddings have been computed based on the OpenFace [5] framework. which consists of four types of features:

- facial landmarks: 2D and 3D coordinates of 68 landmarks points on the face comprised with time stamp, confidence, detection success flag, X, Y, and Z estimated from video.
- Gaze orientation: Time mark, confidence, detection success flag, x0, y0, z0, x1, y1, z1, xh0, yh0, zh0, xh1, yh1, and zh1 are all included in the gaze orientation evaluations for both eyes. Four vectors represent the gaze. The gaze orientation of both eyes is shown by the first two vectors (x0, y0, z0, and x1, y1, z1). The gaze in the head correspondent space is represented by the following two vectors (xh0, yh0, zh0, and xh1, yh1, zh1).
- head pose: the signals collected in this category include head rotation in position and direction, as well as the time stamp, confidence, detection success flag, Rx, Ry, Rz, Tx, Ty, and Tz appear on every line of the head position column. The head turn coordinates (measured in radians) are Rx, Ry, and Rz, while the head position coordinates are Tx, Ty, and Tz.
- AUs: The timestamp, confidence, detection success flag, and a few actual numbers representing the facial action term are all included in each series of the action unit's data, which is collected at a frequency of 30Hz.

The principal components of CLNF embeddings are that (i) they have been devised to detect facial landmarks and tracking, and as a result, can be adopted from facial expression analysis, and; (ii) they conceal the identity of the individuals. Together, these characteristics motivate the adoption of CLNF by the curators of DAICWOZ.

3.3.3. RMSE and MAE for Video Modality Table 2 compares our depression detection performances with three visual-based works reported in [32, 38] and shows the outcomes of every single feature from video modality on the Development set. As you can see, among all the visual-based studies, our approach has the lowest MAE/RMSE.

3.4. Hybrid Modalities

The proposed Hybrid Modality Attention Fusion Network (AFN) integrates features from different modalities, such as visual and audio, using attention mechanisms to focus on relevant information from each modality. For each modality, specific feature extraction layers capture modality-specific information. For example, convolutional neural networks (CNNs) extract visual features, while spectrogram-based techniques are used for audio.

Table 3. Hybrid modality performance on Train and Development set

Partition	dataset	Hybrid modality	method	RMSE	MAE
Our approach	DEV SET	AUDIO and VIDEO	Attention Fusion Network	5,2	3.89
			RF	7.05	5.66
Baseline [41]	TRAIN SET	AUDIO and VIDEO	Attention Fusion Network	2.6	1.91
			RF	-	-

After feature extraction, modality-specific features pass through Long Short-Term Memory (LSTM) layers to capture temporal dependencies within the data. Attention mechanisms are applied independently to each modality's LSTM outputs, enhancing integration by focusing on relevant information.

The attention-weighted features from each modality are combined using fusion techniques like concatenation or element-wise addition to create a unified representation of the input data.

A cross-modal attention mechanism further refines integrated features by attending to informative aspects across all modalities.

Finally, the integrated representation is passed through additional layers for tasks like classification or regression, transforming the fused features into the desired output format, such as the PHQ-8 score.

3.4.1. Hybrid Modalities consideration The PHQ-8 score has been used as a description of the depression severity. The distribution of depression by Train and Development set is seen in Figures 2 and 3. Since the data is significantly biased, it leads to the reporting of imbalanced data classes, which has a strong influence on machine algorithm performance. We conduct RMSE and MAE to obtain considerable performance based on attention fusion networks on both Train and Development sets. The network applied for the multimodal prediction of PHQ-8 scores for determining the connection between the features.

3.4.2. RMSE and MAE based on Multimodality Table 3 shows the performance for both the root mean square error (RMSE) and mean absolute error (MAE) for Development and Train sets using audio and video (multimodality). The baseline paper results in cascading decision-level fusion characteristics in the baseline. The results of the state-of-the-art study on the DAIC-WOZ Training dataset are presented in the second row. In addition, we show how we compare to them. Our model, which combines all characteristics with an attention fusion network, outperforms the competition, particularly on the Train set. We found that the result of the Development set based on our approach is 5.2(RMSE) and 3.89(MAE). In comparison to [38] with 7.05(RMSE) and 5.66 (MAE), our method is inferior. We also present our achievement on the Train set that can dramatically enhance achievement by 2.6 (RMSE) and 1.91(MAE). It can be uncertain because the dataset being used in the paper may be a little different.

4. Conclusions

In this work, we examine the DAICWOZ and apply the LSTM technique to each modality, focusing on the most essential ones, such as audio and video, which are both equally valuable sources of information and might be vital for severity prediction. In this test, we discovered that our video modality technique performed better than two other methods [41, 40]. also, we had better outperform every single audio modality. The use of hybrid modality led us to obtain significantly better improvement on both Train and Development sets compared with baseline [41]. we apply an attention fusion network for audio and video modality. Which causes uncomplicated the networks overall computational complexity and reduces the training and test time. We achieve a 39 percent improvement on the Development set. In comparison to the Development set. experimental results indicate train set can effectively reduce RMSE and MAE rates. For future works, we plan to investigate audio and visual dyadic behaviors based on gender distributions.

References

1. T. Al Hanai, M. M. Ghassemi, and J. R. Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.
2. S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear. Cross-cultural detection of depression from nonverbal behaviour. In *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.
3. S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. F. Cohn. Cross-cultural depression recognition from vocal biomarkers. In *Interspeech*, pages 1943–1947, 2016.
4. S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
5. T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

6. B. Bozkurt, T. Dutoit, B. Doval, and C. d'Alessandro. Improved differential phase spectrum processing for formant tracking. In *Eighth International Conference on Spoken Language Processing*, 2004.
7. H. Dana Mazraeh, M. Kalantari, S. H. Tabasi, A. Afzal Aghaei, Z. Kalantari, and F. Fahimi. Solving Fredholm integral equations of the second kind using an improved cuckoo optimization algorithm. *Global Analysis and Discrete Mathematics*, 7(1):33–52, 2022.
8. T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017. In *Proceedings of the 7th Annual Workshop on audio/visual Emotion Challenge*, pages 27–35, 2017.
9. H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 307–310, 2015.
10. S. DMello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User-Adapted Interaction*, 20(2):147187, 2010.
11. F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.
12. J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik, 2014.
13. S. B. Guze and E. Robins. Suicide and primary affective disorders. *The British Journal of Psychiatry*, 117(539):437–438, 1970.
14. A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei. Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592*, 2018.
15. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
16. K. Honda, T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, S. Takano, Y. Nota, H. Hirata, I. Fujimoto, Y. Shimada, et al. Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling. *Computer methods in biomechanics and biomedical engineering*, 13(4):443–453, 2010.
17. L. G. Kiloh, G. Andrews, and M. Neilson. The long-term outcome of depressive illness. *British J. Psychiatry*, 153(6):752757, 1988.
18. K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.
19. X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42, 2016.
20. C. Mathers, J. Boerma, and D. Fat. Global burden of disease. *Geneva, Switzerland*, 2008.
21. D. D. Mehta, D. Rudoy, and P. J. Wolfe. Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *The Journal of the Acoustical Society of America*, 132(3):1732–1746, 2012.
22. M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 43–50, 2016.
23. T. Nawka, L. C. Anders, M. Cebulla, and D. Zurakowski. The speaker's formant in male voices. *Journal of Voice*, 11(4):422–428, 1997.
24. A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes. Analysis of fundamental frequency for near term suicidal risk assessment. In *In IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 1853–1858. IEEE, 2000.
25. H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
26. K. J. Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015.
27. F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 3–9, 2017.
28. F. Rooholamini, A. Afzal Aghaei, S. M. H. Hasheminejad, R. Azmi, and S. Soltani. Developing chimp optimization algorithm for function estimation tasks. *Computational Mathematics and Computer Modeling with Applications (CMCMA)*, pages 34–44, 2023.
29. S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency, et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, 2014.
30. N. Sebe, I. Cohen, and T. S. Huang. Multimodal emotion recognition. *Handbook Pattern Recognit. Comput*, 4:387419, 2005.
31. Y. Song, L.-P. Morency, and R. Davis. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 237–244, 2013.
32. B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 61–68, 2017.
33. J. Sundberg. Articulatory interpretation of the singing formant. *The Journal of the Acoustical Society of America*, 55(4):838–844, 1974.
34. Z. S. Syed, K. Sidorov, and D. Marshall. Depression severity prediction based on biomarkers of psychomotor retardation. In *Proceedings of the 7th Annual Workshop on audio/visual Emotion Challenge*, pages 37–43, 2017.
35. The WHO World Mental Health Survey Consortium. WHO World Mental Health Survey Consortium Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *JAMA*, 291(21):25812590, 2004.
36. I. R. Titze and B. H. Story. Acoustic interactions of the voice source with the lower vocal tract. *The Journal of the Acoustical Society of America*, 101(4):2234–2243, 1997.
37. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017.
38. M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

39. M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge*, pages 3–10, 2013.
40. J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on audio/vsual Emotion Challenge*, pages 11–18, 2016.
41. L. Yang, D. Jiang, W. Han, and H. Sahli. DCNN and DNN based multi-modal depression recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 484–489. IEEE, 2017.
42. L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 53–59, 2017.

