MultiModal Depression Detection

Anonymous ACL submission

Abstract

The detection of depression through non-verbal cues has gained significant attention. Previous research predominantly centered on identifying depression within the confines of controlled laboratory environments, often with the supervision of psychologists or counselors. Unfortunately, datasets generated in such controlled settings may struggle to account for individual's behaviors in real-life situations. In response to this limitation, we present the Extended D-vlog dataset, encompassing a collection of $1,261$ YouTube vlogs. We extracted features across the auditory, textual, and visual Modalities from this Extended D-vlog dataset. To effectively capture the interrelationship between these features and derive a multimodal representation involving audio, video, and text, we harnessed the TVLT model. Remarkably, the utilization of the TVLT model, in conjunction with video, text, and audio (leveraging wav2vec2 features and spectrograms), produced the most promising results, achieving a remarkable **F1-score of** 67.8**%**.

1 Introduction

Depression, is a prevalent and significant medical condition. It has a harmful impact on one's emotional state, thought processes, and behavior. It manifests as persistent feelings of sadness and a diminished interest in previously enjoyed activities. This condition can give rise to various emotional and physical challenges, affecting one's ability to perform effectively both at work and in personal life. Depression symptoms range from mild to severe and can include persistent sadness, loss of interest in once-enjoyable activities, appetite changes, sleep disturbances, fatigue, psychomotor changes, feelings of worthlessness, cognitive challenges, and, in severe cases, suicidal Thoughts. Symptom severity varies, requiring careful clinical evaluation for diagnosis and treatment (Cleveland Clinic).

According to the Statistics of The World Health Organisation (WHO) (World Health Organization) 3.8% of the world's population experience depression, including 5% of adults less than 60 years of age (4% of men and 6% of women) and 5.7% of Adults above 60 years of age. Approximately 280 million people have depression in which depression is 50% more common in women than men. Depression is 10% more in pregnant women and women who have just given birth (Evans-Lacko et al., 2018). If the depression is left untreated can lead to several serious outcomes such as suicide (Ghosh et al., 2022).[1]

The process of clinically diagnosing depression relies on interviews that incorporate **PHQ-8** (Kroenke et al., 2009) [2] Questionnaires which include questions such as *"Do you experience feelings of sadness, depression, or hopelessness?" and inquire about the duration of these feelings, ranging from 0-1 day to 1-3 days and so forth"*, and using **BDI-II Questionnaires** (Smarr and Keefer, 2011) [3] having questions such as *How often do you have Guilty feelings? with four options as 1) dont feel particularly guilty 2) I feel guilty over many things i have or should have done 3) I feel quilty most of the times 4) I feel guilty all the times.* However, these questionnaires has a limitations, as patients may exhibit hesitance expressing their genuine emotions during these interviews, potentially resulting in an inaccurate assessment of their depression (Yoon et al., 2022). In contrast, the evolution of social media platforms like Twitter, Reddit, and YouTube has provided users with a medium to openly share their thoughts, perspectives, and current life situations. These online platforms contain valuable and diverse emotional information.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

[2] https://www.childrenshospital.org/sites/default/files/2022-03/PHQ-8.pdf

[3] https://naviauxlab.ucsd.edu/wp-content/uploads/2020/09/BDI21.pdf

Majority of methods utilized for Depression detection predominantly focus on analyzing textual information gathered from social media to infer users' emotional states. However, (Yang et al., 2017a), (Yang et al., 2017b), (Gong and Poellabauer, 2017), (Ray et al., 2019) demonstrates that adding modalities, such as utilization of videos to extract facial expressions, body gestures, as well as incorporating audio analysis to detect fluctuations in speech patterns, has the potential to enhance the precision of depression identification.

We utilize the power of **Vision-Language TVLT** (Tang et al., 2022) **Transformer model**, which has demonstrated its power by achieving state-of-the-art performance on various tasks such as video-captioning, Multimodal sentiment analysis and Multimodal emotion recognition. TVLT (Tang et al., 2022) serves as a Encoder-Decoder Modal which take raw video, raw audio and text as input and produce a comprehensive Multimodal representation that can be leveraged for downstream task related to detection and classification. Incorporating additional wav2vec2 (Baevski et al., 2020) features with spectrograms for audio along with video and text, we achieved an impressive accuracy of 67.8% on the Extended D-vlog dataset (Yoon et al., 2022).

Our Contribution are:

- Extended D-vlog dataset (**Original no. of videos: 961**, **Total videos** (after adding 300 videos to the Original dataset): **1261**) which contains videos of various type such as Major depressive disorder, postmortem disorder, anxiety and videos from different age group and gender which was lacking in the original D-vlog dataset.
- TVLT (Tang et al., 2022) model for depression detection, which outperforms baseline models by **4.3%** and establishes a new benchmark, on Extended D-vlog dataset.
- Replacing spectrogram with combination of spectrogram and wav2vec2 (Baevski et al., 2020) features which captures the vocal cues associated with depression more effectively than spectrogram, which further increases the accuracy by **2.2%** resulting in the final F1-score of **67.8** %.

2 Related Work

With the increase in mental health conditions these days, there is an increase in the popularity of the detection of depression. However, very little work is done in creating the dataset to detect depression. Due to privacy concerns, most datasets are only used for their research and are not publicly available. Among the relatively scarce publicly available datasets suitable for analysis, the DAIC-WOZ (Gratch et al., 2014) is one of most famous and used dataset. This dataset encompasses clinical interviews in various formats, including verbal (text) and non-verbal (audio and video). Notably, The DAIC-WOZ (Gratch et al., 2014) dataset relies on self-reporting through the PHQ-8 questionnaire. Another well-known dataset is the Pittsburgh dataset (Keenan et al., 2010), which comprises clinical interviews primarily in audio and video formats. However, this DAIC-WOZ (Gratch et al., 2014) dataset is relatively small, containing only 189 samples, making it a valuable resource for research purposes. The AViD-Corpus (Audio-Video Depressive Language Corpus) is another prominent dataset, with subsets used in AVEC 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) competitions. This dataset includes video recordings of participants engaging in various activities, such as singing and delivering speeches. Notably, the self-reporting in this dataset is conducted in the presence of mental health professionals, including psychiatrists, psychologists, and experienced counselors. The questionnaires used in AViD-Corpus gauge the severity of patients' symptoms over specific periods, with responses recorded on a scale of 0 (not at all), 1 (several days), 2 (more than half the days), and 3 (nearly every day). The overall score is derived by summing the responses. These datasets have been instrumental in gaining insights into depression patterns. Nevertheless, as they are predominantly assembled and curated within controlled laboratory environments, they may not fully encapsulate the typical behaviors exhibited by individuals experiencing depression.

dataset	Modality	# Subjects	# Samples
DAIC-WOZ	A+V+T	189	189
Pittsburg	A+V	49	130
AViD-Corpus	A+V	292	340
D-vlog	A+V	816	961
E-Dvlog	A+V+T	1016	1261

Table 1: Comparision of various Depression datasets with E-Dvlog (Extended D-vlog). Where A: Audio, V: Video, T: Text.

The utilization of social media for depression detection has been increasingly used instead of clinical interviews. Social media datasets can reveal the patient's unusual and atypical behavior, which cannot be seen in the clinical interview conducted under supervision, where the individual may not authentically express their emotion or actual behavior, which is shown in their daily lives (Yoon et al., 2022). Therefore, many approaches have been taken to detect Depression using data from social media sites such as Twitter, Reddit and Facebook. In recent years, depression detection using text from social media has been focused on (Fatima et al., 2019), (Burdisso et al., 2019), (Chiong et al., 2021). Textual-based features focus on the linguistic features of the social media text, such as words, POS, n-gram, and other linguistic characteristics. In (Wang et al., 2013) uses text and tags from micro-blog (Sinba Weibo) used in China. They extracted content behavioral features from the blogs to detect Depression. In comparison, this method focuses on detecting Depression from Social media using text. However, more attention should be paid to video data and multimodal Fusion.

Multimodal Fusion is to combine multiple modalities to predict output. Work is done to detect Depression using multiple fusion. (Haque et al., 2018) in this paper the Authors have uses 3D Facial Expressions and spoken language as features from the dataset to detect Depression. (Yang et al., 2018) use text and video features and hybridizes deep and shallow models for depression estimation and classification from audio, video and text descriptors. (Ortega et al., 2019) proposed an end-to-end deep neural network (DNN) model that integrated three different modalities of speech features, facial features, and text transcription. Each modality is first encoded independently with fully connected layers and then combined into a single representation for estimating the emotional state of subject. To effectively utilize the multimodal fusion data we used Multimodal Transformer to generate multimodal representation for Depression detection.

3 Datasets

The D-vlog dataset (Yoon et al., 2022) is a collection of Depression vlogs of various people posted on YouTube. The D-vlog dataset has around 961 vlogs in total out of which 505 are categorized as depressive vlogs and 465 are categorized as Non-depressive vlogs. However, the D-vlog dataset (Yoon et al., 2022) has some limitations, such as the dataset majorly having Major Depressive Disorder and lacking Other Disorder such as Bipolar Disorder, Postmortem Disorder, and Anxiety with depression. Which will make the dataset more generalized. So, we extended the D-vlog dataset by adding around 300 more vlogs to the D-vlog dataset (Yoon et al., 2022) which now have more vlogs on various depressive disorder from varying age groups and different gender. Figure 1.

3.1 Dataset Collection:

We have collected the dataset vlogs using certain keywords using YouTube API (Yin and Brown, 2018) and downloaded them using the yt-dlp package [4].

Depressive vlogs: 'depression daily vlog', 'depression journey', 'depression vlog', 'depression episode vlog', 'depression video diary', 'my depression diary', and 'my depression story', 'postpartum depression vlogs', 'Anxiety vlogs'.

Non-Depressive vlogs: 'daily vlog', 'grwm (get ready with me) vlog', 'haul vlog', 'how to vlog', 'day of vlog', 'talking vlog', and etc.

We used the same approch to collect the dataset as used in the D-vlog dataset (Yoon et al., 2022). We focused our analysis on vlogs featuring content creators who have a documented history of depression, currently manifesting symptoms of the condition. We specifically excluded vlogs that solely discussed having a bad day without a deeper connection to depressive experiences.

3.2 Dataset Statistics:

The Extended D-vlog dataset has 1261 vlogs with 680 Depressive vlogs and 590 Non-Depressive vlogs as can be seen from the below table.

	Gender	**# Samples**
Depression	Male	273
	Female	406
Non-Depression	Male	232
	Female	350

Table 3: Extended D-vlog Statistics

Extended D-vlog dataset exhibits more representation of Female vlogs as compared to Male vlogs within Depressed catergory, reflecting high

[4] https://github.com/yt-dlp/yt-dlp/wiki/Installation

Self-Reported Questionnaires	# Questions	Content of Questionnaires
PHQ-8	9	sleeping difficulties, excessive guilt, fatigue
Beck Depression Inventory(BDI-II)	21	Mood, self-hate, social withdrawal, fatigability

Table 2: Self-Reported Questionnaires for Depression Assessment.

prevelence of depression among Female. In Non-depressive category similar trend is observe with more female representation than Male vlogs as predominantly "get ready with me vlogs", "Haul vlogs" are uploaded by Females. In Extended D-vlog we added vlogs which have other disorder other than Major Depressive Disorder such as Anxiety disorder, Bipolar Disorder [5],and Postmortem Disorder. The below Figure 1 show the Distribution of various type Depressive vlogs.

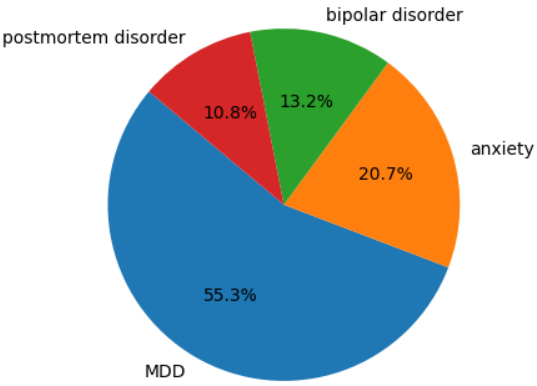


Figure 1: The Above figure shows the distribution of various type Depressive vlogs. where MDD is Major Depressive Disorder, Bipolar Disorder also called as Manic Disorder.

4 Methodology

We used **TVLT (Textless Vision Language Transformer)** (Tang et al., 2022), a minimal end-to-end vision and language Multimodal transformer model that takes raw video, raw audio, and text as input to the transformer model. TVLT (Tang et al., 2022) is a Textless Model, which implicitly does not use text, but with the ASR model (Whisper) (Radford et al., 2023), we can extract text from the audio segments. The TVLT model is more effective for Multimodal classification because the TVLT (Tang et al., 2022) model can capture visual and acoustic information, providing a more comprehensive fused representation of video, audio, and text.

For **Textual Feature**, we make use of the powerful BERT (Kenton and Toutanova, 2019) Language model, a pre-trained model described in to capture important features from text. This means we can understand not only the specific details in the text but also the overall context. These BERT embeddings help us understand text thoroughly, making them perfect for tasks like analyzing sentiment or identifying depression. We apply BERT (Kenton and Toutanova, 2019) to our text, using specific dimensions (dt = 786), and we start with good initial weights using Xavier's method (Kumar, 2017) . This process empowers our model to create detailed and context-rich text representations. These representations form a strong foundation for tasks that rely on textual information.

To extract **Audio features**, we employ a combination of techniques. First, we utilize low-level features like spectrograms, which are generated using the librosa Library (McFee et al., 2015). Additionally, we incorporate features from wav2vec2, as described in (Baevski et al., 2020). These wav2vec2 (Baevski et al., 2020) features encompass various acoustic attributes, including MFCC (Hossan et al., 2010), Spectral (Pachet and Roy, 2007), Temporal (Krishnamoorthy and Prasanna, 2011), and Prosody (Olwal and Feiner, 2005) features which help in identifying pitch, Intonation, Tempo of the audio segment. They excel in capturing both local and contextual information from the raw audio waveform. To create our final audio representation, we compute the average across the spectrogram vector and the wav2vec2 vector. This fusion of spectrogram and wav2vec2 (Baevski et al., 2020) features significantly enhances the model's ability to discern vocal cues, as evidenced in the results section.

Our video processing pipeline involves several essential steps. First, we load the video file using a tool called VideoReader (Frith et al., 2005). Next, we randomly select a subset of frames from the video clip. These frames are then resized and cropped to focus on the subject's frontal view. For extracting **visual features**, we rely on the powerful ViT (Vision Transformer) model introduced

[5] https://www.nimh.nih.gov/health/topics/bipolar-disorder

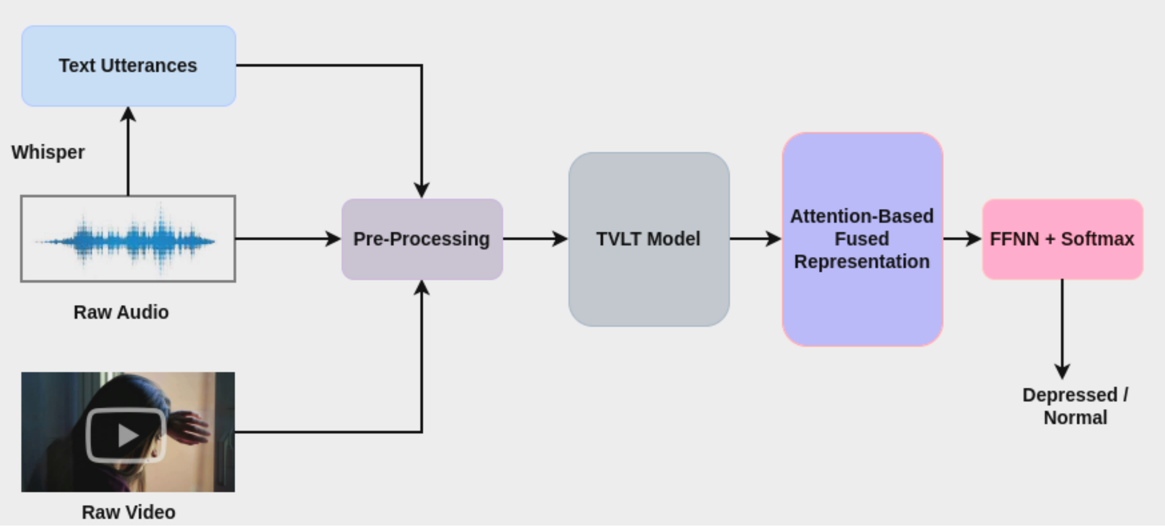


Figure 2: In the Above **Architecture** we leverage three different modalities such as video, audio and text where text is extracted from the audio segment using Whisper ASR Model. we then preprocess all the three modalities and pass to the model where we get the fused representation of all three modalities. These fused representation is then passed to the feed forward Neural Network with sigmoid function to get whether the individual exibits the sign of Depression or is it in Normal state.

in (Dosovitskiy et al., 2020). This model helps us create what we call "vision embeddings." It does this by breaking down each video frame into smaller 16x16 patches. We then apply a linear projection layer to these patches, resulting in a 768-dimensional patch embedding. This vision embedding module is a critical component of our model. It takes each video frame or image and transforms it into a sequence of 768-dimensional vectors. These vectors are rich in both spatial and temporal information, making them invaluable for our model to comprehend the visual content within the input data.

We have implemented the architecture illustrated in Figure 1, where our TVLT (Tang et al., 2022) transformer model comprises a 12-layer encoder and an 8-layer decoder. To obtain the fused representation of all three modalities, we exclusively utilize the encoder portion of the model. These fused representations are subsequently fed into the downstream task for depression prediction. Our model's evaluation is conducted on the Extended D-vlog dataset, which consists of 35,046 video clips collected from 1016 different speakers. For each video clip, we generate text using the ASR model and manually correct any errors to ensure ground-truth transcriptions. In line with previous studies, we employ a 7:1:2 train-valid-test split and evaluate using weighted accuracy (WA) and F1 score metrics. For each downstream task, we introduce a task-specific head (a two-layer MLP) on top of the encoder representation. We train the model jointly using binary cross-entropy loss for these tasks.

$$L(y, \hat{y}) = -\left[y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right] \quad (1)$$

where y: True label and $\hat{y}$: Predicted label

5 Experiments

To obtained fused representation of audio, video and text modalities, we employ pretrained text-based TVLT (Tang et al., 2022) model on video dataset and subsequently fine-tune on the Extended D-vlog dataset.

5.1 Pretrained dataset

- **HowTo100M:** We used HowTo100M, a dataset containing 136M video clips of a total of 134,472 hours from 1.22M YouTube videos to pretrain our model. Our vanilla TVLT is pretrained directly using the frame and audio stream of the video clips. Our text-based TVLT is trained using the frame and caption stream of the video. The captions are automatically generated ASR provided in the dataset. We used 0.92M videos for pretraining, as some links to the videos were invalid to download.

- **YTTemporal180M:** YTTemporal180M includes 180M video segments from 6M

> YouTube videos that spans multiple domains, and topics, including instructional videos from HowTo100M, lifestyle vlogs of everyday events from the VLOG dataset, and YouTube's auto-suggested videos for popular topics like 'science' or 'home improvement'.

We split our dataset into Train, Valid and Test Set in the ratio of 7:1:2. These three set do not overlap i.e no dataset is use in more than one set.

Gender	Train	Valid	Test
Male	354	51	100
Female	530	74	152

Table 4: Number of vlogs in Train, Valid and Test Split of Extended D-vlog dataset

For training the model we have use adam's Optimizer, learning rate of [0.0001, 0.00001], batch size [32, 64]. we trained 4 iterations of the model with different seed value. we trained our model on Nvidia RTX A6000 where each iterations take nearly 3 hours to train. we use binary cross entropy as our loss function for Depression detection or classification task. we reported the F1-scores of the test set in the result section.

6 Result and Discussion

Modalities	F1-scores
T	0.57
A	0.60
V	0.56
V + A	0.631
V + T	0.628
A + T	0.634
V + A + T	**0.656**

Table 5: Results obtained on the Extended D-vlogs dataset via experiments with Video (V), Audio (A), Textual (T).

To analyse the importance of each modality for depression detection, we trained our model on each modality separately and reported the results in the Table 5 above. We found that audio modality have best F1-Score than the other modalities, implies that audio features are more importance than the visual and textual features for Depression detection. This suggest that people with depression have distinct speech features. Even though the audio features are more important than the visual features but when we combine the two modalities we can see that combining the two modalities yield better score than just using audio modality. Also, combination of audio and text modality prove to be better than just using audio as a single modality. Finally, combining all the three modalities we get much better result on relying on just two modalities which tells that combining audio features, visual features and textual features and their relationship is more effective for the Depression detection.

Modalities	F1-scores
V + A + T	0.656
V + A + T(Mask)	0.663
V + A(W2V2+Spect) + T	**0.678**
V(Mask) + A + T	0.661

Table 6: Results obtained on the Extended D-vlogs dataset via experiments with Video (V), Audio (A), Textual (T). T(Mask) is text with word-masking, V(Mask) is Video frames with frame-masking and A(W2V2+Spect) is Audio with wav2vec2 +spectrogram features.

The introduction of random word masking in text modalities proves instrumental in enhancing the model's understanding of textual information. This improvement becomes apparent when analyzing the results, where a subtle yet noteworthy performance boost of 0.007**%** is observed in comparison to not employing word masking in the text. Moreover, the application of frame masking to video data, when combined with audio and text modalities, also contributes to a slight enhancement of 0.005**%**. These findings underscore the efficacy of incorporating diverse modalities in the model. Table 7 provides a comprehensive overview of the results. It unmistakably highlights the significance of leveraging all three modalities—text, video, and audio—in conjunction with wav2vec2 (Baevski et al., 2020) features and spectrograms, as opposed to using spectrograms for audio processing. This approach leads to an impressive F1-score of 67.8**%.**

We have extensively evaluate the TVLT (Tang et al., 2022) model performance on the D-vlog dataset (Yoon et al., 2022) and compared its results with several baseline models. The purpose was to check the effectiveness of the TVLT (Tang et al., 2022) model in evaluating task on Depression detection on D-vlog dataset. The TVLT (Tang et al., 2022) model in isolation was when performed on D-vlog dataset (Yoon et al., 2022), surpasses the Cross

Model Type	Model	Precision	Recall	F1-Score
Fusion Baseline	Concat	62.51	63.21	61.1
	Add	59.11	60.38	58.1
	Multiply	63.48	64.15	63.09
Depression Detector	Cross-Attention	65.4	65.5	63.5
Our Model	TVLT Model	**67.3**	64	**65.6**

Table 7: Comparison of various baseline models with our model on the D-vlog dataset

Attention State-of-the-Art model by 2.2 %. This improvement established the TVLT (Tang et al., 2022) model as the New Benchmark for the D-vlog dataset (Yoon et al., 2022). The result Obtained by TVLT model on D-vlog dataset (Yoon et al., 2022) states that the TVLT (Tang et al., 2022) model was correctly able to understand the characteristics and handling the complexity of the D-vlog dataset (Yoon et al., 2022). The exceptional performance of the TVLT (Tang et al., 2022) model could serve as a catalyst, motivating researchers to explore and develop more advanced techniques in the realm of multi-modal analysis and deep learning.

7 Qualitative Analysis

In this section, we demonstrate how the integration of wav2vec2 (Baevski et al., 2020) features significantly enhances our TVLT (Tang et al., 2022) model's ability to accurately detect depression. By gaining a deeper understanding of vocal cues embedded in the audio data, our model's performance is notably improved. In contrast to the TVLT (Tang et al., 2022) model relying solely on spectrogram data, the inclusion of wav2vec2 (Baevski et al., 2020) features equips our model to make predictions that would be challenging otherwise. The provided table comprises carefully chosen examples where our model correctly identifies depression. In the first instance, the girl's facial expressions exhibit relative consistency i.e their not much change in her facial expression. However, analysis of her audio contains monotone tone, low pitch and filled with cries also the textual utterance is the clear example that she is distress. The TVLT (Tang et al., 2022) model, enhanced with wav2vec2 (Baevski et al., 2020) and spectrogram features, accurately predicts this case, whereas the model using only spectrogram data falters. This underscores the pivotal role of wav2vec2 (Baevski et al., 2020) features in depression detection this is due to the fact that wav2vec2 is able to understand the vocal cues that audio features with spectrogram is failing to understand. In the second example, we observe a girl who simultaneously displays both a smile and tears. The audio content provides a clear indication of her depression, supported by textual information and visual information. However, the model without wav2vec2 (Baevski et al., 2020) features fails to provide an accurate prediction, while the model with wav2vec2 (Baevski et al., 2020) correctly identifies the depressive nature of this example. This highlights the capability of wav2vec2 (Baevski et al., 2020) features to extract crucial information from audio segments, enriching the model's understanding of depression cues. In the third example, the woman displays minimal to no variation in her facial expressions, and her audio content appears unremarkable, lacking any discernible markers typically associated with depression. However, upon analyzing the textual content, we can identify indications of depression, which our model struggles to predict accurately.

8 Summary, Conclusion & Future Work

In this study, we introduced an Extended D-vlog dataset, comprising 1261 videos that include both vlogs by individuals with depression (680 videos) and those without (590 videos). Our objective is to detect depression in non-verbal and non-clinical vlogs. To achieve this, we employed a TVLT model, which is a multimodal transformer (Tang et al., 2022), to create a multimodal representation using text, video, and audio data. We utilized the Vit model for visual embeddings and extracted audio features with wav2vec2 (Baevski et al., 2020) and spectrograms. The TVLT model, incorporating all three modalities, yielded a noteworthy F1-score of 0.656. By introducing text word-masking, we improved the F1-score to 0.663, representing a 0.007% enhancement over the absence of word-

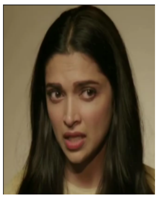
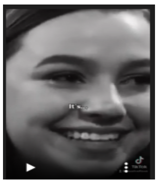
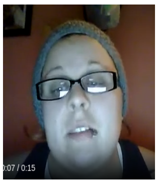
Utterance	Ground Truth	Prediction (w2v2 + spect)	Prediction w/o (w2v2 + spect)	Video frames
I knew what I was feeling, but I don't think I was able to communicate entirely what I was feeling. Like I knew I had this pittish feeling in my stomach. I knew that I'd be scared to wake up. I didn't want to wake up. Yeah, I think waking up was tough because I didn't want to face a day.	Depression	Depression	Normal	
Some days, it's really, really hard to just move. It's... I like it. I, yeah, it's hard to get out of bed. It's hard to even go downstairs to get something to eat.	Depression	Depression	Normal	
No concept of time, no sense of feeling. Have I become cold, dead to the world, where I once mattered? I can't even remember when I was important to someone last. Everything has escaped me. Deeper I fall into a void.	Depression	Normal	Normal	

Table 8: A **Qualitative analysis**, In the given instances, the model, equipped with both wav2vec2 and spectrogram features, effectively detects depression through audio analysis. In the first example, despite seemingly normal facial expressions, the model accurately detects depression. In the second case, the model succeeds in identifying depression even when the individual smiles while crying, whereas the model relying solely on spectrogram data falls short in these situations. In the third scenario, the woman's facial expressions and audio do not exhibit evident signs of depression, while text analysis reveals potential indicators that challenge our model's accuracy, resulting in an incorrect prediction.

level masking. Additionally, with frame-masking, the F1-score reached 0.661.

Our TVLT model, combined with the supplementary wav2vec2 and spectrogram features, outperformed all baseline models on the D-vlog dataset (Yoon et al., 2022) and established a new benchmark on the Extended D-vlog dataset. We believe that our introduced dataset and the multi-modal depression detection model have the potential to play a significant role in the early identification of individuals experiencing depression through their social media presence. This proactive approach aims to ensure timely access to essential clinical interventions for those in need.

In future work, we plan to extend the scope of our research to detect various type of mental health diseases. Additionally, we aim to investigate and incorporate multilingual capabilities into our work frame so that it can be scaled and utilized for various language settings.

9 Limitation:

Our model faces some challenges where it cannot predict some of the depressed classes, such as the case of smiling depression, where the individual conceals their genuine emotions by presenting as cheerful and high-functioning by masking their feelings, making detecting this problem difficult to detect.

10 Ethics Statement:

All the data in the dataset has been sourced from open-access platforms, and none of the videos or text within it contain any offensive or derogatory language aimed at any particular team or entity.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Raymond Chiong, Gregorious Satia Budhi, and Sandeep Dhakal. 2021. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems*, 36(6):99–105.

Cleveland Clinic. Depression.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, and et al. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48(9):1560–1571.

Iram Fatima, Burhan Ud Din Abbasi, Sharifullah Khan, Majed Al-Saeed, Hafiz Farooq Ahmad, and Rafia Mumtaz. 2019. Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, 36(4):e12409.

Simon Frith, Andrew Goodwin, and Lawrence Grossberg. 2005. *Sound and vision: The music video reader*. Routledge.

Saptarshi Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment, and multi-label emotion from suicide notes. *Cognitive Computation*, pages 1–20.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76.

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

Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Md. Afzal Hossan, Sheeraz Memon, and Mark A Gregory. 2010. A novel approach for mfcc feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5.

Kate Keenan, Alison Hipwell, Tammy Chung, Stephanie Stepp, Magda Stouthamer-Loeber, Rolf Loeber, and Kathleen McTigue. 2010. The pittsburgh girls study: overview and initial findings. *Journal of Clinical Child & Adolescent Psychology*, 39(4):506–521.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

P Krishnamoorthy and SR Mahadeva Prasanna. 2011. Enhancement of noisy speech by temporal and spectral processing. *Speech Communication*, 53(2):154–174.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joseph T. Berry, and Ali H. Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Siddharth Krishna Kumar. 2017. On weight initialization in deep neural networks. *arXiv preprint arXiv:1704.08863*.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

Alex Olwal and Steven Feiner. 2005. Interaction techniques using prosodic features of speech and audio localization. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 284–286.

Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. 2019. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*.

François Pachet and Pierre Roy. 2007. Exploring billions of audio features. In *2007 international workshop on content-based multimedia indexing*, pages 227–235. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88.

Karen L Smarr and Autumn L Keefer. 2011. Measures of depression and depressive symptoms: Beck depression inventory-ii (bdi-ii), center for epidemiologic studies depression scale (ces-d), geriatric depression scale (gds), hospital anxiety and depression scale (hads), and patient health questionnaire-9 (phq-9). *Arthritis care & research*, 63(S11):S454–S466.

Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. 2022. Tvlt: Textless vision-language transformer. *Advances in Neural Information Processing Systems*, 35:9617–9632.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. 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.

Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers 17*, pages 201–213. Springer.

World Health Organization. Depression.

Le Yang, Dongmei Jiang, and Hichem Sahli. 2018. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, 12(1):239–253.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017a. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59.

Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. 2017b. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 45–51.

Leon Yin and Megan Brown. 2018. Smappnyu/youtube-data-api.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.