Prediction of Relapse in Adolescent Depression using Fusion of Video and Speech Data

Christopher Lucasius[1,*], Mai Ali[1], Marco Battaglia[2,3], John Strauss[4], Peter Szatmari[2,3,5] and Deepa Kundur[1]

[1]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
[2]Division of Child and Youth Psychiatry, Centre for Addiction and Mental Health, Toronto, Canada
[3]Department of Psychiatry, University of Toronto, Toronto, Canada
[4]Vancouver Island Health Authority, Vancouver, Canada
[5]The Hospital for Sick Children, Toronto, Canada

Abstract

This article presents an innovative approach to predicting depression relapse in adolescents. Adolescentsíntensive use of video and voice-based smartphone apps presents a rich, multimodal dataset that can be utilized for this purpose. This work uses a dataset from the Depression Early Warning study conducted at the Center for Addiction and Mental Health. After using a pre-trained Inception ResNet to generate embeddings of video frames, the proposed framework integrates this with synchronized speech data. These embeddings are fused with audio features, resulting in a multimodal dataset. The combined features are processed through a Long Short-Term Memory model and a fully connected network to predict relapse of depression. An average accuracy of 0.80 highlights the effectiveness of the proposed multimodal approach and underscores its potential to effectively predict depression relapse in adolescents.

Keywords

Depression relapse, Multimodality, Inception ResNet, LSTM

1. Introduction

Depression is a worldwide, prevalent mental health disorder among adolescents. The recognition and treatment of adolescent depression hold paramount significance due to its association with substantial risks, notably suicide, which stands as the fourth leading cause of death within this demographic [1]. Disturbingly, over half of adolescents who commit suicide are reported to have been struggling with a depressive disorder [1]. Beyond this, depression in adolescents causes profound social and educational impairments, underscoring the need for timely intervention. The consequences extend to heightened rates of smoking, substance misuse, and obesity, accentuating the urgency of addressing this mental health concern [2].

Standard mental health diagnoses rely on clinical surveys that may be subject to recall bias. This approach also does not allow for timely interventions [3]. To address these limitations, diverse modalities have been proposed in the literature for timely mental health assessment and prediction. These modalities encompass physiological features such as heart rate and temperature, as well as behavioral features such as voice, facial expression, and gesture. Video chat and gaming are very popular among youth with statistics reaching 87% in this population [4]. However, despite the widespread engagement in these activities, research exploring the use of video and speech modalities for the assessment of depression and prediction of relapse in youth is limited. This work investigates the use of speech and video for depression relapse prediction in adolescents. As far as the authors are aware, it presents the first pipeline for predicting depression relapse in adolescents using fusion of video- and speech-based features.

2. Literature Review

The use of speech and video analysis for depression prediction represents an innovative and promising approach in mental health research. Analyzing speech patterns and facial expressions can provide valuable insights into an individual's emotional and mental state. Below is a review on the use of speech and video for depression prediction.

2.1. Speech-based Depression Prediction

Several studies demonstrated that voice quality contains information about the mental state of a person and vocal

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada
*Corresponding author.
christopher.lucasius@mail.utoronto.ca (C. Lucasius); maia.ali@mail.utoronto.ca (M. Ali); marco.battaglia@camh.ca (M. Battaglia); john.strauss@islandhealth.ca (J. Strauss); peter.szatmari@camh.ca (P. Szatmari); dkundur@ece.utoronto.ca (D. Kundur)

source features can be used as biomarkers of depression severity [5, 6, 7]. The work in [8] is based on a cross-sectional and longitudinal study aimed to explore the potential of voice acoustic features as objective biomarkers for assessing depression severity and treatment effectiveness. The study identified 30 voice acoustic features to be associated with depression such as Mel-cepstral (MCEP), Mel-scale Frequency Cepstral Coefficients deltas (MFCC-deltas) and Harmonic Model Phase Distortion Mean (HMPDM) among others. A neural network model based on Neural Architecture Search (NAS) was developed for predicting depression severity. Grid search was used to obtain the optimal model architecture which consisted of 4 hidden layers with with 32 units each. The model achieved a Mean Absolute Error (MAE) of 3.137 when predicting depression severity based on Hamilton Depression (HAMD) Scale. Additionally, a longitudinal study investigated the changes in voice features after an Internet-based cognitive-behavioral therapy (ICBT) program, revealing four features that significantly decreased: Peak2RMS_kurtosis, MFCC_deltas_10_intercept, MFCC_delta_deltas_4_kurtosis, and MFCC_delta_deltas_9_kurtosis. This indicated their potential correlation with treatment response and improvement in depression. In [9], Vázquez-Romero et.al. proposed a method for automatic classification of depression using speech and ensemble learning with Convolutional Neural Networks (CNNs). In the preprocessing phase, speech files are transformed into sequences of log-spectrograms and randomly sampled to ensure a balance between positive and negative samples. For the classification task, multiple CNNs are trained using different initializations, and their individual predictions are combined using an ensemble averaging algorithm. The predictions are then aggregated for each speaker to obtain a final decision. The performance of the proposed model was evaluated on the DAIC-WOZ dataset and compared against the AVEC-2016 models that use support vector machine (SVM) classifiers and hand-crafted features, as well as the DepAudionet architecture that consisted of a 1D-CNN, Long Short-Term Memory (LSTM) cell, and fully connected layers. The results demonstrated a relative improvement in F1-score of 58.5%, 30.0%, and 10.2% compared to the baseline, DepAudionet, and single 1D-CNN architecture, respectively.

2.2. Video-based Depression Prediction

Behavioral analysis of facial expressions has been studied as a source for eliciting the underlying emotional state [10]. Computer vision methods have been used to analyze facial expressions and gestures to predict the underlying mental health state of users [11]. A framework for estimating depression levels from video data using a two-stream deep spatiotemporal network was introduced in [11]. The framework combined spatial information extracted from the Inception-ResNet-v2 network with a volume local directional number (VLDN) based dynamic feature descriptor to capture facial motions. The VLDN feature map was then fed into a CNN to obtain more discriminative features. Temporal information was obtained using a multilayer Bi-LSTM which integrated the temporal median pooling (TMP) approach on the temporal fragments of spatial and temporal features. The performance of this work was benchmarked against the AVEC2013 and AVEC2014 datasets, and it achieved an MAE of 7.04 and 6.86 on AVEC2013 and AVEC2014, respectively.

Zhou et al. presented a deep regression network called DepressNet which aimed to learn a visually interpretable representation of depression from facial images [12]. Their model is based on a CNN with a global average pooling layer which is first trained with facial depression data, for identifying salient regions of an input image in terms of its severity score based on the generated depression activation map (DAM). The authors proposed a multi-region DepressNet that combines multiple local deep regression models for different face regions to enhance recognition performance. The method achieved an MAE of 6.20 and 6.21 on AVEC 2013 and 2014 datasets, respectively.

2.3. Speech and Video-based Depression Prediction

Physiological and psychological studies have identified differences in speech and facial expressions between patients with depression and healthy individuals, providing potential cues for automatic depression detection [13]. Another related work by [14] presented a depression detection model that utilizes audiovisual features extracted from video logs (vlogs) on YouTube. The model extracts eight low-level acoustic descriptors, including loudness, fundamental frequency (F0), and spectral flux, using the OpenSmile toolkit. These features capture characteristics such as voice intensity and pitch which have been found to be relevant in detecting depression. For visual features, the model utilizes a pre-trained face expression recognition model (FER) to extract emotional information from the vlogs. The proposed eXtreme Gradient Boosting (XGBoost) depression detection model achieved an overall performance with an accuracy of 75.85%, recall of 78.18%, precision of 76.79%, and F1 score of 77.48%. The model's performance was further analyzed based on different modalities where the model trained with audio features performed better than the model trained with visual features. The best performance was achieved by the model trained on the audiovisual features. The work of Othmani et.al. in [15] used deep learning techniques to recognize depression and predict relapse from audio

and visual cues extracted from videos of clinical interviews. It involves a correlation-based anomaly detection framework that compares the audiovisual patterns of depression-free subjects to those of depressed individuals. The correlation between the audiovisual encoding of a test subject and a deep audiovisual representation of depression is computed to monitor depressed subjects and predict relapse. The approach achieves promising results, with an accuracy of 80.99% and 82.55% for relapse depression prediction on the DAIC-Woz dataset.

The existing landscape of research on adolescent depression has made significant strides in understanding the onset and symptoms of depression in this age group. However, there is a notable gap in the ability to effectively predict depression relapse from audio and video modalities. By incorporating synchronized video and speech data, this research captures a broader spectrum of behavioral and emotional cues that might signify impending relapse in adolescents. The synchronization ensures that both modalities are aligned, allowing for a detailed examination of facial expressions, body language, and speech patterns simultaneously.

3. Problem Formulation

Our work aims to classify fused video and speech features for the classification of data that is measured before a relapse event. This entails a binary classification task where the two classes include "relapse sometime in the future" and "non-relapse". This problem is significantly different from detecting the presence of depression or predicting a certain depression rating scale score. The problem of relapse prediction is more complex since it involves the direct prediction of a clinical event within a population of adolescents who are already diagnosed with Major Depressive Disorder.

4. Methods

This work uses a dataset that is collected as part of the depression early warning study that was run in the Centre for Addiction and Mental Health (CAMH). It includes 80 video interviews collected from 52 adolescents aged 12-21 who were all diagnosed with Major Depressive Disorder.

4.1. CAMH Dataset

All participants had an initial baseline visit followed by up to 7 followup visits, each spaced apart by 3-12 months. During each visit, participants were assessed by a trained research coordinator and psychiatrist, providing psychiatric evaluations of their depressive states via the Children's Depression Rating Scale (CDRS). Participants were interviewed by the coordinator during their initial visit and followup sessions. During recorded Zoom sessions, the coordinator asked them 10 open-ended questions about their past activities and mood, resulting in 2-10 minutes of video data per session. This dataset was collected as part of an ongoing research study at CAMH and is unavailable to the public.

4.2. Definition of Relapse

While there are many definitions of relapse in depression, a commonly accepted one is given by [16] which defines a relapse in adolescents as observing a CDRS score of at most 28 during at least 12 weeks of treatment followed by an increase in CDRS to at least 40 for at least two weeks. The first period of 12 weeks corresponds to a remission stage where the depressed adolescent does not exhibit symptoms but has not yet completed treatment. The period of two weeks corresponds to a depressed episode.

In this study, there can be at least a three month break between followup visits. Hence, the timing aspect of Kennard's definition must be accordingly modified to adhere to the provided data. This work proposes a definition of relapse as a period of at least one visit with a CDRS score of at most 40 followed by one visit with a CDRS score of at least 40.

4.3. Pipeline

There are three main stages that make up the methods of this pipeline. The first consists of preprocessing the video and audio data and organizing them such that the two modalities are aligned and the labels are balanced. The second involves training models on random subsets of the training data. In the final stage, the final model that was trained on the training set is evaluated on multiple test sets, and the performance metrics are averaged across each set. A diagram summarizing the pipeline is shown in Figure 1.

4.3.1. Stage 1: Data Preparation

Each video interview is divided into segments where only the participant is speaking. Since the interviews are conducted via Zoom, the videos are also cropped such that only the participant's face is visible. Several spectral features are extracted from the audio data using the Python package libRosa [17]. They include the MFCCs, fundamental frequency, chromagrams, power spectral density, and spectral rolloff. These features are computed over a rolling window that is applied across the video. The amount of overlap is chosen such that the number of windows matches that of the video frames and are evenly spread out across the video.

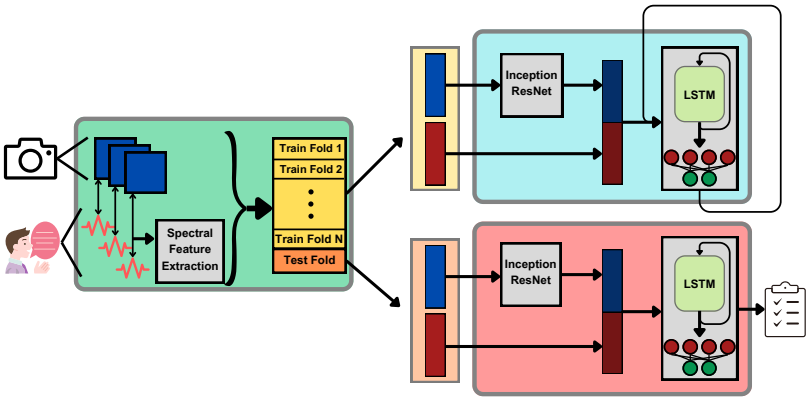

Figure 1: In Stage 1 (green block), signal processing algorithms process audio signals to generate spectral features. Video frames and spectral features are aligned and stored into train and test folds. Train folds are passed through Stage 2 (blue block). Video frames are processed with a pretrained Inception ResNet model. Resulting features (video in blue and audio in red) are fused with the spectral features and then fed through an LSTM and a fully connected network for training. Stage 3 (red block) has the same components of Stage 2, but it is only used on the test fold for evaluation purposes.

In the provided dataset, there is a significant class imbalance where the non-relapse data is heavily over-represented (96.25% non-relapse). In order to not bias the training of the models and the evaluation metrics (described in the next two sections), several training folds are prepared alongside a test fold. The folds are constructed by first randomly selecting a proportion of relapse video clips to use in the test fold. This proportion is chosen to be 30%, and it is computed based on the number of frames within each video clip. A random selection of non-relapse clips are chosen to match the number of frames of the relapse ones (rounded to the nearest whole number of clips). This completes the test fold which is reserved for Stage 3 of the pipeline. The rest of the relapse subjects are assigned to be used by train folds in Stage 2. Non-relapse video clips are randomly sampled without replacement where the number of clips is selected to match the number of frames of the relapse subjects. Each random sample of non-relapse clips makes up another train fold, and this process is repeated until all non-relapse clips are used.

4.3.2. Stage 2: Training of Models

The video frames are fed into an InceptionResNet model that was pre-trained on VGGFace2 [18], a large-scale face dataset. The resulting embeddings from this network are then fused with the spectral features of the audio data. The resulting fused features are then fed into a neural network module (named AudioVisual Network) containing an LSTM and a fully connected network. The LSTM is used to process 16 consecutive frames of features at a time, and the resulting hidden state is then fed into the fully connected network to be classified as either relapse or non-relapse. During the training process, random segments of 16 frames are sampled from the training video clips in order to not bias the training of the network towards a certain class.

The training process is applied to each train fold, and within a given fold, it is repeated for eight epochs. After the AudioVisual Network is trained on a given fold, its saved parameters are used to continue training the network on a new fold. This is repeated until all train folds are exhausted. This allows the network to train on the entire training dataset while still keeping the classes relatively balanced.

4.3.3. Stage 3: Evaluation of Models

After the AudioVisual Network is trained, the architecture (+InceptionResNet), is evaluated on the test fold that was reserved in Stage 1. A receiver operating characteristic (ROC) analysis is carried out on the predictions and ground truth labels. The optimal threshold of the ROC curve is selected by choosing the point that maximizes the difference between the true and false positive rats. This threshold is used to compute the MAE.

The entire process of training the models and evaluating the final one on a test fold is carried out for 10 sets of folds. This is to ensure that the reported metrics are

not biased toward a certain set of subjects. The resulting performance metrics are averaged across all of the test folds.

5. Results and Significance

Table 1 shows the results of evaluating the trained model on the 10 test folds. Each accuracy and MAE measure was reported after finding the optimal threshold of the ROC curve.

An average accuracy of 0.80 shows that video and speech data are relatively promising in the prediction of relapse in adolescent depression. In previous work by Othmani et al. [15], the authors also predicted relapse of depression using video and speech data. Similar to our work, they also yielded accuracies at around 0.8. To the best of our knowledge, this is the only other work that used video and speech to predict relapse of depression. Our work differentiates from Othmani et al. in two significant ways: 1) our study focuses on adolescents and 2) the source of our data is from non-clinical interviews. These interviews allow for more conversational topics that may better mimic a real-life situation in an adolescent's everyday life. While the target population for this work includes adolescents, this framework can be extended to other depressed populations.

Fold	MAE	Accuracy
1	0.077	0.92
2	0.28	0.72
3	0.21	0.79
4	0.23	0.77
5	0.12	0.88
6	0.26	0.74
7	0.17	0.83
8	0.23	0.77
9	0.17	0.83
10	0.27	0.73
Average	**0.21**	**0.80**

Table 1
Results of evaluation of final trained models on test folds

6. Limitations and Future Work

Predicting depression from audiovisual features encounters various challenges. The subjectivity of depression labels and the heterogeneous nature of this condition make it difficult to develop a universally applicable model. Additionally, there may be ethnic and cultural biases in the data that may have impacted the model's generalizability. This work did not consider the context within which interviews were conducted. Furthermore, the exclusion of gender-based analysis is a notable limitation, potentially overlooking important nuances in how depression manifests across different genders.

Future work in predicting depression from audiovisual features will prioritize the development of gender and context aware models. Moreover, given the longitudinal nature of the study, a promising avenue for future work is to exploit the temporal nature of data to track changes in audiovisual features over long periods of time. Employing an overarching time series model could enhance the understanding of the dynamic nature of depression, allowing for the development of more adaptive and personalized prediction models.

Another way to extend this work is to combine other objective sources of data that can be collected simultaneously with video and speech. One such modality includes wearable technologies, and there have been several studies on using them for the prediction of depression [19]. Using similar techniques, it may be possible to fuse audiovisual features and those derived from wearables to create a more robust predictor of adolescent depression relapse. Finally, we intend to evaluate our work using publicly available audio/video depression datasets such as AVEC.

References

[1] World Health Organization, Suicide, 2023. URL: https://www.who.int/news-room/fact-sheets/detail/suicide.

[2] A. Thapar, S. Collishaw, D. S. Pine, A. K. Thapar, Depression in adolescence, The Lancet 379 (2012) 1056–1067. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488279/. doi:https://doi.org/10.1016/s0140-6736(11)60871-4.

[3] N. H. Goldhaber, A. Chea, E. B. Hekler, W. Zhou, B. Fergerson, Evaluating the mental health of physician-trainees using an sms text message–based assessment tool: Longitudinal pilot study, JMIR Formative Research 7 (2023) e45102–e45102. doi:https://doi.org/10.2196/45102.

[4] P. Summerfield, How many kids in canada are connecting with video games?, 2023. URL: https://mediaincanada.com/2023/01/30/how-many-kids-in-canada-are-connecting-with-video-games/.

[5] Q. Zhao, H.-Z. Fan, Y.-L. Li, L. Liu, Y.-X. Wu, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, S.-P. Tan, Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: A cross-sectional study, Frontiers in Psychiatry 13 (2022). doi:https://doi.org/10.3389/fpsyt.2022.815678.

[6] D. Shin, W. I. Cho, C. H. K. Park, S. J. Rhee, M. J. Kim,

H. Lee, N. S. Kim, Y. M. Ahn, Detection of minor and major depression through voice as a biomarker using machine learning, Journal of Clinical Medicine 10 (2021) 3046. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8303477/. doi:https://doi.org/10.3390/jcm10143046.

[7] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Communication 71 (2015) 10–49. URL: https://www.sciencedirect.com/science/article/pii/S0167639315000369. doi:https://doi.org/10.1016/j.specom.2015.03.004.

[8] Y. Wang, L. Liang, Z. Zhang, X. Xu, R. Liu, H. Fang, R. Zhang, Y. Wei, Z. Liu, R. Zhu, X. Zhang, F. Wang, Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study, Frontiers in Psychiatry 14 (2023). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10320390. doi:https://doi.org/10.3389/fpsyt.2023.1195276.

[9] A. Vázquez-Romero, A. Gallardo-Antolín, Automatic detection of depression in speech using ensemble convolutional neural networks, Entropy 22 (2020) 688. doi:https://doi.org/10.3390/e22060688.

[10] P. Ekman, W. V. Friesen, Facial action coding system: Investigator's guide, Consulting Psychologists Press, 1978.

[11] M. Azher Uddin, J. Bibi Joolee, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm | ieee journals & magazine | ieee xplore, 2022. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8976084.

[12] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, IEEE Transactions on Affective Computing 11 (2020) 542–552. doi:https://doi.org/10.1109/taffc.2018.2828819.

[13] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, X. Pan, W. Dang, Deep learning for depression recognition with audiovisual cues: A review, Information Fusion 80 (2022) 56–86. doi:10.1016/J.INFFUS.2021.10.012.

[14] K. Min, J. Yoon, M. Kang, D. Lee, E. Park, J. Han, Detecting depression on video logs using audiovisual features, Humanities and Social Sciences Communications 10 (2023). URL: http://dx.doi.org/10.1057/s41599-023-02313-6. doi:10.1057/s41599-023-02313-6.

[15] A. Othmani, A. O. Zeghina, A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept, Healthcare Analytics 2 (2022) 100090. URL: https://www.sciencedirect.com/science/article/pii/S2772442522000387. doi:https://doi.org/10.1016/j.health.2022.100090.

[16] B. D. Kennard, T. L. Mayes, Z. Chahal, P. A. Nakonezny, A. Moorehead, G. J. Emslie, Predictors and Moderators of Relapse in Children and Adolescents With Major Depressive Disorder, The Journal of Clinical Psychiatry 79 (2018) e1–e8. URL: https://www-psychiatrist-com.myaccess.library.utoronto.ca/jcp/depression/predictors-of-relapse-in-youth-with-major-depressive-disorderhttps://www-psychiatrist-com.myaccess.library.utoronto.ca/jcp/depression/predictors-of-relapse-in-youth-with-major-depre. doi:10.4088/JCP.15M10330.

[17] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, Proceedings of the 14th Python in Science Conference (2015). doi:https://doi.org/10.25080/majora-7b98e3ed-003.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. doi:10.1109/CVPR.2016.308.

[19] L. Sequeira, S. Perrotta, J. LaGrassa, K. Merikangas, D. Kreindler, D. Kundur, D. Courtney, P. Szatmari, M. Battaglia, J. Strauss, Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: A scoping review, Journal of Affective Disorders 265 (2020) 314–324. URL: https://www.sciencedirect.com/science/article/pii/S0165032719310304. doi:https://doi.org/10.1016/j.jad.2019.11.156.