Multimodal Depression Detection in Mother-Child Interactions and Patient-Clinician Interviews

Abstract:

Depression is a recurrent and highly prevalent condition that entails significant personal impairment and suffering, increases risk for suicide, and results in substantial socio-economic costs. Due to its widespread influence over individuals, society, and economy, timely detection and intervention are critical. The principal approaches to depression detection are self-report and clinician interviews. Self-reports are subject to individual differences in literacy, idiosyncractic interpretations of questions, and burden of repeated use. Clinician interviews require a highly-trained examiner, are difficult to standardize, and are prone to biases due to their inherent subjective nature. Computational approaches to depression detection address these limitations. They offer a reliable, valid, and efficient solution using objective measures of behavior.

While significant progress has been made in detecting depression using computational approaches, challenges remain. Most existing datasets fail to include clinically characterized participants, and their demographic variability and interaction contexts are limited. Most also use only a single modality, lack model interpretability, and rely on the same database [1, 2].

We compared three computational approaches to depression detection in well-characterized and diverse participants observed in two different interpersonal contexts. One approach is a classical approach that uses handcrafted features from facial action units, face and head dynamics, speech behavior, prosody, and linguistic usage together with a Support Vector Machine (SVM) classifier. The other two approaches are the deep learning based Multimodal Large Language Model (MLLM) and Multimodal Transformer approaches. Deep learning uses multiple layers to learn representations and has made major advances in a wide range of applications [3]. We compare the classical approach to the two deep learning approaches.

These three approaches are compared in two different contexts. In the depression severity interview context [4, 5, 6, 7], clinicians interviewed participants from a clinical trial for depression treatment using Hamilton Rating Scale for severity assessment. The other context is a mother-child problem-solving interaction, where depression was defined as mothers with a history of treatment for depression and recent or current clinically elevated depression symptoms; non-depressed mothers lacked both lifetime history of treatment for depression and no more than mild depressive symptoms currently [8, 9, 10].

Participants in both patient-clinician interview and mother-child interaction contexts varied in race.

In addition to assessing generalizability between the two contexts in terms of accuracy, we evaluated for each approach whether depression in mothers could be detected from their child's multimodal behavior.

The comparisons between the classical approach and the deep learning approaches resulted in the following findings:

- Finetuning LLMs for depression detection yielded comparable performance to the classical approach, but MLLMs perform poorly than the classical approach.
- The multimodal transformer performed comparable to the classical approach.
- Generalization of depression detection by training on mothers from the mother-child interactions and testing on patients from the patient-clinician interviews modest performance (0.561 accuracy). Similar performance (0.541 accuracy) was found in the vice-versa experiment.
- Our error analysis found that our multimodal transformer approach had lower racial bias compared to the classical approach. In mother-child context, the classical approach incorrectly detected significantly more white individuals to be non-depressed than non-white individuals. Similarly, significantly more non-white individuals were incorrectly detected as depressed than white individuals. These differences were nominal in patient-clinician interviews.

The detection of mother's depression from child multimodal features achieved 0.769 accuracy compared to 0.842 accuracy with mother features. Further findings on comparing mother and child behavior by mother's depression are:

- Prosody features from the classical approach were highly predictive of depression using mother features, and child features. Spectral slope, articulatory effort, and voice quality features were highly associated with mother's depression among mothers and children.
- In facial action units (AUs), AUs 6 and 12 associated with positive affect differed between mothers by depression, while AUs 4 and 15 associated with negative affect differed among children of depressed mothers.
- Depressed mothers also had less head motion acceleration in comparison to the non-depressed mothers, children of depressed mother also had fewer changes in their facial expressions.

- Linguistic analysis found that issues related to home environment were prevalent in the speech of the depressed group, while they had lower references to issues regarding personal hygiene as compared to the non-depressed counterparts.

References:

[1] Jeffrey M Girard, Dasha A Yermol, Albert Ali Salah, and Jeffrey F Cohn. Computational analysis of expressive behavior in clinical assessment. computer, 81423:024140, 2025.

[2] Umut Arioz, Urska Smrke, Nejc Plohl, and Izidor Mlakar. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. Diagnostics, 12(11):2683, 2022.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015

[4] Ellen Frank, Giovanni B Cassano, Paola Rucci, Wesley K Thompson, Helena C Kraemer, Andrea Fagiolini, Luca Maggi, David J Kupfer, M Katherine Shear, Patricia R Houck, et al. Predictors and moderators of time to remission of major depression with interpersonal psychotherapy and SSRI pharmacotherapy. Psychological medicine, 41(1):151–162, 2011.

[5] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. Image and vision computing, 32(10): 641–647, 2014.

[6] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In 2009 3rd international conference on affective computing and intelligent interaction and workshops, pages 1–7. IEEE, 2009.

[7] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. IEEE transactions on affective computing, 4(2):142–150, 2012.

[8] Benjamin W Nelson, Lisa Sheeber, Jennifer H Pfeifer, and Nicholas B Allen. Affective and autonomic reactivity during parent–child interactions in depressed and non-depressed mothers and their adolescent offspring. Research on Child and Adolescent Psychopathology, 49(11):1513–1526, 2021

[9] Jackie A Nelson, Esther M Leerkes, Marion O'Brien, Susan D Calkins, and Stuart Marcovitch. African American and European American mothers' beliefs about negative emotions and emotion socialization practices. Parenting, 12(1):22–41, 2012.

[10] Lisa Sheeber, Jessica Lougheed, Tom Hollenstein, Craig Leve, Kavya Mudiam, Catherine Diercks, and Nicholas Allen. Maternal aggressive behavior in interactions with adolescent offspring: Proximal social–cognitive predictors in depressed and non-depressed mothers. Journal of psychopathology and clinical science, 2023