Literature Review of AI Applications in Clinical Psychology

Heyuan Huang
Yale University
Health Informatics
heyuan.huang@yale.edu

1 Introduction

In 2019, 13% of people in the world had mental disorders, among which 31% have anxiety and 28.9% have depression[16]. During the COVID-19 pandemic, the number of patients increased significantly compared to the limited, almost unincreased mental support medical resources[17]. Automatic and comprehensive AI applications hold significant potential in improving the efficiency and effectiveness of treatment, through automatic mental disorder detection[20], early suicide risk screening[10], auxiliary decision-making [12], and empathetic chatbot companion[19].

In this paper, we aimed to summarize current AI research that can be applied in the clinical psychology field, as well as their impacts and limitations. The whole paper is written in non-technical language, with unavoidable technical terminologies enclosed in square brackets.

2 Methods

Current dominant AI applications in Clinical Psychology can be categorized into 3 main fields: Computer Vision (CV), Natural Language Processing (NLP), and Signal Processing (SP). We searched and read about 20 papers published in main AI conferences, focusing on applications in the clinical psychology field. Search keywords include 'mental', 'emotion', 'depression', 'suicide', etc. We organized these papers according to their AI fields and summarized their approaches to solving clinical problems, with figures for better elaboration. In each method's discussion, we concluded its limitations and implications. We also pointed out promising future AI directions in psychology and called for collaborations at the end of this paper.

3 Computer Vision

Computer vision uses images, such as facial expressions, to recognize people's emotional status and behaviors. In clinical psychology settings, it can be applied to record users' mood swings and behaviors in a certain period and help clinicians better evaluate their patients' disease severity.

3.1 Emotion Recognition

The classical emotional recognition methods[15] focus on constructing and analyzing human facial features to classify human emotions into 7 basic categories, anger, disgust, fear, happy, sad, surprise, and neutral. Figure 1 shows the example images for the 7 basic emotions.

To allow more complicated emotion classifications, Li et al. [13] created the first compound facial expression dataset, RAF-DB, and proposed a deep learning model to classify images into 12 categories, such as happily surprised, disgustedly surprised, sadly angry, fearfully angry, and so on. Figure 2 shows the example images for all 12 compound emotions, which are derived from 6 basic emotions.

Preprint. Under review.

Figure 1: Example images for 7 basic emotions[15]

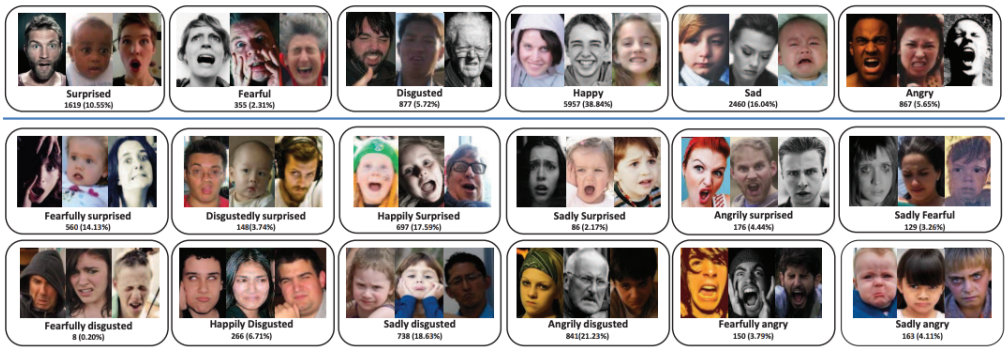

Figure 2: Example images for 6 basic emotions and 12 compound emotions[15]

Lee et al. [11] improves the emotion classification accuracy by using a series of images in a video, instead of one single static facial image. The intuition behind this method is that the model can see and learn more information through an image series, such as background objects and the subject's movement.

3.2 Facial Behavior Recognition

Ekman and Friesen [7] defined a Facial Action Coding System (FACS) to describe all visually distinguishable facial movements. It breaks down human facial expressions into 44 muscle Action Units (AU) in Figure 3. For example, AU 1 is "inner brow raiser", AU 2 is "outer brow raiser", AU 9 is "nose wrinkler" and AU 17 is "chin raiser". Facial expressions can be regarded as combinations of action units, as shown in Figure 4.

Baltrušaitis et al. [2] created an open-source toolkit for facial behavior analysis. The facial action unit recognition can help clinicians evaluate patients' social skills and emotional status during therapy

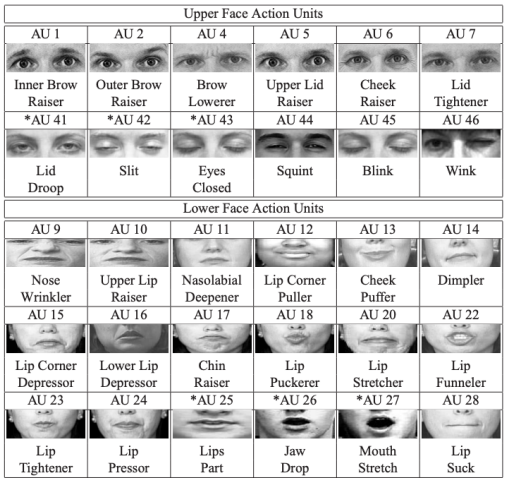

Figure 3: Examples of Facial Action Coding System[21]

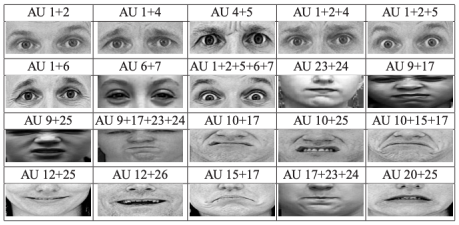

Figure 4: Examples of Action Unit combination[21]

sessions. It also supports head pose estimation and eye-gaze estimation, which can help clinicians evaluate patients' attentiveness.

3.3 Limitations

For real-world clinical applications, there are 3 main concerns that limit computer vision's power.

Image Quality. It is difficult to capture the subjects' whole faces as clearly as the experimental images. Most pictures captured by monitors are low resolution and only contain partial faces. It makes the classification job much harder than it is in the experimental settings.

Privacy. Most people usually don't allow a camera installed in their homes to monitor them all the time, even if it is only used for healthcare purposes. This results in the discontinuity of data collection for a certain user, thus making the tracking mood swings hard and imprecise.

Accuracy. Even in the experimental setting, the state-of-the-art models can not distinguish human emotions perfectly. For example, it can not tell the subtle difference between a fake smile and a real smile, because these two smiles have almost the same facial features. What's more, the models' accuracy is highly dependent on the training data quality and amount. That is, if the training data can not cover all kinds of images in the real world, the models will usually have difficulty in generalizing to unseen data with high classification accuracy.

4 Natural Language Processing

Natural Language Processing utilizes text data to find the underlying human emotions, intentions, personalities, experiences, and other information. In the clinical psychology field, there are 3 main applications, Mental Disorder Detection (MDD), Mental Health Counseling (MHC), and Suicide Detection.

4.1 Mental Disorder Detection

Social media posts are publicly available and contain many user self-disclosed information. Self-Reported Mental Health Diagnoses (SMHD) [5] is a Reddit Post dataset, containing millions of Reddit posts from mental disorder diagnosed users and a matched control group, which has 335,952 control users and 116M control posts. It labels each patient's posts with their diseases, including ADHD, Anxiety, Autism, Bipolar, Depression, Eating disorder, OCD, PTSD, and Schizophrenia. Song et al. [20] compared the similarity of text semantic meanings between social media posts and the disease diagnostic criteria (i.e., symptoms) in Diagnostic And Statistical Manual Of Mental

Disorders, Fifth Edition (DSM-5)[1]. The higher the similarity score is, the more probable the user will be to have the symptom. By finding typical symptoms from a user's post, the model can classify a user as a possible disease patient. In this way, the classification model can learn clinical knowledge and then can classify posts into disease categories. However, the classification accuracy of the state-of-the-art models on general posts is not high enough for clinical use and still has plenty of room for improvement.

4.2 Mental Health Counseling

To address the expensive therapy cost and shortage of well-trained mental health professionals, researchers are developing virtual assistants and conversational agents to mimic professionals' behaviors and learn their expert knowledge to generate empathetic and actionable replies to help seekers. Sharma et al. [18] improved the empathy ability of conversational agents so that they can rewrite a low-empathy reply to a high-empathy reply, without impairing conversation quality regarding text fluency and context information. Figure 5 shows how the conversational agent removes the original low-empathy text and inserts a more supportive reply to the help seeker.

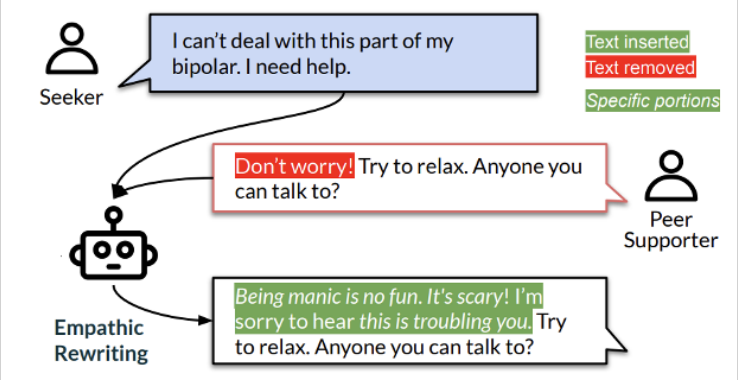

Figure 5: Example of empathic rewriting[18]

By enabling the agent with the ability to understand humans and learn professional knowledge, we can improve the quality of online peer support to be similar to trained professionals' replies. In this way, we can significantly reduce the burden on trained professionals and scale up the accessible therapy through the Internet.

4.3 Suicide Detection and Resource Recommendation

Similar to mental disorder detection, Izmaylov et al. [10] developed a model to automatically detect suicide risk in online chat sessions. As the sessions go on, more information will be learned by their model to make suicide risk evaluations. Their model can conduct early risk detection with about 60% to 80% of the whole conversation messages. The improvement direction in this field is to use messages as few as possible to find the suicide signal at an early stage.

Dang et al. [6] developed a recommendation system based on users' past Reddit posts to point users to the correct subreddit communities (e.g., depression, anxiety) they belong to.

4.4 Limitations

Privacy. Compared with facial images, public social media posts are more accessible. However, researchers still don't have access to other kinds of text data most of the time, such as users' daily text messages and chats. These private data can contain more precise information than public posts, which users might manipulate before posting to maintain a good public image.

Potential Implications. Mental health conversational agents are still under examination regarding their response quality, health outcomes, transparency, ethics, and cultural heterogeneity. Cho et al. [4] conducted a comprehensive survey of conversational agents studied from Computer Science and

Medicine perspectives. It finds that computer science research focuses more on technical developments, while medicine research focuses more on the health outcomes of participants. Currently, researchers can not explain complex model behaviors very well. This black-box nature of complicated large language models prevents the models from being trusted and used in large-scale clinical practices. The development and implementation of mental health conversational agents need collaboration from stakeholders in all sections (e.g., law and policy, medicine, statistics, and computer science), to make it safe, reliable, and under control.

Accuracy. The state-of-the-art models still can not recognize sarcasm or metaphor in plain texts perfectly and their disease classification accuracy on general social media posts is not high enough for clinical use. Similar to the generalization issue of Computer Vision models, most natural language models can not generate out-of-domain words or answers. What is worse is that these language models might generate hallucinated responses, which are wrong, irrelevant to context, and might cause harm to users.

5 Signal Processing

Signal Processing analyzes users' audio, vital, or other behavioral signal data to gain insights about user conditions, such as emotional conditions and stress conditions.

5.1 Speech Emotion Recognition

Different emotional audio's sound wave has different acoustic features. Zou et al. [23] extracted multiple acoustic features and trained a model to classify audio into 4 emotion categories, happy, sad, angry, and neutral, with accuracy ranging from 82% to 64% for each class. Similarly, Li et al. [14] trained a model to classify audio into 5 categories, happy, sad, angry, disgusted, and neutral, with accuracy ranging from 62% to 35% for each class. It is still difficult to distinguish audio emotions using only acoustic information.

5.2 Sensor Monitoring

Compared with conventional questionnaire-based evaluation, mobile mental health allows remote, continuous, and non-intrusive evaluation, by using wearable devices and smartphone sensors to collect users' physiological and behavioral data. Fazeli et al. [8] proposed a dynamic stress detection system from physiological signals (e.g., heart rate) collected by smartwatch. Figure 6 shows the signal waves of their collected data, which are fluctuant and continuous. By processing and analyzing sensor data, this automatic monitoring system can help clinicians evaluate patients' daily stress and anxiety levels in a more convenient and precise way.

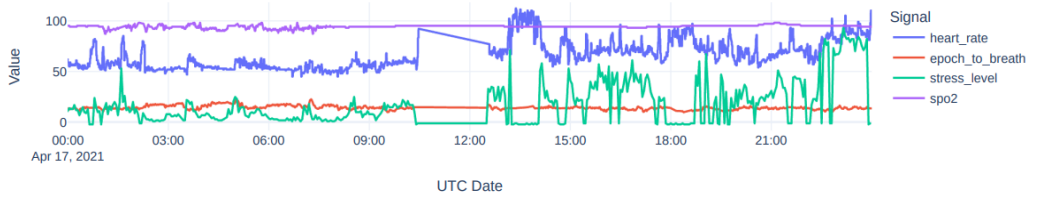

Figure 6: Example of physiological signals collected by smartwatch[8]. spo2 is oxygen saturation.

5.3 Limitations

Data Quality. In the real world, there are many noises such as background noise, and multiple speakers' noise, when recording audio of a user. These noisy data increase the burden of denoising without impairing the original user's signal quality and missing subtle features. In addition, recording devices (e.g., smartphones, professional recording devices, etc) can determine the collected data quality.

Privacy. It is usually not allowed to record users' daily utterances continuously, thus making it infeasible for real-world applications.

Accuracy. The wearable devices and smartphone sensors can collect numerous data of users, but these data need to be labeled by participants with their recalled stress and anxiety levels. This manual process contains a lot of human error as subjects may not be able to precisely recall the exact time and level of their stress feelings. Models trained on these data will have unavoidable inaccuracy and subject distribution bias.

6 Conclusion and Future Work

As discussed before, any modality data (text, image, audio, sensor signal) contains special information about users but also has its intrinsic limitations. A trend in this field is to fuse multiple modalities' data to extract more comprehensive information. For example, Chen et al. [3] uses textual, acoustic, and visual information in videos to classify characters' emotions.

Flessa and Huebner [9] defined the lifecycle of innovative technologies in the healthcare system, which goes through innovation, promotion, and adoption. Currently, most of the computer science research is still in the innovation stage to improve the accuracy and output quality. Westfall et al. [22] also pointed out that it takes 17 years on average for only 14% of the research findings to be adopted into clinical practices. The AI applications, especially in the clinical psychology area, still have a long way to go.

Although AI technologies have enormous potential to facilitate clinical procedures and develop more personalized treatments, they still need to be regulated under interdisciplinary collaborations to examine their reliability, and transparency before market readiness.

7 Appendix

The main AI conferences we refer to are listed below.

4 conferences for Natural Language Processing include ACL (Annual Meeting of the Association for Computational Linguistics), EMNLP (Conference on Empirical Methods in Natural Language Processing), NAACL (The North American Chapter of the Association for Computational Linguistics), COLING (International Conference on Computational Linguistics).

3 conferences for Computer Vision include CVPR (IEEE Conference on Computer Vision and Pattern Recognition), ICCV (IEEE International Conference on Computer Vision), and ECCV (European Conference on Computer Vision).

1 conference for Signal Processing includes ICASSP (IEEE International Conference on Acoustics, Speech, and Signal Processing).

References

[1] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. *Diagnostic and Statistical Manual of Mental Disorders*, 5(5), 2013. doi: https://doi.org/10.1176/appi.books.9780890425596.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016. doi: 10.1109/WACV.2016.7477553.

[3] Feiyu Chen, Jie Shao, Shuyuan Zhu, Heng, and Tao Shen. *Multivariate, Multi-frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation*. URL `https://openaccess.thecvf.com/content/CVPR2023/papers/Chen_Multivariate_Multi-Frequency_and_Multimodal_Rethinking_Graph_Neural_Networks_for_Emotion_CVPR_2023_paper.pdf`.

[4] Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Chandra Guntuku. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives, 2023.

[5] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. SMHD: a large-scale resource for exploring online language usage for multiple

mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1126.

[6] Hy Dang, Bang Nguyen, Noah Ziems, and Meng Jiang. Embedding mental health discourse for community recommendation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 163–172, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.codi-1.22. URL https://aclanthology.org/2023.codi-1.22.

[7] Paul Ekman and Wallace V. Friesen. Facial action coding system: a technique for the measurement of facial movement. 1978. URL https://api.semanticscholar.org/CorpusID:141196166.

[8] Shayan Fazeli, Lionel Levine, Mehrab Beikzadeh, Baharan Mirzasoleiman, Bita Zadeh, Tara Peris, and Majid Sarrafzadeh. Passive monitoring of physiological precursors of stress leveraging smartwatch data. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2893–2899, 2022. doi: 10.1109/BIBM55620.2022.9995354.

[9] Steffen Flessa and Claudia Huebner. Innovations in health care—a conceptual framework. *International Journal of Environmental Research and Public Health*, 18(19):10026, 2021. doi: https://doi.org/10.3390/ijerph181910026. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8508443/.

[10] Daniel Izmaylov, Avi Segal, Kobi Gal, Meytal Grimland, and Yossi Levi-Belz. Combining psychological theory with language models for suicide risk detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2430–2438, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.184. URL https://aclanthology.org/2023.findings-eacl.184.

[11] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks, 2019.

[12] Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.577. URL https://aclanthology.org/2023.acl-long.577.

[13] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. doi: 10.1109/CVPR.2017.277.

[14] Yang Li, Constantinos Papayiannis, Viktor Rozgic, Elizabeth Shriberg, and Chao Wang. Confidence estimation for speech emotion recognition based on the relationship between emotion categories and primitives. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7352–7356, 2022. doi: 10.1109/ICASSP43922.2022.9746930.

[15] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28 (5):2439–2450, 2019. doi: 10.1109/TIP.2018.2886767.

[16] World Health Organization. World mental health report: Transforming mental health for all, Jun 2022. URL https://www.who.int/publications/i/item/9789240049338.

[17] Damian F. Santomauro, Ana M. Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M. Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O. Amlag, Aleksandr Y. Aravkin, Bree L. Bang-Jensen, Gregory J. Bertolacci, Sabina S. Bloom, Rachel Castellano, Emma Castro, Suman Chakrabarti, Jhilik Chattopadhyay, Rebecca M. Cogen, James K. Collins, and Xiaochen Dai. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312): 1700–1712, Oct 2021. doi: https://doi.org/10.1016/S0140-6736(21)02143-7.

[18] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450097. URL https://doi.org/10.1145/3442381.3450097.

[19] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach, 2021.

[20] Hoyun Song, Jisu Shin, Huije Lee, and Jong Park. A simple and flexible modeling for mental disorder detection by learning from clinical questionnaires. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12190–12206, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.681. URL https://aclanthology.org/2023.acl-long.681.

[21] Yingli Tian, Takeo Kanade, and Jeffrey Cohn. *Facial Expression Recognition*, pages 487–519. 01 2011. doi: 10.1007/978-0-85729-932-1_19.

[22] John M. Westfall, James Mold, and Lyle Fagnan. Practice-based research—"blue highways" on the nih roadmap. *JAMA*, 297(4):403, Jan 2007. doi: https://doi.org/10.1001/jama.297.4.403.

[23] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. Speech emotion recognition with co-attention based multi-level acoustic information, 2022.