Dear Editors and Reviewers:

On behalf of all the contributing authors, I would like to express our sincere appreciations of your letter and reviewers' constructive comments concerning our article entitled "**Automatic Recognition of Depression Based on Audio and Video: A Review**" (Manuscript NO.: **90160, Minireviews**). These comments are all valuable and helpful for improving our article. According to the editor and reviewers' comments, we have made extensive modifications to our manuscript. The reviewer comments are laid out below in *italicized font* and specific concerns have been numbered. Our response is given in normal font. In this revised version, changes to our manuscript were all highlighted within the document by using yellow-colored text.

Reviewer #1:

1 Comment: *Automatic Recognition of Depression Based on Audio and Video: A Review This paper summarizes a recent literature survey on automatic depression estimation (ADE) methods. The inclusion focuses on those based on extracting and classifying relevant features from audio and video data, by means of deep learning. It was hypothesized that such schemes would alleviate the problems found in traditional evaluation by human physicians. Its content was divided into datasets, current limitations and prospectives. The manuscript was well written and structured. The studied area would be of interest to the World Journal of Psychiatry readers. The references list is relevant and up-to-date. The narratives, insights, and discussions on the topic, as presented by the authors, are appropriate and scientifically sound.*

1 Reply: Thank you very much for your recognition of our work. We hope that readers can gain a deeper understanding of recent methods for automatic estimation of depression through this review. We have also included discussions on limitations and future research directions in the article, hoping to inspire readers. Your comments are highly valuable for improving the quality of our manuscript. In accordance with your suggestions, we have made improvements to the manuscript and hope to receive your approval.

2 Comment: *Graphical representation of various elements, e.g., facial action units (AU) and their characterization, could help the readers in broader fields to grasp how FACS operates. Please consider.*

2 Reply: Thank you very much for your reminder. It is indeed challenging to imagine the operation of AUs solely through text. Therefore, we have inserted a footnote in the Introduction section, directing readers to a blog. Firstly, this blog provides a detailed introduction to all AUs. Secondly, it uses animations to illustrate the facial movements corresponding to each AU. We believe that this blog can assist readers in quickly understanding AUs.

3 Comment: *General overview of prominent ADE methods (in general, e.g., biological measurements and classifications, and not limited to deep learning) and their conventional counterparts, could be systematically grouped. Subsequently, the authors could clarify what types of ADE were focused here.*

3 Reply: Thank you for your suggestion. We have incorporated an introduction to traditional methods in the Introduction section. In this part, we first outline the main steps and drawbacks of traditional methods. Subsequently, this leads to the introduction of depression detection methods based on deep learning models.

4 Comment: *Prior to discussing the datasets, a section describing how deep learning-based ADE plays its part in psychiatric diagnosis and how they were incorporated into modern practice, should be added. Please consider.*

4 Reply: Thank you for your suggestion. During the process of retrieving relevant literature, we found only a small number of reports on the practical application of ADE. This has also triggered a discussion on the interpretability of ADE models. Therefore, we have placed this discussion in the Discussion section. In this part, we have added an introduction to the retrieved reports on practical applications and, in turn, introduced another future direction for ADE - interpretability.

5 Comment: *Short paragraphs pertaining detailed insight and authors' own critiques/ opinions on existing audio-based and video-based, and their fusion depression estimation methods should be given at the end of the respective sections.*

5 Reply: Thank you for your feedback. After the section introducing depression estimation methods, we have added relevant summaries and comments. These summaries introduce and summarize the construction process and key issues of the relevant methods.

6 Comment: *Relevant citations of related works should be added in appropriate places in the Discussion section.*

6 Reply: In accordance with your suggestion, we have inserted necessary references in the Discussion section.

7 Comment: *In conclusion and discussion, the "lack of exploration of the body expressions of individuals with depression," as stated by the authors is rather superficial. Are recognizing and classifying "body expressions" not already the prime areas of investigations in the most recent research? Please elaborate.*

7 Reply: I apologize for the lack of clarity in our expression. In the original manuscript, we intended to convey to the readers that "compared to depression detection methods based on facial expressions, methods based on body expressions are much fewer." In our search process, researchers have predominantly focused on facial expressions. Simultaneously, studies on body expressions of individuals with depression are largely based on private databases, which explains why there are fewer detection methods based on body expressions. We have rephrased the relevant section in the Discussion and added some further discussion.

We tried our best to improve the manuscript and made some changes marked in yellow in revised paper which will not influence the framework of the paper. We appreciate for Editors/Reviewers' warm work earnestly, and hope the correction will meet with approval. Once again, thank you very much for your comments and suggestions.