

Reply to the comments of Reviewer 1

We thank the reviewer for additional comments and requests for further clarification. All comments have been addressed and appropriate changes have been made in the manuscript. Our responses are given in boldface.

Thank you for addressing some of my concerns and providing a point-by-point response to my comments.

Although some of the points are addressed in your response, I would like some of the provided clarifications to be included in the text. In particular:

- which bandpass filter you used to filter BVP (as questioned in my previous point 8);

A 4th order Butterworth bandpass filter is used for signal denoising. It was stated in Figure 3 and now we also mentioned in the text on page 6, line 179.

- a more detailed explanation about the GSR signal preprocessing (as questioned in my previous point 10)

A paragraph regarding the preprocessing of GSR signal is expanded in the text (Page 6, lines 186 – 194).

- how co-linearity was evaluated (questioned in my previous point 15)

As requested, additional information is provided to describe the methodology in the manuscript. (Page 8, lines 236 – 238)

- which classifier was used in the sequential forward selection algorithm (questioned in my previous point 17)

Bootstrap-aggregated decision trees were used in the feature selection. The information is added to the manuscript (Page 9, lines 262-263).

Additionally, the following concerns remain open:

1. Concerning the participants in your study, were they healthy, or did they suffer from diabetes? In case they were healthy, how do you expect the algorithm to behave in patients with diabetes? Will it generalize well?

Participants in our study were combination of healthy subjects and subjects with Type 1 diabetes. Acute psychological stress inducements that are accepted by the Institutional Review Boards are limited both in type and intensity. To have a large enough dataset, we have used the data from both healthy and diabetes groups.

The algorithm is basically developed based on the body movement (3D accelerometer signal), body temperature, skin conductivity of the body, and cardio vascular activity of the body (BVP signal). These variables are signals directly recorded by sensors and regardless of the glycemic medical condition of subjects. Hence, they indicate the state of the body namely, activation of sympathetic and parasympathetic nervous system (presence of acute psychological stress), physical activity, and other states of the body that are not covered in this manuscript (such as sleep characteristics). Therefore, the acute psychological stress detection model is quite general as it has been tested with samples from both healthy and diabetes subjects. Of course, the effect of acute psychological stress on blood glucose levels will be different for subjects without and with diabetes. Glucose concentration estimations for people with Type 1 diabetes is addressed in our paper: Sevil M, M Rashid, I Hajizadeh, M Park, L Quinn, A Cinar. Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management, IEEE Trans on Biomedical Engineering, 2021, 68(7):2251-2260. DOI:10.1109/TBME.2020.3049109. PubMed ID: 33400644

2. In lines 286-288, please clarify how many classification tasks and what classes have you considered in each classification task ("Figure 5a shows the RNN architecture for the classifications of APS types, device types, PA types, and joint occurrence of PA and APS.")

Various classification tasks and classes were stated in Table 4. As requested, the text is modified and the different types of classification tasks and classes are stated after the sentence mentioned by the reviewer (starting line 287). APS classes include: Non-Stress, Emotional Anxiety Stress, and, Mental Stress. AP classes based on equipment type are: Sedentary State, Stationary Bicycling, and Treadmill Run. AP types are Sedentary State, Aerobic Exercise and Anaerobic Exercise. The jointly estimated states are centered on APS and PA: Sedentary State and Non-Stress, Sedentary State and Mental Stress or Emotional Anxiety Stress, Aerobic Exercise or Anaerobic Exercise and Non-Stress, and Aerobic Exercise or Anaerobic Exercise and Mental Stress or Emotional Anxiety Stress.

3. In my previous question 21, you mentioned that a "random number of samples from each subject is included in all three splits". By knowing this, I would conclude that your good results are, in fact, just a consequence of overfitting due to the fact that the algorithm is being tested on the same subjects included in training. This can, thus, be a major flaw of your approach. Could you please comment or justify this?

First, we should clarify that no duplication and repetitive sample are added to either training, validation, or testing split. What we mentioned in our response was the fact that observations from all subjects with diabetes and healthy subjects are randomly selected from sessions they performed (without replacement). In order to ensure the richness of training and validation data, having a high variety of samples in all splits is imperative.

Second, testing algorithm on the same subject necessarily doesn't achieve high scores just because data is taken from same subjects. The samples in the training, validation, and

testing splits are taken from different sessions of different subjects. Neither samples from the same session, nor any duplication samples are added to each one of three splits. It should be noted that the variation of biosignals during different clinical sessions is significantly different and having samples of different sessions of the same subject doesn't contribute to the estimation of samples from other classes (clinical session experiments).

Third, based above reasons, each class of the samples in the testing and validation split (for any specific subject), is predicted by samples-classes of all other subjects with the same label and samples of the same subject with different labels.

Additionally, demographic information (features) from training, validation, and testing split are excluded to test the generic behavior of the algorithm. This approach avoids misleading results obtained in high variance models that basically estimate classes based on demographic information.

The extracted features on different timestamps from a session from specific subjects are different as many unexpected variations and oscillations can be observed in different episodes. Additionally, what is generalizable is the overall behavior of 3D accelerometer signal, BVP signal, interbit intervals during different physical activity and physiological stress over all subjects.

Text has been added to clarify the formation of training, validation and testing datasets (page 11, lines 323-327).