Dear Editor, Dear Reviewers,

We sincerely want to thank you for you deep, detailed and insightful feedback concerning our manuscript titled "Anomaly Detection in Multi-Wavelength Photoplethysmography using Lightweight Machine Learning Algorithms". The comments are all valuable and constructive for revising and improving our paper.

We have addressed all comments by modifying the manuscript and the models accordingly. The corresponding response and adaptations are individually described in more depth below. Finally, the modifications are highlighted in the manuscript.

Yours sincerely,

Vlad-Eusebiu, Joan, Angel, Juan Carlos, Johan and Bruno

Reviewer 2

This paper presents a study of anomaly detection algorithms to detect various types of artifact in multiple wavelength PPG signals. Using a limited true dataset with data augmentation, the authors describe the identification of salient features and the training and testing of various ML models to separate artifactual time windows from clean time windows of data. I believe this paper could benefit from several major and a few minor revisions, as detailed below.

1. **Throughout the paper, the terms "multi-wavelength", "multi-channel", "multi-site", and "multi-sensor" are used casually without clear delineation or definition. This is very confusing to understand and should be distinguished early on clearly. For example, are there multiple sensors for each wavelength as well as multiple channels per sensor?**

   Response: Thank you for your comment. There was indeed a need to discuss this topic more clearly in the beginning. Our device uses multiple LEDs and PDs. Each source-detector pair represents a channel. However, we used only a number of three measurement channels in our analysis (1 PD and 3 wavelengths).

   Modifications: We have added a new subsection in Section 3 titled "MW-PPG Measurement and Artifact Detection." In this subsection, we provide an explanation of the distinctions between multi-sensor, multichannel, multi-wavelength, and multi-site configurations. We will refer to this subsection throughout the text whenever we feel that additional background information is necessary to aid the reader's understanding. Additionally, we provide a detailed explanation of how we conducted the anomaly detection experiments, specifically at the channel-level and sensor-level, along with the motivations behind these approaches.

2. **How was the data collected? Was there IRB approval? What were the age ranges of the subjects etc.? Was it done fully in a controlled setting? How was true artifact annotated? None of this is included.**

   Response: Thank you for your comment. The data was collected using the MW-PPG described in Section 4. Due to the small-scale experiment, and since we use de-identified data that target anomaly detection that is not subject-dependent, we consider that experiments do not pose potential risks or ethical considerations. We adhere to ethical principles and obtain informed consent from the participants (one of them is also the author). The experiment proposes to address artifact detection, such as gross body movement or another type of artifact that might not be subject-dependent. The data was recorded from two young participants (24, 26 years) in a non-controlled setting. The artifacts were labeled manually with a custom-made GUI on the measured recordings. Even if most of the channels are affected by the artifacts, the data is labeled channel-wise since we are channel-related artifacts in the recordings.

   Modifications: We added in Subsection 4.1 the mention that data was measured in a non-controlled setting. We also mentioned the age and some details about the subjects. In Subsection 4.2 we described how the true artifacts were annotated. In Subsection 4.4 we added a few details on how signal segmentation deals with the prior annotated signal.

3. **How was ground truth of artifact determined in the data? Was it done by visual inspection? Was it done for each channel? What were the criteria? In what window size? This was left unclear throughout.**

   Response: Thank you for your comment. The labeling process involved visual inspection and

the use of a custom-made GUI tool. Each channel was annotated individually. The ground truth for artifacts was determined visually, taking into consideration various factors. These factors included the computation of the heart rate using an open-source package called HeartPy, as well as the identification of gross motion artifacts and irregular waveform shapes that were not PPG-related. The heart rate computation served as a primary criterion in determining the presence of artifacts. The annotation is not related to the window size. We added a few details on how signal segmentation deals with the prior annotated signal.

Modifications: We modified Subsection 4.2 to include such details that were indeed not specified. In Subsection 4.4, we added a few details on how signal segmentation deals with the prior annotated signal.

4. **The data augmentation process needs to be explained in more detail.**

   Response: Thank you for your suggestion. That was indeed a need to add more details about how synthetic data is generated.

   Modifications: We made modifications to Subsection 4.3, where we described the channel-wise operations performed for data augmentation. Additionally, we relocated a figure from the former Subsection 3.2.3 to this subsection, as it was generated using data augmentation specifically for sensor-related artifacts.

5. **Why were the tsfresh and TSFEL packages chosen to be used in the first place?**

   Response: Thank you for your remark. We decided to use these packages since both TSFEL and tsfresh offer a wide range of predefined features that can be extracted from time series data. Moreover, when studying the literature about anomaly detection in multivariate time series sensors, we noticed the frequent usage of these tools. We considered it worth comparing the two in the context of MW-PPG, since they were used multiple times for other types of sensor data (IMU sensors, force sensors). They have the versatility of computing features that capture various aspects such as statistical properties, spectral characteristics, and other time series-related features.

   Modifications: We added three new references (Subsection 4.4) that we considered relevant for both TSFEL and tsfresh. There are examples of how activity recognition was performed based on these feature extractors, or even a more recent study, how blood pressure could be estimated from PPG signals using one of these tools.

6. **Is all of Section 4 done using the entire dataset? Or just the training data? This is not clear. If Section 4 is performed on the entire dataset, does this not introduce bias into Section 5's analysis since the features were selected by including the test data?**

   Response: Thank you for your remark. We used for feature selection only the training data set to prevent any leakage of the test set in the training process.

   Modifications: We modified Subsection 4.4.2 and specify that the feature selection is done on the training set and not on the entire dataset.

7. **Are the y-axis labels in Figures 12-14 Entropy or Information Gain?**

   Response: Thank you for your remark. The axis should be information gain, not entropy.

Modifications: We eliminate some of the figures to reduce the length of the article, and we change the x-axis label of the figure that remained in Subsection 5.1

8. **The threshold setting process used to arrive at a subset of features needs to be better explained.**

   Response: Thank you for your remark. The threshold is determined based on a recursive search method that recursively eliminates features from the feature set. The algorithm is based on a Random Forest. We used WEKA to achieve all of this.

   Modifications: We added in Subsection 4.4.2 more details on how the threshold is chosen and the pseudocode of the method.

9. **The motivation for the entire PCA section (Section 5.2) is not clear or compelling, and this analysis may be better served in the Supplementary Material. Also it is not clear if the PCA was performed separately on normal and anomalous windows or altogether.**

   Response: Thank you for your comment. We understand your perspective, and we agree that the details of the PCA analysis and its results could be better suited for the Appendix section. Regarding the PCA procedure, it was performed separately on the normal and anomalous windows to evaluate their respective contributions to the overall variance.

   For the normal windows, PCA was applied to explore the inherent variability and structure within the normal PPG signals. This analysis aimed to identify the principal components that explain the majority of the variance in the normal signal patterns

   Similarly, PCA was also applied to the anomalous windows to investigate their unique characteristics and distinguish them from the normal windows. By examining the principal components of anomalous windows, we can gain insights into the patterns and features that differentiate them from the normal windows.

   Modifications: We decided to move the PCA subsection in the Appendix section. We consider that it provides some insights into how different feature sets discriminate between anomalous and clean PPG windows.

10. **The terms "custom" and "proposed" features are both used and the delineation is unclear.**

    Response: Thank you for your remark. Indeed, a clear delimitation was not made between the custom and proposed features. The proposed features are the ones explained in Subsection 4.5. Based on this, we create a custom feature set that combines off-the-shelf features (TSFEL, tsfresh) and the proposed features (data-aware features).

    Modifications: We have provided a more detailed explanation of the process involved in creating the custom feature set and have explicitly listed the features included in each set. Additionally, we have included Table 3 (Subsection 5.1), which clearly distinguishes between the proposed features and the features comprising the custom feature set.

11. **Can the authors provide more clarity on what the features in each of the final datasets in Section 5 actually consist of? How many features are in each feature set, and how are they defined? Otherwise, it is difficult to reproduce these results.**

    Response: Thank you for your constructive remark. We have put a large amount of thought

into this remark. Since another reviewer suggested reducing the size of the text, we decided to prioritize this remark and to focus on the reproducibility of the experiments. We reduced the manuscript size after reading all the questions and remarks from all the reviewers. We explained better the feature sets and the proposed features along with the feature extraction and feature selection methods.

Modifications: We added Table 3 in Subsection 5.2, which lists the features in each final feature set. We also included some extra information about what channels are used when performing sensor-level anomaly detection. Moreover, we added some details on how the custom feature set is created.

12. **Are all of the results reported in Section 5 on the test set? Can the number of windows in the training, validation, and test sets be noted for each window size in a table?**

    Response: Thank you for your comment. Indeed, the results reported in Section 5 are on the test set.

    Modifications: We added at the beginning of Section 5 a table that lists the number of training, validation and test samples for each window size.

13. **What about F1 score in Figure 22?**

    Response: Thank you for pointing out the missing F1 score in Figure 22. We apologize for the oversight.

    Modifications: We updated Figure 22 to include the F1 score for better completeness and clarity.

14. **There is no discussion of limitations of the study in Section 6. For example, one major one is that this only includes data from 2 subjects.**

    Response: Thank you for your remark. The artifacts investigated in our study primarily consist of gross body movement, which can be considered universally present across individuals, as well as sensor-related artifacts and contact force artifacts. Indeed, contact force artifacts can vary depending on individual factors such as vascular anatomy or tissue temperature. Additionally, different sensor configurations, including variations in component placement, may introduce additional hardware-dependent artifacts. We appreciate your comment, as it helped us recognize the significance of these factors in artifact occurrence and detection.

    Modifications: We have included a paragraph in Section 6 to address several limitations that were not addressed in this study.

15. **Define 'lightweight' ML earlier in the paper. Currently the definition is in Section 3.3 but the term is referenced in Section 2 as well.**

    Response: Thank you for your remark. It is indeed important to ensure that the definition is clearly stated and referenced earlier in the paper.

    Modifications: We have revised the paper accordingly to address this inconsistency. We moved the definition of 'lightweight ML' in Section 1.

16. **Can the authors also label the AC and DC components of the signal in Figure 1?**

Response: Thank you for your remark. We appreciate your attention to detail and your contribution to improving the visual presentation of our work.

Modifications: We have updated Figure 1 to include labels for the AC and DC components of the signal. This additional labeling provides clarity and enhances the understanding of the figure.

17. **The reference to [16] towards the end of Section 2 seems placed a bit late in the section. It should be moved up a few paragraphs earlier when all of the other previous studies are mentioned.**

    Response: Thank you for your suggestion. We agree that the reference to [16] would be better placed earlier in Section 2, along with the other references to previous studies.

    Modifications: We moved the reference few paragraphs earlier, along with other previous studies mentioned.

18. **Figure 16 needs an x-axis label.**

    Response: Thank you for pointing out the missing x-axis label in Figure 16. We apologize for the oversight.

    Modifications: We decided to move the figure to the Appendix due to space constraints. We updated the figure to include a clear and descriptive x-axis label to improve the clarity and understanding of the plot.

19. **Consider renaming LF/HF ratio since that has a very specific definition in the very related domain of heart rate variability.**

    Response: Thank you for your suggestion. We agree that using a more distinct and specific term would help avoid any confusion with the LF/HF ratio commonly used in the heart rate variability domain.

    Modifications: We used an alternative name that accurately reflects the concept we are referring to in our study while ensuring clarity and avoiding any potential misunderstandings.