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 1

1. **The manuscript reports experiments and analysis of lightweight anomaly detection algorithms photoplethysmography. The manuscript has a lot of descriptive content and it is also very long to comprehend the research work done on anomaly detection in multi-wavelength PPG technology. 26 pages are many for a paper that simply covers a few ensemble learning methods, such as decision trees, Random Forest, SVM, a custom SVM algorithm, and an autoencoder towards machine learning algorithms, and that are state-of-the-art for the use case, in a top quality journal such as Sensors! The F1 scores and accuracy for the custom SVM algorithm are impressively high for various feature sets experimentally investigated in the study and these results can be interesting in the context of use case that use hand crafted feature generation based classifiers. The authors needs to shorten the manuscript by focusing the main experiments with the selected anomaly detection algorithms and present the most striking results in the main paper to 16 pages. Anything that does not fit in 16 pages can be made available online as Supplemental Material online.**

   Response: Thank you for your constructive comment. There was indeed a need to shorten the text size and keep a balance between descriptive content and results.

   Modifications: We have successfully condensed the article to 17 pages by reevaluating its structure and focusing on highlighting the most salient contributions and striking results of our research. We included the rest of the content that we considered necessary as Appendix. We struggled to keep the limit of 16 pages, but it was necessary to address other comments from the other reviewers.

2. **In addition the English and presentation style of the manuscript should be improved, for example using some concept diagrams, more equations, and succinctly described methods.**

   Response: Thank you for your remark. Your suggestions will help us present the research findings in a more concise and visually appealing manner. Due to the increased text size, we understand the difficulty in fully understanding the content.

   Modifications: We worked on incorporating tables, additional equations, and clearer descriptions of the methods to enhance the clarity and understanding of the content. We tried to apply your suggestions whenever possible, aiming to succinctly describe the concepts and use more reader-friendly formulations.

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.

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 3

The study focus is to test effect of temporal window sizes, signal feature sets and different ML algorithms (DT, RF, SVM, and RCA) on accuracy and F1 of the anomaly detection in MW-PPG signals. The study is interesting and has applications especially in wearables where light weight and optimized algorithms are needed for PPG analysis.

The authors identified the optimal signal feature sets, window sizes for different algorithms and demonstrated that simultaneous analysis of multichannel signal outperforms single channel signal analysis.

The article is well written and includes all needed information to understand the study and its results. However, the article is a bit lengthy therefore I'd suggest to reduce it by moving the non-necessary content to Appendix or removing it completely. There are many repetitions, some figures are not needed, and some well known algorithms or approaches are unnecessarily explained in details. An interested reader without proper background would easily find needed information in the literature.

1. **P.1 par2: "other valuable information resides in the PPG pulse shape" Please write which additional info can be extracted from PPG signals (e.g. respiration rate)**

   Response: Thank you for your valuable feedback. We appreciate your suggestion to clarify the additional information that can be extracted from PPG signals. In addition to heart rate, respiration rate, and SpO2, blood pressure or blood glucose can be estimated. We included this information in the revised version of the manuscript to provide a more comprehensive understanding of the valuable information that can be derived from PPG signals.

   Modifications: We explicitly mentioned in Section 1 the respiration rate and other two markers that are a hot research topic nowadays.

2. **The descriptions of PPG artifacts, ML algorithms and evaluation metrices are very detailed, which is an advantage for a newcomer in the field. While they are unnecessary for experienced reader. I'd suggest the authors to reduce the section to the minimum necessary information, while the extensive descriptions can be moved to Appendix.**

   Response: Thank you for your feedback. We agree that providing a concise and focused presentation of the essential information is important. Since another reviewer had the same comment, we decided to comply with his suggestion, and we reduced the size of the manuscript to 17 pages.

   Modifications: We have successfully condensed the article to 17 pages by reevaluating its structure and focusing on highlighting the most salient contributions and striking results of our research. We included the rest of the content that we considered necessary in Appendix section.

3. **P.8 par. 4: RCA – write full name, "robust collaborative autoencoder"**

   Response: Thank you for your remark. Since we trained the autoencoder only on clean PPG windows, we decided to drop the RCA abbreviation. Both autoencoders and RCA are based on the same basic principle of encoding and decoding, but RCA introduces additional mechanisms to handle noise and corruption in the input data. In response to the confusion between RCA and the autoencoder, we apologize for any inaccuracies in our previous statement. We must clarify that we unintentionally mixed up the concepts.

   Modifications: We decided to drop the RCA abbreviation for the autoencoder.

4. **P9. Par. 3: Provide information about models and manufacturers, actual distances between LEDs and LDs, signal acquisition.**

   Response: Thank you for your feedback. This information will help to provide a more comprehensive understanding of the experimental setup and data collection procedures.

   Modifications: In Section 4.1, we will provide additional information regarding the models and manufacturers of the components used, the actual distances between LEDs and PDs, as well as details about the signal acquisition process.

5. **P9. Par. 4: "recorded data measured from two subjects" – Provide more information about the subjects (skin color, age, gender). Including only two subjects does not cover population variation, therefore the study results should be discussed considering the small sample size. A statement about the medical ethics committee approval is missing, since the study was performed on human subjects.**

   Response: Thank you for your comment.

   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).

   Modifications: TBA

6. **P.9 par. 4: Please provide information how the artifacts were introduced to the signal.**

   Response: Thank you for your valuable comment. In our experimental setup, we artificially introduced different types of artifacts by manipulating specific parameters during the data acquisition process. For instance, we introduced motion artifacts by intentionally moving the finger during signal recording. We also introduced contact force artifacts by applying varying levels of pressure to the sensor.

   Modifications: TBA

7. **P. 15: "Peak variance" calculation – How noise effects the peak location determination?**

   Response: Thank you for your comment. The temporal distribution of peaks can be influenced by the presence of noise, with the extent of this influence depending on the amplitude of the noise. Some types of anomalies (motion artifacts) might lead to increased variability in peak variance. Such anomalies cause unpredictable patterns in the occurrence of peaks. Contact force artifacts, or signal channel saturation, could lead to a decrease in peak variance.

   Modifications: No modifications are done regarding this comment.

8. **P. 15: "Low-Frequency/High=Frequency (LF/HF) ratio" – "=" symbol should be replaced by "-"**

   Response: Thank you for your remark. We decided to rename the feature since another reviewer suggested renaming the LF/HF feature to a more suggestive name that does not create any confusion with other concepts.

Modifications: We have updated the concept name to "Frequency band ratio," and it is consistently referred to by this name throughout the text.

9. **P.16 Fig 16: What is the x-axis? Is it a window number? What are the features in the graphs (peaks/dips)?**

   Response: Thank you for your comment. The x-axis represents the window number or, more specifically, the LF/HF ratio of a specific window.

   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.

10. **P.16: "Number of peaks" is not explained.**

    Response: Thank you for your feedback. We apologize for the lack of explanation regarding the term.

    Modifications: We modified Subsection 4.5 and included a clear definition of "Number of peaks" and ensured that it is properly explained.

11. **TSFEL and tsfresh feature sets are not explained**

    Response: Thank you for your remark. In response to the reviewer's request, we have included a table in the manuscript that provides a detailed explanation of all three feature sets used in our study. This table serves as a comprehensive reference, highlighting the key characteristics and components of each feature set. By including this information, we aim to enhance the clarity and understanding of the feature extraction process in our research.

    Modifications: In Subsection 5.1, we have incorporated a table that elucidates the three feature sets determined in our study, namely the custom feature set, TSFEL feature set, and tsfresh feature set. This table provides a comprehensive overview of each feature set, outlining their distinctive characteristics and components.

12. **P.17 par. 4: The "random forest" method is described two times, but once it should be the "decision tree" method.**

    Response: Thank you for your remark. Indeed, there was an incorrect reference to the "random forest" method instead of the intended "decision tree" method. We apologize for the confusion caused by this oversight.

    Modifications: In Subsection 5.2 we provide the appropriate correction to accurately reflect the parameters of "decision tree" method.

13. **P.21 Table 2: Explain the difference between custom feature set and the other two.**

    Response: Thank you for your comment. In response to the reviewer's request, we have included a table in the manuscript that provides a detailed explanation of all three feature sets used in our study.

    Modifications: In Subsection 5.1, we have incorporated a table that elucidates the three feature sets determined in our study. This clarifies the difference between the custom feature set and the other two sets.

14. **The authors should discuss effect of a very small subject group size (two volunteers) on the study results. How does it affect them? What would they expect if a darks or obese volunteer would participate?**

Response: Thank you for your remark. Indeed, it has been seen in previous research that inaccurate PPG measurements occur more frequently in dark skin as compared to light skin, and more in obese subjects. We believe that better sensor selection, optimal wavelengths, and sensor configurations should be studied.

In the current study, we are not focusing on detecting anomalies in the PPG pulse shape or disease-related anomalies but more general and subject-independent artifacts such as body gross movement, signal dropout, signal attenuation due to contact force or low temperature. Therefore, if such volunteers were included in the test set, assuming that the current sensor configuration is not performing well in measuring the PPG pulse from such subjects, we could detect this as an anomaly. This is because such samples would already be considered anomalies in the training set, either due to dropout, contact force, or low temperature, as observed in other subjects that have lighter skin and low adipose tissue.

However, we acknowledge the very small subject group size could represent a drawback due to the limited generalizability of the results. Anomaly detection algorithms rely on learning patterns and identifying deviations from those patterns. With only two subjects, the algorithm may not have enough diverse data to accurately capture the full range of motion artifacts and contact force variations that can occur across different individuals. However, we tried to leverage that by applying data augmentation methods, such as scaling, warping, and noise addition, to introduce different instances of artifacts based on measured data.

Modifications: We have put a large amount of thought into this remark and decided to add in Section 6 a limitation of this study.

15. **How the results of this study would impact the wearables design?**

Response: Thank you for your comment. We believe that the feature wearables design will shift towards a multichannel and a multi-wavelength (MW) approach. This shift can be due to the enhanced signal quality that is offered by a multichannel and MW-PPG system. However, by only increasing the number of measurement channels, without other channel fusion algorithms, this advantage cannot be explored. For instance, to make a PPG system more suitable for measuring dark skin tones and obese individuals, powerful signal processing techniques should be used (signal decomposition, independent component analysis). However, when applying these techniques, the assumption is that all channels contain relevant PPG information and that the motion artifacts are present in all channels to some extent. This thing might not be true for all the channels, and that is why we believe an anomaly detection stage would improve the overall accuracy of the system and avoid distorted signal reconstruction.

Moreover, for the current wearable devices that already employ MW-PPG for SpO_2, heart rate and respiration rate estimation, the classical signal quality index (SQI) can be replaced by an anomaly detection stage that takes into account all the channels and has an increased accuracy. Several studies have demonstrated that HR measurements from wearable devices are often less accurate during physical activity or cyclic wrist motions. We believe that an anomaly detection stage increases the confidence level of the extracted parameters.

Modifications: In Section 6, we presented a practical use case on how PPG devices can benefit from anomaly detection in MW-PPG signals.