**Rebuttal letter on:**

**Review of Contribution of Werth et al, Special issue of EHD**

Dear reviewers,

Thanks a lot for putting such effort in reviewing our manuscript in detail. We appreciate all the technical and methodical detailed question as they prompted us to check every step of our work again. We considered the majority of the mentioned points in our work as you will find by our answers. Nevertheless, for the mostly non-technical audience we left some details unmentioned. On many points we added additional information in the manuscript and your questions will generally be an excellent checklist for any further technical publication.

**Reviewer point 1**

In general this is a very nice article regarding HRV and sleep staging. However, we would like to question whether it would not be better to enlarge the review part of the article in favor of the study that is presented. The authors already proved to be able to write a good review regarding HRV before and I would be in favor to write another review like this as this was the scope of the special edition. Of course the study by itself is still interesting, however different issues need to be discussed before publication because it is a rather technical part in the article and there are different issues to be discussed as described further. We see two possibilities:

1. Enlarge the review part and describe the study as an example of using HRV in sleep studies , or
2. Restrict the EHD paper to the review part and publish the study elsewhere as a technical paper.

The remarks on the technical part of the study are hereunder (mainly remarks from the engineering department).

**Our rebuttal to point 1:**

*Generally we agree with the fact that this special issue is focusing on reviews rather than technical publications. Nevertheless, the original topic of our special issue part was exactly presented in our previous review on unobtrusive measurement techniques. Rather than copying the review and thereby self-plagiarize our work, we discussed with the Editor that a slightly more technical submission would benefit all parties.*

*As the special issue is intended not for a technical audience but rather medical, we focused on a more general description in combination with broader background explanations. Thereby, we omitted some technical details. After reading the remarks on the technical part we realized that we should reconsider to mention certain technical specifications.*

*To address the reviewers concern on the “review topic” we rewrote the introduction by adding a new review on automated sleep staging for term and preterm infants. This also includes a new Table (Table 2) with an overview of 12 papers on automated sleep analysis in preterm infants. .*

*The new part of introduction can be found on page 6. The new Table 2 can be found in chapter Tables and Figures.*

*As part of the new review we added a remark and citation to the co manuscript in this special issue of the Leuven group. We think this will enrich this overview part majorly as it becomes the most up to date review with publication. This will also increase the duration of actuality of the publication.* ***If this is acceptable is to decide by the Editor****. We kindly ask to consider our thoughts on this.*

*The part we are talking about reads as follows on page 9:*

“A more complete overview on sleep state classification based on EEG is given in another article of this special issue by Anneleen Dereymaeker et al. “*Review of sleep-EEG in preterm and term neonates*” [enter final title/ citation].”

**Reviewers points on the technical part:**

**Remarks on the technical part**

Since this paper is meant to be a review journal, it does not really fit in with this concept. This seems like a more conventional technical paper and while the authors are promoting the use of HRV for sleep state classification, which is an appropriate topic, there is no evidence of a review on this topic here. Even in the introduction, apart from a list of references, nothing is expanded on in the context of HRV and sleep staging, and even more importantly on its role in understanding more about early development. Can the authors elaborate more on this since this paper was not meant to be a technical paper?

Below we summarize our remarks on the technical part of the study hoping these remarks can help the authors when publishing their study as a technical paper. The authors are advised to publish their technical study in a more technical or biomedical engineering journal.

The authors study show how Active sleep and Quiet sleep can be separated making use of HRV measures only extracted from the ECG of only 8 normal preterm infants. The authors limit their study to standard HRV features as given by the task force for adult ECG and use standard methods for building their classifier focusing on binary classification only of AS versus QS differentiation.

Main weaknesses of the study are:

**Reviewers point 2:**

The dataset is too small to make general conclusions

**Our rebuttal to point 2:**

*We agree that the number of participants are not ideal for general conclusions. Therefore, we did not claim having created a ready to use algorithm but in our opinion these results show at least a foundation for sleep classification. We are also in the process of gathering more data for a more solid classification. Nevertheless, to stress that this is not yet a robust classifier we added some additions to the conclusion chapter.*

*On page 21 we added a new heading “Future recommendations” there we added following line:*

“For a stable performance based only on HRV and over a wide range in the preterm infant population, additional data is needed…”

*The conclusion chapter on page 22 reads now as follows:*

“This study shows that using a nonlinear SVM classifier approach for HRV features provides good results for preterm infant sleep state analyses of AS and QS. While our findings cannot yet be seen as robust due to the limited trial population, the classifier performance can compete with the literature [23,25,51]. Merging the different vital sign approaches e.g. respiration [12] and activity [66] with HRV will most likely lead to a robust, unobtrusive automated methodology for continuous preterm infant sleep monitoring.”

**Reviewers point 3:**

The features are based on the Task force for adult ECG and no attempts were made to optimize these for preterms. E.g. the LF and HF and VLF measures could have been adapted to differences in respiratory and heart rate in neonates, in order to better differentiate AS from QS. Although you use the Lomb periodogram, the frequency bands are different in literature for preterm babies (at least LF = [0- 0.2] Hz, HF = [0.2-2] Hz) due to the high respiratory rate.

**Our rebuttal on point 3:**

*The reviewers are right on the rather generic frequency ranges. We did a small literature survey beforehand and came to the conclusion that to date no general consent on preterm infant frequency ranges is established. As we could not investigate deeper into that matter at that point, we stick to the Task Force which was updated for use in term babies. Regardless, it is true that it is not optimal to disregard specifically preterm frequency ranges and therefore we added two additional frequency ranges. pHF1 (0.45 – 0.7Hz) and pHF2 (0.7- 1.5Hz). The feature total\_power and connecting features were also extended to this range.*

*It reads now in HRV features on page 10:*

“We selected 20 HRV features combining 18 commonly used features in adult sleep analysis [14,16,30] with two preterm infant specific frequency domain features. The adult frequencies are mostly divided into very low frequency (VLF), low frequency (LF) and high frequency (HF). We refer to Table 2 for details. To create the two additional premature frequency ranges we extended the HF feature to three HF features: HF, pHF1 and pHF2 with a standard adult frequency range from 0.15 to 0.4Hz for HF, and additional ranges of 0.45-0.7Hz for pHF1 and 0.7-1.5Hz for pHF2 [31] to accommodate the increased cardiorespiratory rates in preterm infants …”

*In the discussion on page 17 it reads now:*

“… The features we chose are derived from adult sleep analysis [49,50] as the main objective was to proof the feasibility of separating sleep states in preterm infants based on HRV. We also added two additional features with increased frequency ranges to accommodate the general higher cardiorespiratory rates in preterm infants. The new used frequency ranges for preterm infants (See **Table 1**) were proposed in 2008 by Indic et al. [31]. The new feature pHF1 consistently appeared in the top feature subsets (#2 onwards, AUC: 0.866) underlining the importance and influence on sleep state separation and also the assumption of Indic et al. that higher cardiorespiratory rates in preterm infants need adapted frequency ranges seems to be valid. This is not surprising as other groups already explained [6] and demonstrate [12,51] the benefit of respiration analysis for sleep staging.

In the top feature five feature subsets, which all resulted in comparable performance (AUC: 0.85-0.87), total power and SDNN were always present. pHF1, NN20, NN30 and pNN20 were present three out of five times. The time domain feature BpE appeared only in the top feature subset.”

*And* in the chapter Methodological limitations on page 20:

“…. As the neonatal heart and breathing rate differs from adults the recommendations of the Task Force may not be applicable in preterm infants. However, until recommendations for neonatal standardized analytical methods are made, many fetal and neonatal studies uses the recommendations of the Task Force [64,65]. We followed this approach by using mainly the adult recommendations from the Task Force. Nevertheless, to account for the increased cardiorespiratory rates in preterm infants we used two additional frequency domain features, pHF1 and pHF2 [31].”

**Reviewers point 3:**

The authors’ method has very limited applicability since it can only be applied on labelled AS and QS epochs. Apart from all difficulties in labelling the epochs the method cannot be used in clinical practice. For automated sleep state classification, more classes need to be identified. In addition to AS and QS additional classes such as wake sleep, REM Sleep, indeterminate sleep,… Why did the authors not consider AS versus non-AS as goal? Or QS versus non-QS?

**Our rebuttal to point 3:**

*We understand the concern of the reviewers and we have to agree but also disagree at the same time with the reviewer’s point of view. We agree, that a holistic algorithm in a clinical setting needs to run stable on all possible stages or at least minimize the false positive results due to false classification of minority classes such as wake. Nevertheless, we want to point out that this study was meant to showcase that the majority classes AS and QS can be separated based on the rather unobtrusive measure HRV alone. In addition, we explained in the introduction the importance of AS and QS for brain development determination. We focused on development and on determining phases where the infants should not be disturbed (QS) to allow an optimal environment for undisturbed development. This justifies an algorithm focusing only on AS and QS. In the manuscript we only constitute our work as a foundation for use in clinical practice. We do not claim to have a clinical ready algorithm.*

*We do have the classification between all states on our agenda for future research. To address these points we added a few sentences to the conclusion and the future recommendations.*

*It reads now on page 21:*

“… For possible clinical application and a more holistic view on preterm infant sleep, classification between all states should be considered….”

**Reviewers point 4:**

The authors do not take into account any differences in maturation of the sleep states (QS and AS and other states) with gestational age. It is clear that the sleep state characteristics for the early preterm infants (less than 30 weeks GA) clearly deviate from those of the moderately preterm infants (32-35 weeks GA) and the late preterm infants (34-36 weeks GA) . These differences in maturation are expected to influence the performance of the classifier. They included a factor C to compensate for age, however a thorough discussion is missing.

**Our rebuttal to point 4:**

*The reviewers are correct, with their statement about change in sleep state characteristic. What is changing is the sleep pattern and the deviation in the features itself (considered in the point below). The change of sleep pattern as of percentage of total sleep per state is what we are aiming for in the long run. We tried to take this into account with the mentioned adapted C value pairs. Due to the medical audience we did not specify more detailed which C value pairs we used as this would not generate any benefit for the clinical audience. Please see also the comments further below.*

**Reviewers point 5:**

The authors do not take into account any differences in maturation of the ANS and CNS with gestational age. It is clear that the HRV features for the early preterm infants (less than 30 weeks GA) are expected to deviate from those of the moderately preterm infants (32-35 weeks GA) and the late preterm infants (34-36 weeks GA). These differences in maturation and their effect on the selected features are expected to influence the performance of the classifier. Any discussion is missing.

**Our rebuttal to point 5:**

*This is unfortunately true. As we mentioned in the discussion, due to the small dataset it was not viable to create different feature sets for different age clusters including altering feature modalities. Nevertheless, it is generally more desirable to create an algorithm independent on age or developmental status. The closer an algorithm is fit to a specific class the more error prone it becomes.*

*To make that point more pronounced we edited the part of the “Discussion” in “Sleep state distribution. This is also related to another point of the reviewers further down about the state distribution clustering.*

*It reads on page 18 now as following:*

“The age clustering for compensating changing state distributions increased the performance. Nevertheless, while the performance improved for most subjects (4) some did not change (2) and one patient actually decreased. This could be interpreted as hitting the right cluster improves the classification while a false clustering can lead to a decrease of performance. In our opinion, only clustering on age is the reason for the possibility of a decrease in performance. Commonly, it is postulated that sleep state change over gestational age [7]. This is certainly correct, but age is only one indicator for neural development. The coupling between age and development (sleep state distribution) can change with neural miss-development. Therefore, we suggest that the correct clustering has to be determined and based not only on age but rather overall condition including weight and size and other biomarkers.

Nonetheless, also biomarkers which are not directly linked to neural development can give miss information and potentially create false sleep staging which consequently would lead to false results of development monitoring.

Generally, it can be said that sleep state separation without additional background information should be aspired.”

*And on page 20 in Methodological limitations:*

“…Note that the methodology of gestational age clustering in this study was used only for consideration of the sleep state distribution, not for age determined training sets. In our case age clustering for different classification was not feasible due to the small dataset. Clustering the data would have further decreased the training data…”

*Also on page 21 in Future recommendations we added:*

“… Also, if enough data is at hand we suggest clustering by age in combination with developmental biomarkers….”

**Reviewers point 6:**

The authors used 3 annotators for labelling the sleep states. However, they only kept those epochs where at least 2 annotators agreed and throwed away more than one fifth of the available epochs. Which ones were thrown away? The datasets might easily be biased since the discarded epochs are not selected ad random. This questions the generalization properties of the derived classifier. Moreover, the interrater agreement was only moderate for QS labelling. The authors could have included the interrater uncertainty into the labelling of their datasets. The authors are recommended to perform a statistical analysis of the discarded epochs in order to deduce systematic properties of the removed epochs.

**Our rebuttal to point 6:**

*Yes, are aware that we discarded data points which are not used for training and testing. To appoint this we added a sentenced about the introduced biased to good epochs in the discussion.*

*It reads now on page 20:*

*“As we necessarily had to discard certain epochs of disagreement in the annotation, we biased the classification in favor of good epochs. Nonetheless, we want to point out that in this study we meant to showcase that the majority classes AS and QS, which are most important for development monitoring, can be separated based on the rather unobtrusive measure HRV alone. Future studies will have to attend the integrated problems of sleep staging.”*

*Nevertheless, the discarding was necessary as partly only two observers would have annotated the data. When they disagreed we had to discard the epochs as we use a supervised learner. We also strongly agree on implementing observer uncertainty. Unfortunately, we did not have consistent numbers of observers annotation per epoch to be able to create an epoch wise stable uncertainty value. As the QS is already the minority class it is not advisable to add additional counter weights on AS due to uncertainty. We are at the moment working on an uncertainty labeling for each epoch. We hope to implement this in future work. We want to thank the reviewers for mentioning that point. We addressed this in future recommendation.*

*The added part on page 21 reads:*

“…For future work we also suggest to increase the stability by using the uncertainty of the observer annotations, especially for transition states which are often difficult to manually observe.”

**Reviewers point 7:**

In addition, the paper contains several typing errors and incorrectly formulated sentences. The authors are advised to thoroughly screen their manuscript and improve the language where needed. E.g.:

p.15, line 5 from bottom: rewrite ``Nevertheless, …. Dataset’’.

p.15, line 3 from bottom: correct ``Neural misdevelopment …’’

p.29, caption Fig.1: rewrite ``Quiet sleep …published in [7]’’

p.20, caption Fig.2: correct typing errors such as iteration, chosen, performance, method,…

p.30, caption fig.3 and fig.3: ROC means ``Received-Operating-*Characteristic*’’, please correct. Also, add hyphens

**Our rebuttal to point 7:**

*We want to thank the authors for pointing out incorrectness in language. We corrected the mentioned parts and screened the document. We suggest that, if accepted, we could send the final document to a professional institute for a language check.*

*One small remark from our side is that it is called Receiver Operating Characteristic.*

**Reviewers point 8:**

In addition to the general comments above, the methodology is unclearly explained. In each step of the approach, questions arise. In particular, more specific comments are given below:

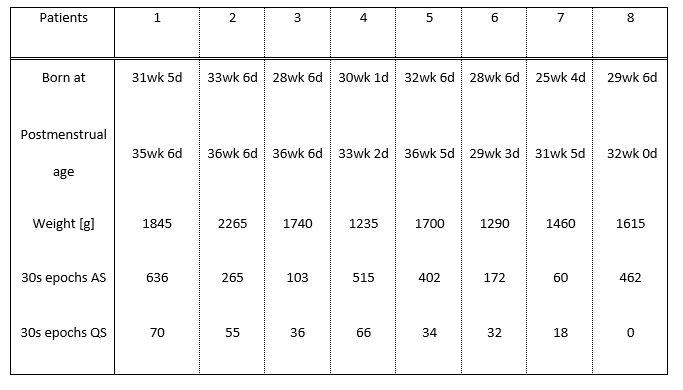
Annotation:

Only 8 recordings were included. It is known that QS/AS patterns as well as the ANS characteristics differ according to the GA. It is unclear whether differences in maturation might influence the results. To clarify, add a table with individual demographic characteristics of the 8 infants. The datasets is unbalanced: 2617 AS epochs versus 312 QS epochs. It is unclear how this imbalance has been taken into account in the construction of the classifier.

**Our rebuttal to point 8:**

*Actually, we had the patient information in a table but during our internal discussions decided against it as our neonatologist did not found it to be of interest for the clinical audience. The necessary information is given in mean age and standard derivation as well as epoch distribution, both on page 9. In addition, we have a limitation on tables which seems to be reached with the addition of the automated sleep staging table (Table 2). We see that this could add more clarity, especially for more technical readers. If the editor in Chief decides on adding this table to the appendix, we can provide it.*

*The table could look like the following:*



*The dataset is indeed unbalanced as expected as there should be more AS then QS. That is why we choose the SVM classifier as it is relatively robust against imbalance. We used the C value adaptation to level for the imbalance as we knew the expected imbalance of the developmental stage. Due to added features we could increase the performance especially via the age clustering.*

*On page 12 in Material and methods it reads now:*

“As the data is unbalanced for AS and QS states the C value can additionally be multiplied with value pairs representing the distribution of a dataset to compensate for the class imbalance. In preterm infants the class distribution change over time and development, therefore, we clustered the participants into four different GA ranges and computed the class weight pair for each cluster not on the actual data distribution per cluster, but on the gross expected distribution known from literature [7] as seen in Figure 1 . The subjects were ranked into the following clusters in GA: 31-32wk (n=2), 33-34wk (n=1), 35-36wk (n=2) and 37-39wk (n=3). This clustering should not be mistaken for an age clustering to create different feature/training sets which is not feasible due to the limited population size”

*Also in the Result paragraph on page 14 we added:*

… As the sleep states were unbalanced we adapted the classifier parameter C with weighting factors pairs calculated from the expected distribution which grossly correspond with the four age cluster (see Material and Methods). Using the different class weights we achieved a better performance of the sleep staging with an AUC of up to 0.87± 0.42 (Figure 4).The optimal subset features resulted from the wrapper analysis for both classification types were BpE, NN20, SDNN, pNN20 and total power.”

*In Discussion on page 16 it reads now:*

“…The classifier performance with receiver operating curves resulted in a mean value for the AUC of 0.85 (Figure 3) and under consideration of sleep state distribution an AUC of 0.87 (Figure 4), indicating that HRV features are valuable for automated sleep state separation in preterm infants.”

**Reviewers point 9:**

Reliable and robust R-peak detection is very important for HRV analysis and feature extraction. The authors used an in-house algorithm described in Ref. [21]. However, Ref. 21 focuses on PPG signals which are different.

What about apneas and possible non-stationary transitions in the RR signal?

**Our rebuttal to point 9:**

*We did not consider apneas in the R-peak detection/feature creation due to several reasons. First the definition of apnea is still not clear. The definition changed form 1956 with 2 min of cessation of breathing to 20 seconds including bradycardia in 1978 leading to a definition today with >10s plus bradycardia and desaturation.   
Secondly, the detection of apnea is still an unsolved problem which several groups are working on. For us it was not feasible to incorporate the latest apnea detection algorithms to rule out false R peak detections. Surely, we integrated some errors in the R peak detection by not considering this. This just means that we have more room for improvement in the future when the apnea detection problem is solved.*

**Reviewers point 10:**

Please provide more details on the methodology used here for R-peak detection.

**Our rebuttal to point 10:**

*To achieve an accurate and robust R peak detection for preterm infants we tried several different in-house and external R peak detectors. Several once specifically made for preterm infants. To our own surprise the R peak detector of Wijshoff et al. ,while not intended for this patient population ,worked most accurate and robust even in extremely noise distorted signal sections. Maybe it would be an interesting idea for Ralph Wijshoff to publish a comparison study of his algorithm on different populations. We will suggest that to him.*

*As good R peak detection is worthy of an own publication and has been done in the past, we did not want to dive too deep into that topic. Still, as we fully agree that a proper R peak detection is essential for the base of HRV analysis we added slightly more information on choosing the R peak detector.*

*The passage on page 10 reads now as follows:*

“From the ECG signal the R peaks were detected with an in-house developed algorithm [29]. As to date there are no validated preterm infant R peak detection algorithms, the in-house algorithm was chosen after an extensive manual comparison between several R peak detection algorithms and bio signal toolboxes as to date there are no validated preterm infant R peak detection algorithms. Subsequently, the R-R intervals were used for HRV feature analysis.”

*And we added a short comment to the end of future recommendation on page 21:*

“Finally, as the R peak detection is essential for correct classification with HRV, we suggest to investigate and validate R peak detection algorithms specifically for preterm infants. To our knowledge, no validated algorithm of that kind exists to date.”

**Reviewers point 11:**

A sliding window of 1, 2.5 and 5 min is used. It is unclear why these lengths are selected. Please, explain. Also, the authors centered the window around each 30-s epoch. Each window is 2 to 10 times larger than the center 30-s epoch and can easily include QS and AS and other epochs. However, the derived features are labelled according to the center epoch although they pick up a lot of information outside the epoch of interest. How does this influence /affect the performance of the AS/QS classifier? How do the authors handle the epochs close to the boundaries of their dataset? How do the authors correct for discarded epochs to be considered as missing values? Please, explain clearly.

**Our rebuttal to point 11:**

*The reviewers are correct on this. The labels can overlap. We did a boundary decision on a majority vote with threshold 50%. Thereby, the overlapping can be maximal 2.5min. Considering the duration of each sleep state and the more gradual change of states, this overlap can be seen as smoothing or low pass filtering. This will decrease the performance only minimal. As this is a signal processing specific problem we did not include it in this more clinical manuscript.*

**Reviewers point 12:**

HRV features:

In addition, the values of LF,HF and other features change substantially among the different time windows. It is odd that the values for the features show no statistical significance between different window lengths when they seem at first sight to vary quite a lot. Can you comment on that? How is it possible there is no statistical difference among the three?

**Our rebuttal to point 12:**

*This is due to the dependence of the summation used in features which links them to the window length itself. Nevertheless, as this was confusing, unnecessary information we discarded most of the information and present now a new condensed Table 3. As we anyway only considered 5min windows we discarded all other window sized.*

*In the Result section on page 14 it reads now:*

*“…We used 1, 2.5 and 5 min windows to investigate effects of window length effects in preterm infant sleep staging. The combination of different window sized did not increase the classifier performance. In contrary, the analysis of only 30s epoch based HRV features lead to a lowered sleep staging performance of 0.71. Hence, we used the 5-min window length for sleep state separation and classifier methodology recommended by the Task Force [30].”*

**Reviewers point 13:**

Feature selection:

Simple selectors such as CFS and FSMC are combined only with a linear classifier while a more sophisticated mixed-wrapper method is combined with a nonlinear classifier. The authors conclude that sleep separation with a linear kernel is not possible, however this conclusion may be due to the use of too simple feature selection methods. Did the authors combine mixed-wrapper feature selection with a linear SVM? Why not trying a state of the art method such as Linear Discriminant Analysis? Please, comment and better motivate the choice for a nonlinear kernel.

**Our rebuttal to point 13:**

*This is actually a good point. We did not considered the use of the more complex wrapper including the thorough brute force approach for a linear kernel. For a following technical paper we will definitely consider this approach to be able to compare the linear and nonlinear classification in all detail. For this publication this comparison would exceed the scope. As mentioned earlier the SVM was chosen due to its good performance for unbalanced data. LDA is definitely another option and will be considered in future work.*

**Reviewers point 14:**

RBF tuning? How did you define sigma squared?

**Our rebuttal to point 14:**

*We tried two additional common nonlinear kernels (poly and hyperbolic) which did show lesser performance. We did not mentioned this but instead just stated the chosen kernel considering the audience. The sigma square was found with a grid search. It was again decided to leave this out as an explanation for sigma squared would be necessary which again would not be of interest for the target audience. As we agree with the reviewers that this value is necessary for a complete picture we added now the ½\*s^2 value with a minimal explanation.*

*On page 12 it reads now:*

“The applied SVM used the radial basis function (RBF) kernel for non-linear approximation. The RBF kernel also uses a free parameter γ which determines the inverse influence of the support vectors. We used a γ value of 0.2.”

**Reviewers point 15:**

The SVM classifier operates using distance metrics, did the authors normalise/standardize the features before using the classifier? They presumably did, but it should have been mentioned in the methods.

**Our rebuttal to point 15:**

*Right, we forgot to mention the normalization step. Thanks for pointing out. We added this.*

*Feature selection, page 11 reads now:*

Supervised learning for classification with a support vector machine (SVM) needs a selection of normalized input features to avoid overfitting resulting in a decrease of performance….

**Reviewers point 16:**

Can the authors also motivate their choice for the mixed-wrapper as feature selection method? Why is a brute-forward search needed as first step? Why not applying a forward selection search directly? Or a backward selection search? Or a stepwise selection search combining a forward and backward selection search step? These methods are more commonly used and are controlled by the selection criterion used to remove or add a feature. Which selection criterion did the authors use in their paper? Please, add more details or references. Furthermore, the authors should better clarify how they selected after the brute force forward search ``the 10 remaining generation of subset combinations’’ as input to the next step? Finally, can the authors argue why their method is still optimal?

***Our rebuttal to point 17:***

*The reviewers have a good point here. Other methods are more common than the brute force method. We chose the brute force method as we already expected difficulties due to the small data size in addition with the discarding of epochs. We wanted to compensate this by using the most complete method for the subset search. Nothing beats a brute force method (not talking about computation time) to find the truly best feature subset up to a specific point. The method is still optimal as we have not too many features as to blow up the computation time. In the future we are planning to add more features such as e.g. breathing features. Then we definitely will have to switch to another approach. The criterion for the next best feature is the increase in AUC of the SVM test output. This are all valid and good questions. Nevertheless, they will not fit to the scope of audience. We will certainly discuss this in future technical publications. And we want to thank the reviewer again to have brought this to our attention.*

*On page 12 it reads now:*

*“*The remaining 10 generation of subset combinations were analyzed with a sequential forward search (best first), where only the next single best feature was added to the previous subset combination (105 iterations).”

**Reviewers point 17:**

Classifier:

SVM is chosen and the C value is experimentally set to default value 1. How exactly was this done? This is quite an important choice of value to not expand further upon. Did the authors try to optimize the hyper parameters? Is imbalance of the datasets taken into account?

**Our rebuttal to point 17:**

*The C value was found by a parameter search from 1\*10^-5 to 1\*10^5. With the new implemented features, we also dd another grid search with a new value for C of 3.6.*

*‘Page 12 reads now:*

“For this study the C value was set to value 3.6 after a parameter search as a good balance between speed, accuracy and generalization.”

**Reviewers point 18:**

The authors apply a LOOCV which is fine if all the folds are similar. However, any information is missing here. Please, add.1

**Our rebuttal to point 18:**

*The reviewers are right that the folds should be of equal length. Due to the nature of the data the folds are not of equal length but as we train on x out of x+1 folds, the length of the training set changes only marginally (std of the exchanged fold) which is again permitted. We could have used same length folds by using classic n fold cross validation but for our small dataset it was more important that we separate the patients data to avoid biased by training and testing on the same physiological background. As this is in, our opinion, a too detailed information for physicians we will leave it out while of course consider it for a technical publication. We want to thank the reviewers to make us aware of this point.*

**Reviewers point 19:**

**RESULTS**

Using a factor C, the authors try to take imbalance in the sleep states into account. This imbalance corrects for differences in GA leading to differences in distribution between QS and AS epochs. It is unclear how to the authors deduced this optimal C from figure 1. How did the authors quantify classifier stability? How did they tune C? Please, discuss better the relation with age clustering.

**Our rebuttal to point 19:**

*Indeed the reviewers are right as we did not have any method to quantify the classifier stability. Therefore, we deleted the part about stability.*

*The text in the discussion on page 18 reads now:*

“… To increase the classification performance the misclassification penalty was calculated on the class distribution based on the distribution cluster belonging to the GA…”

*As it appears to be not clear how we incorporated the state distribution over age into the classification, we added additional information in the introduction.*

*As mentioned earlier, on page 12 it reads now:*

“As the data is unbalanced for AS and QS states the C value can additionally be multiplied with value pairs representing the distribution of a dataset to compensate for the class imbalance. In preterm infants the class distribution change over time and development, therefore, we clustered the participants into four different GA ranges and computed the class weight pair for each cluster not on the actual data distribution per cluster, but on the gross expected distribution known from literature [7] as seen in Figure 1 . The subjects were ranked into the following clusters in GA: 31-32wk (n=2), 33-34wk (n=1), 35-36wk (n=2) and 37-39wk (n=3). This clustering should not be mistaken for an age clustering to create different feature/training sets which is not feasible due to the limited population size”

*Also on page 15 in the Result part it reads now:*

“…As the sleep states were unbalanced we adapted the classifier parameter C with weighting factors pairs calculated from the expected distribution which grossly correspond with the four age cluster (see Material and Methods). Using the different class weights we achieved a better performance of the sleep staging with an AUC of up to 0.87± 0.42 (Figure 4).The optimal subset features resulted from the wrapper analysis for both classification types were BpE, NN20, SDNN, pNN20 and total power.”

**Final note to the reviewers**

We hope we could answer all questions and remarks to your full satisfaction. Your input actually raised a lot of intense questions which strengthened our view on this topic but also lead to a better performance output as we implemented suggested items such as two new frequency features. We have to thank you for this support through tough criticism.

We are looking forward to good future collaboration on preterm infant sleep analysis.

In representation of all authors and with best regards,

Jan Werth