General: Thank you for your feedback. The main motivation/intuition behind our approach is the excellent performance we achieved with U-Nets for image segmentation. The theoretical/conceptual contribution is to exploit the similarity between sleep staging – and time series segmentation in general – and image segmentation. Still, U-Time is not simply a 1D U-Net (e.g., because of the *segment classifier*, dilated convolutions, normalization, NN-upsampling, ...).

We did not run our DeepSleepNet implementation as a baseline on all datasets, because of lacking compute resources and, given the extensive experiments reported in the supplement, because we were (and still are) sure that U-Time would clearly outperform it. We now understand the value of explicitly stating the additional baseline results and are currently running the corresponding experiments. We would like to stress that we are comparing a system developed on a single design dataset (S-EDF-39) across several datasets. In contrast, the baselines come from papers that propose one system for one dataset. Thus, the cited results are prone to (unintentional) method overfitting. This is why we regard our results as a significant advance to automated sleep staging resaerch and clinical practice.

The N1 class is hard because it is rare and difficult to clearly separate from other classes, even for human experts as can be seen in Table 1. N1 lies between wake and deeper non-REM sleep, and the transitions are gradual.

Rev. 1: Yes, we used only one channel for a fair comparison to existing literature. We are currently evaluating the use of multiple channels and would be happy to report results in the supplement. We will include a supplementary figure showing an examplary confidence score (softmax) output for a $C = 3$ multi-channel input, see Figure 1 below and our response to Rev. 2.

Details of the layers can be found in the supplement Table S.2. We will add another, larger figure illustrating the U-shape and the three parts of the architecture. We will make the code available for maximum reproducibility. The classes in our application are already very imbalanced, so U-Time may also work for very rare events.

Rev. 2: We are happy to restructure the manuscript according to your suggestions. Regarding confidence scores, qualitative analysis, and interpretability: It is particularly interesting to inspect the outputs of U-Time when the (freely adjustable) segmentation frequency is set to match to the input signal frequency, see Figure 1 below. The sleep stage scores indeed show human interpretable patterns even on short timescales. We believe that this special property of U-Time will allow for a better analysis of sleep stage transitions in healthy and diseased populations. We will discuss this in the main text including cases where the model fails to predict the true sleep stages and add the figure below to the supplementary material.

AASM stands for *American Academy of Sleep Medicine* (see line 129, sorry, we forgot to add the abbreviation). T is the number of fixed-length connected segments (each typically 30s) input to the model (line 68), which we will recall in line 169. B is the batch size, which we will introduce properly.

Rev. 3: Regarding architecture choice: The U-Net part of our architecture is based on an architecture that worked extremely well in medical image segmentation across a wide range of problems. There is a paper at the upcoming MICCAI showing this, which we cannot cite without revealing authors of the current submission (we can share a preprint with the area chair). In our submission (lines 91–97), we discuss the sizes of the kernels in relation to the time windows the layers see – an important design criterion. We selected them based on our physiological understanding of sleep staging. We regard it as a big advantage that we did not extensively tune our architecture to the tasks. The fact that this was not necessary demonstrates the soundness of the basic approach and the robustness of the implementation. It is important for us that our results are not artefacts resulting from (unintentional) overfitting through architecture / hyperparameter tuning. In contrast, as shown in the supplement, we unsuccessfully tried to tune competing methods to reach the performance of U-Time. Tuning the U-Time architecture and hyperparameters may improve the results. However, we assume that adding different channels or other input modalities is more important.

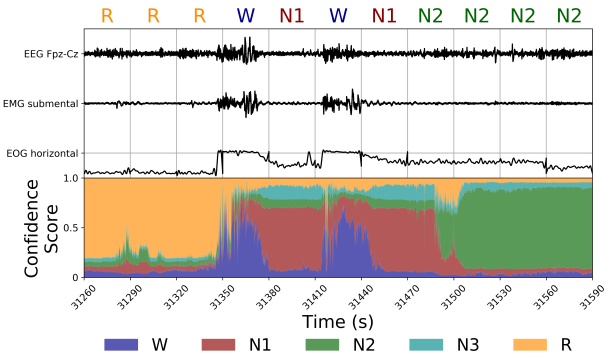

Figure 1: U-Time confidence scores (softmax output) over $T = 11$ segments (30s each) for $C = 3$ input channels (EEG, EMG and EOG). The freely adjustable segmentation frequency is set to match the input signal frequency.

Confusion matrices (CMs) for U-Time on all datasets are given in the supplement, and we are happy to add CMs for other methods. We will improve the colour palette as suggested. See the new figure above for an illustration of the confidence scores (which we will explain better in the main text). We used 5-fold cross-validation on the design dataset to choose between two loss functions (cross entropy and the dice loss).