Physician and AI in the loop:

Enhancing performance of automated sleep scoring algorithm using uncertainty estimates.

M. Bechny[1,2], J. van der Meer[3], M. H. Schmidt[3], C. L. A. Bassetti[3], A. Tzovara[1,3], F. D. Faraci [2]

[1] University of Bern, Institute of Informatics, Cognitive Computational Neuroscience group (CCN), Bern, Switzerland;
[2] University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Department of Innovative Technologies (DTI), Institute of Digital Technologies for Personalized Healthcare (MeDiTech), Lugano, Switzerland;
[3] University of Bern, Bern University Hospital - Inselspital, Department of Neurology, Sleep Wake Epilepsy Center | NeuroTec, Bern, Switzerland

UNIVERSITÄT BERN

Scuola universitaria professionale della Svizzera italiana
SUPSI

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HÔPITAL UNIVERSITAIRE DE BERNE

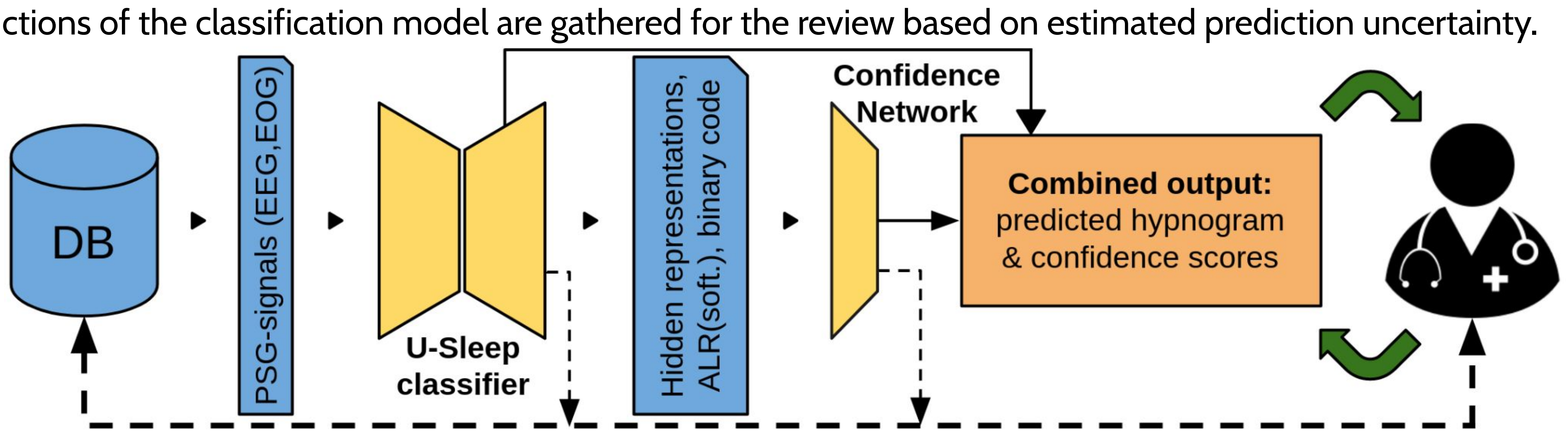
Introduction

Deep Learning algorithms are the most promising solution to support physicians in the time-consuming task of sleep scoring. However, they can hardly outperform the interscorer agreement of about 82%[1] as a consequence of the human-introduced noise in the labels used for training these algorithms. As a result, physicians have to review and eventually correct all the predictions they disagree with. In our research, we optimized approaches to assess prediction uncertainty in order to implement an efficient pipeline where physicians can interact with the scoring algorithm and review the most uncertain predictions.

Pipeline

It may take up to 2 hours for an experienced physician to score a single night of sleep into 5 stages: Wake, N1, N2, N3, REM. Deep Learning algorithms help to make this task more efficient as they provide accurate predictions. ***As the performance of scoring algorithms can hardly surpass data quality - and about 20% of interscorer disagreement can be expected for hypnograms - the physicians have to subject the proposed predictions to a careful revision.*** Our research addresses this challenge and implements a pipeline where the least-confident sleep stage predictions of the classification model are gathered for the review based on estimated prediction uncertainty.

Our pipeline (on the right) consists of two main parts: (i) the U-Sleep[2] classifier, augmented with (ii) the LSTM-based confidence neural network[3] trained to estimate prediction uncertainty using hidden representations extracted from (i). The confidence score is available for each sleep epoch and can also be averaged for each patient, providing subject-wise confidence.


DB

PSG-signals (EEG,EOG)

U-Sleep classifier

Hidden representations, ALR(soft.), binary code

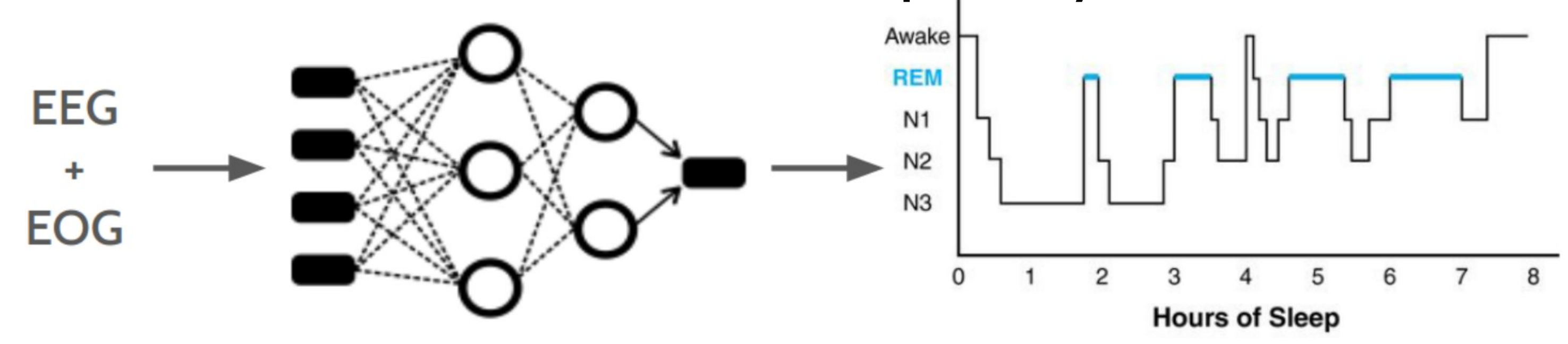
Confidence Network

Combined output: predicted hypnogram & confidence scores

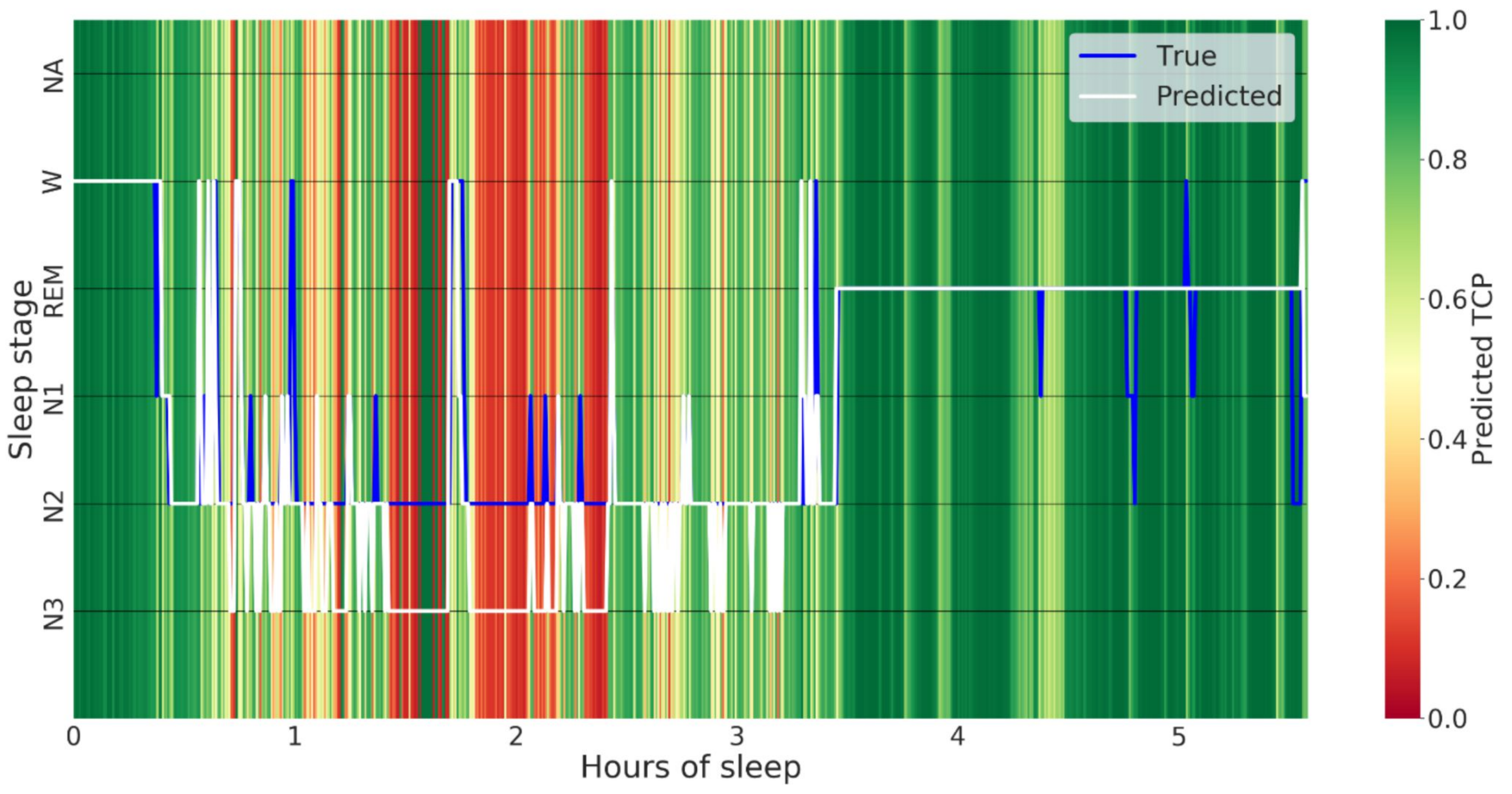
Methods & results:

Automated sleep scoring

U-Sleep[2], a state-of-the-art deep learning based sleep scoring algorithm, has been adopted. The architecture was trained on 12 open-access databases and fine-tuned on the Bern Sleep Data Base (BSDB) covering a broad spectrum of sleep disorders. **The U-Sleep reached (82.5, 82.8, 75.0)% in accuracy, weighted F1, and Cohenn's κ on the BSDB test data, respectively.**



Confidence-supplemented hypnogram

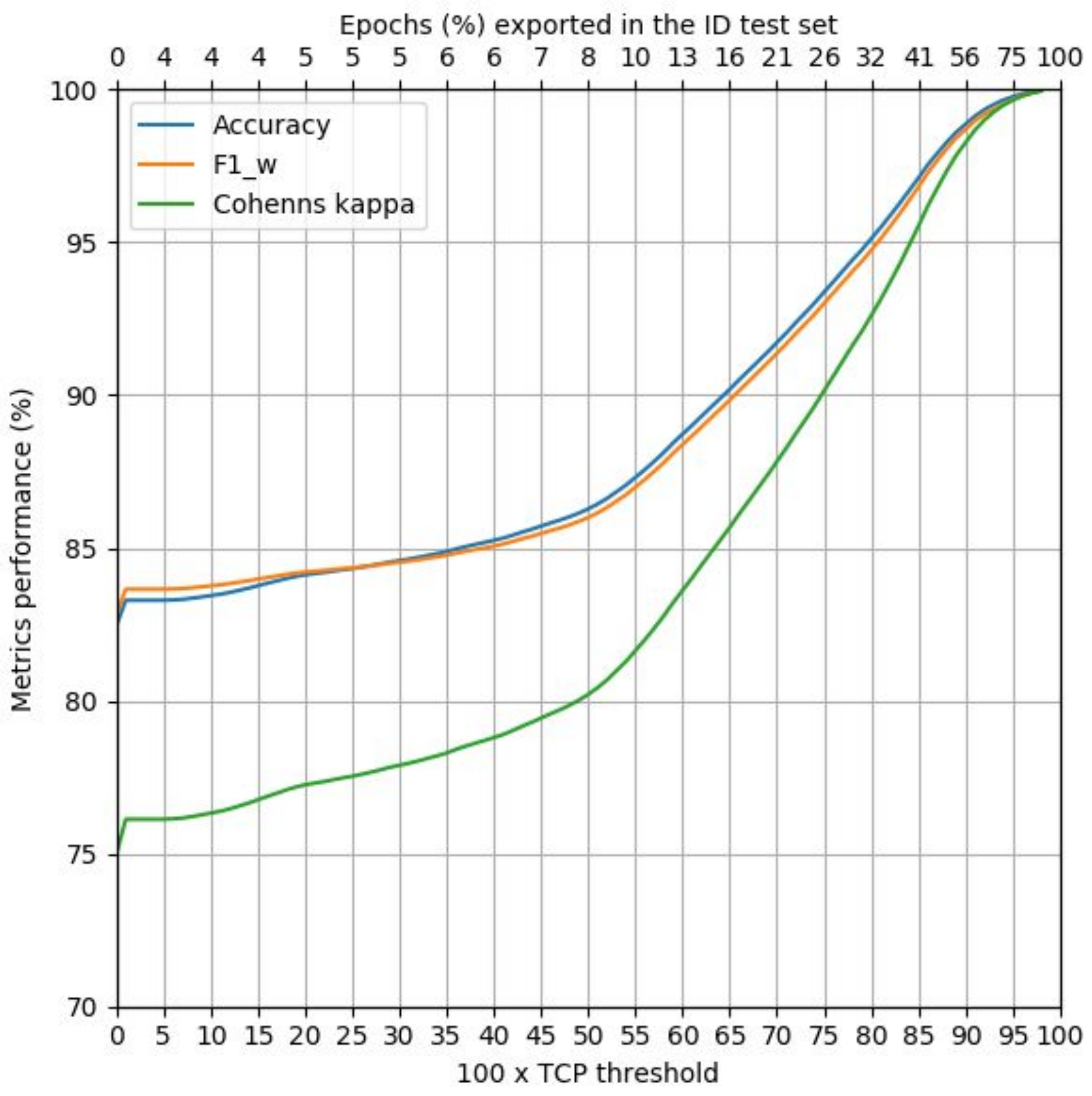


Uncertainty estimation

To estimate the prediction uncertainty, we compared several approaches based on transformations of the U-Sleep-softmax (e.g., entropy, max) with an *auxiliary LSTM-based confidence neural network* trained on hidden representations extracted from U-Sleep and estimating the so-called *TCP score*[3]. To evaluate the best approach of confidence estimation, we considered uncertain predictions as those being misclassified by U-Sleep and we compared various confidence metrics in terms of their ability to identify them. **The TCP score reached AUROC 85.7% and AUPRC 63.1% and outperformed all the softmax-based approaches.** The second best approach - softmax-maximum - reached AUROC 76.5% and AUPRC 42.9%.

Performance boost

We have evaluated the positive impact of our pipeline in gathering uncertain predictions for physician review based on different distributional thresholds of the confidence score. **For example, one can expect that by using the TCP-threshold of 0.8, about 1/3 of predictions will need to be inspected, which, when corrected, leads to about 95% in accuracy and weighted F1-score and to about 93% in Cohen's κ.**



Utility of the approach

The statistical analysis has shown that the estimated TCP-score is significantly higher for the correctly classified epochs, and also, that the on-subject average confidence score positively correlates with the on-subject classification performance achieved. Both, irrespective of subjects' sleep-disorder status. These findings suggest that our pipeline can be used to gather both, likely-misclassified sleep stage predictions, but also subjects on which U-Sleep performed with difficulties.

References

1] H. Danker-Hopfe et al., *Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard.* Journal of sleep research, 18(1):74–84, 2009.
2] M. Perslev et al.:. *U-sleep: resilient high-frequency sleep staging.* NPJ digital medicine, 4(1):1–12, 2021.
3] Ch. Corbiere et al.: *Confidence estimation via auxiliary models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6043-6055, 2022.