Explaining uncertainty in AI for clinical decision support systems*

Elisabeth R.M. Heremans[1][0000−0001−6860−2356] and Maarten De Vos[1,2][0000−0002−3482−5145]

[1] STADIUS Center for Dynamical Systems, Signal Porcessing and Data Analytics - Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
elisabeth.heremans@kuleuven.be
[2] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

Abstract. How can we explain the uncertainty of clinical machine learning models to doctors? In this extended abstract, we propose uncertainty explanations for AI-enhanced clinical decision-making. In the last years, both explainability and uncertainty estimation have gained significant attention from the AI community. This has happened as a response to the black box criticism that AI faces because of its increasing application in high-stakes decision-making processes. We are particularly interested in the application of uncertainty explanations in healthcare settings, demonstrating it on the use case of automated sleep analysis. We show a proof-of-concept of the idea and identify the directions to further investigate in a follow-up study.

Keywords: Explainability · Uncertainty estimation · Sleep staging.

1 Introduction

The advent of machine learning has introduced a new era in medicine, marked by a wide range of advances in computer-assisted diagnosis and remote patient monitoring. While machine learning models have achieved accuracy levels on par with human experts in controlled tasks in many medical applications [7, 9, 14–16, 20], their integration into clinical practice remains relatively limited. The slow adoption of AI in the clinic can be partly explained by a lack of trust from physicians in these automated decision systems. Guarantees of reliability, accompanied by explanations and uncertainty estimation, are crucial to convince clinicians to incorporate these models into healthcare [4, 11, 12]. Beyond explaining model decisions and estimating their uncertainty, we propose using explainability techniques to also explain the uncertainty of model decisions.

* Supported by a PhD fellowship from the Research Foundation - Flanders (FWO) for E. Heremans (FWO project number 1SC2921N) and by the Flemish Government (AI Research Program).

1.1 Explainability

The success of deep learning models on increasingly difficult tasks comes at a cost of increasing model complexity. As a result, these models are not understandable in human terms, explaining why they are often described as 'black boxes'. Recognizing this lack of transparency, the field of explainable artificial intelligence (XAI) has gained significant attention, as evident in recent review studies [2, 3]. Post-hoc explainability methods are especially popular, as these provide explanations without requiring any changes or compromises to be made regarding the prediction power of the models. Among the many possibilities for post-hoc explanations are feature importance methods, which highlight the most relevant input features for the model to make a prediction. Examples include perturbation-based methods like feature occlusion [19], gradient-based methods like integrated gradients [13], and others like SHAP [13] and LIME [18].

1.2 Uncertainty Estimation

Predictions by any AI system are unavoidably subject to noise and model inference errors. To cope with uncertainty in high-stakes fields such as medicine, it is thus of crucial importance to represent it in a trustworthy manner. A variety of uncertainty estimation methods have been studied in the context of machine learning and deep learning [1, 11, 22]. Examples include ensemble-based methods and Bayesian statistical methods. In a healthcare setting, these uncertainty estimation methods can be used to simply communicate the level of uncertainty associated with a prediction, or to abstain from making a prediction in samples of high uncertainty, deferring the decision to a clinical expert in those cases.

2 Uncertainty Explanations

To our best knowledge, explaining the uncertainty of AI decisions is a largely unexplored area. Simply quantifying the level of uncertainty does not always give enough information, as a user may wonder why this uncertainty arises: is the sample noisy, does it contain features of multiple classes, or is it out-of-distribution? We envisage exploiting post-hoc explainability methods to explain the uncertainty associated with those individual predictions. In healthcare settings, this could help doctors to understand why an AI abstains from making a prediction or indicates a high level of uncertainty. We have performed proof-of-concept experiments for this idea on sleep staging, a clinical annotation task known for its high levels of uncertainty and relatively low inter-rater agreement [8, 17].

3 Proof-of-concept Experiment

Sleep staging consists of classifying 30-second segments of physiological recordings, usually including electroencephalography (EEG) signals, as one of the five sleep stages {W, N1, N2, N3, R} [5]. Fig. 1a shows an example of an EEG signal

in the time domain with the associated sleep stages. This 5-minute signal is a sequence of 10 30-second segments of the classes Wake, N1 and N2. We transform this signal to a time-frequency representation as shown in Fig. 1b, before presenting it to our sleep staging model, a state-of-the-art deep neural network called SeqSleepNet [15]. Fig. 1a shows the predicted sleep stages, with prediction errors for two segments.

To take into account the time dependencies in our multivariate time series data, we chose the explainability method Dynamask [6], a perturbation-based feature importance method based on dynamic masks. Fig. 1c shows how Dynamask explains the model's prediction. Decisions are mostly based on local patterns (called K-complexes and sleep spindles in sleep literature). Then, we demonstrate how Dynamask explains the uncertainty of the predictions for this sequence (see Fig. 1d). The uncertainty explanation is clearly different from the prediction explanation. It shows for which features the perturbation decreases the prediction entropy (a measure of prediction uncertainty). In other words, the mask highlights features that make the prediction more uncertain.

This example shows that features which make the uncertainty high are highly concentrated around seconds 180 to 240. This is an area where classification errors and sleep stage transitions occur. Our uncertainty explanations are hence consistent with the general knowledge that uncertainty is higher on sleep stage transitions [10, 21]. Looking more closely at segment [210 s, 240 s] in the time domain (Fig. 1e) and in the frequency domain (Fig. 1b), we see high frequency noise. The uncertainty explanations show that these high frequencies make the model uncertain. This makes sense, given that these higher frequencies are likely the result of muscle movement. This may be a reason to classify this segment as Wake instead of N1. We contrast the time segment in Fig. 1e to segment [120 s, 150 s] in Fig. 1f, where noise levels are much lower.

4 Future Work

Our proof-of-concept experiment shows some promising results, however, it is just preliminary work of an open research challenge. First, we need to assess the quality of these explanations. This is a challenging task, as the true feature importance is not known. A quantitative way of measuring the quality of explanations is by masking out the features that supposedly make the model more uncertain, which should result in a drastic change in the prediction uncertainty. A qualitative evaluation can be performed with clinicians, to see if they find the explanations plausible. Second, we want to compare different post-hoc explanation strategies combined with different uncertainty estimation methods. A third challenge relates to the time-series aspect of this problem. We have evaluated the effect of all features on all predictions in a full time sequence, but we could also evaluate the effect of all input features on a single prediction or the effect of the input features associated with a single segment on the prediction associated with it. Which of these uncertainty explanations is most useful remains to be investigated.

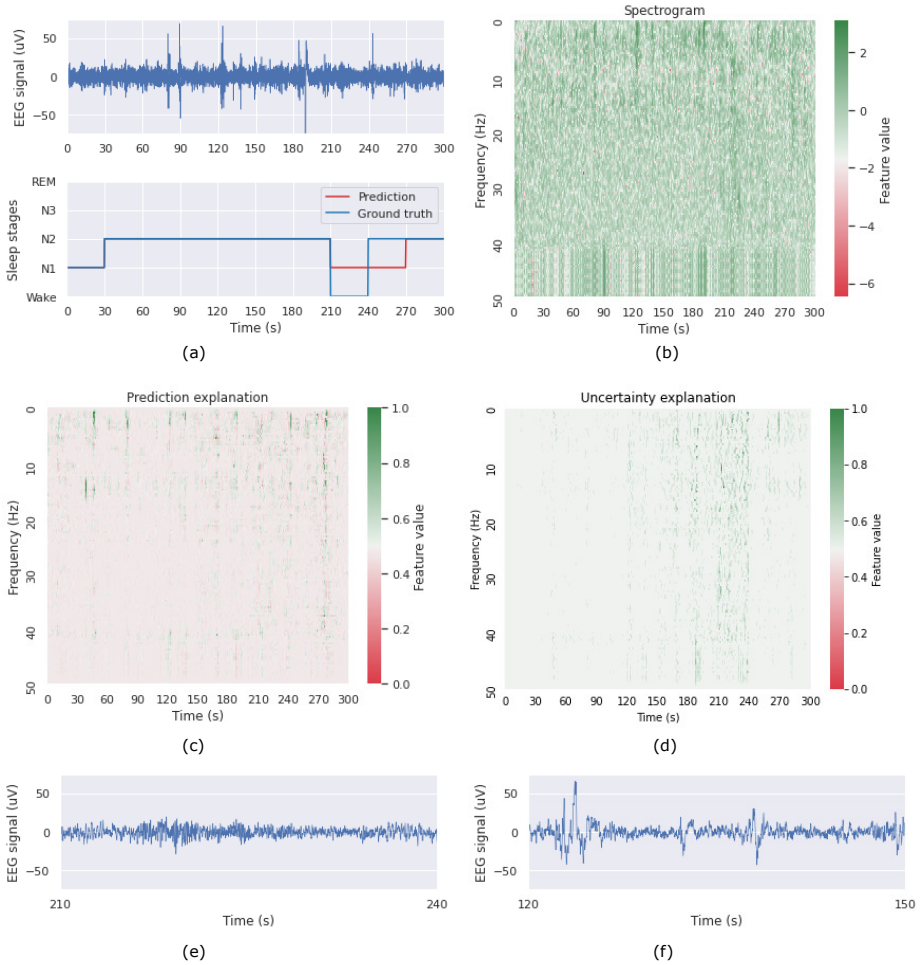


Fig. 1. (a) 5-minute EEG signal and the corresponding 10 sleep stages (both ground truth and model predictions). (b) Spectrogram of the EEG signal. (c) Explanations of the model predictions through feature importance (using the method Dynamask [6]). High feature values are associated with features that are important for making the predictions. (d) Explanations of the uncertainty of the predictions. High feature values indicate features that increase the entropy of the prediction. (e) and (f) Two segments in detail. Segment [210 s, 240 s] is corrupted by high-frequency noise, while segment [120 s, 150 s] shows a clean EEG signal.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. Information Fusion **76**, 243–297 (nov 2020). https://doi.org/10.1016/j.inffus.2021.05.008, http://arxiv.org/abs/2011.06225 http://dx.doi.org/10.1016/j.inffus.2021.05.008
2. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (sep 2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (jun 2020). https://doi.org/10.1016/J.INFFUS.2019.12.012
4. Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence 2019 1:1 **1**(1), 20–23 (jan 2019). https://doi.org/10.1038/s42256-018-0004-1, https://www.nature.com/articles/s42256-018-0004-1
5. Berry, R., Brooks, R., Gamaldo, C., Harding, S., Lloyd, R., Quan, S., Troester, M., Vaughn, B.: AASM Scoring Manual Updates for 2017 (Version 2.4). Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine **13** (2017). https://doi.org/10.5664/jcsm.6576
6. Crabbé, J., van der Schaar, M.: Explaining Time Series Predictions with Dynamic Masks. ICML (jun 2021), https://arxiv.org/abs/2106.05303v1
7. Egger, J., Gsaxner, C., Pepe, A., Pomykala, K.L., Jonske, F., Kurz, M., Li, J., Kleesiek, J.: Medical deep learning—A systematic meta-review. Computer Methods and Programs in Biomedicine **221**, 106874 (jun 2022). https://doi.org/10.1016/J.CMPB.2022.106874
8. van Gorp, H., M Huijben, I.A., Fonseca, P., G van Sloun, R.J., Overeem, S., van Gilst, M.M., author Hans van Gorp, C.: Certainty about uncertainty in sleep staging: a theoretical framework. Sleep **2022**, 1–8 (jun 2022). https://doi.org/10.1093/SLEEP/ZSAC134, https://academic.oup.com/sleep/advance-article/doi/10.1093/sleep/zsac134/6604464
9. Guillot, A., Thorey, V.: RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1441–1451 (2021). https://doi.org/10.1109/TNSRE.2021.3098968
10. Heremans, E.R., Seedat, N., Buyse, B., Testelmans, D., van der Schaar, M., De Vos, M., M Heremans, E.R., De Vos are with Leu-ven, M.K.: U-PASS: an Uncertainty-guided deep learning Pipeline for Automated Sleep Staging (jun 2023), https://arxiv.org/abs/2306.04663v1
11. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine **4**(1), 4 (2021). https://doi.org/10.1038/s41746-020-00367-3, https://doi.org/10.1038/s41746-020-00367-3
12. Kundu, S.: AI in medicine must be explainable. Nature Medicine 2021 27:8 **27**(8), 1328–1328 (jul 2021). https://doi.org/10.1038/s41591-021-01461-z, https://www.nature.com/articles/s41591-021-01461-z

13. Lundberg, S.M., Allen, P.G., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems **30** (2017), https://github.com/slundberg/shap
14. Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., Igel, C.: U-Sleep: resilient high-frequency sleep staging. npj Digital Medicine **4**(1), 72 (dec 2021). https://doi.org/10.1038/s41746-021-00440-5, http://www.nature.com/articles/s41746-021-00440-5
15. Phan, H., Andreotti, F., Cooray, N., Chén, O.Y., De Vos, M.: SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**(3), 400–410 (sep 2018). https://doi.org/10.1109/TNSRE.2019.2896659, http://dx.doi.org/10.1109/TNSRE.2019.2896659
16. Phan, H., Chén, O.Y., Tran, M.C., Koch, P., Mertins, A., De Vos, M.: XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging. IEEE Transactions on Pattern Analysis and Machine Intelligence (jul 2020). https://doi.org/10.1109/TPAMI.2021.3070057, http://arxiv.org/abs/2007.05492 http://dx.doi.org/10.1109/TPAMI.2021.3070057
17. Phan, H., Mikkelsen, K.: Automatic Sleep Staging: Recent Development, Challenges, and Future Directions (nov 2021), https://arxiv.org/abs/2111.08446v1
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **13-17-August-2016**, 1135–1144 (aug 2016). https://doi.org/10.1145/2939672.2939778, https://dl.acm.org/doi/10.1145/2939672.2939778
19. Suresh, H., Hunt, N., Johnson, A., Celi, L.A., Szolovits, P., Ghassemi, M.: Clinical intervention prediction and understanding with deep neural networks. In: Machine Learning for Healthcare Conference. pp. 322–337. PMLR (2017)
20. Yang, S., Zhu, F., Ling, X., Liu, Q., Zhao, P.: Intelligent Health Care: Applications of Deep Learning in Computational Medicine. Frontiers in Genetics **12**, 444 (apr 2021). https://doi.org/10.3389/FGENE.2021.607471/BIBTEX
21. Zhang, L., Fabbri, D., Upender, R., Kent, D.: Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. Sleep **42**(11), 1–10 (nov 2019). https://doi.org/10.1093/SLEEP/ZSZ159, /pmc/articles/PMC6802563/ /pmc/articles/PMC6802563/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6802563/
22. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A Review of Uncertainty Estimation and its Application in Medical Imaging (feb 2023), https://arxiv.org/abs/2302.08119v3