Identifying the neural predictors of subjective sleep quality from sleep EEG recordings in trans-diagnostic psychiatric conditions

Camille Challier, Gonçalo Braga & Wesley Monteith-Finas
EPFL CS-433, 21/12/2023

***Abstract*—This study explores the interplay between objective polysomnography sleep metrics and subjective sleep quality using machine learning techniques. Our multidimensional analysis encompasses feature extraction, the application of diverse ML models, and the incorporation of temporal insights through U-Time, a temporal convolutional neural network.**

Despite challenges in generalization stemming from a limited sample size and high-dimensional feature space, our models attempt to provide insights into potential associations between specific sleep parameters and subjective experiences. The complexity of accurately predicting subjective sleep quality is highlighted, paving the way for future investigations. The study's limitations call for larger datasets, advanced regularization techniques and further exploration of temporal aspects.

I. Introduction

Sleep is vital and plays a key role in emotional regulation, memory consolidation and general daily cognition functions [1]. Sleep disorders on the other hand, are highly prevalent and associated with a multitude of medical issues including hypertension, obesity, impaired immune function, mood disorders and neurodegeneration[1]. A problem often reported in the literature is the apparent disconnection between the objective and subjective sleep quality[2][3]. A night of sleep perceived as poor can affect the patient's day even in the presence of objectively good polysomnographic measures of sleep quality.

In this study we attempt to find that missing link by leveraging machine learning algorithms. We are further interested in quantifying the importance of different parameters for subjective sleep quality prediction.

Data from 43 patients of a clinical trial by the Human Sleep Psychopharmacology Laboratory (UZH), investigating the sleep-promoting effects of prospective and established sleep medications in humans, amounted to 129 recorded nights. This data presented itself in two modalities: raw Polysomnography recordings (EEG, ECG & EMG [1]) and hand labeled sleep stages across each night. The response variables were ten different scores on a morning questionnaire (MQ) filled post sleep monitoring, addressing the patient's experienced quality of sleep (see Annex VIII).

II. Methods

A. Sleep Stages

The sleep stages data is a categorised time series of different lengths, split into epochs of 30 seconds, corresponding to the sleep stages for each patient for each night (Awake, N1, N2, N3, and REM). Given the difficulty of using time series directly for prediction, different features were extracted from this data.

1) Time related features: The cumulative time spent in each stage, the total sleep time and the patient's counts in a particular stage (Fig 1-a,b).

2) Sleep stage latencies: Time before the 1^{st} occurrence of a sleep stage as well as time before falling asleep (defined as the first three consecutive non-awake epochs[4]) (Fig 1-a)

3) Continuities: Sleep stage bidirectional transitions i.e. how many times did the subject go from a certain sleep stage to another (Fig 1-c).

[1]Electroencephalogram, Electrocardiogram, Electromyography

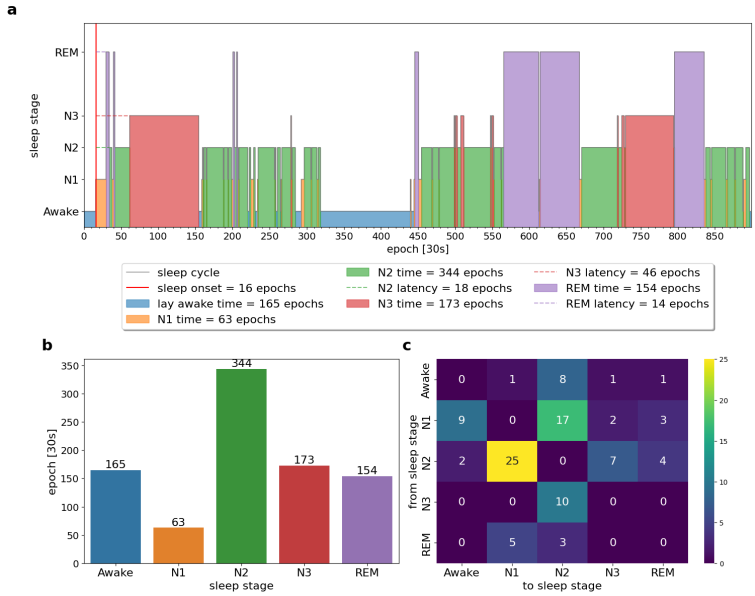

Figure 1: Different sleep cycle extracted features of the same night. **a** Sleep cycle visualisation with different sleep stages progression and corresponding latencies. **b** Cumulative time spent in each sleep stage (in 30s epochs). **c** Bidirectional transitions from and to each sleep stage.

4) Awakening: The stage the patient awoke from, which in turn also defines whether the wake up was natural or caused by the research team.

5) Resampled time series: As the raw time series differed in length, an inconvenient aspect for the use as a feature, they were resampled to the duration of the shortest night above a reasonable threshold of 7h.

6) Frequency domain: Frequency domain approaches were used to express temporal information in a more compact manner and to capture stage oscillations. We chose to use the Fourier transform and Power spectral density [5][6].

B. Polysomnography

1) Preprocessing EEG data: Prior to feature engineering, raw Polysomnography data (EEG, ECG & EMG) was processed with the MNE library [7] to enhance its quality by addressing artifacts and noise. We excluded the initial 60 seconds of each recording due to a significant presence of artifacts, potentially attributed electrode cap adjustments. We downsampled from 512 Hz to 100 Hz to reduce the number of data points [8] and standardised the ECG and EMG recordings. A commonly manually performed pre-processing step is the removal of eye blinks and eye movements, muscular activity, electrical noise and other noise artefacts. This couldn't be done due to a lack of specific knowledge, so an approach with Independent Component Analysis (ICA) was attempted [9]. This method proved unfruitful since too many components were capturing noise.

Afterwards the means of each EEG channels were subtracted and we re-referenced the data to the common average [9]. We band pass filtered the signal (f_{low}=0.5Hz and f_{high}=49.9Hz) to minimize the low-frequency drift and the line noise respectively[10].

Fourier Transform	Time	ECG
Mean	Energy	Mean RR [ms]
Maximum	Fisher information	STD [ms]
Minimum	Hjorth activity	Mean HR Kubios
Peak frequency	Hjorth complexity	Mean HR
Power	Kurtosis	STD HR
Power ratio	Hjorth mobility	Min HR
Spectral entropy	Line length	Max HR
Variance	Mean	RMSSD [ms]
Value range	Median	pNN50 [%]
DWT	Minimum	NN50
Mean	SVD entropy	**Patient Info**
Maximum	Zero crossing	Age
Minimum	ZC derivative	Sex
Bounded variation	**EMG Tib**	BMI
Power	Movement	**Others**
Power ratio		Night length
Spectral entropy		
Variance		

Table I: Polysomnography features extracted by categories: Fourier and Discrete Wavelet Transform (DWT), Time, tibialis Electromyography, ECG and Patient information.

2) Features extracted for the full night: There are two categories of ML methods that automate clinical EEG analysis. The first are Feature-based methods, which use handcrafted computed features. The second are End-to-end approaches, that employ learned features [11]. Taking into consideration the size of our data, we choose to focus on features-based methods.

From the EEG signals, features describing time and frequency structure were generated[11]. Frequency-based features computed were either based on the Discrete Fourier Transform computed with a Fast Fourier Transform (FFT) or on the Discrete Wavelet Transform (DWT) and calculated in the standard brain frequency bands [2]. From these transforms, a set of other parameters were extracted (see Table I), according to the most commonly employed features in the literature [11][12][13][14][15].

In a similar way, time related features like Hjorth parameters, the Fisher information, as well as numerous signal characteristics and analysis metrics were derived [13] [11]. The ECG recording was used to compute cardiac variability metrics and the EMG tibialis signal to infer movement during the night. Refer to Table I for further details.

In the end, a total of 1211 features were computed from the Polysomnography data.

3) Features extracted by epoch: For a more fine grained analysis, ie. to capture time evolution and reduce the impact of noise, we also generated the EEG features mentioned above per epochs of different size. This approach sought to produce multidimensional datasets, effectively treated as images for training convolutional neural networks (CNNs). First, we downsampled each night's data to the smallest duration, approximately 7h20 minutes, ensuring uniformity across all nights. Subsequently, epochs of approximately 13 seconds were created, resulting in a dataset with dimensions of 1920 epochs multiplied by the number of EEG channels (26) and the extracted features (42 per channels). This particular dataset was utilized to train the U-Time model (see II-E2). The selection of epoch sizes was made to effectively capture the temporal evolution and ascertain the architectural requirements for both models.

[2] δ: 0–2, 2–4Hz; θ: 4-8Hz; α: 8–13Hz; β: 13–18, 18–24, 24–30Hz; and γ 30–50Hz

C. Subjective Regression to Classification

Discrete MQ scores that spanned in between 1 to 100 were considered to be semi-arbitrary and thus grouped in 4 categories using a quantile approach (Fig mq in Annex). Indeed, with the sample size being small (n=129), a small variation between two patient's answers who slept well or un-well, ultimately conveys the same information. This allows us to conveniently remodel the problem into a 4-class classification task, class 0 representing the lowest sentiment (restless, tired, empty) to class 3 representing the highest sentiment (calm, well-rested, full of energy) for MQ 4, 6 and 8. The opposite being true for MQ 5, 7, 9 and 10 (class 0 highest sentiment and class 3 lowest sentiment).

D. Machine Learning Models

1) List of models: Various algorithms were employed in this study: Linear and Logistic Regression (resp. LinR and LR) and their regularized derivatives, in an attempt to prevent over-fitting, such as Lasso (L1), Ridge (L2) and ElasticNet (EN) which penalize the weights by adding resp. the L_1, L_2 or a combined norm. Ensemble methods such as Random Forests (RF) and AdaBoost (AB) with a Decision Tree (DT) base learner, were also used, as highly popular in other similar studies [11] [10]. Additionally we also tried K-Nearest Neighbors (KNN), a single DT, and Support Vector Machines (SVM). Most of these were tested both for regression and classification.

2) Evaluation metrics: Two metrics were mainly employed to evaluate the performance of our models. For the regression setting (MQ1-3) the Root Mean Squared Error ($RMSE$) was used: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$. For the classification setting (MQ4-10), the $F1_{weighted}$ score was used, which calculates the F1 score "for each label, and finds their average weighted by support (the number of true instances for each label)." [16] This metric takes into account multi-class imbalance. It is important to note that for a 4-class classification, a "chance level" $F1$ score is $F1_{chance} = 0.25$ $F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. Both of these metrics were paired with a (stratified in the case of classification) 5-fold Cross Validation (CV) in order to get an estimate of the test error.

3) Hyper-parameter tuning: Hyper-parameters (HP) of each model were tuned using a Bayesian optimization algorithm available in the *Scikit-Optimize* library [17], in order to *maximise* the stratified 5-fold CV $F1_{weighted}$ score for classifications (respectively *minimise* $RMSE$ for regression).

E. Deep Learning Models

More elaborate models leveraging state of the art neural network architectures were constructed using *PyTorch* [18] and its ecosystem tools in the form of *Skorch* [19], a high level library that provides compatibility with *Sci-kit Learn* [16].

1) Feed-forward neural network: A simple feed-forward neural network architecture was implemented. This neural network was structured with three hidden layers, comprising 512, 256, and 64 neurons, respectively. To enhance model generalization and prevent overfitting, dropout regularization was applied at each layer with a probability of 0.2 ($p_{dropout}$).

2) UTime: According to the literature, more elaborated Temporal Convolutional Neural Networks (TCNNs) could try and address the aforementioned problem. One such TCNN is U-Time[20], based on the popular U-Net architecture, originally designed for sleep stage classification on EEG data sets (architecture in Annex 8). This model leverages multiple convolution blocks, batch norm and max-pooling layers for down-sampling (Encoder), transposed-convolution blocks, batch norm layers, concatenation of feature maps and a point wise convolution for up-sampling (Decoder). U-time was designed to process the same type of data as our polysomnographic channels' traces,

and therefore our hypothesis was that it could extract meaningful features from our data and that we could use it's output for a simple 1 layered feed-forward classifier. Therefore only the *Segment Classifier* part was modified. We trained this model using time, DFT (fft) and DWT features extracted from the EEG signal (see II-B3). To compare U-Time's performance with other models, and knowing that this is a computationnaly intensive model, we only trained it on our best classifiable MQ ie. MQ4.

F. Preprocessing

Once features have been extracted from these two types of data, we prepared it for training by processing it. We excluded variables that were not relevant for the given prediction task (identification variables like Participant, Group, Night, Drug). We also dropped short nights (threshold set at about 4h or 500 epochs of 30s). We one-hot encoded categorical variables and standardised the remaining features to make sure all features would contribute equally to the model prediction. We also removed constant variables with no variability. Considering the substantial quantity of extracted features in the polysomnographic data, we removed correlated features exceeding a threshold of 90 %. At the end of this preprocessing pipeline, the extracted features data of the polysomnogram were reduced to 925.

G. Dimensionality reduction

Dimensionality reduction techniques were employed in order to visualise the extracted features data in lower dimensional sub-spaces (2D and 3D) and to try and visualise patterns by overlaying a color corresponding to an experimental condition, or a MQ. The three employed techniques were Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). The data used for each procedure was the features extracted from the polysomnographyc signal throughout the full night.

H. Feature Importance

Model-hyperparameter-features combinations that achieved the best test score to predict each MQ were trained on the entire data, and each individual feature importance to the model's prediction was assessed. For ensemble methods (RF or AB), the metric used was the default feature importance attributes. They are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. For models utilising weights (LR, L1, L2 or EN), the metric corresponds to the mean absolute weights across every class.

III. Results

A. Models

Starting with the regression MQs, our best model for MQ1 was ABR (rmse= 11.8±4.1) followed closely by RFR (rmse=11.9±4.5), both of them on the continuity/changements/durations subset. Their train error was comparable (Δ=0.4). For MQ2 RFR performed best on both the resampled time series (rmse=11.94±0.56) and the frequency domain subset (rmse=11.96±0.64) and again the train errors were comparable (Δ=0.054). Finally MQ3 was better predicted with RFR on the resampled time series subset(rmse=25.4±11.0). All scores are test scores unless stated otherwise.

Pursuing with the classification MQs, all our best models except MQ5, used the EEG extracted features in their best predictions. MQ4 was best predicted by a RFC ($F1_{test}$=0.434±0.086), with our top 5 models having a high mean train score ($F1_{train}$=0.949). L2 was our best model for MQ5 on the continuity/changements/duration subset ($F1_{test}$=0.386±0.074). RFC predicted best MQ6 ($F1_{test}$=0.396±0.065, $F1_{train}$=1) followed by ABC ($F1_{test}$=0.39±0.12, $F1_{train}$=0.715). The best model for MQ7 was NNC ($F1_{test}$=0.422±0.089, $F1_{train}$=0.701). For MQ8 L1 performed best ($F1_{test}$=0.39±0.06 with $F1_{train}$ generally

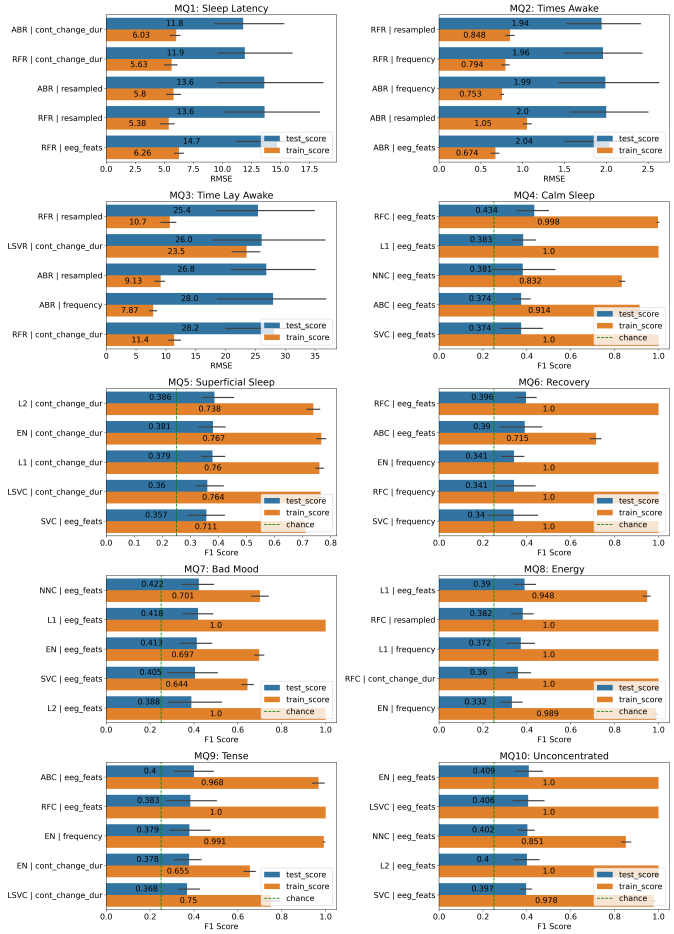

Figure 2: Top 5 best models for each MQ: 10 sub-plots corresponding to each MQ value (name on top). Corresponding model and feature combination are on the y-axis. For regression MQ1-3, RMSE on x-axis. For classification MQ4-10 F1 weighted on the x-axis, green dotted line corresponds to chance F1. Test score in blue, train score in orange.

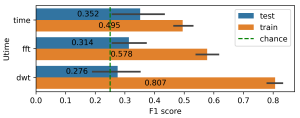

Figure 3: U-Time performance on MQ4 Calm Sleep, with different EEG extracted features used as channels (y-axis).

close to 1 in the top models). MQ9 was better explained with ABC ($F1_{test}$=0.40±0.12 with very high $F1_{train}$ >0.968 for our top 3 models). Finally MQ10's best model was EN ($F1_{test}$=0.409±0.085 with $F1_{train}$=1 for our top 2 models). Note: All scores described above are in fact $F1_{weighted}$, written synthetically as F1.

B. U-Time

The U-Time TCNN was applied to the MQ4 and yielded its best result for the EEG extracted time features ($F1_{weighted}$=0.352±0.089).

C. Feature Importance

In Fig. 4 one can observe the feature importance in each of our best models. For MQ 2 and 3, the most crucial features correspond to the sleep stage around the middle of the night and also late portions for the Times Lay Awake. The feature that most significantly explained sleep latency (MQ1) was the total sleep time, far more influential than the calculated sleep latency itself. Similarly, the prediction of Calm sleep (MQ4) appears to be primarily influenced by the bounded variation of the Discrete Wavelet Transform (DWT) of channel O2 (in δ and θ frequency bands). Similarly, MQ5 Superficial sleep could be explained by a set of different features including N3 deep sleep latency or different sleep stages transitions. Extracted from ECG signals, the heart rate demonstrates a substantial contribution to predicting

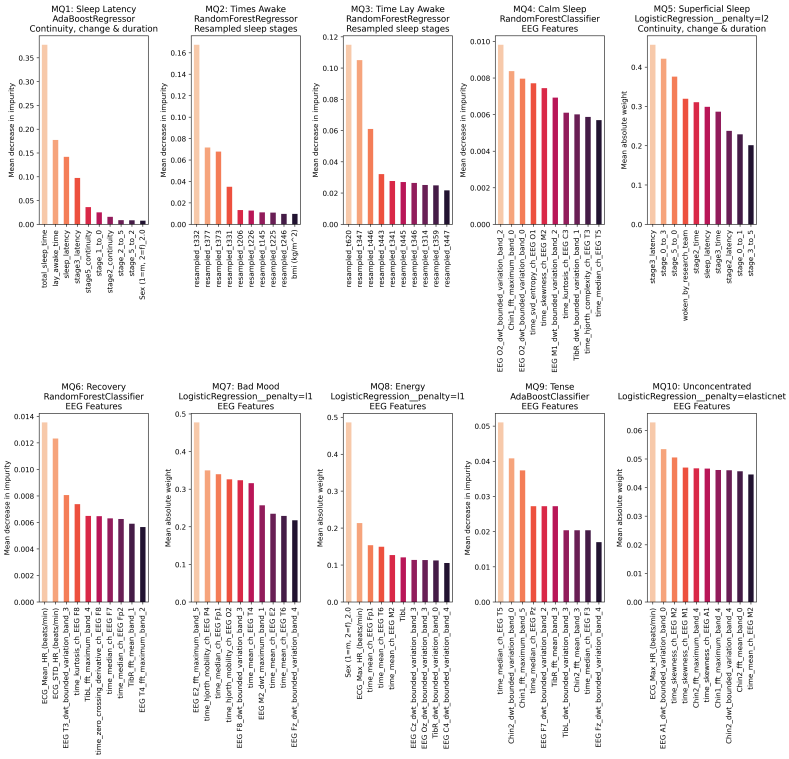

Figure 4: Top 10 important features for each best interpretable model predicting an MQ. Titles are MQ names, the model and the features used. Feature names on the x-axis. Importance metric on the y-axis. Color represents ranking within the top 10 features.

Recovery (MQ6) and Unconcentrated (MQ10). For MQ7, 8 and 9, the most crucial features are respectively: the maximal band of the FFT of the right EOG electrode E2 in the β frequency band; the Sex; and the median of EEG channel T5 signal. In 15 we can observe that the female gender exhibits a negative correlation with elevated daily energy levels.

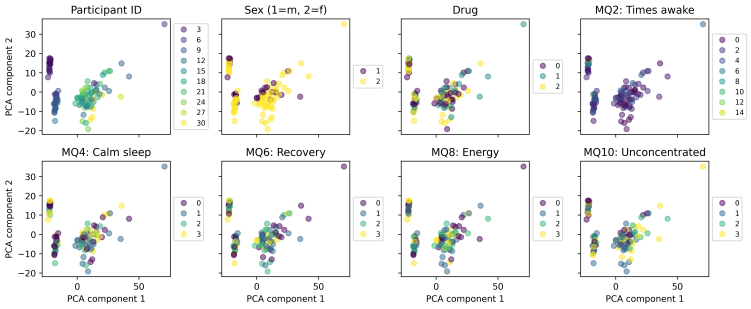

Figure 5: Representation of the 2 first components of a PCA dimensionality reduction on the features generated with polysomnographic data. Color represents a discrete range of values for top-left and top-right plots, but classes for all others.

IV. Discussion

A. Performance

From our results we can see that neither the classification nor the regression task at hand can be predicted well. The test $F1$ is low and the test $RMSE$ is high. This is further emphasised when looking at the confusion matrices and predicted versus ground truth identity plots in Annex 6 7. The best classifiable MQ is MQ4, calm sleep, using a RF on the EEG extracted features. We can also see that EEG extracted features give rise to an overall better performance for classifiable MQs, but sleep stages data works better for regression MQs. Our second best classifiable MQ (7) was best predicted by a NN, showing the relevance of deep learning for this task.

B. Over-Fitting

With our results, it is evident that the vast majority of our best (optimised) models over-fit the training data. This means that our models are learning, but are too flexible, and cannot be generalised to unseen data. In order to battle this problem we tried to impose more regularizing parameters when HP tuning with Bayes Search. For RF and AB this means decreasing the number of trees and their max depth. For LR, L1, L2 and EN, it means increasing the penalisation coefficient α (resp. decrease $C = \frac{1}{\alpha}$). Similarly grouping extracted features (EEG, sleep cycles time series, ...) decreased the number of predictors available, which reduced the flexibility of our models. Even so, these measures did not prevent over-fitting. This phenomenon is most likely due to the low sample size ($n_{predictors} >> n_{samples}$) and it could have been prevented with access to more data, but getting clinical data is financially and time consuming. Plus, for ethical reasons (see VI), this sort of data cannot be share online on large databases. For future prospects, a data augmentation technique could be an answer to this limitation.

C. Dimensionality Reduction

From Fig 5 one can observe 3 clusters from the 2D projected PCA data. The same observation was made both with t-SNE and UMAP embeddings (2D and 3D). There seems to be a correlation with the participant ID and these clusters. Small IDs top-left, intermediate bottom-left and high right. This notable difference based on the ordering of each patient may be caused by experimental setup variations across time, or an error in our preprocessing. Similarly, there seems to be a distinction between male and female participants, males being higher than females on PCA dimension 2, within each cluster. Nevertheless, clusters do not seem to represent each experimental drug condition, nor each MQ value. Hence we did not try an unsupervised learning method on these lower dimensions.

D. U-time

U-time did not out-perform our best model (RFC) on the most predictable MQ (4: Calm sleep). This may be caused by the fact that our classifier block was too simple. Additionally, due to it's time complexity HP were not tuned as extensively as other models. Potentially with greater computational power, for HP tuning, U-time performance could drastically improve.

E. Features Importances

Had our models performed more effectively, we could have established a correlation between the signals of the Occipital Lobe O2 channels and the pivotal role of occipital lobes in dreaming. This, in turn, would have explained the impact on the perception of calm sleep. As seen in 13 an extended duration before entering deep sleep is correlated with very superficial sleep. If the patient is awakened by the research team, a sense of shallow sleep is often reported upon waking.

V. Conclusion

In conclusion, this study explores the intricate relationship between objective sleep metrics and subjective sleep quality. While our models exhibited challenges in generalization, likely due to the limited sample size and high-dimensional feature space, they shed light on potential associations between specific sleep parameters and subjective experiences. The performance, though not optimal, underscores the complexity of predicting subjective sleep quality accurately. Future endeavors should focus on obtaining larger datasets, implementing advanced regulaatrizion techniques, and exploring additional temporal aspects, as highlighted by models like U-Time.

VI. Ethical risks

One big ethical concern faced in this study was tied to the data itself. Indeed neural data is highly individual and some argue that in the future it "might be as identifying as a fingerprint"[21][22]. This poses confidentiality issues and possible privacy breaches for the patients that were followed in this study. With current technology, there is no risk of this happening but the repercussions would be pretty severe, would it be the case. The patients health status, sleep disorders, drugs prescribed and mental health could hypothetically be deducted from the data. There is no way to quantify this risk but one could quantify the risk of the data leaking. To take it into account the data was anonymized, never pushed to a repository (educational or private), Google Colab wasn't been used, and data was transfered through a password-protected cloud storage that was only available to predetermined individuals. If the geographical distance between the research facility and EPFL had been smaller, it would have been more ideal not to resort to a cloud storage.

VII. Acknowledgements

We would like to warmly thank PhD Laura Katharina Schnider for her guidance, and the Human Sleep Pharmacology Laboratory for making their data available to us as well as ML4 science for the opportunity to do a real world, applied project.

References

[1] S. L. Worley, "The Extraordinary Importance of Sleep," *Pharmacy and Therapeutics*, vol. 43, no. 12, pp. 758–763, Dec. 2018.

[2] K. A. Kaplan, J. Hirshman, B. Hernandez, M. L. Stefanick, A. R. Hoffman, S. Redline, S. Ancoli-Israel, K. Stone, L. Friedman, J. M. Zeitzer, and Osteoporotic Fractures in Men (MrOS), Study of Osteoporotic Fractures SOF Research Groups, "When a gold standard isn't so golden: Lack of prediction of subjective sleep quality from sleep polysomnography," *Biological Psychology*, vol. 123, pp. 37–46, Feb. 2017.

[3] L. E. Cudney, B. N. Frey, R. E. McCabe, and S. M. Green, "Investigating the relationship between objective measures of sleep and self-report sleep quality in healthy adults: A review," *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, vol. 18, no. 3, pp. 927–936, Mar. 2022.

[4] I. Feinberg and T. C. Floyd, "Systematic trends across the night in human sleep cycles," *Psychophysiology*, vol. 16, no. 3, pp. 283–291, May 1979.

[5] A. Gabryelska, B. Feige, D. Riemann, K. Spiegelhalder, A. Johann, P. Białasiewicz, and E. Hertenstein, "Can spectral power predict subjective sleep quality in healthy individuals?" *Journal of Sleep Research*, vol. 28, no. 6, p. e12848, Dec. 2019.

[6] D. Stoffer and D. Tyler, "Spectral Analysis for Categorical Time Series: Scaling and the Spectral Envelope," *Biometrika*, vol. 80, pp. 611–622, Sep. 1993.

[7] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, 2013.

[8] K. G. van Leeuwen, H. Sun, M. Tabaeizadeh, A. F. Struck, M. J. A. M. van Putten, and M. B. Westover, "Detecting abnormal electroencephalograms using deep convolutional networks," *Clinical Neurophysiology*, vol. 130, no. 1, pp. 77–84, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138824571831349X

[9] J. S. George, J. Strunk, R. Mak-McCully, M. Houser, H. Poizner, and A. R. Aron, "Dopaminergic therapy in Parkinson's disease decreases cortical beta band coherence in the resting state and increases cortical beta band power during executive control," *NeuroImage: Clinical*, vol. 3, pp. 261–270, Jan. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158213001034

[10] M. Aljalal, S. A. Aldosari, M. Molinas, K. AlSharabi, and F. A. Alturki, "Detection of Parkinson's disease from EEG signals using discrete wavelet transform, different entropy measures, and machine learning techniques," *Scientific Reports*, vol. 12, no. 1, p. 22547, Dec. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-26644-7

[11] L. A. Gemein, R. T. Schirrmeister, P. Chrabaszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball, "Machine-learning-based diagnostics of EEG pathology," *NeuroImage*, vol. 220, p. 117021, Oct. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811920305073

[12] S. Pravin Kumar, N. Sriraam, P. G. Benakop, and B. C. Jinaga, "Entropies based detection of epileptic seizures with artificial neural network classifiers," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3284–3291, Apr. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741740900832X

[13] L. Logesparan, A. J. Casson, and E. Rodriguez-Villegas, "Optimal features for online seizure detection," *Medical & Biological Engineering & Computing*, vol. 50, no. 7, pp. 659–669, Jul. 2012. [Online]. Available: https://doi.org/10.1007/s11517-012-0904-x

[14] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084–1093, May 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417406000844

[15] H. Cai, X. Sha, X. Han, S. Wei, and B. Hu, "Pervasive EEG diagnosis of depression using Deep Belief Network with three-electrodes EEG collector," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2016, pp. 1239–1246. [Online]. Available: https://ieeexplore.ieee.org/document/7822696

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] T. Head, MechCoder, G. Louppe, I. Shcherbatyi, fcharras, Z. Vinícius, cmmalone, C. Schröder, nel215, N. Campos, T. Young, S. Cereda, T. Fan, rene-rex, K. K. Shi, J. Schwabedal, carlosdanielcsantos, Hvass-Labs, M. Pak, SoManyUsernamesTaken, F. Callaway, L. Estève, L. Besson, M. Cherti, K. Pfannschmidt, F. Linzberger, C. Cauet, A. Gut, A. Mueller, and A. Fabisch, "Scikit-optimize/scikit-optimize: V0.5.2," Zenodo, Mar. 2018.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Dec. 2019.

[19] M. Tietz, T. J. Fan, D. Nouri, B. Bossan, and skorch Developers, *skorch: A scikit-learn compatible neural network library that wraps PyTorch*, Jul. 2017. [Online]. Available: https://skorch.readthedocs.io/en/stable/

[20] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging," Oct. 2019.

[21] "BRAIN 2.0 Neuroethics: Enabling and Enhancing Neuroscience Advances for Society — BRAIN Initiative," https://braininitiative.nih.gov/vision/nih-brain-initiative-reports/brain-20-neuroethics-enabling-and-enhancing-neuroscience.

[22] "Rommelfanger Lab," https://neuroethicslab.com/.

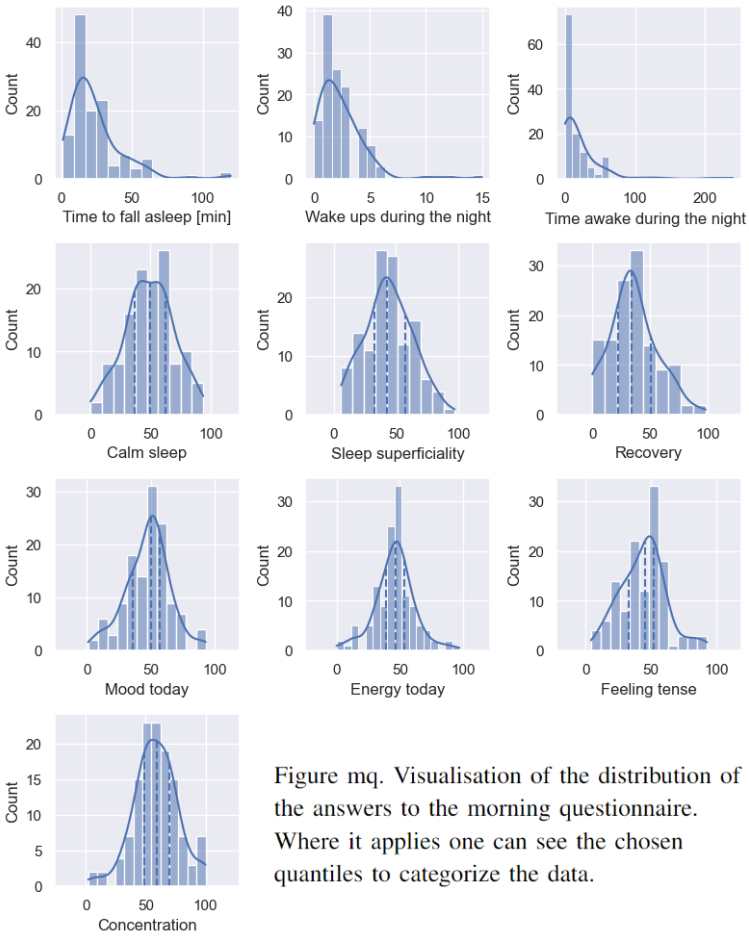

Figure mq. Visualisation of the distribution of the answers to the morning questionnaire. Where it applies one can see the chosen quantiles to categorize the data.

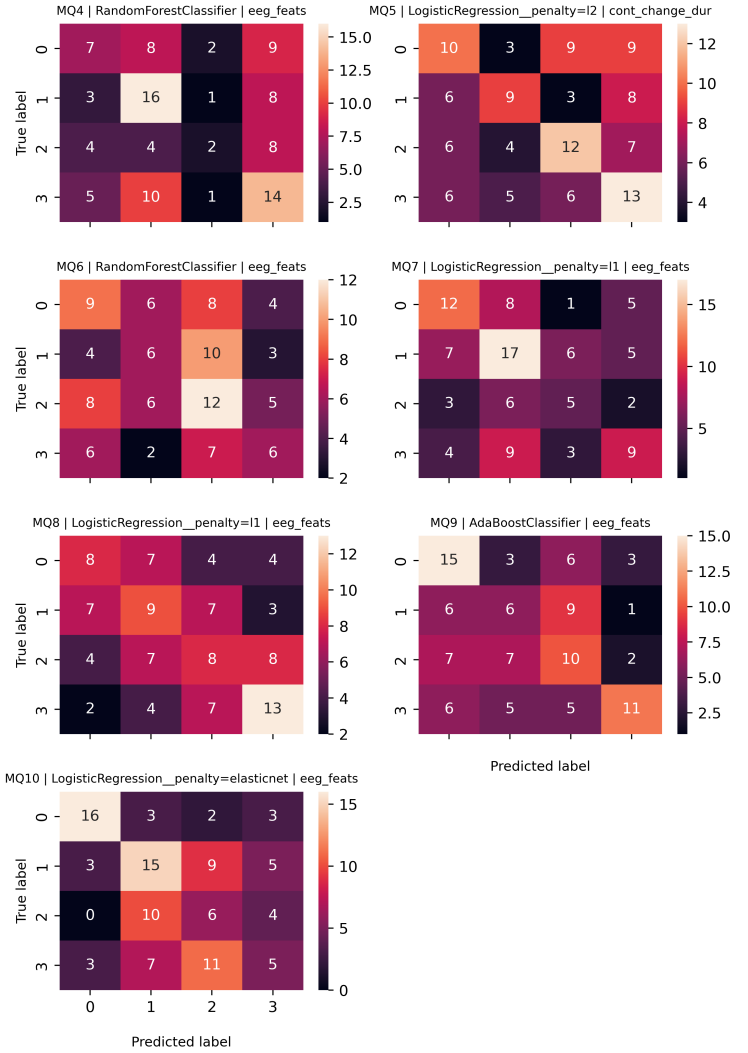

Figure 6: Confusion matrix of each best models found for MQ 4 to 10.

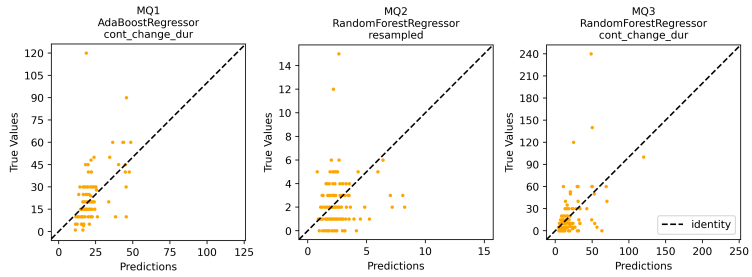

Figure 7

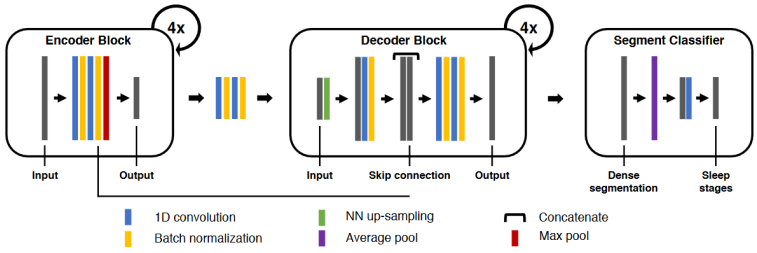

Figure 8: U-time original architecture. Source: Perslev et al., 2019 "U-Time: A Fully Convolutional Network for Time Serie Segmentation Applied to Sleep Staging" [20]

MQ	MQ Question	Scale
MQ1	How long did it take you to fall asleep?	number of minutes
MQ2	How many times did you wake up?	number of times
MQ3	How many minutes were you laying awake during the night?	number of minutes
MQ4	How calm was your sleep last night?	0-100 [restless to calm]
MQ5	How superficial was your sleep?	0-100 [deep to superficial]
MQ6	How rested do you feel today?	0-100 [tired to well-rested]
MQ7	How good in a mood do you feel today?	0-100 [good to bad mood]
MQ8	How much energy do you have today?	0-100 [indifferent and empty - full of energy]
MQ9	How tense do you feel?	0-100 [calm to tense]
MQ10	How concentrated do you feel?	0-100 [concentrated to unconcentrated]

Table II: Note: The original questions were asked in German.

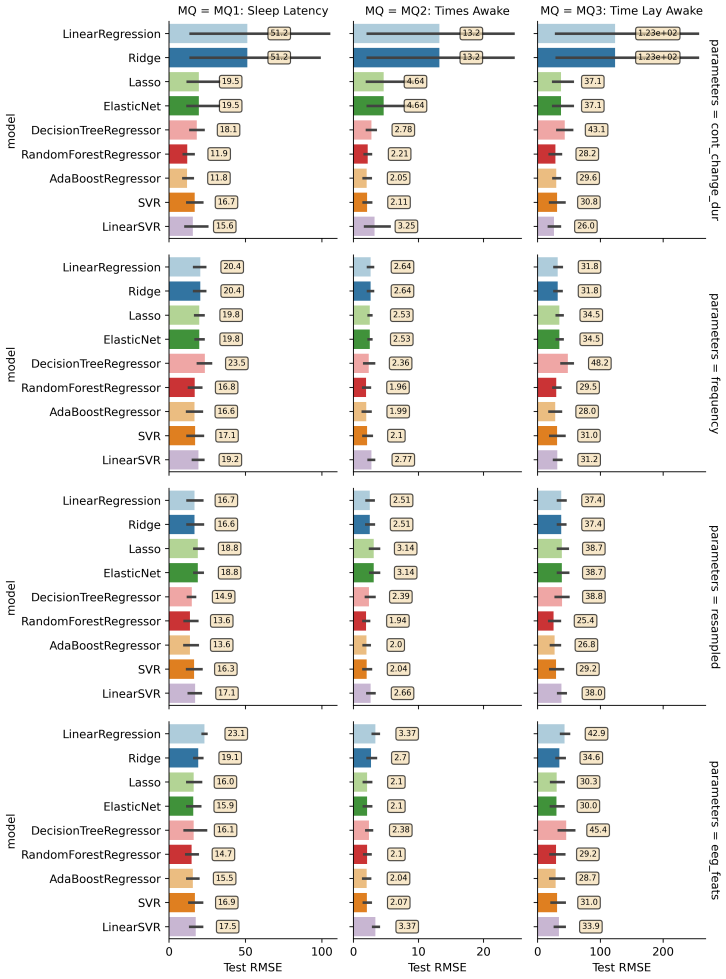

Figure 9: Test F1 score of all models tested for regression MQ 1 to 3.

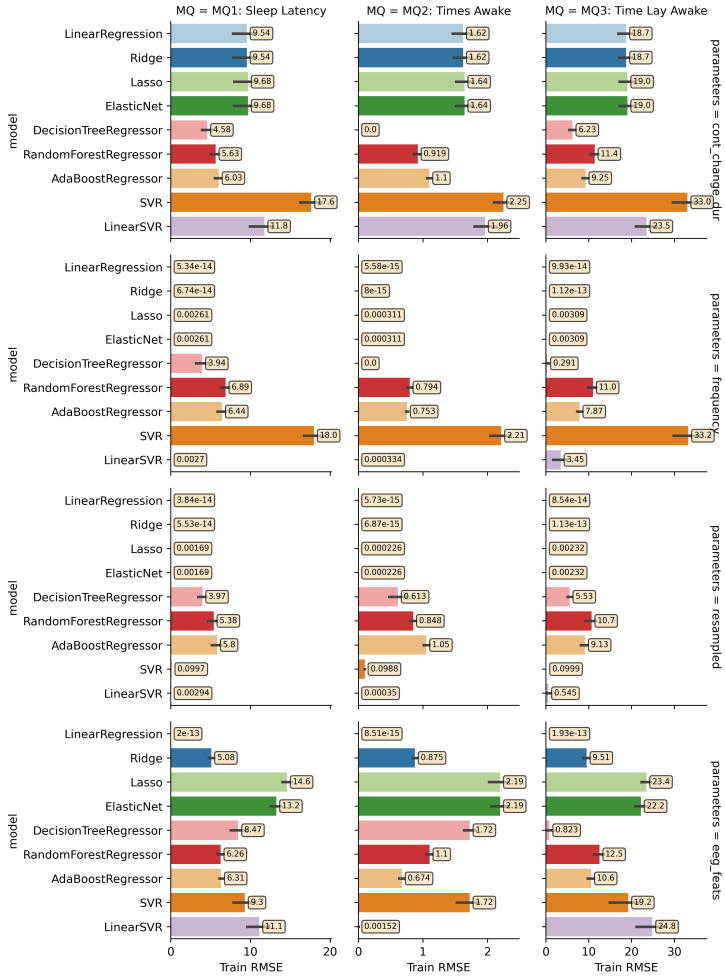

Figure 10: Train F1 score of all models tested for regression MQ 1 to 3.

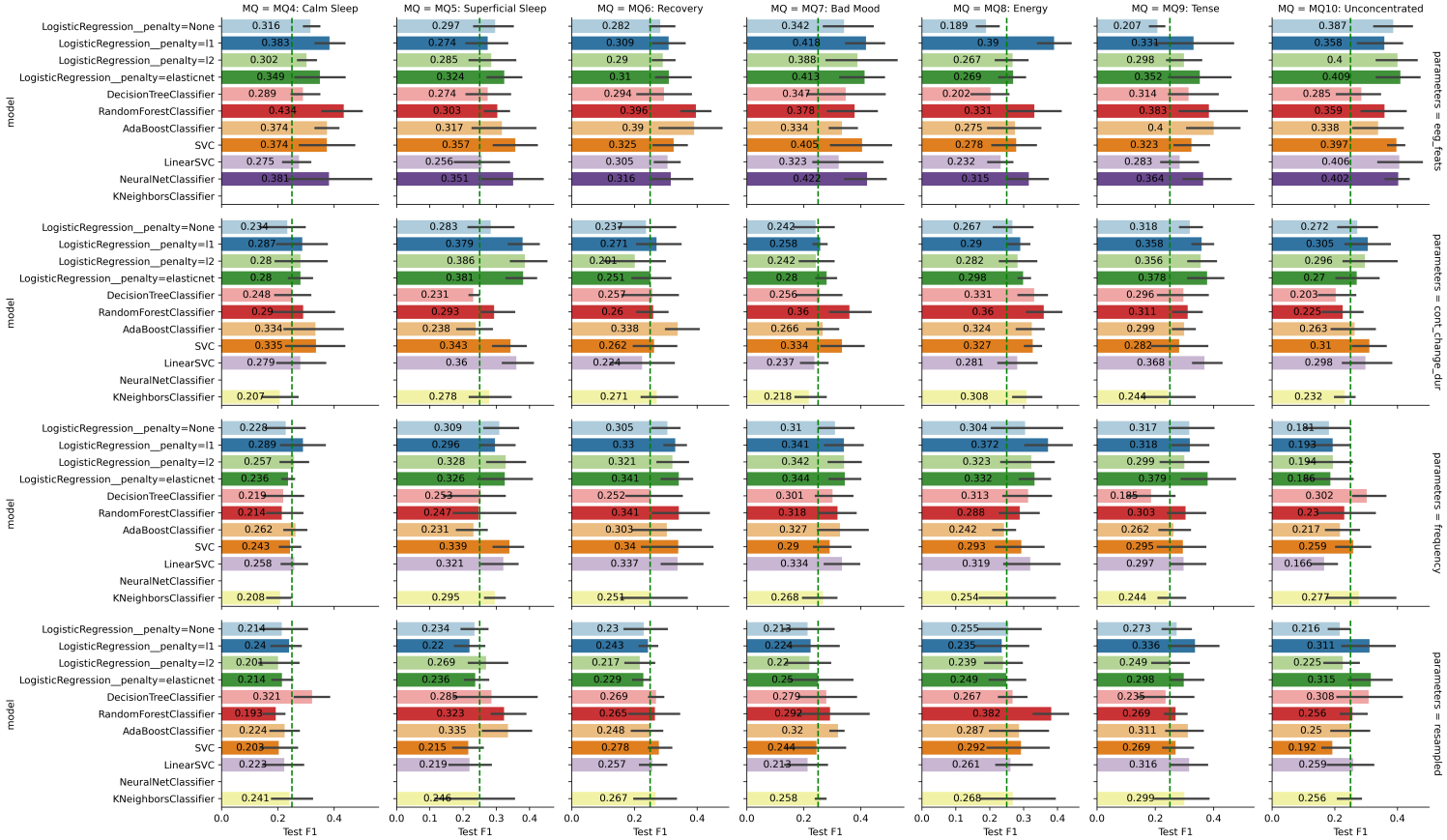

Figure 11: Test F1 score of all models tested for Classification MQ 4 to 10.

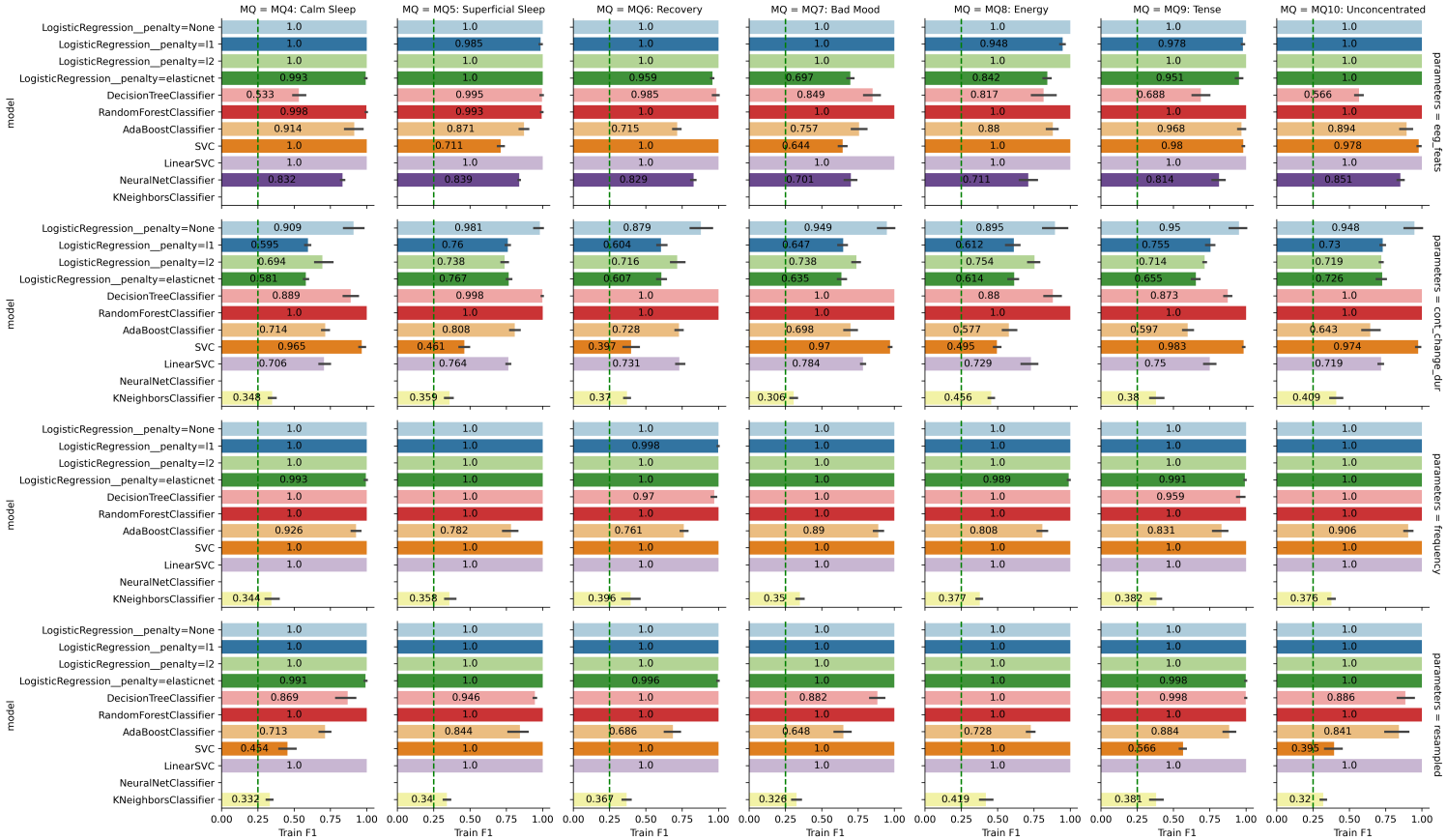

Figure 12: Train F1 score of all models tested for Classification MQ 4 to 10.

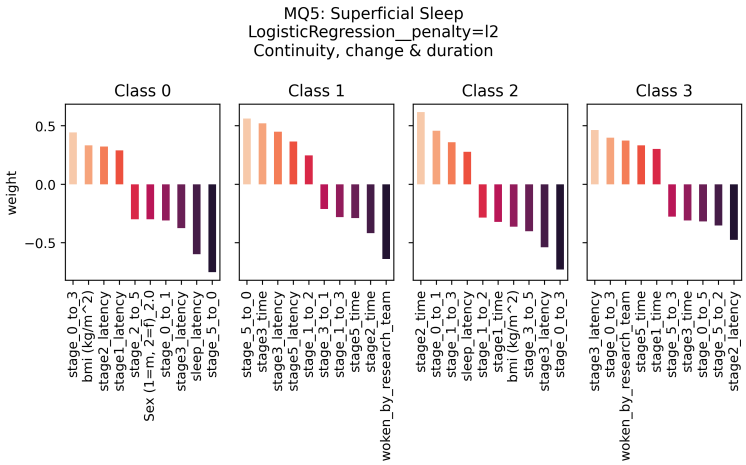

Figure 13: Weights across 4 classes for MQ5, L2 penalty

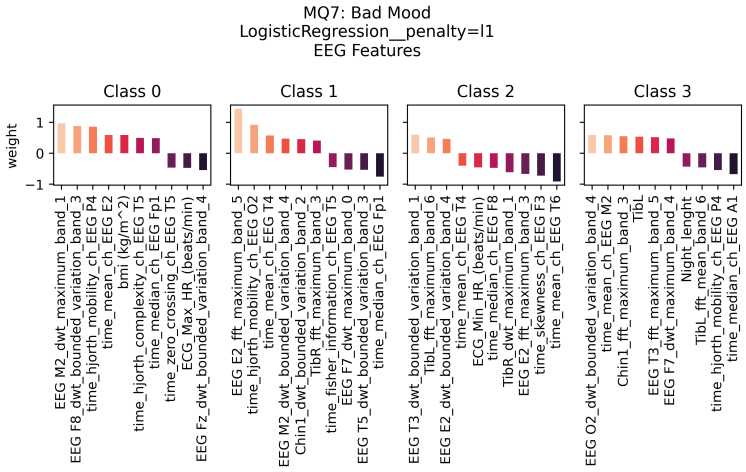

Figure 14: Weights across 4 classes for MQ7, L1 penalty

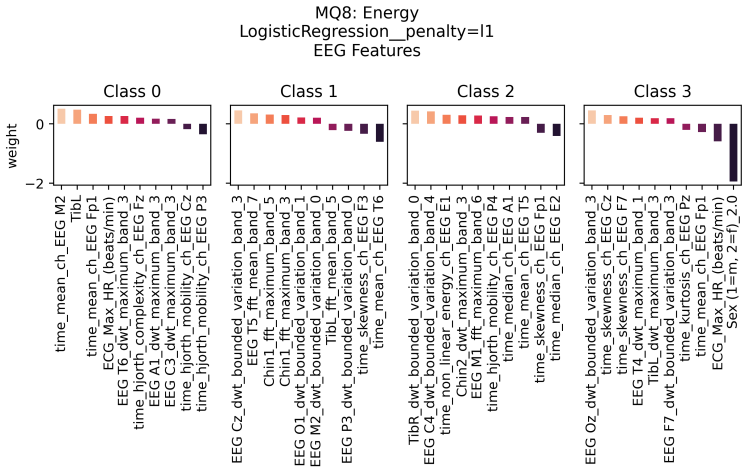

Figure 15: Weights across 4 classes for MQ8, L1 penalty

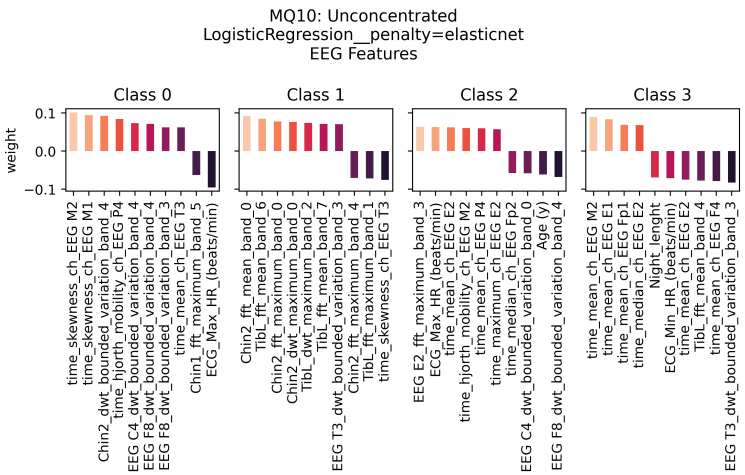

Figure 16: Weights across 4 classes for MQ10, EN penalty