Exploring the effect of caffeine consumption on EEG sleep signals using ML

by Philipp Thölke

Index

Introduction - Motivation

- the quality of sleep has a direct impact on health
    - lack of sleep and sleep disorders can lead to deterioration of sleep-related brain processes
    - bad sleep quality increases the risk for depression, weight gain, hypertension, cardiovascular diseases and diabetes
- caffeine as a psychostimulant and antagonist to adenosine reduces the natural circadian sleep pressure by attaching to adenosine receptors
    - feeling of higher alertness and invigoration
- a better understanding of the effects of caffeine on brain activity during sleep is a major health concern due to high levels of consumption in the population

Introduction - Previous work

- caffeine affects spectral power, the amount of contribution to the signal by different frequencies, of the EEG during sleep
    - decreases in delta power and increases in beta have been found by multiple studies
    - has been detected in humans and Cynomolgus monkeys (suggested to be representative due to diurnal nature and similar proportion of sleep stages compared to humans)
- sleep variables are disturbed by the ingestion of caffeine before sleep
    - increase in sleep latency, decrease in efficiency, decrease in total sleep duration and shift in sleep stage distribution
- an increase in resting brain entropy due to caffeine has been found in an fMRI study
    - entropy is a measure of complexity, indicating how unpredictable or random a signal is
    - higher resting brain entropy after caffeine ingestion might indicate an increase of information processing capacity

Introduction - Hypotheses

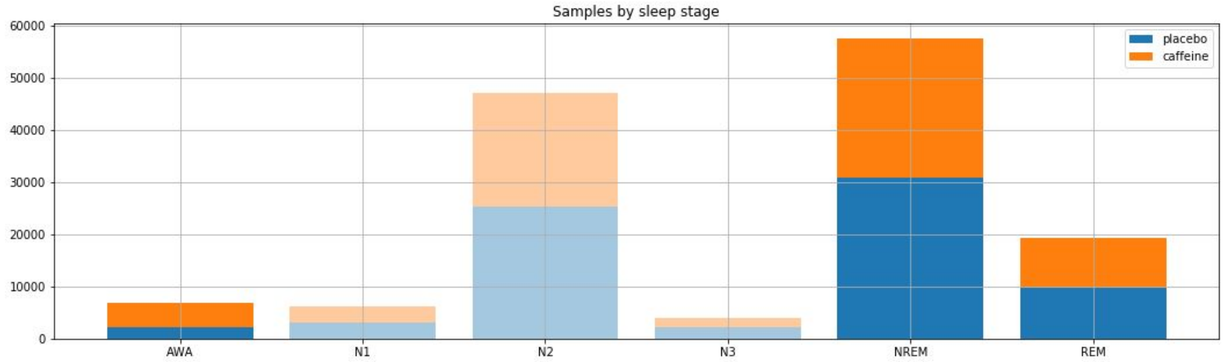
- caffeine might also increase brain entropy in electrophysiological recordings, not only in the BOLD signal
- exploration of EEG features using a data-driven approach may quantify the separation between caffeine and placebo
- new biomarkers could be identified, helping the understanding of the influence of caffeine on cerebral dynamics during sleep

Methods - Data acquisition

- sleep EEG from 40 participants
  - from 28 to 58 years old, mean age 35.3 ± 14.3 years
  - 21 male, 19 female
- randomized, double-blind, cross-over design
  - subjects spent two nights at sleep laboratory, receiving 200 mg caffeine or placebo (lactose) capsule before regular bedtime
- no caffeine consumption after noon on day of recording
- recording was done with 20-electrode EEG cap at 256 Hz
  - arranged in international 10-20 system
  - referential montage with linked ears

Methods - Preprocessing

- artifact removal, spindle and slow wave extraction
- data divided into 20s epochs
- segmentation into sleep stages using hypnogram (AWA, N1, N2, N3, REM)
    - N1, N2, N3 were combined into single NREM stage
    - AWA stage only contains wake epochs after sleep onset (no data for two subjects, excluded)



Methods - Feature extraction

Power spectral density (PSD)

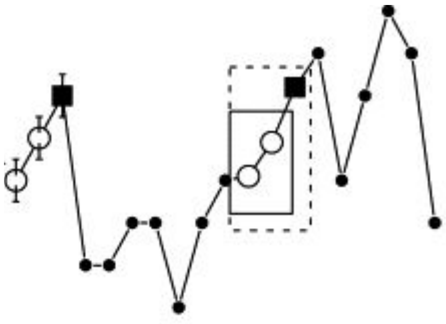
- computation using Welch's method of averaged periodograms
  - Hamming window on six segments without overlap (Bartlett's method)
- extraction of power bands from six frequency intervals by averaging
  - delta: 0.3 - 4.0 Hz
  - theta: 4.0 - 8.0 Hz
  - alpha: 8.0 - 12.0 Hz
  - sigma: 12.0 - 16.0 Hz
  - beta: 16.0 - 32.0 Hz
  - low gamma: 32.0 - 50.0 Hz

Methods - Feature extraction

Spectral entropy (SpecEn)

- PSD extracted in the same way as for power bands
- Shannon entropy was applied to full power spectrum (instead of frequency bands)
- estimation power spectrum complexity (higher entropy means higher complexity/more randomness)
- Shannon entropy looks at the relative power of each frequency bin
  - uniform distribution over all bins corresponds to maximal entropy
  - 100% power in one frequency bin means no complexity/randomness → minimal entropy

Methods - Feature extraction



Sample entropy (SampEn)

- estimates entropy by evaluating probability of temporal continuation of a window somewhere else in the signal
    - negative logarithm of probability that if two windows of size m match by a distance threshold r times the standard deviation of the signal, the two windows also match with size m+1
- similar to approximate entropy (ApEn; often used for biomedical time series)
- advantages of SampEn over ApEn:
    - does not count self-matches → less bias
    - not dependent on signal length due to normalization term
    - higher consistency concerning parameter choice
- parameter choice: m=2, r=0.2 * signal standard deviation

Methods - Feature extraction

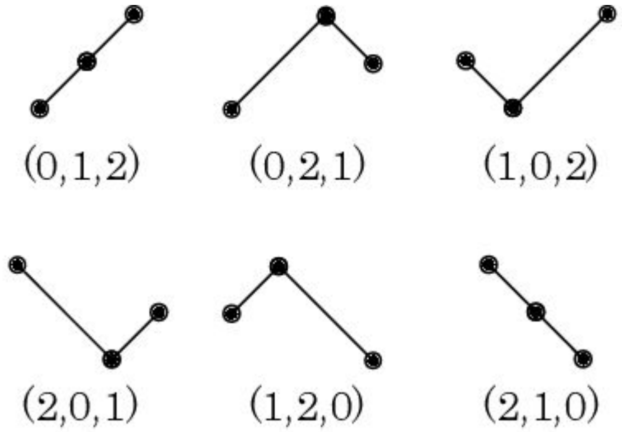
Spectral sample entropy (SpecSampEn)

- very similar to SpecEn but using SampEn on the power spectrum instead of Shannon entropy
- able to look at similar frequency powers at once
- looks for similar patterns across frequencies in the power spectrum
- parameters for SampEn: m=2, r=0.2 * signal standard deviation

Methods - Feature extraction

Permutation entropy (PermEn)

- evaluates times series by computing Shannon entropy on the probability distribution of different ordinal patterns of length n with a delay of τ
- occurrences are counted for n! different ordinal patterns
- looks at diversity in the ordering of values in the signal
  - random ordering leads to high PermEn
- chosen parameters were embedding dimension n=3, sample delay τ=1



Methods - Feature averaging

- low comparability between machine learning performances across stages when sample count varies strongly
- classifier performance may suffer from skewed class distributions
- features were averaged subject-wise, leaving two samples per subject
    - 80 samples in NREM and REM stages (40 caffeine and 40 placebo)
    - 76 samples in AWA stage (38 caffeine, 38 placebo)
- feature-wise z-transformation of samples using mean and standard deviation across electrodes

Methods - Statistical analyses

- assessing the statistically significant differences between caffeine and placebo conditions for each electrode
- placebo samples were subtracted subject-wise from the caffeine condition
- two-sided paired permutation-based pseudo t-tests (tmax correction) applied to all extracted features
- exhaustive permutations with the number of permutations n=1000
- significance was evaluated at $p<0.05$

Methods - Single-feature machine learning

- single-feature, single-electrode classification between caffeine and placebo condition
- four different algorithms for more reliability:
  - support vector machine (SVM), linear discriminant analysis (LDA), Gaussian process, perceptron
- permutation tests (n=1000) were applied to estimate the confidence of classifier scores (significant at $p<0.05$)
  - retraining the classifier n-1 times with permuted labels to determine if the classifier learned a feature-label dependency or is guessing randomly
- 10-fold cross-validation in each permutation, score is averaged across test folds

Methods - Multi-feature machine learning

Single-electrode classification

- determine overall effect of caffeine on the features for each electrode
- classifiers were trained on all extracted feature combined (10 features)
  - 6 PSD bands, SpecEn, SampEn, SpecSampEn, PermEn
- four different algorithms for more reliability:
  - support vector machine (SVM), linear discriminant analysis (LDA), random forest, multilayer perceptron (MLP)
- grid search was applied on data from 25% of the subjects per electrode
- permutation tests (n=1000) with 10-fold cross-validation on the remaining 75% of the subjects (significant at $p<0.05$)
- an ensemble classifier from the four classifiers' predictions using majority vote

Methods - Multi-feature machine learning

Multi-electrode classification

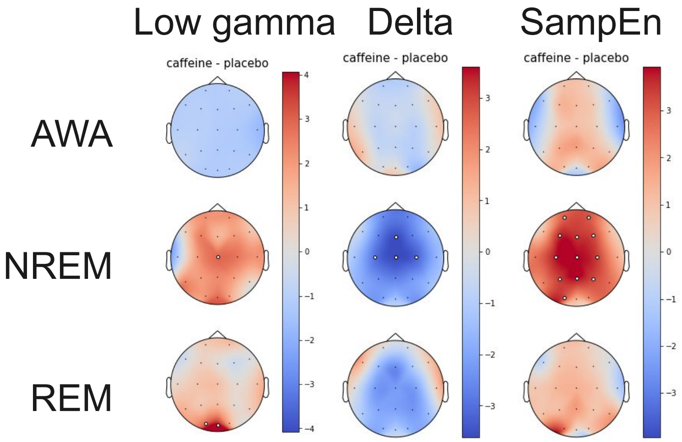
- a random forest was trained on combined features from all electrodes
  - 10 extracted features * 20 electrode = 200 total features
- feature importance was used to estimate the effect of caffeine on different electrodes and features
  - importance for one feature can be calculated by averaging the height of the feature in all decision trees in the random forest
  - higher features (closer to the root) contribute more to the predictive decision of decision trees
- training of the random forest was repeated 1000 times
  - determine variance in feature importance and classification accuracy
  - 7-fold cross-validated hyperparameter grid search inside a leave 5 subjects out cross-validation (left out subjects different in each iteration of training a random forest)

Results - Statistical analyses

- few significant electrodes in AWA stage
  - cluster of significant decreases of sigma power in frontal electrodes
- most statistically significant electrodes in NREM
  - decrease in delta power in central electrodes
  - strong increase in SpecEn, SampEn and SpecSampEn over many electrodes
- occipital region shows significant differences in REM
  - increase in sigma, beta and low gamma power in the visual system
- SpecEn increase in occipital region in REM

Results - Statistical analyses

- increased SpecEn, SampEn and SpecSampEn in NREM
- decreased delta band in NREM
- increase in sigma, beta and low gamma band in occipital during REM

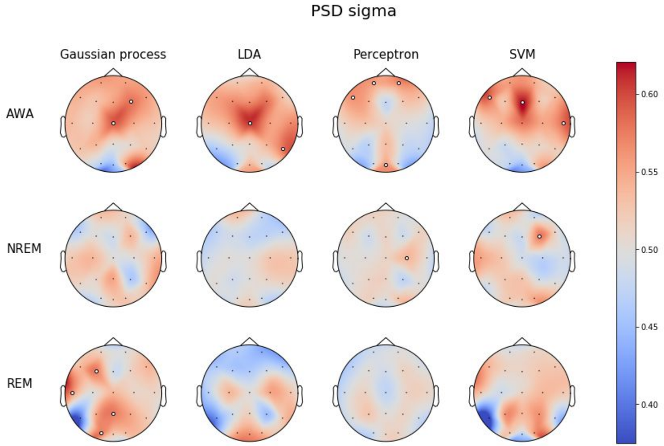


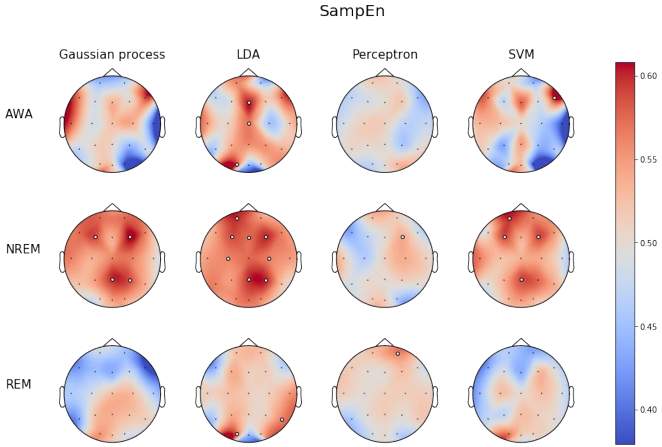
Results - Single-feature machine learning

- significant electrodes in sigma band in AWA across classifiers
  - similar but less strong results in beta band
- good performance on occipital electrodes for PermEn in AWA
  - significant only for two classifiers, effect still visible in the others
- significant decoding accuracies across classifiers for SpecEn, SampEn and SpecSampEn in NREM
- central alpha band scores in NREM show significance across two classifiers
- significant classifiers for occipital electrodes using SpecEn in REM
- theta band yields good scores in REM

Results - Single-feature machine learning

- sigma power in AWA yields significant scores
- good classification results on SpecEn, SampEn and SpecSampEn in NREM
- SpecEn yields significant occipital classifiers in REM





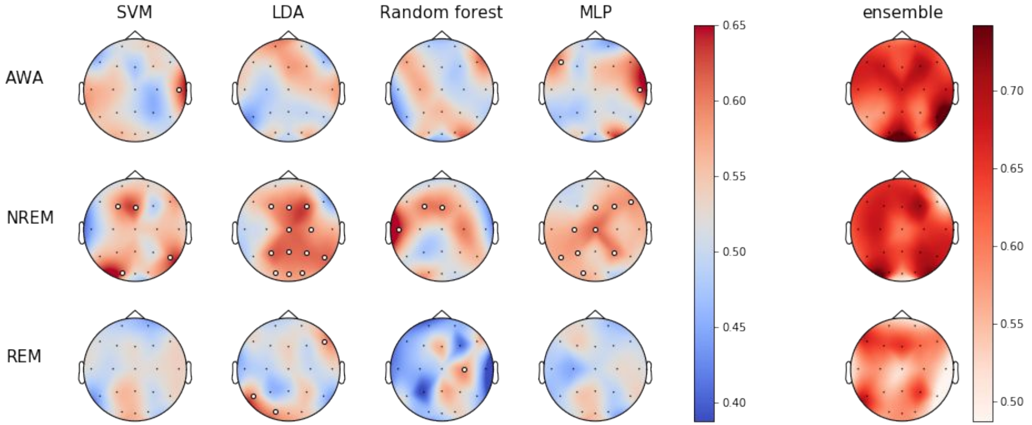
Results - Multi-feature machine learning

Single-electrode classification

- most significant classifiers in NREM but also some in AWA and REM
- LDA and MLP are the most successful classifiers
- overall decoding accuracy improves with ensemble classifier
- ensemble performance in AWA and NREM are similar, REM ensemble is slightly worse
- single classifier scores reach 65% accuracy in some electrodes, ensemble accuracy reaches 74%

Results - Multi-feature machine learning

- ensemble strongly increases classification score
- NREM performs best with single classifiers
- similar ensemble performance in AWA and NREM



Results - Multi-feature machine learning

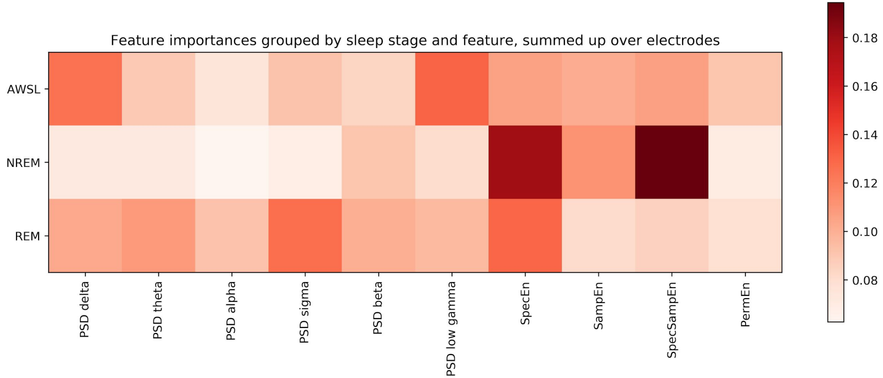
Multi-electrode classification

- best classification accuracy in NREM stage (≈60%), followed by AWA (≈55%)
- score on REM stage is close to the level of random guessing (50%)
- grouped electrode feature importances differ between sleep stages
  - delta and low gamma bands most important in AWA but other features still play a role
  - strong peaks in feature importance for SpecEn and SpecSampEn in NREM
  - sigma band and SpecEn show slightly higher importance than other features in REM
- feature importance variance increases for on average more important features

Results - Multi-feature machine learning

Multi-electrode classification

- best classification accuracy was achieved in NREM
- PSD bands have higher feature importance than entropy for AWA and REM
- entropy is most important in NREM, especially SpecEn and SpecSampEn



Discussion - agreements in statistics and ML

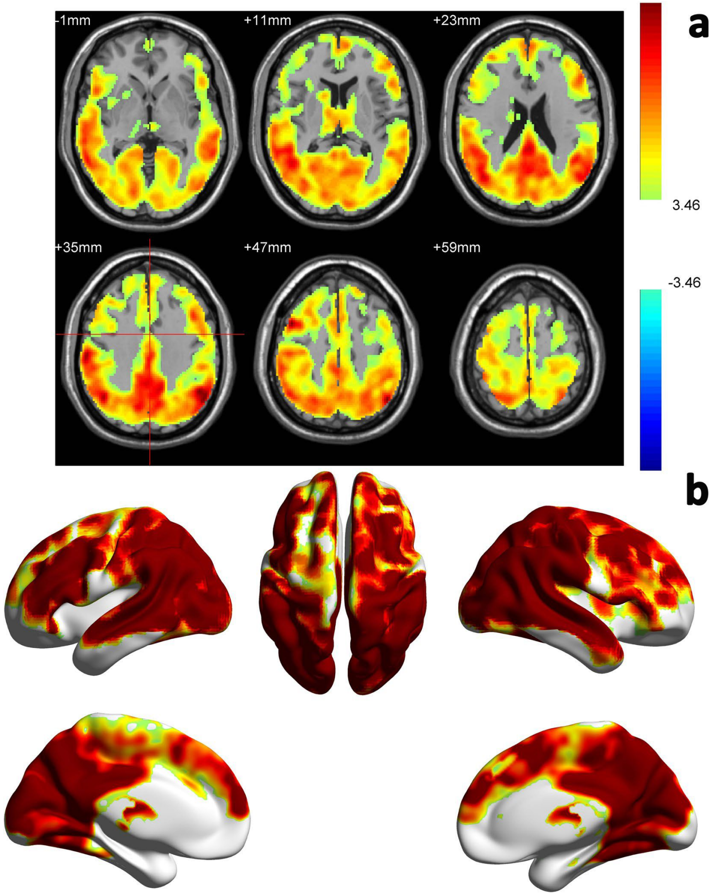
- high importance of SpecEn, SampEn and SpecSampEn in NREM could be observed throughout all methods, making the result more robust
- a significant decrease of the sigma power in AWA visible in statistical analyses
    - high importance of sigma power in AWA in the different machine learning methods
- change in delta power in NREM can be seen in the statistical analysis and single feature, single-electrode classification but not in the feature importance approach with the random forest
    - lacking importance in random forest might be due to a much stronger importance of entropy features

Discussion - Spectral power

- changes in spectral power are consistent across statistical and machine learning analyses
- results match findings from previous studies comparing spectral power in caffeine and placebo conditions

Discussion - fMRI brain entropy

- fMRI study showing an increase in sample entropy in resting brain induced by a 200 mg caffeine dose
- brain entropy increase in large portion of the cerebral cortex, DMN, visual cortex and motor network
- the fMRI results closely match the increase of sample entropy observed in AWA, NREM and REM
    - only statistically significant difference in NREM, might be due to a larger amount of samples in NREM stage leading to less noise after averaging
- increased brain entropy could also be observed in electrophysiological recordings, not only in the BOLD signal



Discussion - SpecSampEn

- SpecSampEn is a new method to measure brain entropy using spectral analysis
- highly important feature in the random forest multi-feature, multi-electrode approach in NREM
- closely matches the results from SpecEn
- might be able to capture more complex patterns in spectral power
    - sample entropy is a more complex measure than Shannon entropy

Discussion - grid search

- using electrode-wise grid search for single-electrode classification yields an increase in variance for decoding accuracy
  - might be due to small sample count, scores determining hyperparameter sets are not expressive enough
  - single grid search for all electrodes would probably increase performance but not variance
- removing grid search from multi-feature, single-electrode classification strongly reduces performance
  - could probably be adjusted for by carefully choosing different fixed hyperparameters
- scores in single-feature, single electrode classification do not increase much if grid search was added for hyperparameter selection

Discussion - 400 mg group

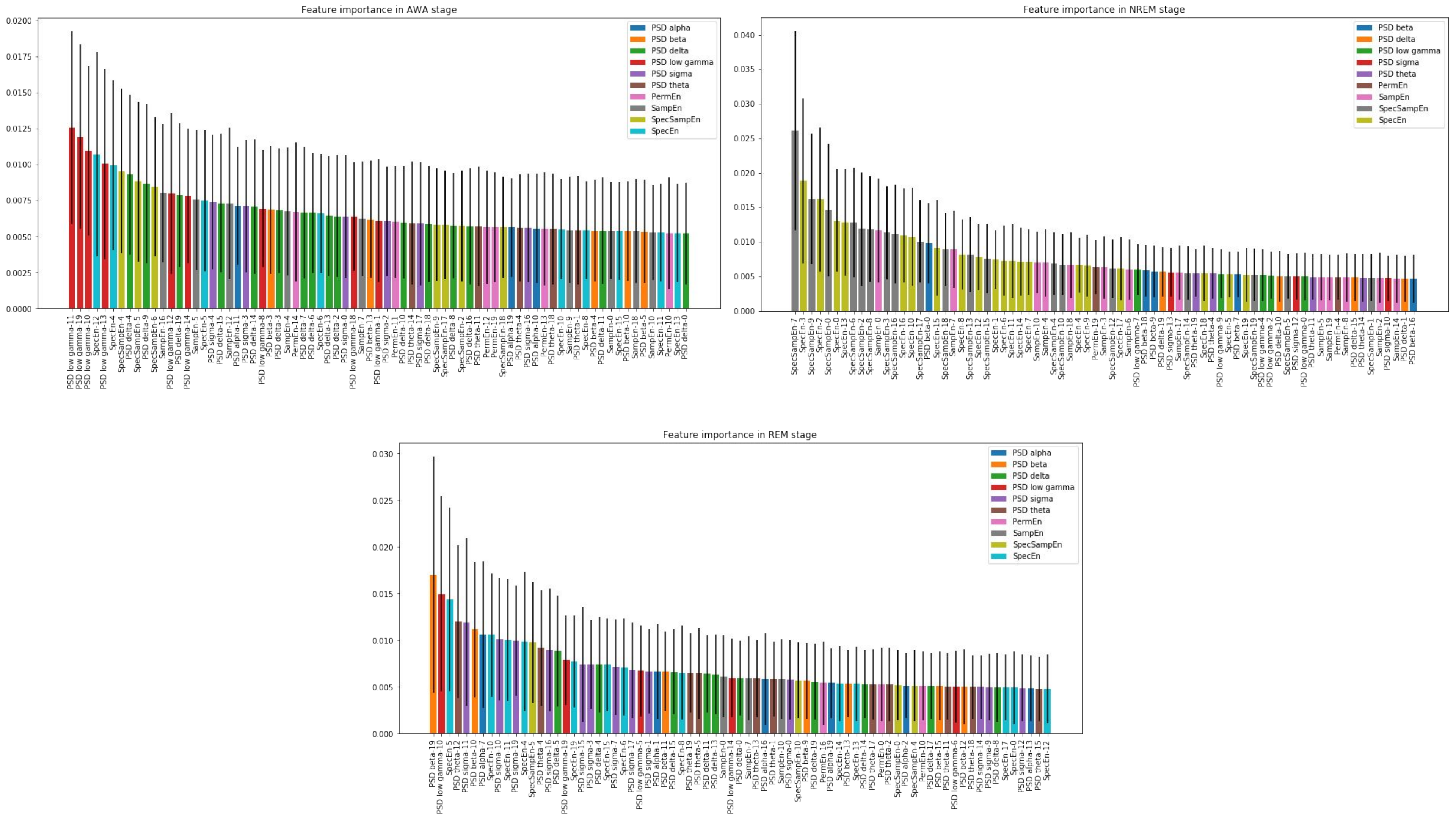
- in the same study, another group of subjects were recorded with a 400 mg dose of caffeine before sleep
    - only 20 subjects, decreased robustness of analyses
- overall similar results as 200 mg group, difficult to distinguish differences due to low subject count
    - stronger decrease in delta band, visible across statistical and machine learning analyses
    - SampEn statistics and machine learning performances show a weaker entropy increase for 400 mg than for 200 mg

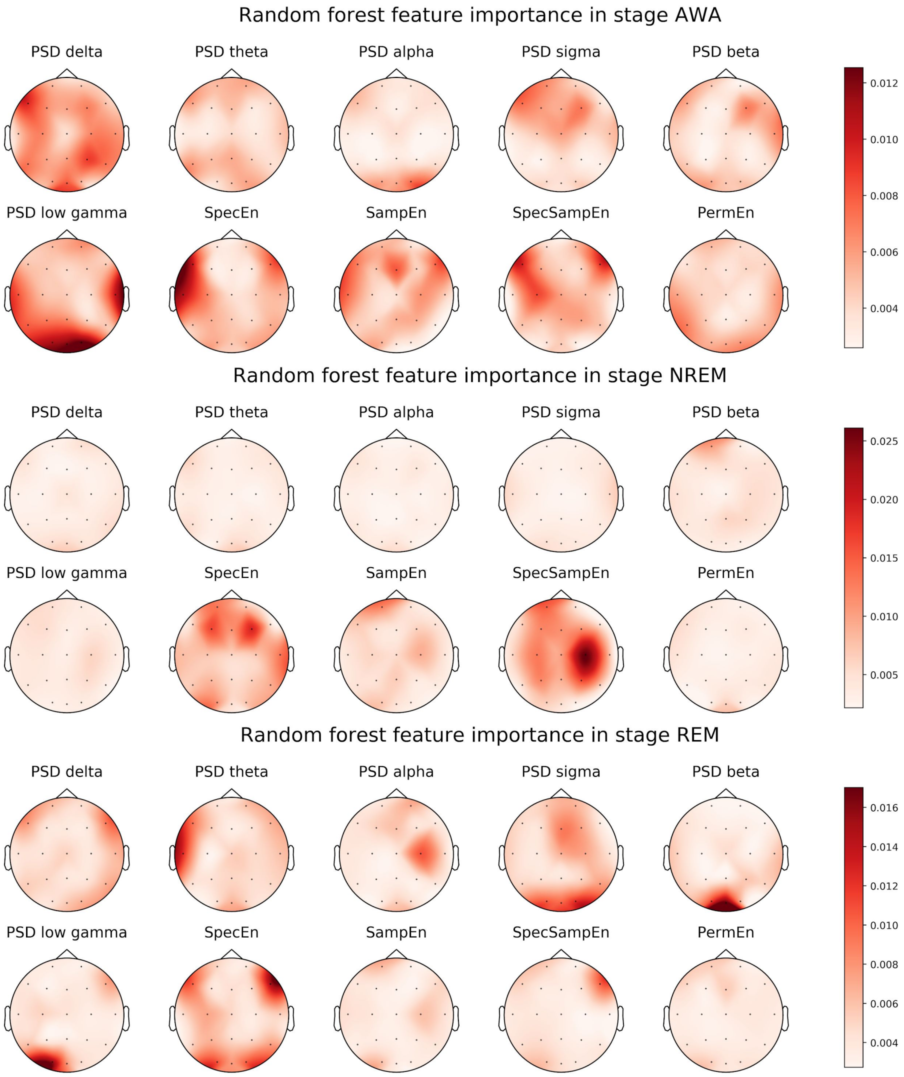
Future work

- analysing difference between 200 mg and 400 mg doses
    - transfer learning, testing on 400 mg data
    - t-SNE showing placebo, 200 mg, 400 mg classes for features
- connectivity analysis (comparison to MEG paper on caffeine)
- comparing PSD vs entropy features in sleep stages (class separation, t-SNE, clustering)
- spectral entropies on frequency intervals (instead of complete spectrum)
- deep learning on raw EEG signal (CNN, LSTM, domain adaptation network)
- removing eye movement artifacts from EEG in REM stage
    - using PCA or ICA to remove artifact components in combined EEG and EOG
    - maybe convolutional autoencoder with modified loss function to exclude eye movements
- further analyses of SpecSampEn, comparison to SpecEn

Thank you for your attention!

Bonus figures


Feature importance in AWA stage
PSD alpha
PSD beta
PSD delta
PSD low gamma
PSD sigma
PSD theta
PermEn
SampEn
SpecSampEn
SpecEn
Feature importance in NREM stage
PSD beta
PSD delta
PSD low gamma
PSD sigma
PSD theta
PermEn
SampEn
SpecSampEn
SpecEn
Feature importance in REM stage
PSD alpha
PSD beta
PSD delta
PSD low gamma
PSD sigma
PSD theta
PermEn
SampEn
SpecSampEn
SpecEn


Random forest feature importance in stage AWA
PSD delta
PSD theta
PSD alpha
PSD sigma
PSD beta
PSD low gamma
SpecEn
SampEn
SpecSampEn
PermEn
0.012
0.010
0.008
0.006
0.004
Random forest feature importance in stage NREM
PSD delta
PSD theta
PSD alpha
PSD sigma
PSD beta
PSD low gamma
SpecEn
SampEn
SpecSampEn
PermEn
0.025
0.020
0.015
0.010
0.005
Random forest feature importance in stage REM
PSD delta
PSD theta
PSD alpha
PSD sigma
PSD beta
PSD low gamma
SpecEn
SampEn
SpecSampEn
PermEn
0.016
0.014
0.012
0.010
0.008
0.006
0.004