Study of a deep learning model for temporal sleep stage classification

Bachelor Thesis Defense

Maëlys Solal[1], Dr Alexandre Gramfort[2] and Dr Olivier Pallanca[3]

[1]École polytechnique

[2]Inria Paris Saclay, Parietal team

[3]Laboratoire d'Informatique de l'École polytechnique (LIX), DaSciM team

January - March 2021

Table of Contents

Table of Contents

Sleep stages

- Sleep cycle = mix between REM sleep and NREM sleep, lasts 90 to 120 minutes and occurs 4-6 times per night
- 5 main sleep stages: Wake (W), Rapid Eye Movement (REM) $\simeq$ paradoxical sleep, Non REM1 (N1) $\simeq$ light sleep. Non REM2 (N2) $\simeq$ deeper sleep, Non REM3/4 (N3/4) $\simeq$ deep sleep.
- Have distinct characteristics as seen in the analysis of brain waves

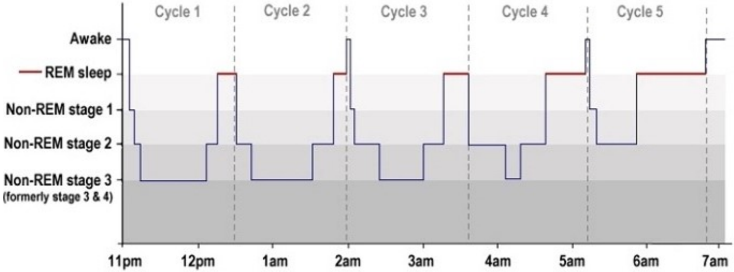


Figure: A typical hypnogram showing sleep stages and cycles, by Luke Mastin

Sleep stages

Polysomnography = sleep study

- Biophysical changes during sleep
- Includes EEG (brain's electrical activity), EOG (eyes), EMG (muscles), and ECG (heart)

Sleep stage classification = sleep scoring

- Visual investigation of the PSG, labelling 30s time segments with sleep stages
- According to a precise set of rules
- Usually done manually by sleep experts (scorers)

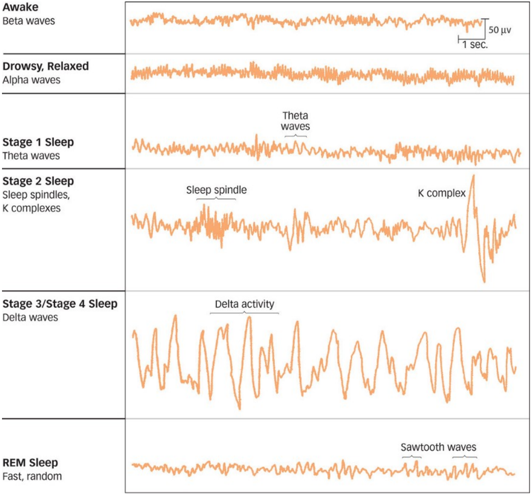


Figure: Brain waves during the different sleep stages, from MacMillan Learning

Motivations for sleep stage classification

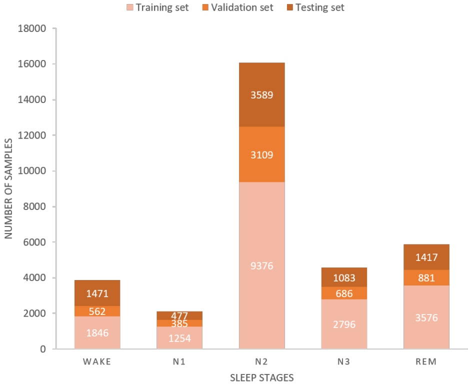


Figure: Sleep stages imbalance in MASS dataset

From a clinical point of view:

- Used as a preliminary examination for clinical diagnosis of sleeping disorders
- Manual scoring is tedious

From a statistical learning point of view:

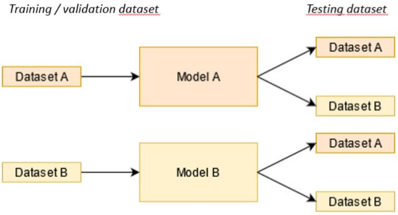
- Multiclass classification with imbalanced classes as shown in Figure 3
- Domain adaptation (i.e. differences of raw data between datasets)
- In general, quite noisy data (especially clinical data)

Table of Contents

Study of a deep learning model for sleep scoring

- Our model of study: deep convolutional neural network, performs temporal sleep stage classification using multivariate and multimodal time series, by Chambon et al. in 2018[a]
- Study the transferability of the model i.e. performance depending on the training/validation sets and testing set

[a] Stanislas Chambon et al. "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series". In: (2018).



Model perfomance depending on the training and testing set

Training/validation \ Testing	Dataset A	Dataset B
Dataset A		
Dataset B		

Figure: Schematic representation of our experiment

Data

- Datasets
  - Montreal Archive of Sleep Studies (MASS)[1]
  - SleepPhysionet[2]
  - Clinical dataset
- Varied number and types of EEG/EMG/EOG channels
  - Figure 5 shows the variety of EEG channels
  - For comparing the performance of the model across datasets: should have similar EEG/EMG/EOG channels

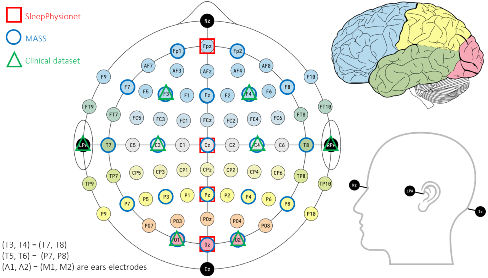


Figure: EEG electrodes positioning used in each dataset

[1] Christian O'Reilly et al. "Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research". In: (2014).

[2] B. Kemp et al. "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG". In: (2000), Ary L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: (2000).

Organising the datasets: BIDS standard

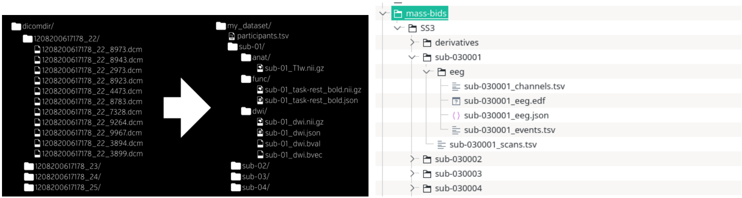


Figure: BIDS standard

- Neuroimaging data is often complicated to arrange
  - Comes from various experiments / medical examinations
  - Outputs multiple files for a single patient
  - Little consensus about how to organise files
- BIDS standard proposes a way to organise this type of data
  - BIDS = Brain Imaging Data Structure[3]
  - subject > session > type / experiment
  - Compatible with existing software
  - Captures metadata

[3] Krzysztof J. Gorgolewski et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". In: (June 2016). DOI: 10.1038/sdata.2016.44. URL: https://www.hal.inserm.fr/inserm-01345616, Stefan Appelhoff et al. "MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis". In: (2019). URL: https://doi.org/10.21105/joss.01896, Cyril Pernet et al. "BIDS-EEG: an extension to the Brain Imaging Data Structure (BIDS) Specification for electroencephalography". In: (Jan. 2018).

Table of Contents

Main steps, code structure

Main steps for running our experiment:

1. Loading the datasets
2. Preprocessing the raw signals and extracting the 30s windows from events
3. Splitting the dataset into training, validations and testing sets
4. Creating / loading our model, training and testing
5. Visualising the results

Code structure, inspired by the `braindecode`[a] library

- `datasets` to load the datasets that were previously converted to BIDS
- `datautil` to take care of splitting the datasets
- `models` to load and save our models
- `visualisation` for visualising results and the 30s windows

[a] Robin Tibor Schirrmeister et al. "Deep learning with convolutional neural networks for EEG decoding and visualization". In: (). URL: http://dx.doi.org/10.1002/hbm.23730.

Description of the model

- General feature extractor, denoted by $Z : \mathbb{R}^{C \times T} \rightarrow \mathbb{R}^{D}$, where C is the number of input channels, T is the number of time steps and D is the size of the estimated feature space
- **Linear spatial filtering**: to estimate virtual channels
- **Convolutional layers**: to capture spectral features
- **Separate pipelines**: to handle several modalities at the same time
- Performs **temporal** sleep staging, that is, takes into account the temporal context of the sample of interest to predict a label

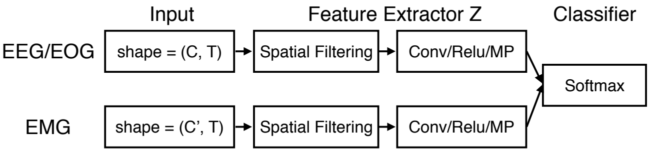


Figure: Network general architecture

Description of the model

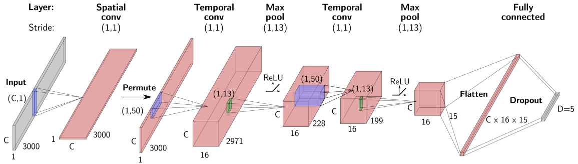


Figure: Schematic representation of the sleep staging model's architecture

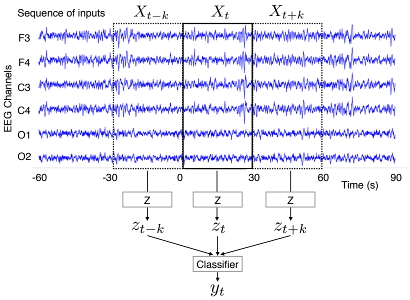


Figure: Schematic representation of the time distributed multivariate network

Pre-processing, training

Preprocessing steps, using `mne-python`[a], same steps as Chambon et al.

- Low-pass filtered at 30Hz: to mitigate the impact of higher frequency noise
- Downsampled to 100Hz: SleepPhysionet was sampled at 100Hz
- Convert signals from V to μV: very small amplitude brain waves
- Cropped 30 minutes of wake events at the beginning and the end of the night
- Divided our signal in 30s samples (windows) corresponding to one specific sleep stage
- Standardised the windows (zero mean, unit variance): cope for the varying recording conditions

[a] Alexandre Gramfort et al. "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroscience* (2013). URL: `https://www.frontiersin.org/article/10.3389/fnins.2013.00267`.

Training specification

- Implemented with `PyTorch`
- 60 subjects for each experiment, splitted using stratified cross-validation: roughly 60% of events in training set, 20% in validation set and 20% in testing set
- Weights initialised with a normal distribution ($\mu = 0$, $\sigma = 0.1$)
- Loss function: categorical cross entropy
- Optimizer: AdamOptimizer
- Minimisation: stochastic gradient descent, learning rate $\texttt{lr} = 5 \times 10^{-4}$, batch size 8, 10 training epochs

MNE

MEG + EEG ANALYSIS & VISUALIZATION

Table of Contents

Typical results for our experiment

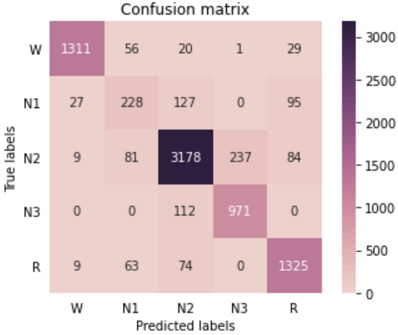


Figure: Confusion matrix for MASS dataset

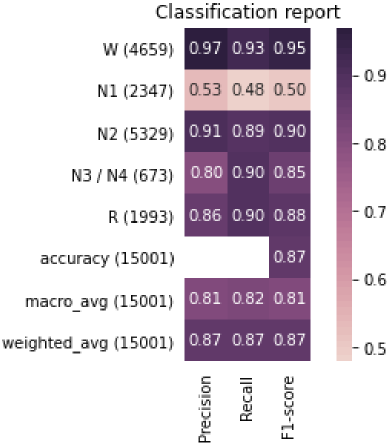


Figure: Classification report for MASS dataset

Results of our main experiment and Discussion

Training set \ Testing set	MASS	SleepPhysionet
MASS	**0.802**	**0.487**
SleepPhysionet	**0.560**	**0.634**

Training set \ Testing set	MASS	Clinical
MASS	**0.817**	**0.390**
Clinical	**0.530**	**0.635**

Table: Balanced accuracy for main experiment

- Clinical-Clinical and SleepPhysionet-SleepPhysionet lower than expected
  - Pre-processing steps
  - Model initially benchmarked using MASS, model biased towards MASS?
  - Hyperparameters are not optimal
- Model transfers better from SleepPhysionet/Clinical to MASS than from MASS to SleepPhysionet/Clinical
  - Specificity in the MASS dataset?
- Domain adaptation remains one of the big challenges of sleep scoring algorithms
- Increasing number of EEG channels and adding EOG and EMG data greatly improves performance

Limitations and Further Works

- SleepPhysionet scored according to the Rechtschaffen and Kales guidelines, MASS and Clinical scored according to the AASM guidelines
  - N4 sleep stage
  - Transition rules
- Tune hyperparameters and improve the model's architecture to get better performance on Clinical and SleepPhysionet
- Core differences in terms of data
  - Clinical datasets are in general more difficult to work with

Table of Contents

Perspectives

Acknowledgements

Annex

Row: MASS — Column: MASS

	W	N1	N2	N3	REM
W	0.87	0.04	0.02	0.00	0.08
N1	0.10	0.41	0.27	0.00	0.22
N2	0.00	0.02	0.89	0.07	0.02
N3	0.00	0.00	0.10	0.90	0.00
REM	0.01	0.02	0.04	0.00	0.94

Row: MASS — Column: SleepPhysionet

	W	N1	N2	N3 / N4	REM
W	0.90	0.02	0.00	0.00	0.08
N1	0.53	0.15	0.07	0.00	0.24
N2	0.14	0.24	0.42	0.05	0.15
N3 / N4	0.10	0.10	0.31	0.46	0.03
REM	0.32	0.12	0.05	0.00	0.50

Row: SleepPhysionet — Column: MASS

	W	N1	N2	N3 / N4	REM
W	0.82	0.10	0.04	0.01	0.04
N1	0.22	0.12	0.16	0.00	0.49
N2	0.04	0.00	0.57	0.04	0.34
N3 / N4	0.22	0.00	0.14	0.52	0.12
REM	0.09	0.06	0.08	0.00	0.77

Row: SleepPhysionet — Column: SleepPhysionet

	W	N1	N2	N3 / N4	REM
W	0.87	0.06	0.06	0.00	0.01
N1	0.32	0.20	0.43	0.00	0.05
N2	0.03	0.04	0.88	0.02	0.03
N3 / N4	0.00	0.00	0.53	0.47	0.00
REM	0.12	0.11	0.25	0.00	0.52

Table: Table of confusion matrices, comparing the datasets MASS and SleepPhysionet

MASS (row) / MASS (column)

	Precision	Recall	F1-score
W (1417)	0.95	0.87	0.91
N1 (477)	0.58	0.41	0.48
N2 (3589)	0.91	0.89	0.9
N3 (1083)	0.79	0.9	0.84
REM (1471)	0.83	0.94	0.88
accuracy (8037)			0.87
macro_avg (8037)	0.81	0.8	0.8
weighted_avg (8037)	0.87	0.87	0.86

MASS (row) / SleepPhysionet (column)

	Precision	Recall	F1-score
W (4659)	0.61	0.90	0.73
N1 (2347)	0.17	0.15	0.16
N2 (5329)	0.82	0.42	0.56
N3 / N4 (673)	0.54	0.46	0.50
R (1993)	0.36	0.50	0.42
accuracy (15001)			0.54
macro_avg (15001)	0.50	0.49	0.47
weighted_avg (15001)	0.58	0.54	0.53

SleepPhysionet (row) / MASS (column)

	Precision	Recall	F1-score
W (4659)	0.64	0.82	0.72
N1 (2347)	0.21	0.12	0.15
N2 (5329)	0.84	0.57	0.68
N3 / N4 (673)	0.76	0.52	0.61
R (1993)	0.41	0.77	0.54
accuracy (15001)			0.62
macro_avg (15001)	0.57	0.56	0.54
weighted_avg (15001)	0.68	0.62	0.62

SleepPhysionet (row) / SleepPhysionet (column)

	Precision	Recall	F1-score
W (4659)	0.78	0.87	0.82
N1 (2347)	0.4	0.2	0.27
N2 (5329)	0.69	0.88	0.77
N3 / N4 (673)	0.76	0.47	0.58
REM (1993)	0.76	0.52	0.62
accuracy (15001)			0.7
macro_avg (15001)	0.68	0.59	0.61
weighted_avg (15001)	0.68	0.7	0.68

Table: Table of classification reports, comparing the datasets MASS and SleepPhysionet

		MASS					Clinical				
		W	N1	N2	N3	REM	W	N1	N2	N3	REM
MASS	W	0.93	0.04	0.01	0.00	0.02	0.86	0.01	0.04	0.00	0.08
	N1	0.06	0.48	0.27	0.00	0.20	0.80	0.14	0.02	0.00	0.04
	N2	0.00	0.02	0.89	0.07	0.02	0.52	0.11	0.32	0.01	0.04
	N3	0.00	0.00	0.10	0.90	0.00	0.26	0.02	0.38	0.25	0.09
	REM	0.01	0.04	0.05	0.00	0.90	0.38	0.19	0.05	0.00	0.38
Clinical	W	0.80	0.00	0.08	0.00	0.12	0.94	0.00	0.02	0.00	0.03
	N1	0.07	0.00	0.39	0.00	0.55	0.49	0.01	0.22	0.01	0.28
	N2	0.01	0.00	0.87	0.01	0.12	0.08	0.00	0.76	0.12	0.04
	N3	0.01	0.00	0.87	0.11	0.00	0.03	0.00	0.11	0.86	0.00
	REM	0.02	0.00	0.11	0.00	0.87	0.20	0.00	0.19	0.00	0.61

Table: Table of confusion matrices, comparing the datasets MASS and Clinical

Row: MASS — Column: MASS

	Precision	Recall	F1-score
W (1417)	0.97	0.93	0.95
N1 (477)	0.53	0.48	0.5
N2 (3589)	0.91	0.89	0.9
N3 (1083)	0.8	0.9	0.85
REM (1471)	0.86	0.9	0.88
accuracy (8037)			0.87
macro_avg (8037)	0.81	0.82	0.81
weighted_avg (8037)	0.87	0.87	0.87

Row: MASS — Column: Clinical

	Precision	Recall	F1-score
W (3560)	0.46	0.86	0.6
N1 (389)	0.06	0.14	0.08
N2 (4094)	0.59	0.32	0.41
N3 (1702)	0.9	0.25	0.4
REM (1879)	0.53	0.38	0.44
accuracy (11624)			0.48
macro_avg (11624)	0.51	0.39	0.39
weighted_avg (11624)	0.57	0.48	0.46

Row: Clinical — Column: MASS

	Precision	Recall	F1-score
W (1417)	0.93	0.8	0.86
N1 (477)	1	0	0
N2 (3589)	0.69	0.87	0.77
N3 (1083)	0.86	0.11	0.2
REM (1471)	0.59	0.87	0.71
accuracy (8037)			0.7
macro_avg (8037)	0.81	0.53	0.51
weighted_avg (8037)	0.76	0.7	0.65

Row: Clinical — Column: Clinical

	Precision	Recall	F1-score
W (3560)	0.78	0.94	0.86
N1 (389)	0.18	0.01	0.01
N2 (4094)	0.82	0.76	0.79
N3 (1702)	0.73	0.86	0.79
REM (1879)	0.74	0.61	0.67
accuracy (11624)			0.78
macro_avg (11624)	0.65	0.64	0.62
weighted_avg (11624)	0.76	0.78	0.76

Table: Table of classification reports, comparing the datasets MASS and Clinical