Self-supervised Deep Learning and EEG Categorization

Mats Svantesson

FACULTY OF MEDICINE AND HEALTH SCIENCES
Linköping University Medical Dissertation No. 1901, 2024
Department of Clinical Neurophysiology in Linköping
Department of Clinical and Experimental Sciences
Linköping University
SE-581 83 Linköping, Sweden

www.liu.se
Self-supervised deep learning and EEG categorization

Mats Svantesson

Department of Clinical Neurophysiology in Linköping
Department of Biomedical and Clinical Sciences
Faculty of Medicine and Health Sciences
Linköpings universitet, SE-581 83 Linköping, Sweden
Linköping 2024
1 Abstract

Deep learning has the potential to be used to improve and streamline EEG analysis. At the present, classifiers and supervised learning dominate the field. Supervised learning depends on target labels which most often are created by human experts manually classifying data. A problem with supervised learning is intra- and interrater agreement which in some instances are far from perfect. This can affect the training and make evaluation more difficult.

This thesis includes three papers where self-supervised deep neural networks were developed. In self-supervised learning, the input data to the networks themselves contain structures that are used as targets for the training and no labeling is necessary.

In paper I, deep neural networks were trained to increase the number of, or to recreate missing EEG-channels. The performance was at least on the same level as that of spherical interpolation, but unlike in the case of interpolation, missing data does not have to be identified manually first.

Papers II and III involved developing deep neural networks for clustering analysis. The networks produced two-dimensional representations of EEG data and the training strategy was based on the principle of t-distributed stochastic neighbor embedding (t-SNE).

In paper II, comparisons were made to parametric t-SNE and EEG-features obtained from time-frequency methods. The deep neural networks produced more distinct clustering when tested on data annotated for epileptiform discharges, seizure activity, or sleep-wakefulness.

In paper III, the newly developed method was used to compare annotations of epileptiform discharges. Two experts performed independent annotations and classifiers were trained on these, using supervised learning, which in turn produced new annotations. The agreement when comparing two sets of annotations was not larger between the two experts than between an expert and a classifier. The analysis showed that differences in the annotations by the experts influenced the training of the classifiers. However, the clustering analysis indicated that although it was not always the exact same waveforms that were assessed as epileptiform discharges, they were often similar.

The work thus resulted in different methods to process and analyze EEG data, which may have practical usefulness. Traditional agreement scores only assess the exact agreement. However, they reveal nothing about the nature of disagreement. Cluster analysis can provide a means to perform this assessment.
2 Svensk sammanfattning

Artificiell intelligens (AI) har fått mycket uppmärksamhet de senaste åren. Den har börjat smyga sig in i våra liv och arbetet med att utveckla den pågår inom de flesta områden. Den dominerande metoden för att skapa artificiell intelligens i dagsläget kallas djupinlärning och bygger främst på s.k. artificiella neurala nätverk.

En vanlig uppgift man vill lösa med AI inom sjukvården är att klassificera patienter som sjuka eller friska. Inom EEG-analys kan det t.ex. vara att detektera anfallsaktivitet i EEG under epileptiska anfall. Den vanliga strategin är då att låta EEG-tolkare bedöma EEG och markera när anfall uppträder. Detta används sedan som facit när man tränar och utvärderar nätverken. Tyvärr finns det fall då det är svårbedömt om det är någon anfallsaktivitet och man kan då få en variation i hur olika EEG-tolkare gör bedömningar. Denna variation kan påverka resultatet när man tränar nätverken men det blir också svårare när man utvärderar dem eftersom vad som anses rätt och fel varierar.

Denna avhandling inom klinisk neurofysiologi och EEG studerade denna variation, men framför allt vidareutvecklade metoder för att analysera EEG utan att använda någon föregående expertbedömning.


De två efterföljande arbetena handlade främst om att utveckla nätverk som skapade "bilder" av EEG där korta EEG-segment representerades som punkter. Deras placering i förhållande till varandra i bilden reflekterade hur lika eller olika segmenten var. Detta innebär att t.ex. punkter som motsvarade anfallsaktivitet hamnade nära varandra. Kategorier i data bildar alltså kluster i bilderna och kan då lättare identifieras.

I andra arbetet jämfördes med några andra alternativ för klusteranalyser och den utvecklade metoden producerade kluster av viktiga EEG-mönster och som var lättare att identifiera.

I tredje arbetet förfinades metoden ytterligare och användes för att studera skillnader mellan hur neurofysiologer och klassificerare bedömer EEG. Här visades att det fanns skillnader i bedömningarna och att detta kunde påverka klassificerarna, men även att om bedömningarna inte var exakt samma så var de oftast ganska lika.

Avhandlingen visar potentialen för flera metoder att användas för olika typer av EEG-analys, både kliniskt och inom forskning.
3 Acknowledgement

The work was funded by Saad Nagi (grants RÖ-974228, RÖ-962769, and RÖ-941377), Magnus Thordstein (grant RÖ-986017), and Håkan Olausson (grants LIO-936017 and RÖ-941359).

I would like to thank for the support during the work from: The main supervisor Magnus Thordstein and co-supervisor Håkan Olausson at the Department of Clinical Neurophysiology in Linköping. Co-supervisor Anders Eklund at the Department of Biomedical Engineering, Linköping University. Greta Gustafsson, head of the department, Hans Lindehammar, altmeister, and all the other colleagues in all professions at the Department of Clinical Neurophysiology in Linköping.
4 List of papers


5 Acronyms

AASM American Academy of Sleep Medicine
ACC Accuracy
AI Artificial intelligence
AUC Area under the curve
BCI Brain computer interface
CNN Convolutional neural network
DNN Deep neural network
ED Epileptiform discharge
EEG Electroencephalography or -gram
EMG Electromyography
EOG Electrooculography
GAN Generative adversarial network
GPU Graphics processing unit
ICU Intensive care unit
IFCN International Federation of Clinical Neurophysiology
ILAE International League Against Epilepsy
LEP Laser evoked potential(-s)
LLM Large language model
LTM Long-term monitoring
MCC Matthew's correlation coefficient
MRI Magnetic resonance imaging
PDR Posterior dominant rhythm
PSG Polysomnography
ReLU Rectifying linear unit
REM Rapid eye movement
ROC Receiver operating characteristic
SELU Scaled exponential linear unit
SVM Support vector machine
t-SNE t-distributed stochastic neighbor embedding
6 Contents
1 Abstract.................................................................................... iii
2 Svensk sammanfattning........................................................... v
3 Acknowledgement ...................................................................... vii
4 List of papers ........................................................................... ix
5 Acronyms ................................................................................. xi
6 Contents............................................................................................ 13
7 List of figures .............................................................................. 16
8 Introduction................................................................................... 17
  8.1 Prologue..................................................................................... 18
    8.1.1 Classifiers for laser evoked potentials .................... 18
    8.1.2 Intrarater variability ................................................. 19
    8.1.3 Illustrating the problem of rating variability ............ 20
  8.2 Outline of the thesis .............................................................. 21
9 EEG .............................................................................................. 23
  9.1 Summary ................................................................................ 23
  9.2 Neurophysiological background ........................................... 23
  9.3 Recording and basic signal processing .................................... 24
  9.4 EEG interpretation .................................................................. 26
    9.4.1 Characterization of EEG signals ......................... 26
    9.4.2 EDs ................................................................................. 28
    9.4.3 Seizure activity ......................................................... 29
    9.4.4 Sleep stages ............................................................ 29
  9.5 The Temple University EEG Corpus ..................................... 30
10 Agreement scoring ....................................................................... 33
  10.1 Summary ............................................................................... 33
  10.2 Metrics ................................................................................. 33
  10.3 Interrater agreement in EEG ................................................... 35
    10.3.1 Agreement scores for EDs ....................................... 35
    10.3.2 Agreement scores for seizure activity ..................... 37
10.3.3 Agreement scores for sleep staging ......................... 38
10.3.4 Intrarater agreement ............................................... 39
10.3.5 Characteristics of agreement of EDs ....................... 39
11 Deep learning ............................................................ 43
11.1 Summary .................................................................. 43
11.2 The principle of DNNs ............................................. 44
11.3 Layers ..................................................................... 45
11.4 Network architectures: Classifiers and autoencoders .... 48
11.5 Loss function and backpropagation ......................... 50
11.6 Normalization ......................................................... 50
11.7 Supervised and self-supervised learning .................... 51
11.8 Data representations ............................................... 53
11.9 t-SNE ...................................................................... 55
11.10 Evaluation metrics .................................................. 56
11.11 Deep learning and EEG ......................................... 57
11.11.1 Supervised learning and EEG ............................ 58
11.11.2 Self-supervised learning and EEG .................... 61
11.11.3 t-SNE and EEG .............................................. 63
11.12 Deep learning from a practical point of view .......... 63
12 Aims .......................................................................... 65
12.1 Initial aims ............................................................. 65
12.2 Realigned aims ....................................................... 66
13 Ethical considerations .................................................. 67
14 Results ......................................................................... 69
14.1 Paper I – Upsampling and reconstruction of EEG channels ............................................. 69
14.2 Paper II – t-SNE encoders ..................................... 69
14.3 Paper III – Intrarater agreement and classification EDs 69
14.4 Additional examples .............................................. 70
14.4.1 Channel reconstruction ..................................... 70
14.4.2 Comparison of some approaches for t-SNE encoders

15 Discussion

15.1 Signal reconstruction

15.2 t-SNE encoders

15.3 EEG interpretation and scoring systems

15.4 Is it reasonable to expect classifiers to achieve high scores when experts disagree?

15.5 Deep learning and EEG

\hspace{1em} 15.5.1 General reflections from the work

\hspace{1em} 15.5.2 Network architecture and waveform analysis

\hspace{1em} 15.5.3 Normalization and activation functions

\hspace{1em} 15.5.4 Training schedule

15.6 Future developments

\hspace{1em} 15.6.1 Artificial general intelligence

\hspace{1em} 15.6.2 Impediments for AI in healthcare

15.7 Ethical considerations and risks

15.8 Limitations of the papers and the thesis

16 Conclusions

17 Appendix

\hspace{1em} 17.1 Accuracy model

\hspace{1em} 17.2 Simulation for Cohen’s kappa and Gwet’s AC1

\hspace{1em} 17.3 Channel reconstruction

\hspace{1em} 17.4 Comparisons of clustering approaches

\hspace{2em} 17.4.1 Approach 1 – Semi-supervised learning

\hspace{2em} 17.4.2 Approach 2 – Supervised learning

\hspace{2em} 17.4.3 Approach 3 – Self-supervised simple encoder

\hspace{2em} 17.4.4 Approach 4 – Self-supervised general encoder

\hspace{2em} 17.4.5 Approach 5 – Self-supervised cluster encoder

\hspace{1em} 17.5 Illustration of the effect of normalization strategies

18 References
7 List of figures

Fig. 1. Ordinal quality scale for single LEPs...............................................19
Fig. 2. Summary of data, theoretical model, and results of
simulation for rating variability..........................................................21
Fig. 3. Illustration of the electrode placement according to the
international 10–20 system and visualization of EEG. ......................24
Fig. 4. Examples of different EEG signals.............................................28
Fig. 5. Comparison of Cohen's kappa and Gwet's AC1 in a binary
categorization....................................................................................35
Fig. 6. Illustration of the general structure for neural networks
and the loss function.........................................................................45
Fig. 7. Illustration of three of the most common layers......................46
Fig. 8. Illustration of activation functions...........................................48
Fig. 9. Examples of DNN designs..........................................................49
Fig. 10. Illustration of the basic GAN..................................................52
Fig. 11. Illustration of five different data representation of a 15-s
EEG-signal.........................................................................................54
Fig. 12. Illustration of dimensional reduction.................................56
Fig. 13. Number of publications in PubMed for EEG and deep
learning...............................................................................................58
Fig. 14. The average spectra of ten missing signals..........................71
Fig. 15. Example of ten reconstructed missing signals ....................72
Fig. 16. Examples of different clustering approaches.......................74
Fig. 17. Distributions of the EEG per cluster..................57
Fig. 18. Alternative color coding.........................................................75
Fig. 19. Examples illustrating the effect of normalization.................85
Fig. 20. Illustration of two autoencoder architectures....................98
Fig. 21. Illustration of the basic principle of the t-SNE encoder. 99
Fig. 22. Confusion matrix for test data of the sleep stage classifier.
.................................................................................................100

All figures were originally created for the thesis.
8 Introduction

*The world is everything that is the case.*

—Ludwig Wittgenstein, “Tractatus Logico-Philosophicus”

The aim of the work presented here was to assess the problem of categorizing EEG data and to facilitate this, develop a tool using artificial neural networks that could aid in studying how EEG patterns compare.

The traditional way to categorize EEG, e.g., epileptiform discharges (EDs) and seizure activity, is arbitrary and we do not know if it is optimal. Furthermore, when using the categorization system, it must be interpreted, and this interpretation may vary. There may thus be both intra- and interrater variability.

Classifiers are usually developed using supervised learning based on data categorized by experts. If the categorization system is suboptimal, the classifiers will thus replicate this suboptimal interpretation of the data. In addition, rater variability may have a negative impact on learning and make evaluation of the classifiers more difficult.

Clustering analysis is a way to visually assess data and identify categories. There are methods to perform clustering analysis of data which are un- or self-supervised. A problem for EEG in this setting, common to machine learning, is how the data should be represented to make data sizes manageable and to achieve the best results. In many instances of machine learning, substantial preprocessing of the EEG is performed to generate a new representation of the data. Deep neural networks (DNNs) usually solve this problem since the they can take the raw EEG as input and transform the data into a “better” representation within themselves, which is learned during training.

An overarching idea in this work is to use DNNs trained using self-supervised strategies to generate new representations of EEG that may be more suitable for clustering analysis compared to the raw data or commonly used preprocessed representations. In the context of studying different ways of categorizing the same data, using a self-supervised method could reduce bias or uncertainty compared to supervised learning based on a certain categorization system using data with rating variability.
Another motivation for working on developing self-supervised methods are the large language models (LLMs) that currently are emerging and the basis for these is self-supervised learning (Rothman, 2022). If this represents a benchmark for a successful path to developing artificial intelligence, then it is important to find the corresponding way of developing applications for EEG.

A short prologue based on work which formed the seed of this thesis is given first to illustrate the problem of supervised learning and rating variability.

8.1 Prologue

The motivation for the thesis originates from trying to create classifiers for laser evoked potentials (LEPs). These classifiers sadly turned out to perform poorly. A subsequent analysis of the manually scored LEP data used to develop the classifiers revealed a high intrarater variability. It was speculated that this was a large contributor to the poor performance of the classifiers.

8.1.1 Classifiers for laser evoked potentials

LEP was introduced in 2016 at the Department of Clinical Neurophysiology, Linköping University Hospital. It is a method used to assess the function of the pain pathways (Valeriani et al., 2012). By selectively stimulating pain receptors in the skin with short laser pulses, evoked cortical potentials can normally be registered by scalp electrodes. The signals are propagated through the peripheral thin nerve fibers, the spinothalamic tract in the spinal cord, the thalamus, and finally the cerebral cortex. Injury or disease at any level of the pain pathway can affect the LEPs, producing low amplitude, longer latency, or abolish the LEPs altogether.

The result from single laser stimulations may not be visible due to background activity or noise (Fig. 1; it is also possible that there is no cortical response). The average of several stimulations is therefore used. However, when working with the method, it was noted that clearly visible potentials from single stimulations were often produced by healthy subjects (Fig. 1 F). It was also noted that the averages often converged quickly and that the number of stimulations could possibly be reduced. A hypothesis was that the presence of these distinctly visible responses with normal latency could indicate a low probability of disease of the pain pathways (as detectable by standard LEP).
To assess this, a technique to detect and classify responses from single stimulation had to be developed. An arbitrary discrete quality scale 0–5 was constructed (Fig. 1 A–F): 0) no visible response, 1) possible but very uncertain response, 2) possible but somewhat uncertain response, 3) vague response, 4) distinct but blunt response, and 5) distinct and sharp response.

**Fig. 1.** Ordinal quality scale for single LEPs. Examples of scored potentials randomly selected from each class. Even in this small selection of six stimulations the rating can be questioned—How large difference is there between the class 3 and 4 potentials?

LEP data were manually scored by the author according to the scale. Classifiers were then developed using DNNs which were trained to replicate the scoring. The performance of the classifiers was disappointing with accuracies mostly in the range 65–70%.

### 8.1.2 Intrarater variability

The suboptimal results could of course be due to several reasons, e.g., the small sizes of the datasets or suboptimal network architectures. However, when visually assessing the scorings produced by the
classifiers, they looked reasonable, a human might as well have made them. At least part of the problem could thus be the manual scoring.

To test the intrarater variability, batches of 500 responses were scored twice by the author, and the resulting agreement between scorings of the same data was around 65%.

The question was how this affects the learning and evaluation of the networks. If the scoring is inconsistent, it will likely be harder for the networks to learn, but it will also mean that evaluating will be harder since there will be “errors” in the ground truth. It was speculated that the detrimental effect in learning from random errors would be less than for systematic errors, i.e., randomness would be averaged out while systematic errors could be learned.

8.1.3 Illustrating the problem of rating variability

When data are scored, each example of the data is labeled according to which category it belongs to. Intra- and interrater variability thus introduces variation in the scoring, i.e., label noise.

For purely random label noise, simpler datasets, and some form of imagined ideal classifier, it is possible to construct a simple model for the accuracy. If the categories exist and have distinctive features that the classifier can detect it is assumed that the classifier will recognize the categories and assign the most probable label. This would be the label that most frequently is associated with the category during training. However, in the real case, the data used to evaluate the classifier will also have label noise and the resulting classification error will then be proportional to the amount of label noise. In the model, the accuracy when testing the classifier using noisy or noise-free labels can be compared (Fig. 2 B; red = noisy, green = noise-free).

A simple single-channel EEG dataset is used to illustrate the phenomenon. It has three categories: 1) EDs, 2) muscle activity, and 3) alpha rhythms (Fig. 2 A). There is a uniform distribution of examples across the categories. This dataset was created manually but due to its simplicity, it is assumed that the label noise is very low to start with and it was possible to train networks to an almost perfect accuracy using the original labels (c.f., clustering analysis in Fig. 19). For more information on the theoretical model for this dataset (Fig. 2 B), see Appendix 17.1.
DNNs were trained with different degrees of artificially induced label noise and evaluated with both corresponding level of label noise and noise-free labels. The results followed the theoretical model relatively well. As expected, when evaluating with noise-free labels a shift from almost no correctly predicted examples to almost all occurred when more than one third of the training examples were correctly labeled (Fig. 2 C; green curve). However, when evaluating with noisy labels, the error was proportional to the noise (Fig. 2 C; red curve).

This property of resistance to random label noise has been shown previously for image data (Rolnick et al., 2017), it has here now been demonstrated for EEG data, and it is probably a general property. Purely random label errors may not be common in real classification problems, but if so, may thus not be a major issue. However, the example above demonstrate that a non-optimal accuracy may be due to label noise, which can be hard to detect and result in the faulty conclusion that a classifier performs poorly. In practice, data usually are more complex, and the label noise may be due to a combination of random and systematic errors or variation. For instance, the rater may have a threshold for a certain category, but the threshold may drift over time. Or the rater may make mistakes and some types of mistakes may be more common than others.

8.2 Outline of the thesis

A brief introduction to the method EEG is given in chapter 9. The principles for agreement scoring and the results of studies of inter-rater agreement for EEG is reviewed in chapter 10 as an introduction
and to assess the problem of interrater variability of EEG interpretation. An overview of the basic principles of deep learning, t-SNE, and research where deep learning is applied to EEG is provided in chapter 11. The aims of the research studies and how these changed during the work are accounted for in chapter 12. The ethical considerations are detailed in chapter 13. A summary with results of the included paper of the thesis and a couple of additional examples are provided in chapter 14. The discussion is in chapter 15. The conclusions are presented in chapter 16. Some more technical details have been referred to an appendix, chapter 17.

The thesis is formally on the subject of clinical neurophysiology but is in practice interdisciplinary. Given the formal subject, it has been refrained from delving into many technical and mathematical details. Chapter 9 on EEG is, in addition to providing description of EEG categories that are central in the thesis, intended as a short introduction aimed for, e.g., data scientists and engineers. Chapter 11 on deep learning is relatively extensive with the intention that it to some capacity can serve as an introduction to the subject for the interested clinical neurophysiologist.

The intersection of different fields may cause readers with different expertise to experience a lack of references at some point or be interested to learn more. Therefore, three general references are provided here, for the basic concepts and further reading. For EEG, an advanced reference is “Niedermeyer’s Electroencephalography” edited by Schomer and Lopez da Silva (Oxford University Press, 7th ed, 2018). Interrater agreement is discussed extensively in “Handbook of Interrater Reliability” by Kilem L. Gwet (Advanced Analytics LCC, 4th ed, 2014). Finally, for deep learning, the book “Deep Learning” by Goodfellow et al. (Massachusetts Institute of Technology, 2016) is recommended.
9 EEG

9.1 Summary

EEG is an acronym for electroencephalography. It is a method for recording the electrical fields from large populations of synchronized neurons of the cerebral cortex. The activity is recorded using electrodes, often placed on the scalp according to the international 10–20 system. The signals are usually band pass filtered and can be referenced in different ways, which will affect how the signals appear.

Many signal patterns have been identified and defined, the most important for this thesis being EDs, seizure activity, and sleep stages. In the clinical context, EEG is mainly assessed visually and interpreted according to a learned tradition. The interpretation is usually summarized as a written report.

9.2 Neurophysiological background

This section on the origin of the signals recorded on EEG is based on the book “Electrical Fields of the Brain” by Nunez and Srinivasan (Oxford University Press, 2nd ed, 2006), which is written from a physicist’s point of view and is recommended.

The recorded activity is due to variation in the post-synaptic membrane potentials of pyramidal neurons in the cortex. The field of a single neuron is very weak, so it is only when a large enough mass of neurons has synchronized activity that a signal can be recorded. In addition, there are several anatomical conditions that must be met, the most important being the orientation of the neurons and their closeness to the recording electrodes. Each pyramidal cell forms a dipole, and together with nearby parallel synchronized pyramidal cells form a dipole layer. The signal strength will depend on the angle and the distance between the layer and the recording electrode. This means that most of the recorded activity is from the gyri of the cortex just below the cranium near the scalp.

The spatial resolution is relatively low, which is due to smearing of the signals because of differing characteristics of the tissue, e.g., conductivity and anisotropy. An example is the cranial bone which has a low conductivity. Defects in the bone may result in localized higher conductivity, shunting signals, producing a smearing effect. There is also a spatial low pass filtering effect due to the cranial bone. It has been estimated that a cortical area of 6 cm² or more is necessary to produce a recordable signal on the scalp (Cooper et al., 1965). Since synchronized high frequency activity usually has a smaller spatial
distribution, this means that higher frequencies will be recorded to a lesser extent compared to lower frequencies.

In some instances, it is of interest to find the source of EEG signals, e.g., presurgical investigation of epilepsy. This is referred to as solving the inverse problem, or more commonly, source localization (Eom, 2023). However, the same electrical field recorded by the electrodes can theoretically be generated by a very large number of combinations of different sources. The number of solutions can be reduced by making assumptions, e.g., that the source consists of one dipole. There are several algorithms and techniques for source localization, which involve modeling the anatomy of the head and can include a combination with MRI.

9.3 Recording and basic signal processing

In this work, only scalp recordings were used, and the term EEG will throughout the thesis refer to scalp-EEG. The scalp electrodes are most often placed in a standardized way. The international 10–20 system (Jasper, 1958) is probably the most common and is used for all EEG material in the included papers of the thesis (Fig. 3). The system is based on relative divisions of 10 or 20% of distances between fixed anatomical points and is thus independent of head size and shape. This will fairly well result in a consistent placement of the electrodes relative to the cortex.

![Fig. 3. Illustration of the electrode placement according to the international 10–20 system and visualization of EEG. Electrodes have uneven numbers on the left side and even numbers on the right side. Fp = frontopolar, F = frontal, C = central, T = temporal, P = parietal, O = occipital, and A = aurical. (A) View of electrode positions from above with the anterior direction up. (B) Example of electrode order when visualizing the EEG referenced to the common average.](image)
During the recording, all electrodes are referenced, usually to a separate electrode. The electrodes can be re-referenced in several ways, and these are often referred to as montages. Different montages will have different effects on the signal and the same activity may appear differently or only in some montages (Fisch, 1999). In bipolar montages, the signals are the difference between neighboring electrodes. This will accentuate local activity but may be more sensitive to random signals as muscle activity and low amplitudes may be incorrectly interpreted as suppressed activity when the activity is very similar for the compared electrodes, e.g., when there is conduction on the skin between electrodes or when the activity is synchronized over larger areas. Another common montage is to subtract the average of all electrodes (CAR). This montage causes many waveforms of interest to have the same polarity independent of location, but one pitfall is to confuse signals due to the reference as local activity, e.g., this may mask a cortical area with suppressed activity. This montage was used throughout the thesis.

The resulting signals recorded on the scalp that are due to cortical activity are weak, usually below 1 mV and must be amplified. The electrodes are exposed to other electrical sources that may cause artifacts in the recording, e.g., muscle activity and eye movements. The International Federation of Clinical Neurophysiology (IFCN) and the International League Against Epilepsy (ILAE) provide recommendations for standard (minimum) recordings of EEG (Peltola et al., 2023). A sampling frequency of 256 Hz is recommended, but many modern systems can sample at higher rates. The signals should be bandpass filtered using high pass at 0.3 Hz and low pass at 70 Hz. In practice, a notch filter may be necessary to reject AC noise or setting the low pass filter below the frequency of the AC noise. They recommend the 25-electrode IFCN montage (Seeck et al., 2017), an extension of the 10–20 system, which is otherwise recommended as a secondary choice.

The EEG is visually presented in a two-dimensional form where the electrodes form one dimension and time the other dimension. The spatial information will thus be in a more or less discontinuous format. The electrode order usually follows a standardization. For example, the CAR montage may have the order Fp1, F7, T3, T5, Fp2, F8, T4, T6, F3, C3, P3, O1, F4, C4, P4, O2, Fz, Cz, Pz, A1, and A2 (Fig. 3). This order thus forms longitudinal arrays, alternating the sides (left lateral, right lateral, left parasagittal, right parasagittal, midline, ears), and make side comparisons easier.
According to the minimum recommendation (Peltola et al., 2023), a standard exam of wakefulness should be ≥20 minutes and a sleep exam should be ≥30 minutes. The basic provocations are intermittent photic stimulation and hyperventilation, and for sleep exams, sleep deprivation. This is to increase the probability of activity associated with epilepsy. Sleep may also be induced pharmacologically, especially for children, where melatonin is recommended. It is not uncommon that the patient becomes drowsy or even fall asleep during a standard exam. There are also long-term EEG-monitoring (LTM). Some include simultaneous video recording, e.g., to study epilepsy. It is common in intensive care units (ICUs) to monitor the cerebral function. LTMs can last for several days, sometimes weeks or even months in the ICU setting.

9.4 EEG interpretation

In clinical routine work, EEG is assessed by visual inspection. Over the years, several signal patterns have been identified for both normal and pathological conditions. Most patterns are difficult to quantify objectively, and the visual assessment is learned in a teacher-student manner.

In the clinical context, the impression of an EEG exam is summarized in a written report. Kane et al. (2017) suggest a standardized report containing patient information, recording conditions, description of the recording, and the interpretation of it.

9.4.1 Characterization of EEG signals

By tradition, the activity is divided into frequency bands (Schomer and Lopez da Silva, 2018): 1) delta < 3.5 Hz, 2) theta 4–7.5 Hz, 3) alpha 8–13 Hz, 4) beta 14–30 Hz, and 5) gamma > 30 Hz. The EEG of a normal adult during wakefulness is dominated by alpha and beta activity. Delta and theta activity occur normally during sleep, but also in various pathological conditions.

Activity can be categorized as regular or irregular. The posterior dominant rhythm (PDR) in the alpha band and sleep spindles are examples of regular activity, while the examples of theta and delta activity in Fig. 4 are of more irregular character. Activity can, relative some time perspective, be stationary, i.e., have a similar character for a longer period (Fig. 4; Alpha rhythm and Delta activity). It can have a transient character (Fig. 4: Sleep spindles, K-complex, and Epileptiform discharges). The activity can also be characterized as continuous or discontinuous. If the amplitude is less than 10 µV, the activity is referred to as suppressed and if suppression occur for more than
10% of the time, the activity is called discontinuous (Hirsch et al., 2021). If suppression constitutes 50–99% of the time the activity is referred to as burst-suppression, and if more than 99%, it is suppressed. These latter patterns are typically seen in cases of severe cortical injury or during deep sedation.

The general character of the activity is often referred to as the **background activity**. Kane et al. (2017) define it as an underlying activity in the EEG from which focal or transient activity can be distinguished. This means that, e.g., EDs and seizure activity, are not part of the background activity (c.f., the definition of EDs in section 9.4.2). In the definition of background activity, it is also stated that specific rhythms, such as the PDR, are not part of the background activity.

The categorization of patterns does not only depend on the waveform characteristics, e.g., frequency and rhythmicity of the PDR or sleep spindle. The spatial localization and distribution also may matter, i.e., the PDR is seen in the posterior electrodes, while the sleep spindles are seen in the central electrodes. Abnormalities can be focal, lateralized, or general. This information can have a clinical significance. In some cases, a focal abnormality can increase the indication for additional investigations, e.g., performing an MRI. In the context of epilepsy, determining the type of epilepsy may in part depend on EEG-findings and their spatial characteristics (Scheffer et al., 2017), which influences the choice of treatment (Kanner et al., 2018). As mentioned earlier, EEG can sometimes be used in the localization of the source of epileptic seizures (Eom, 2023).

As is well-known to clinicians and will be illustrated in this thesis, interpretation can have an uncertainty. Waveforms may be ambiguous and in many cases the interpretation of an EEG section is influenced by what the interpreter has viewed in other parts of the EEG. This means that a set of suspect waveforms can be assigned a higher probability of being pathological than each waveform assessed independently. In addition, clinical information in the referral may influence how the signals are interpreted.

A final characteristic of the EEG and condition for the interpretation which is mentioned in this section is the large variation in time perspective. For example, an ED may be of 100 ms, a seizure 10 s, and the whole recording 20 minutes.
Fig. 4. Examples of different EEG signals. Transient patterns that only constitute part of the signal are indicated with asterisks (*). The duration of each example is 10 s, except for seizure activity which is 20 s. N.B., some patterns also depend on spatial localization and distribution.

9.4.2 EDs

EDs represent highly synchronized cortical activity and are associated with epilepsy (Pillai and Sperling, 2006). They may appear during seizures (ictal) but are mainly seen in between seizures (interictal). In
the revised glossary by the ICFN (Kane et al., 2017), epileptiform patterns are defined as fulfilling 4 or more of 6 criteria (quoted exactly):

1. Di- or tri-phasic waves with sharp or spiky morphology (i.e. pointed peak).
2. Different wave-duration than the ongoing background activity, either shorter or longer.
3. Asymmetry of the waveform: a sharply rising ascending phase and a more slowly decaying descending phase, or vice versa.
4. The transient is followed by an associated slow after-wave.
5. The background activity surrounding epileptiform discharges is disrupted by the presence of the epileptiform discharges.
6. Distribution of the negative and positive potentials on the scalp suggests a source of the signal in the brain, corresponding to a radial, oblique or tangential orientation of the source (see dipole). This is best assessed by inspecting voltage maps constructed using common-average reference.

9.4.3 Seizure activity
Seizures are the hallmark of epilepsy (Fischer et al., 2014). During a seizure, cortical areas display synchronized activity with repetitive discharges and which when recorded on EEG may appear as characteristic rhythmic patterns. Aimed for EEG terminology in critical care, Hirsch et al. (2021) defines electrographic seizure activity as EDs appearing at more than 2.5 Hz for more than 10 seconds, or any pattern lasting more than 10 seconds that have an evolution. Evolution refers to a change of the pattern during the seizure, which can be in frequency, waveform morphology, or location. C.f., seizure activity in Fig. 4, where amplitude increases while the frequency decreases as the activity progresses.

9.4.4 Sleep stages
The state of consciousness has a natural physiological variation, i.e., wakefulness and sleep. The EEG changes character depending on the state and based on this, different sleep stages have been defined. The American Academy of Sleep Medicine (AASM) Manual for the Scoring of Sleep and Associated Events defines five different stages (Iber et al., 2007):
1. Wakefulness (W)
2. Non-REM sleep stage 1 (N1)
3. Non-REM sleep stage 2 (N2)
4. Non-REM sleep stage 3 (N3)
5. REM sleep (R)

and in sleep staging, the EEG is scored in 30-s epochs. The EEG in wakefulness can have a large variation, but which is mostly due to artifacts. In a restful state and with closed eyes, a characteristic trait is a PDR in the alpha band. In the AASM manual, stage W is characterized by PDR, artifacts from eye movements (blinking, reading, or other rapid movements), and more muscle artifacts compared to sleep. With closed eyes, PDR should be seen for more than 50% of an epoch. Stage N1 is characterized by slow eye movements, theta activity replacing the PDR for more than 50% of an epoch, and vertex waves may occur. Stage N2 is characterized by the occurrence of characteristic waveforms, K-complex and sleep spindles. Deep sleep, stage N3, occur when more than 20% of an epoch is dominated by low frequency background activity ($\leq 2$ Hz and amplitude $\geq 75$µV). In REM-sleep, the background has a lower amplitude, consist of irregular predominantly theta activity, the tonic muscle activity decrease, and there are rapid eye movements.

From the above, it is apparent that it is necessary to, in addition to EEG, record eye movements through electrooculography (EOG) and electromyography (EMG). Due to its multimodal character, this type of sleep recording is referred to as polysomnography. It often includes several other measurements: electrocardiography, body position, body movement, sound (snoring), breathing (air flow, movement of the thorax and abdomen).

9.5 The Temple University EEG Corpus

There are several public EEG databases. One of the largest is the Temple University Hospital EEG Corpus (Obeid and Picone, 2016). At the associated website, it is stated that it consists of 26,846 clinical EEGs (https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml; accessed on March 4, 2024). Most recordings are from adults, have more than twenty electrodes, and a sampling frequency of 250 Hz. Many recordings include a written report, but the quality of these have a large variation, from very extensive and detailed to as short as “Normal”.

The database contains several annotated datasets, e.g., seizures, artifacts. The dataset TUH EEG Events Corpus (TUEV) was
downloaded on May 19, 2020. It contains a six-class annotation (spike and sharp wave, generalized periodic EDs, periodic lateralized EDs, eye movement, artifact, and background). In the authors' opinion, some annotations were peculiar. For example, in many instances, generalized periodic discharges were estimated as focal and EDs were very blunt. Of course, to some extent this reflects one of the problems addressed in the thesis, interrater agreement. However, it was hard to see how to use the dataset, since the annotations were relatively sparse compared to the amount of EEG data and there was overlap between categories.
10 Agreement scoring

10.1 Summary

Chapter 10 provides an overview of agreement assessment and results from previous studies on EEG regarding EDs, seizure activity and sleep staging. There are several metrics to assess agreement between raters, but the most common used in articles reviewed in section 10.2 are the percent agreement, Cohen’s kappa, Fleiss’ kappa, and Gwet’s AC1. There is a large variation in agreement scores between the different studies presented in section 10.3. Overall, the agreement is moderate for EDs but substantial for seizure activity. The agreement is substantial for wakefulness, fair to moderate for sleep stage N1, substantial for sleep stages N2 and N3, and substantial to almost perfect for REM-sleep.

10.2 Metrics

The most basic measure of agreement is the percent agreement, which simply is the percentage of the data on which the experts agree regarding the classification. One problem with this metric is that experts may agree by chance. For example, if a dataset contains two categories and an equal amount of each category, the chance of agreement if the experts classify by tossing coins is 0.5. Another example is if the dataset contains 99 examples of category 1 and only one example of category 2, then if rater 1 rates correctly and rater 2 just rate all as belonging to category 1, the percent agreement would be 99 even though rater 2 may not be able to identify any examples of category 2.

Cohen (1960) developed an agreement coefficient, now referred to as Cohen’s kappa, to adjust for the agreement by chance. An extension of Cohen’s kappa to multiple raters is the Fleiss’ kappa (Fleiss, 1971), which thus produces one value for the agreement of a group of raters. Cohen’s definition of chance agreement is based on the marginal distributions (see Appendix 17.2). As Brennan and Prediger (1981) note, Cohen’s definition is mathematically valid if the category prevalence is fixed (known) but may not be valid if the prevalence is free (unknown). Another important point they make is that the formulation of Cohen’s kappa is valid if ratings are independent (items being categorized, the raters, and the categories) and this is probably most often not a realistic assumption.

Cohen (1960) developed an agreement coefficient, now referred to as Cohen’s kappa, to adjust for the agreement by chance. An extension of Cohen’s kappa to multiple raters is the Fleiss’ kappa (Fleiss, 1971), which thus produces one value for the agreement of a group of raters. Cohen’s definition of chance agreement is based on the marginal distributions (see Appendix 17.2). As Brennan and Prediger (1981) note, Cohen’s definition is mathematically valid if the category prevalence is fixed (known) but may not be valid if the prevalence is free (unknown). Another important point they make is that the formulation of Cohen’s kappa is valid if ratings are independent (items being categorized, the raters, and the categories) and this is probably most often not a realistic assumption.

There are thus theoretical issues with using the Cohen’s kappa in many instances. It has been criticized for having several problems and these have been referred to as kappa paradoxes (Gwet, 2014a). One paradox regards producing unexpectedly low values in some
cases when the percent agreement is high. Another paradox regards producing higher scores for asymmetrical compared to symmetrical marginal distributions (in a binary categorization problem, the marginal distributions are asymmetrical if the raters hold opposite positions regarding which category is in majority; in the symmetrical case, both agree on which category is in majority). To address the paradoxes, Gwet (2008) proposes an agreement coefficient, referred to as Gwet’s AC1. In devising the coefficient, Gwet assumes that chance agreement occurs if at least one rater makes a random rating and that only a subset of the ratings is random. That is, Gwet (2014b) assumes that items can be easy or hard to classify, that agreement by chance may occur for some hard items, and this idea is incorporated into the coefficient.

To help interpreting agreement scores, Landis and Kock (1977) suggested an ordinal translation (Tab. 1).

<table>
<thead>
<tr>
<th>Kappa score</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00–0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21–0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41–0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61–0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81–1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

This interpretation has been used for various of the agreement measures, despite them having different characteristics and producing very different scores for the same data.

The use of the interpretation is discouraged for Gwet’s AC1 by Vach and Gerke (2023). Similar to their work, a simulation is provided here to illustrate the dependence of Cohen’s kappa and Gwet’s AC1 on the prevalence of categories for binary categories (see Appendix 17.2 for the details). The simulation was performed for 50, 70, 90, and 98 percent agreement. The distribution of the categories varied from uniform to the extreme of having almost only one category present in the data. For a more uniform distribution, the scores are close but as the distribution become more skewed, they diverge with Cohen’s kappa decreasing and Gwet’s AC1 increasing (Fig. 5). This illustrates the paradox of Cohen’s kappa producing very low scores for high percent agreement and skewed distributions (Fig. 5 D). Gwet’s AC1 is possibly a better alternative but does it always produce realistic values, or is it overly optimistic?
Fig. 5. Comparison of Cohen’s kappa and Gwet’s AC1 in a binary categorization. The y-axis is the value of the respective measure. The x-axis is the relative skewness of the distribution between the two categories, i.e., a value of 0 is a uniform distribution and a value of 1 represents that there is only one category present in the data. The simulations were performed in steps of 0.02 across the x-axis (from 0.00 to 1.00) where the value of each step presented in the subfigures is the average of 100 simulations. (A) 50 percent agreement. (B) 70 percent agreement. (C) 90 percent agreement. (D) 98 percent agreement.

10.3 Interrater agreement in EEG

Paper II of the thesis involves annotation of EDs, seizure activity, and sleep staging. Paper III involves annotation and classification of EDs. Section 14.4.2 of this thesis involves annotation and classification of sleep stages. Reviewing interrater variability will be limited to these categories, but with emphasis on EDs since interrater comparisons are only made for this type of activity (paper III). Searches were performed on PubMed using different combinations of the terms “eeg”, “interrater”, and “agreement”, with one of the terms “epileptiform”, “seizure”, “sleep”, or “sleep staging”.

10.3.1 Agreement scores for EDs

Summary of different studies assessing interrater agreement for EDs, including references, are given in Tab. 2 and Tab. 3. From the
reviewed articles, the median (min–max) agreement score for single EDs is 0.56 (0.30–0.81) and for EDs in whole EEGs 0.65 (0.16–0.83). The type of agreement metric varies, but most often Cohens' kappa or Gwet’s AC1 is used. The number of raters vary from 3 to 35 in the case of single EDs and from 2 to 49 for whole EEGs.

In clinical practice, every ED is not accounted for in an EEG but rather it is assessed whether EDs are present or not, and if so, a rough estimate of the amount is made (e.g., rare, occasional, frequent, or abundant; Hirsch et al., 2021). However, when developing a tool for automatic detection of EDs, it might be important to have a reasonable sensitivity and specificity for detecting single EDs. In the study by Black et al. (2000) raters agreed on individual EDs in 39 percent of cases but in 85 percent whether EDs were present in whole EEGs. When assessing individual EDs, Jing et al. (2019) found a percent agreement of 72 and an AC1 score of 0.49, but when assessing the presence of EDs in whole EEGs, the values increased to 81 and 0.69, respectively. These studies suggest that agreement is more likely to be higher for rating whole EEGs compared to individual EDs. The median score calculated from Tab. 2 is higher for whole EEGs but the difference is not significant (Mann-Whitney U test, statistic=48.0, p=0.61), but the sample size is small, and the scores are based on different type of calculations. The median score for studies using Cohen’s or Fleiss’ kappa is 0.69 (0.16-0.83) and for Gwet’s AC1 is 0.63 (0.30-0.82). But, in the former case seven out of ten studies compared EDs for whole EEGs compared to two out of seven in the latter case.

### Tab. 2. Summary of interrater agreement for individual epileptiform discharges from previous studies. First assessment only. If not otherwise indicated scores are Cohen’s kappa. In some studies, agreement scores are presented as percent, and these have been converted to a relative value (0.00–1.00). Raters: number of raters; Subjects: number of subjects; Data: data size per subject; #EDs: number of epileptiform discharges per rater. Study: reference.

<table>
<thead>
<tr>
<th>Score</th>
<th>Raters</th>
<th>Subjects</th>
<th>Data</th>
<th>#EDs</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33^b</td>
<td>18</td>
<td>200</td>
<td>30 s</td>
<td>16–195</td>
<td>Bagheri et al. (2017)</td>
</tr>
<tr>
<td>0.63^c</td>
<td>13</td>
<td>1,528</td>
<td>20 min</td>
<td>344–537</td>
<td>Beuchat et al. (2021)</td>
</tr>
<tr>
<td>0.40^d</td>
<td>2</td>
<td>7</td>
<td>15-30 min</td>
<td>2,199–2,640</td>
<td>Dümppelmann and Elger (1999)</td>
</tr>
<tr>
<td>0.30^b/0.69^c</td>
<td>19</td>
<td>200</td>
<td>30 s</td>
<td>16–195</td>
<td>Halford et al. (2017)</td>
</tr>
<tr>
<td>0.81^b</td>
<td>35</td>
<td>200</td>
<td>30 s</td>
<td>6–212</td>
<td>Halford et al. (2018)</td>
</tr>
<tr>
<td>0.49^e</td>
<td>8</td>
<td>1,051</td>
<td>30–60 min</td>
<td>13,262</td>
<td>Jing et al. (2020)</td>
</tr>
<tr>
<td>0.79^f</td>
<td>5</td>
<td>50</td>
<td>295±192</td>
<td>733–1618</td>
<td>Wilson et al. (1996)</td>
</tr>
</tbody>
</table>

a) Gwet’s AC1, b) Gwet’s AC2, c) Fleiss’ kappa, d) percent agreement, e) Inter-reader correlation, it is unclear if it is a chance corrected score or ordinary correlation.
### Tab. 3. Summary of intrarater agreement for epileptiform discharges in whole EEGs from previous studies. First assessment only. If not otherwise indicated scores are Cohen’s kappa. In some studies, agreement scores are presented as percent, and these have been converted to a relative value (0.00–1.00). Raters: number of raters; Subjects: number of subjects; Data: data size per subject; Study: reference.

<table>
<thead>
<tr>
<th>Score</th>
<th>Raters</th>
<th>Subjects</th>
<th>Data</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.36e</td>
<td>3</td>
<td>74</td>
<td>10 min</td>
<td>Abend et al. (2011)</td>
</tr>
<tr>
<td>0.60e</td>
<td>4</td>
<td>72</td>
<td>10 min</td>
<td>Abend et al. (2017)</td>
</tr>
<tr>
<td>0.52e</td>
<td>3</td>
<td>90</td>
<td>5 min</td>
<td>Ahrens et al. (2021)</td>
</tr>
<tr>
<td>0.69c</td>
<td>3</td>
<td>100</td>
<td>-</td>
<td>Azuma et al. (2003)</td>
</tr>
<tr>
<td>0.39d</td>
<td>3</td>
<td>106</td>
<td>20 min</td>
<td>Black et al. (2000)</td>
</tr>
<tr>
<td>0.82g</td>
<td>49</td>
<td>37</td>
<td>10–60 s</td>
<td>Gaspard et al. (2014)</td>
</tr>
<tr>
<td>0.47h/0.68i</td>
<td>6</td>
<td>22</td>
<td>-</td>
<td>Hussein et al. (2014)</td>
</tr>
<tr>
<td>0.16</td>
<td>16</td>
<td>82</td>
<td>10 s</td>
<td>Mani et al. (2012)</td>
</tr>
<tr>
<td>0.38g</td>
<td>2</td>
<td>143</td>
<td>33 min</td>
<td>Nguyen et al. (2010)</td>
</tr>
<tr>
<td>0.74–0.77e</td>
<td>3</td>
<td>27</td>
<td>&gt;20 min</td>
<td>Piccinelli et al. (2005)</td>
</tr>
<tr>
<td>0.69g</td>
<td>3</td>
<td>30</td>
<td>30 min</td>
<td>Reus et al. (2022)</td>
</tr>
<tr>
<td>0.83</td>
<td>2</td>
<td>93</td>
<td>-</td>
<td>Stroink et al. (2006)</td>
</tr>
<tr>
<td>0.65g</td>
<td>16</td>
<td>37</td>
<td>10–60 s</td>
<td>Zhuo Ding et al. (2019)</td>
</tr>
</tbody>
</table>

a) Gwet’s AC1, c) Fleiss’ kappa, d) percent agreement, e) unspecified kappa score, g) Randolph’s free-marginal multitrer kappa, h) focal EDs, i) multifocal EDs.

### 10.3.2 Agreement scores for seizure activity

The results of the review of agreement for seizure activity are provided with references in Tab. 4. The median agreement is 0.76 (min: 0.38, max: 0.97). As in the case of EDs, there was a variation in which metric that was used. The number of raters vary from 3 to 49 and the number of subjects is in the range 12–2,711. There is a variation in the data sizes, in some cases smaller sections of events, other longer recordings of continuous data.
Agreement scores for sleep staging

The AASM Manual for the Scoring of Sleep and Associated Events was published in 2007 (Iber et al.), where the currently used sleep stages were introduced: wakefulness (W), non-REM stage 1 (N1), non-REM stage 2 (N2), non-REM stage 3 (N3), and REM-sleep (R). A summary of published studies, including references, regarding sleep scoring agreement is provided in Tab. 5. Only studies using the scoring system proposed in the AASM manual and that report chance corrected agreement scores for all sleep stages on a group level are included. The median (min–max) scores of the sleep stages are W 0.78 (0.58–0.89), N1 0.40 (0.16–0.69), N2 0.65 (0.5–0.72), N3 0.67 (0.49–0.79), and R 0.82 (0.63–0.92). Most use full over-night polysomnographies (PSGs), but a few use a selection of epochs from PSGs, and a median of 50 (10–1,066) PSGs are used. The number of raters varies from 2 to 10.

Li et al. (2022) conducted a meta-analysis to assess interrater reliability of sleep stage scoring. The articles span from 1989 to 2019 and thus include some articles published before the AASM scoring system. From 101 articles they identify 11 articles that satisfy their selection criteria, which totals 23 raters, 219 PSGs, and 202,304 sleep epochs. The calculated Cohen’s kappa values are W 0.70, N1 0.24, N2 0.57, N3 0.57, and R 0.69. The selected articles include Deng et al. (2019) of Tab. 5.

Duce et al. (2014) report intrarater agreement with Cohen’s kappa values 0.87, 0.51, 0.66, 0.60, and 0.92. Somaskandhan et al.
(2023) report intrarater percent agreement of 89–90, 44–63, 77–81, 92–93, and 93–93 percent for two raters.

### Tab. 5. Summary of interrater agreement for sleep staging from previous studies.

All studies use Cohen’s kappa except the study by Deng et al. (2019), where Fleiss’ kappa is used. Sleep stages are W, N1, N2, N3, and R. PSG: polysomnography. Study: reference.

<table>
<thead>
<tr>
<th>W</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>R</th>
<th>Raters</th>
<th>PSGs</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.81</td>
<td>0.40</td>
<td>0.65</td>
<td>0.77</td>
<td>0.82</td>
<td></td>
<td>1,066</td>
<td>Cesari et al. (2021)</td>
</tr>
<tr>
<td>0.85</td>
<td>0.40</td>
<td>0.68</td>
<td>0.63</td>
<td>0.82</td>
<td>2</td>
<td>86</td>
<td>Choo et al. (2023)</td>
</tr>
<tr>
<td>0.86</td>
<td>0.46</td>
<td>0.72</td>
<td>0.73</td>
<td>0.91</td>
<td>2</td>
<td>72</td>
<td>Danker-Hopfe et al. (2009)</td>
</tr>
<tr>
<td>0.89</td>
<td>0.45</td>
<td>0.72</td>
<td>0.79</td>
<td>0.87</td>
<td>7</td>
<td>40</td>
<td>Deng et al. (2019)</td>
</tr>
<tr>
<td>0.74</td>
<td>0.47</td>
<td>0.68</td>
<td>0.60</td>
<td>0.92</td>
<td>2</td>
<td>10</td>
<td>Duce et al. (2014)</td>
</tr>
<tr>
<td>0.78</td>
<td>0.31</td>
<td>0.60</td>
<td>0.67</td>
<td>0.78</td>
<td>9</td>
<td>15</td>
<td>Magalang et al. (2013)</td>
</tr>
<tr>
<td>0.76</td>
<td>0.41</td>
<td>0.67</td>
<td>0.74</td>
<td>0.86</td>
<td>10</td>
<td>50</td>
<td>Nikkonen et al. (2023)</td>
</tr>
<tr>
<td>0.84</td>
<td>0.69</td>
<td>0.65</td>
<td>0.63</td>
<td>0.63</td>
<td>10</td>
<td>70</td>
<td>Younes et al. (2018)</td>
</tr>
<tr>
<td>0.58-</td>
<td>0.10-</td>
<td>0.50-</td>
<td>0.49-</td>
<td>0.66-</td>
<td>5</td>
<td>30</td>
<td>Zhang et al. (2015)</td>
</tr>
<tr>
<td>0.76</td>
<td>0.30</td>
<td>0.58</td>
<td>0.68</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 10.3.4 Intrarater agreement

Direct literature searches regarding intrarater agreement in EEG surprisingly yielded no relevant hits for EDs, seizure activity, or sleep staging. However, some studies assessing intrarater agreement turned up in the material reviewed above. Nguyen et al. (2010) report that two raters of EDs scored kappa values of 0.60 and 0.80, respectively. For sleep stages W, N1, N2, N3, and R, Duce et al. (2014) report intrarater agreement with Cohen’s kappa values of 0.87, 0.51, 0.66, 0.60, and 0.92 for two raters. Somaskandhan et al. (2023) report intrarater percent agreement of 89–90, 44–63, 77–81, 92–93, and 93–93 for two raters.

### 10.3.5 Characteristics of agreement of EDs

It has been suggested that the consensus of a group of experts could be used as gold standard for defining EDs for automated detection (Wilson et al., 1996). This presupposes that the agreement will converge given a large enough group of experts. If that is the case, the number of experts that is needed is unknown. The quality of the experts reasonably is an important factor. The agreement is probably favored if the experts base their judgement on the same principles.

Intuitively it seems reasonable that the average of a group of experts could reduce variability in categorizations. Nonetheless, there is always the possibility that there is one or a limited number of experts that perform better, in the sense of rating in a more useful way. Including more, and inferior, experts would then deteriorate the result. On the other hand, higher agreement does not guarantee that we are
closer to some absolute truth, so including somewhat disagreeing experts may reduce the risk of extreme opinions. Wilson et al. (1996) imply that five experts are enough for a consensus, but the study only uses five experts, and the claim is therefore difficult to assess. In the study by Halford et al. (2017), the agreement decrease steadily as more raters are added to the group. This is possibly an artifact of the method used to analyze since raters are effectively added in decreasing order of agreement. However, by analyzing the change in consensus when changing the number of experts, they conclude that the optimal number of experts could be in the range six to ten. Bagheri et al. (2017) use a high-dimensional model to predict interrater agreement and the performance levels out at around eleven raters, which might give an idea of what number of raters is needed for a stable group consensus.

Education might provide a common model for rating, which in combination with training can lead to a more methodical and streamlined rating. Halford et al. (2017) find that higher agreement is associated with American Board of Clinical Neurophysiology board certification. In another study by Halford et al. (2018), a comparison between 19 academic and 16 private practitioners shows higher performance for academics with ≥ 1.5 years of epilepsy fellow training and board certification. Beuchat et al. (2021) compare rating by technicians and neurophysiologists and find more frequent identification of EDs (and seizures) for the technicians.

A few studies show an effect on agreement from consensus discussions and creating guidelines. In a study by Azuma et al. (2003), three raters assess 100 consecutive EEGs and achieve a kappa value of 0.69 for EDs. Afterward the raters develop a common guideline for interpretation, and the kappa value increase to 0.87 when an additional 50 EEGs are assessed. However, in a study by Stroink et al. (2006), the kappa score decreased from 0.83 to 0.63 after a consensus discussion regarding the distinction between other EEG characteristics that was scored (abnormal background pattern vs. focal non-epileptiform abnormalities), but the difference could also be attributed to differences in the scored material of the two sessions.

There are some studies that imply that experts may use a common internal model for rating. By comparing different subgroups of a larger group of raters, Halford et al. (2017) conclude that there does not seem to be subgroups with different opinions. For a fixed number of raters per subgroup they find that the most agreeing subgroup scored significantly higher compared to the most agreeing subgroup of the remainders. They interpret this as to indicate that the
remaining raters disagree significantly more with each other compared to the raters of the high scoring subgroup and so probably does not represent another opinion. Bagheri et al. (2017) apply a model based on 5,538 signal features to predict interrater agreement. They find that local normalization of the data can increase the performance and they interpret this as if waveform morphology is more important than amplitude. Jing et al. (2020) suggest that varying agreement is due to experts using different thresholds for EDs rather than different internal “models” for scoring. By applying multivariate statistical models to the waveforms assessed by raters, they show that a general model with individually fitted thresholds to each rater perform comparable to models fitted specifically to each rater.
11 Deep learning

11.1 Summary

Deep learning is a form of artificial intelligence most commonly based on large artificial neural networks, here referred to as deep neural networks (DNNs). The basic principles and their design are outlined in sections 11.2–11.4. In its simplest form, DNNs take some form of data as input, processes it, and produces an output. The DNNs typically consist of many connected layers, each processing the data as they are fed through the DNNs. There are several types of layers, each of which can perform more or less different types of operations on the data. The composition of all the layers of DNNs forms an architecture and some architectures may be more efficient in general or be more suitable for certain tasks. For time series and images, convolutional neural networks (CNNs) are commonly used as they are efficient for such data.

The principles for training DNNs are described in sections 11.5–11.6. The DNNs are developed by training them to achieve some objective. Many layers have parameters that can be adjusted during training, and the adjustments represent the DNNs’ learning. During training, the loss function evaluates values produced by the DNNs, compares it to some target values, and in most cases the objective is to minimize the difference between the two sets of values. The DNNs and the loss function together form a large function. The minimum is found by calculating the gradient of this function with respect to the parameters of the network, and in small steps changing the parameters in the direction in which the gradient decreases the most. It is important to avoid very small or large values in the gradient. The input data therefore must be preprocessed, and the DNNs’ architectures have to include structures and certain layer types to counteract large differences in values.

The main training paradigms discussed here are supervised and self-supervised learning, discussed in section 11.7. The most common example of supervised learning is classification where the objective for the DNNs is to label data correctly. The correct labels, i.e., the ground truth, are often created by manual labeling of data by humans. In self-supervised learning, intrinsic structures in the input data are used to achieve the objective and no labels are needed. An example is data compression where the DNNs reduce the size of the data while preserving important information in the data.

The same data can be transformed and represented in different ways, some transformations preserve information of the original data,
while other will reduce the information. EEG data in its original form is a time representation. The most common transformation of EEG data is into a frequency representation. A written report of an EEG exam is also a form of representation, but heavily compressed. In DNNs, data are successively processed into different representations, the possible representations are the DNNs latent space. t-distributed stochastic neighbor embedding (t-SNE) is a method to reduce high-dimensional data to an arbitrary lower number of dimension but is most often used to visualize data in two dimensions.

Supervised learning dominates in deep learning studies in EEG analysis. In studies of deep classifiers for EDs or seizure activity, the median accuracy is 0.9 and 0.98, respectively. For sleep staging, the median accuracies are W 0.93, N1 0.58, N2 0.87, N3 0.85, and R 0.92. There are several studies using self-supervised approaches, the most common being for pretraining classifiers, generating artificial EEG for data augmentation, and signal processing.

11.2 The principle of DNNs

Some of the basic principles of deep learning date back to the earlier parts of the 20th century (Russell and Norvig, 2016). There are some more recent developments that have contributed to the development of deep learning during the last decades, but it foremost correlates with the increasing capacity of computers which is necessary to implement large DNNs and work with large datasets. The term deep refers to the large number of layers and parameters, compared to, e.g., the perceptron of the 1950s (Rosenblatt, 1957).

The architecture of DNNs can be very varied but the general principle is that they are composed of stacked layers (Fig. 6 A). Data is feed into the first layer of a DNN, the layer performs some form of operation on the data, then feedforwards the result to the next layer, and this process is iterated through the DNN until the last layer produces the result which is the output of the DNN.
Fig. 6. Illustration of the general structure for neural networks and the loss function. (A) Network with N layers. (B) The loss function illustrated in one dimension.

Each layer has a set of parameters that affects how it processes the input data. The values of the parameters are learned during training. The learning and training can be formulated as an optimization problem where the objective is to find a solution given the training data, the DNNs architecture, and some form of boundary conditions, all forming a system. Boundary conditions are usually defined by a loss function. In most cases the objective is to minimize the loss. By computing the gradient of the whole system, successively lower values for the loss can be found, and if the solution converges, eventually a local minimum (Fig. 6 B).

11.3 Layers

There are many types of layers. To confuse matters more, it can also vary which components of the DNNs that are referred to as layers, and what is referred to as layers can consist of several layers. In this section, only a brief overview will be given of the most important layers used in this work.

The most basic layer is the fully connected layer (Fig. 7 A) (Goodfellow et al., 2016). It consists of a set of units referred to as nodes. Each node receives all values from the preceding layer, each value will first be weighted and then summed in the node. The result is then feedforward to all nodes of the next layer. The weights are the parameters that the layer learns during training.

Another important layer is the convolutional layer (Fig. 7 B) (Goodfellow et al., 2016), where each layer consists of a set of filters. The filters are usually small compared to the data size, and so the filtering is local across the filtered dimension and independent of location. That is, the filters have a limited field of view and will extract the same type of feature independent of position in the data. The filters have a set of parameters which determines the filtering effect, and
these parameters are learned during training. Since there usually are many filters per layer, each layer can analyze several different features in the data. An important effect of stacking convolutional layers and pooling layers (see below) is that the field of view will increase for each additional layer, and in the ideal case, this will result in a capacity to analyze increasingly complex data features.

Fig. 7. Illustration of three of the most common layers. (A) Fully connected layers. (B) Convolutional layers with filters of size three, i.e., the filters analyze three values at a time. (C) Max pooling layer with a window of two and strides of two.

The convolutional layer is suitable for analyzing time series, and hence for many types of neurophysiological data. Compared to fully connected layers, convolutional layers are much more efficient as only a number of small filters are learned. The action of the layer will be determined by the DNNs architecture and how its trained, but the idea is that it reduce data to a set of features representing its character and that are independent of the position in the data. In the case of neurophysiological data, position is normally in the time dimension. The spatial dimensions of EEG, i.e., the electrode positions, are usually arranged into one dimension, the electrode order will be arbitrary and so the spatial field will become discontinuous. The features may therefore be specific to certain locations, using convolutions may still produce usable results but cannot be motivated by the location invariance argument. Nonetheless, there are studies using spatial convolutions (e.g., Ng et al. 2023).

A layer used in conjunction with convolutional layers in this work is the max pooling layer (Fig. 7 C) (Ranzato et al., 2007). This layer compares values within a narrow window and outputs the maximum value. The idea is that large values represent salient features in
the data while small values can be discarded. The window strides across the data and by using a stride larger than one, the result is a decrease in data size, i.e., downsampling. These layers will thus extract the most interesting features and compress the data; this can enhance the effect of the convolutional layers (Scherer et al., 2010).

In some cases, deconvolutional (or transpose convolutional) layers are used. This type of layer reverses the effect of convolutional layers. It also consists of a set of filters and the result is upsampling while preserving the continuity of the data. As in the case of convolutional layers, the filters have parameters that are adjusted during training.

The presented layers above only perform linear operations and so in themselves have limited processing capacity. By inserting nonlinear functions after each layer, the resulting structure can perform nonlinear operations and become more powerful. These functions are referred to as activation functions (Fig. 8). Strictly speaking, they are also layers, but this is not always clear in the literature, and in some cases, they are rather regarded as parts of a layer. The short overview provided in the following will be limited to activation functions used in this work. The rectifying linear unit (ReLU) will output zero for all input values below zero, and otherwise output the input linearly (Fig. 8 A). If the input to this activation is normalized to a zero mean, half of the data will thus be discarded (set to zero) and as with the max pooling layer described below, this can be beneficial for extracting useful-while discarding non-useful information. The ReLU can have problems with a vanishing gradient (section 11.5), and Leaky ReLU is a modification which in addition produces a small linear output for negative input values (Fig. 8 B) and may increase performance (Maas et al, 2013). This activation was used in paper I. Performance is usually increased for these activations by using batch normalization layers described below. The scaled-exponential linear unit (SELU) has similarities with the Leaky ReLU, but the negative input values instead follow an exponential function, and the activation has self-normalizing properties (Fig. 8 C) (Klambauer et al., 2017). This was used as the standard activation in most of the DNNs in paper II and III.
The sigmoid (Fig. 8 D) and softmax (Fig. 8 E) activations both produce output values in the interval from zero to one. The sigmoid does so for a single input. The softmax takes multiple inputs, produce an output of the same size, and which is normalized to have the sum of one. Both functions can thus be used to produce output with the character of probabilities and are most often used in classification tasks. The sigmoid produces one probability value and is therefore used for binary classification. The softmax produces a probability distribution and is suitable for multiple-category classification. The softmax is used extensively in the transformer model, where data is scored in the multi-head attention layers (Vaswani et al., 2017). Sigmoid and softmax activations were used at strategic positions of the encoders in paper II. In paper III and in the thesis, the activations were used in classifiers. The softmax in conjunction with the ReLU also served an important function in the cluster networks developed in paper III.

The final type of layers presented here are normalization layers. This type can, as the name imply, normalize the data within the DNNs, e.g., transform the data to have zero mean and a standard deviation of one. For many years the dominating type has been batch normalization, where the normalization is based on a whole batch of training data (Ioffe and Szegedy, 2015). During recent years, with the advance of the transformer model, layer normalization has gained popularity, where normalization is performed independently per example of the batch (Ba et al., 2016).

11.4 Network architectures: Classifiers and autoencoders

In this section, the basic principle for designing classifiers and autoencoders are outlined. DNNs are thus built of many layers. As mentioned, there will usually be a multi-parameter layer first, e.g., convolutional, or fully connected, followed by an activation. So, often there are repeating series of layers. A common combination is a
convolutional-, batch normalization-, ReLU-, and max pooling layer (Fig. 9 A). This combination can then in turn be repeated several times, forming a section which analyze by convolution and downsamples the data. Similarly, a section of repeating fully connected-, batch normalization-, and ReLU layers can be built.

Fig. 9. Examples of DNN designs. (A) An example of order for layer types that form units that often are repeated several times. (B) Illustration of the general structure of a classifier. It starts with a section of series of convolutional layers that can have the form of that in A. The following section consists of a series of fully connected layers. The last fully connected layer has number of nodes equal to the number of categories when using the softmax layer to generate scores for the categories. (C) Illustration of an autoencoder. The first half consist of series of convolutional layers with max pooling that successively downsamples the data. The second half reverse this, deconvolute and upsample the data, in the basic case to recreate the original data.

The basic multi-category classifier starts with a convolutional section, which is followed by a fully connected section that ends with a fully connected layer with the number of nodes equal to the number of categories, and finally a softmax activation (Fig. 9 B). The convolutional section will break down the data into a set of features that characterize the data, and the fully connected section will then analyze the set and decide which category the data most likely belong to.

While DNNs usually have many parameters, i.e., a large capacity to learn, an important principle to promote learning is to have bottlenecks in the architecture. These are structures that limit information flow, and the idea is that this will force the DNNs to reduce the data into a more compact representation while still retaining enough information. The filters of the convolutional layers essentially are bottlenecks. The autoencoder is an example where there is a bottleneck built into the overall architecture (Fig. 9 C). It starts with a
convolutional section which downsamples the data, resulting in a heavy compression. The compression is then reversed by a section of deconvolutional layers and in the basic case the intent is to recreate the input data as close as possible.

11.5 Loss function and backpropagation

As stated previously, the training data, DNNs, and the loss function constitute a system that produces a value, which is the loss. The objective in most instances is to minimize the loss. The loss is most often constructed to have zero as the lowest possible value. Basic examples of loss functions are mean absolute error, cross entropy, or mean squared error. However, the loss function and resulting system can be very elaborate, where the system may consist of several DNNs (e.g., see cycle GANs in section 11.7 or the loss function for the cluster encoder in paper III).

The principle for finding the solution to minimizing the loss, is by computing the gradient of the whole system and adjust parameters in the direction where the gradient is decreasing fastest (Fig. 6 B). However, the system is usually very complex and the whole gradient cannot be computed directly. The solution to this problem is by using the chain-rule for the derivative and is often referred to as backpropagation (Rumelhart et al., 1986). During training, the data is first feedforwarded through the DNNs. The gradient is then computed starting at the last layer, adjusting the weights, then repeating this for each layer backwards through the DNNs until the first layer is reached and all parameters have been adjusted. In many cases the loss will not become zero but will converge to a value which corresponds to a local minimum.

The architecture and depth of the DNNs mean that values will be multiplied several times as data passes through the layers, larger values will thus then tend to grow, while smaller values tend to shrink, and the effect can be exponential. This may result in loss of important information contained in data having small numerical values. This also creates problems with calculating the gradient in the form of a so-called exploding or vanishing gradient (Pascanu et al., 2013). If possible, it is important to avoid large differences in numerical values, in both the original data used as input, as well as its representation within the DNNs.

11.6 Normalization

To counteract the problem of vanishing gradient, the input data is often normalized. The approach can differ, but one usually strives to
keep most of the data varying within -1 to 1 and have a zero mean. A simple way to accomplish this is to divide the data with the standard deviation or a percentile and then subtract the mean. However, there are several ways to perform this. An example from clinical neurophysiology could be a polysomnography where it probably is best to normalize the EMG and EEG separately. Furthermore, multi-channel data such as EEG can be normalized more globally or more locally. For instance, take a dataset of several EEGs where the DNNs only analyze segments of 10 s of EEG at a time. There are then several options, e.g.: 1) the whole set of EEGs can be normalized together, 2) each EEG can be normalized individually, 3) each segment can be normalized individually, 4) each channel can be normalized individually, 5) each channel of each segment can be normalized individually. Bagheri et al. (2017) improve the performance of their model to predict agreement regarding EDs by local normalization. In the case of EDs, it may be that morphology is more important than amplitude. Though, this is probably not a universal result, and the best option will depend on the data and the categorization problem. An obvious example where amplitude reasonably is important would be to detect suppression.

In the DNNs, the activations will affect the distribution of values in the feedforward process, and some may be more prone to contribute to a vanishing gradient. For instance, the sigmoid activation (Fig. 8 D) produces weak gradients and in addition the mean will be non-zero since values only vary between zero and one (LeCun et al., 2012). As mentioned in section 11.3, there are different types of normalization layers that can be inserted into the DNNs to counteract vanishing gradients. Another counteractive measure is to insert residuals where data is forwarded while skipping some layers (He et al., 2016). In addition to preserving structures of the data, this can prevent values from becoming too small.

11.7 Supervised and self-supervised learning

The classifiers in sections 8.1.1 and 8.1.2 are examples of supervised learning. In supervised learning the data consists of a set of examples of pairs of data $x$ and labels $y$. During training the classifiers try to predict the correct labels and the parameters of the classifiers are successively adjusted to align the predictions with the true labels through the backpropagation routine. Take a binary classification problem as example: The labels $y$ can be represented by 0 and 1. The DNNs take the data $x$ as input and generate an output value between 0 and 1 through a sigmoid function. Rounding this value will produce the
predicted classification $\hat{y}$. The loss function can be the mean absolute of the error $\hat{y} - y$ or the binary cross-entropy and during training the goal is to minimize the error.

Another strategy for training is self-supervised learning (In this thesis, no distinction is made between unsupervised and self-supervised methods, all are referred to as self-supervised). In this case there are no labels, and the learning is instead based on intrinsic structures in the training data itself. An example is the basic autoencoder described above. This type of DNNs recreate the input $x$ as output $\hat{x}$ and so the loss function can be the mean absolute of the error $\hat{x} - x$. In this case, no labeling is required since the data itself is the ground truth.

The basic generative adversarial network (GAN) is an interesting case of learning strategy (Goodfellow et al., 2014). It consists of two networks, one network generates artificial data, and the other is a classifier that discriminates between real and artificial data (Fig. 10). In essence, they learn by competing. The idea is that the competition eventually will lead to the generation of very realistic artificial data. The loss function for the generative network incorporates information on how the discriminator classifies its artificial data. The loss function for the discriminative network involves classifying real and artificial data; in principle, this is supervised learning, but in the basic case, labeling of the data need no external assessment and can be made automatic. Since the generative network never encounters the real training data, memorization of the data is avoided, but the method instead suffers from the risk of mode collapse, where the generated data will have little or no variation (Kossale et al., 2022).

![Fig. 10. Illustration of the basic GAN. A generative DNN (Generator) produces some form of artificial data. In its most simple form, it is driven using noise as input (it acts as random sampling of the latent space). A classifier (Discriminator) is trained to discriminate between real and artificial input data. The result of classification of artificial data is used as feedback in the training of the generative DNN.](image)
One further development is the cyclic GAN, where two simpler GANs are combined into a circuit (Zhu et al., 2017). In this case, two generative networks send artificial data to each other, and two discriminative networks are used to promote them to generate realistic data. This setup makes it possible to transform data between different domains, i.e., some structures are preserved while some characteristics change during transformations. In the original article by Zhu et al. (2017), they demonstrate this by transforming images, e.g., changing the season of a landscape between summer and winter, or changing the character of a motive between photo and oil painting. For EEG data, the two domains can be, e.g., normal and pathological EEG, or scalp- and subdural-EEG.

One disadvantage of supervised learning is that it requires labeled data. Though large amounts of data may be available, it is most often not labeled, and labeling can be time-consuming. If labeled data is scarce, results can be improved by using DNNs trained for some other purpose. It can be classifiers trained for other problems or DNNs can be pretrained using a self-supervised strategy. The DNNs can then be finetuned using the labeled data and supervised learning. The idea is that the DNNs in the previous training has learned the features of the data, and the finetuning then serves to identify and focus the DNNs on the feature that are most important to the classification task. The concept is sometimes referred to as transfer learning. Many of the more advanced DNNs have training paradigms that are typically a combination of self-supervised and supervised learning. A current example are the LLMs that are under development (Rothman, 2022). These are often pretrained self-supervised, e.g., predicting the next or a missing word in a sentence. They are then finetuned to a more specific task using supervised or reinforcement learning.

11.8 Data representations

Data with many dimensions can be difficult to assess visually. High-dimensional data can also be difficult to manage due to practical circumstances such as computing power and memory capacity. There are several methods that can transform data into new representations or reduce the number of dimensions. Different representations may have different advantages and drawbacks. Reducing the number of dimensions will inevitably lead to loss of information. Sometimes loss of information can be beneficial and aid analysis if the lost information is uninteresting. Choosing the optimal method to process the data can thus be challenging.
Without going into the mathematical formalism, given certain conditions, data can be transformed between different representations where information and some mathematical properties are preserved. To the neurophysiologist, the most familiar transformation is that from time to frequency, the most common being the Fourier transform. Neurophysiological data are often in the form of time series (Fig. 11 A) and this data type can thus be transformed into a frequency representation (Fig. 11 C). There are also ways of creating time-frequency representations (Fig. 11 B). The dualistic nature of the relationship between time and frequency means that there is an opposing relationship between them, e.g., increasing the frequency resolution will decrease the time resolution. An example of a time-frequency method is the short-time Fourier transform (STFT) (Sejdić et al., 2009).

![Fig. 11. Illustration of five different data representation of a 15-s EEG-signal (F4-electrode in average reference). (A) Time representation. (B) Time-frequency representation using short-time Fourier transform. (C) Power spectral density. (D) Text description and binary scoring based on visual assessment.](image)

One of the more basic ways of processing time series data is by filtering and thereby reducing the contribution of some frequencies to the signals. The main reason for doing this is to reduce low- and high-frequency noise, i.e., reduce uninteresting signal components. The operation will not reduce the data size. Transforming data may not always solve the problem of managing data sizes and more direct methods for reducing sizes may be needed. A common practice is to reduce sizes by downsampling or taking averages over certain
intervals and dimensions. Heuristic methods can also be used. The descriptions and categories used in clinical praxis is also a form of dimensional reduction, e.g., a 24-hour EEG recording can be represented in a few sentences (c.f., Fig. 11 D).

In the context of deep learning and EEG, there are many examples where the data are extensively preprocessed, transformed into new representations, e.g., time to frequency, before they are used as input to DNNs (Roy et al., 2019). Since DNNs identify and learn data features during training, preprocessing may affect the potential for learning negatively if it removes important information. This type of preprocessing is often referred to as feature engineering.

DNNs will transform input data into new representations within themselves. I.e., if values produced from some input data are extracted between two layers in a DNN, then these values are a representation of the input data. All information may be preserved, but most likely the information will to some degree be filtered and compressed. The total possible ways that data can be represented constitutes the DNN’s latent space.

11.9 t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is a method for reducing high-dimensional data into a low-dimensional representation. The most common use is for visualizing data. It is based on matching the probabilities of data being close (neighbors) in both a high- and a low-dimensional representation (van der Maaten and Hinton, 2008). The high-dimensional data is assumed to have a Gaussian neighbor distribution, while the corresponding low-dimensional representation follows a t-distribution. The use of different distributions is assumed to create favorable conditions to separate the data into clusters in the low-dimensional representation. The method is non-parametric, i.e., it does not create a model for the mapping between the representations.

A further development of the method is parametric t-SNE, which as the name implies generates a model for the mapping (van der Maaten, 2009). In its original implementation, stacked restricted Boltzmann machines are trained using a greedy layer-wise strategy to learn the mapping between representations. This allows for the analysis of new data without refitting the model.

By choosing the low-dimensional representation to have two (or maybe three) dimensions, it is possible to produce a visualization of the high-dimensional data (Fig. 12). For instance, ten second segments of 21-channel EEG sampled at 250 Hz each have 52,500
dimensions. Performing t-SNE on the segments will reduce them to points, a standard exam of 20 minutes can be represented by 120 points (240 values), and this allows for visualization of large amounts of data in a compact format. Close points will represent EEG-segments of similar character, and this can possibly provide opportunities to discover new patterns or aid in determining boundaries between categories.

**Fig. 12.** Illustration of dimensional reduction. A constructed example showing EEG-signals of three different general characters and how these are mapped into points, where the distances between the points depend on how similar the EEG-signals are. (A) Single-channel EEG signals of 1 s duration, or 250 sample points. Each point is a dimension, and each signal thus has 250 dimensions. The signals have been placed symbolically in three groups of similar character. (B) The corresponding two-dimensional representation of the signals.

### 11.10 Evaluation metrics

The most commonly used metrics found in the studies reviewed in section 11.11.1 are the accuracy (ACC), area under the curve (AUC), and the F1-score. The ACC is simply the number of correctly predicted examples divided by the total number of examples. It is equivalent to the percent agreement. The AUC is often based on the receiver operating characteristic (ROC). The ROC is a graph of the true positive rate vs. the true negative rate. Similar to the percent agreement, imbalanced datasets can produce high scores for the ACC and AUC by chance. This can be compensated by weighting the categories. The balanced-ACC is calculated by taking the average of the true positive and negative rates. In the case of AUC, using precision and recall instead of the true positive and negative rates, can be an alternative. Since the calculation of precision and recall also involves the false
positive and negative rate, respectively, this will to some extent com-
pensate for imbalanced datasets. An alternative is the F1-score based
on the Fβ-score (Chinchor, 1992) which is the harmonic mean of pre-
cision and recall according to

$$Fβ = \left(1 + \beta^2\right) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

and the F1-score thus has $\beta = 1$. The properties of precision and recall
will produce a compensation for imbalanced datasets. As can be seen
from the definition of $Fβ$, values of $\beta < 1$ will emphasize precision
while $\beta > 1$ will emphasize recall. Another alternative, used in paper
III, is the Mathew’s correlation coefficient (MCC) (Chicco and
Jurman, 2020)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

where $TP$ is the true positive rate, $TN$ is the true negative rate, $FP$ is
the false positive rate, and $FN$ is the false negative rate. Other metrics
like the F1-score can still be inflated in some instances of unbalanced
datasets, and the MCC can possibly be more stable to such effects
(Chicco and Jurman, 2020).

11.11 Deep learning and EEG
There is an increasing number of publications per year regarding EEG
and deep learning (Fig. 13). To my knowledge, the only comprehen-
sive review of the field as of February 2024 is by Roy et al. (2019).
They review 154 articles published from January 2010 to July 2018.
Classification of EEG data constitutes $86\%$ of the studies. The most
common types are sleep staging, seizure detection, brain-computer
interfacing (BCI), and cognitive and affective monitoring. Studies for
BCI dominate, but there is an increasing proportion of studies in epi-
lepsy. The remaining studies concerned processing of EEG or gener-
ation of data. Craik et al. (2019) review classification studies and iden-
tify 90 published articles with a distribution of $22\%$ BCI, $16\%$ emotion
recognition, $16\%$ mental workload, $14\%$ seizure detection, $10\%$ event
related potentials, $9\%$ sleep staging, and $13\%$ miscellaneous studies.
However, from Fig. 13 it can be deduced that the reviews are already
obsolete.
Supervised learning and EEG

The impression is that EEG-classification studies still dominate research and a small review of the categories used in this thesis is provided here. PubMed was used with the main search terms “eeg” and “deep learning” which were combined with either “epileptiform activity”, “epileptiform discharges”, “seizure detection”, or “sleep staging”. In the case of seizure detection, there were several hundreds of hits, and the review was therefore limited to studies published from 2022 to early 2024. The studies presented here were then selected from this in steps. First, it would seem likely from the title that the study concerned EEG using deep learning and for the respective classification type. Second, if passing the first step, this would be further confirmed in the abstract. Third, passing the second step, the article was assessed and accepted if the results were presented adequately, i.e., with a concrete value for a commonly used performance score.

For EDs, the most common performance scores are ACC and AUC, to a lesser extent the F1-score. The results with references are presented in Tab. 6. The median ACC is 0.90 (min: 71; max: 1.00), the median AUC is 0.94 (0.76, 1.00), and the median F1 is 0.94 (0.75, 0.97). The size of the data used for the studies varies, some use short EEG-examples, some long EEG-recordings, and in some cases the exact size is unclear. The number of subjects range from a few subjects to several thousands of subjects.

Nhu et al. (2022) review deep learning approaches for detecting EDs published between 2012 and 2022. Their selection process results in 23 studies, all of which use supervised learning, the most common metric is AUC, and the corresponding median is 0.94.
Tab. 6. Summary of studies of classification of epileptiform discharges. ACC: accuracy; AUC: area under the curve; Study: reference.

<table>
<thead>
<tr>
<th>ACC</th>
<th>AUC</th>
<th>F1</th>
<th>Description of data</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.89</td>
<td>0.90</td>
<td>0.88</td>
<td>46 subjects with epilepsy, intracranial 13,959 IEDs</td>
<td>Abou Jaoude et al. (2020)</td>
</tr>
<tr>
<td>0.91</td>
<td>-</td>
<td>-</td>
<td>18 subjects with temporal lobe epilepsy</td>
<td>Antonaides et al. (2017)</td>
</tr>
<tr>
<td>0.79</td>
<td>0.94</td>
<td>1.00</td>
<td>Examples: 5,568 with spikes, 11,136 without spikes</td>
<td>Cheng et al. (2022)</td>
</tr>
<tr>
<td>0.97</td>
<td>0.95</td>
<td>-</td>
<td>38 subjects with focal epilepsy + 232 controls</td>
<td>Chung et al. (2023)</td>
</tr>
<tr>
<td>0.95</td>
<td>-</td>
<td>0.95</td>
<td>300 subjects, 72 hours per subject, hemorrhagic contusions</td>
<td>Faghilpirayesh et al. (2021)</td>
</tr>
<tr>
<td>0.80</td>
<td>-</td>
<td>-</td>
<td>100 subjects/EEGs</td>
<td>Fürbass et al. (2020)</td>
</tr>
<tr>
<td>0.95</td>
<td>0.95-0.97</td>
<td>0.92-0.97</td>
<td>100 subjects/EEGs</td>
<td>Geng et al. (2021)</td>
</tr>
<tr>
<td>0.99</td>
<td>-</td>
<td>-</td>
<td>70 subjects/EEGs, centrotemporal spikes</td>
<td>Jeon et al. (2022)</td>
</tr>
<tr>
<td>0.98</td>
<td>-</td>
<td>-</td>
<td>4 children with BECTS, 20 minutes recording each</td>
<td>Jiang et al. (2023)</td>
</tr>
<tr>
<td>-</td>
<td>0.95</td>
<td>-</td>
<td>5 EEGs</td>
<td>Johansen et al. (2022)</td>
</tr>
<tr>
<td>0.77</td>
<td>0.76</td>
<td>-</td>
<td>Hybrid system: Encevis, SpikeNet, and Persyst.</td>
<td>Kural et al. (2022)</td>
</tr>
<tr>
<td>0.93</td>
<td>0.93</td>
<td>-</td>
<td>Evaluation on 30 EEGs with IED and 30 without</td>
<td>McDougall et al. (2023)</td>
</tr>
<tr>
<td>0.95</td>
<td>0.94</td>
<td>-</td>
<td>TUEP, subjects: 100 epilepsy + 100 controls</td>
<td>Medvedev et al. (2019)</td>
</tr>
<tr>
<td>0.79</td>
<td>-</td>
<td>1.00</td>
<td>12 subjects, 433 hours</td>
<td>Medvedev et al. (2019)</td>
</tr>
<tr>
<td>0.78</td>
<td>0.88</td>
<td>-</td>
<td>12 subjects, intracranial EEG</td>
<td>Nejedly et al. (2023)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>0.80-0.97</td>
<td>TUEV + 257 Routine EEGs</td>
<td>Nhu et al. (2023)</td>
</tr>
<tr>
<td>0.80</td>
<td>0.86</td>
<td>0.84</td>
<td>4 EEGs, 4,615 EEGs</td>
<td>Thomas et al. (2020)</td>
</tr>
<tr>
<td>0.75</td>
<td>0.81-0.83</td>
<td>-</td>
<td>TUEP, 257 Routine EEGs</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.76</td>
<td>-</td>
<td>0.94</td>
<td>4,132 subjects, 4,615 EEGs</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>0.75</td>
<td>554 EEGs</td>
<td>Thomas et al. (2020)</td>
</tr>
<tr>
<td>0.95</td>
<td>0.97</td>
<td>0.95</td>
<td>484 EEGs containing IEDs</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.95</td>
<td>0.97</td>
<td>0.95</td>
<td>5 EEGs, focal epilepsy</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.71</td>
<td>-</td>
<td>0.71</td>
<td>11 children, 4 hours EEG per child</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.97</td>
<td>-</td>
<td>-</td>
<td>30 subjects, 1-13 hours recording per subject</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.78</td>
<td>-</td>
<td>-</td>
<td>30 EEGs with IEDs</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.78</td>
<td>-</td>
<td>-</td>
<td>1,750 EEGs with IEDs + 3,708 EEGs with no IEDs</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.95</td>
<td>-</td>
<td>-</td>
<td>50 EEGs, focal epilepsy</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.71</td>
<td>-</td>
<td>-</td>
<td>11 children, 4 hours EEG per child</td>
<td>Thomas et al. (2021)</td>
</tr>
<tr>
<td>0.87</td>
<td>-</td>
<td>-</td>
<td>30 EEGs with IEDs</td>
<td>Thomas et al. (2021)</td>
</tr>
</tbody>
</table>

Wang et al. (2023D) report a Cohen’s kappa of 0.83 for their experts, which can be compared to an AUC of 0.99 and an F1-score of 0.75 in their classification of EDs. Thomas et al (2020) report a percent agreement of 81 for their experts and their classifier achieve an F1-score of 0.84.

In the case of seizure classification, the most used metrics are ACC and F1-score. The results and references are presented in Tab. 7. The median ACC is 0.98 (min: 0.68; max: 1.0) and the median F1-score is 0.99 (0.7, 1.0). Most studies use public databases.
Table 7. Summary of studies of classification of seizure activity. ACC: accuracy; Data: database, with more information below the summary; Study: reference, where ‘-’ indicates that the result belong to the last previous study above.

<table>
<thead>
<tr>
<th>ACC</th>
<th>F1</th>
<th>Data</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>CHB-MIT</td>
<td>Abdallah et al. (2023)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.98</td>
<td>BED</td>
<td>Ahmad et al. (2023)</td>
</tr>
<tr>
<td>0.95</td>
<td></td>
<td>CHB-MIT</td>
<td>Dissanayake et al. (2022)</td>
</tr>
<tr>
<td>0.96</td>
<td></td>
<td>Siena</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>BED</td>
<td>Duan et al. (2022)</td>
</tr>
<tr>
<td>0.87</td>
<td></td>
<td>CHB-MIT</td>
<td>Feici et al. (2023)</td>
</tr>
<tr>
<td>0.96</td>
<td></td>
<td>BED</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>BED</td>
<td>Hassan et al. (2022)</td>
</tr>
<tr>
<td>0.98</td>
<td></td>
<td>CHB-MIT</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td></td>
<td>BED</td>
<td>Hilal et al. (2022)</td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>BED</td>
<td>Ibrahim et al. (2022)</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>BED</td>
<td>Islam et al. (2022)</td>
</tr>
<tr>
<td>0.85</td>
<td></td>
<td>30 epilepsy + 71 controls; 8.5 min/subject</td>
<td>Khan et al. (2023)</td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>CHB-MIT</td>
<td>Mir et al. (2023)</td>
</tr>
<tr>
<td>0.99</td>
<td></td>
<td>CHB-MIT</td>
<td>Nemati and Mesghini (2022)</td>
</tr>
<tr>
<td>0.99</td>
<td></td>
<td>BED</td>
<td>Pan et al. (2022)</td>
</tr>
<tr>
<td>0.96</td>
<td></td>
<td>CHB-MIT</td>
<td>Pameley et al. (2022)</td>
</tr>
<tr>
<td>0.93</td>
<td></td>
<td>Siena</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>BED</td>
<td></td>
</tr>
<tr>
<td>0.97</td>
<td></td>
<td>CHB-MIT</td>
<td>Patro et al. (2023)</td>
</tr>
<tr>
<td>0.96</td>
<td>0.96</td>
<td>CHB-MIT</td>
<td>Qui et al. (2023)</td>
</tr>
<tr>
<td>0.99</td>
<td></td>
<td>CHB-MIT</td>
<td>Rokhkar and Tiwars (2023)</td>
</tr>
<tr>
<td>0.98</td>
<td>0.97</td>
<td>6 subjects with focal epilepsy</td>
<td>Shah et al. (2022)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>TUH, 30 subjects</td>
<td>Shankar et al. (2022)</td>
</tr>
<tr>
<td>0.98</td>
<td></td>
<td>CHB-MIT</td>
<td>Shi et al. (2023)</td>
</tr>
<tr>
<td>0.98</td>
<td></td>
<td>Siena</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>CHB-MIT</td>
<td>Srinivasan et al. (2023)</td>
</tr>
<tr>
<td>0.96</td>
<td></td>
<td>TUSZ</td>
<td>Statsenko et al. (2023)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>Private dataset</td>
<td>Tang et al. (2024)</td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>CHB-MIT</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>TUEP</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>BED</td>
<td>Tripathi et al. (2023)</td>
</tr>
<tr>
<td>0.94</td>
<td></td>
<td>CHB-MIT</td>
<td></td>
</tr>
<tr>
<td>0.98</td>
<td></td>
<td>CHB-MIT</td>
<td>Wang J et al. (2023A)</td>
</tr>
<tr>
<td>0.99</td>
<td></td>
<td>CHB-MIT</td>
<td>Wang J et al. (2023C)</td>
</tr>
<tr>
<td>0.85</td>
<td></td>
<td>CHB-MIT</td>
<td>Wei and Mooney (2023)</td>
</tr>
<tr>
<td>0.87</td>
<td></td>
<td>CHB-MIT</td>
<td>Wong et al. (2023)</td>
</tr>
<tr>
<td>0.68</td>
<td></td>
<td>CHB-MIT</td>
<td>Yildiz et al. (2022)</td>
</tr>
<tr>
<td>0.75</td>
<td></td>
<td>TUH</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.7</td>
<td>9 subjects from CHB-MIT + 14 subjects from TUH</td>
<td>Zhao et al. (2022)</td>
</tr>
<tr>
<td>0.94</td>
<td>0.84</td>
<td>BED</td>
<td>Zhang et al. (2023A)</td>
</tr>
</tbody>
</table>

AU: Ankara University Hospital Dataset; 27 TLE patients
BED: Bonn Epilepsy Dataset; 500 x 23.6 s, 100 ictal, one-channel
CHB-MIT: Children’s Hospital Boston-Massachusetts Institute of Technology; 22 subjects, > 664h
Siena: Siena Scalp EEG database; 14 subjects, 5-23 h/subject
TUEP: Temple University Hospital EEG Ictal Corpus; 100 epilepsy patients + 100 control subjects
TUH: Temple University Hospital EEG Epilepsy Corpus; 1229 seizures, 10,874 subjects
TUSZ: Temple University Hospital EEG Seizure Corpus; 674 subjects, multi-class seizure types
UPenn: Pennsylvania University; 8 subjects, 653 s seizure, 7,664 s non-seizure

Sleep stage classification is most often evaluated using ACC or F1-score. The results with references are presented in Tab. 8. For the ACC the medians are W 0.94 (min: 0.82; max: 0.98), N1 0.55 (0.40, 0.94), N2 0.87 (0.79, 0.92), N3 0.84 (0.75, 0.99), R 0.92 (0.80, 0.98), and the median number of PSGs is 3,342 (4, 15,804). For the F1-score the medians are W 0.89 (0.80, 0.94), N1 0.45 (0.20, 0.67), N2 0.87 (0.68, 0.92), N3 0.84 (0.33, 0.89), R 0.86 (0.66, 0.91), and the
The median number of PSGs is 91 (19, 5,463). In the study by Abou Jaoude et al. (2020), they report an overall Cohen’s kappa values of 0.75±0.11 for the experts and 0.78 for their method.

<table>
<thead>
<tr>
<th>Type</th>
<th>W</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>R</th>
<th>PSGs</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>0.94</td>
<td>0.53</td>
<td>0.82</td>
<td>0.70</td>
<td>0.78</td>
<td>6,341</td>
<td>Abou Jaoude et al. (2020)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.85</td>
<td>0.58</td>
<td>0.89</td>
<td>0.86</td>
<td>0.92</td>
<td>15,804</td>
<td>Biswal et al. (2018)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.98</td>
<td>0.94</td>
<td>0.92</td>
<td>0.99</td>
<td>0.98</td>
<td>602</td>
<td>Cho et al. (2022)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.94</td>
<td>0.45</td>
<td>0.87</td>
<td>0.78</td>
<td>0.85</td>
<td>153</td>
<td>Korkalainen et al. (2020)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.89</td>
<td>0.47</td>
<td>0.87</td>
<td>0.80</td>
<td>0.91</td>
<td>891</td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.92</td>
<td>0.66</td>
<td>0.89</td>
<td>0.94</td>
<td>0.87</td>
<td>100</td>
<td>Kwon et al. (2022)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.82</td>
<td>0.40</td>
<td>0.79</td>
<td>0.92</td>
<td>0.80</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.94</td>
<td>0.45</td>
<td>0.91</td>
<td>0.75</td>
<td>0.92</td>
<td>15,684</td>
<td>Olesen et al. (2021)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.94</td>
<td>0.62</td>
<td>0.86</td>
<td>0.83</td>
<td>0.92</td>
<td>5,793</td>
<td>Zhang et al. (2023B)</td>
</tr>
<tr>
<td>ACC</td>
<td>0.93</td>
<td>0.63</td>
<td>0.87</td>
<td>0.85</td>
<td>0.93</td>
<td>14,118</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.89</td>
<td>0.43</td>
<td>0.87</td>
<td>0.85</td>
<td>0.83</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.84</td>
<td>0.20</td>
<td>0.80</td>
<td>0.33</td>
<td>0.66</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.80</td>
<td>0.42</td>
<td>0.83</td>
<td>0.78</td>
<td>0.79</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.80</td>
<td>0.44</td>
<td>0.84</td>
<td>0.73</td>
<td>0.68</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.86</td>
<td>0.60</td>
<td>0.92</td>
<td>0.83</td>
<td>0.91</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.92</td>
<td>0.55</td>
<td>0.90</td>
<td>0.65</td>
<td>0.87</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.81</td>
<td>0.22</td>
<td>0.68</td>
<td>0.64</td>
<td>0.82</td>
<td>3,082</td>
<td>Huttunen et al. (2021)</td>
</tr>
<tr>
<td>F1</td>
<td>0.90</td>
<td>0.65</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>173</td>
<td>Lal et al. (2023)</td>
</tr>
<tr>
<td>F1</td>
<td>0.89</td>
<td>0.39</td>
<td>0.87</td>
<td>0.87</td>
<td>0.83</td>
<td>39</td>
<td>Lv et al. (2022)</td>
</tr>
<tr>
<td>F1</td>
<td>0.92</td>
<td>0.43</td>
<td>0.83</td>
<td>0.88</td>
<td>0.78</td>
<td>153</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.93</td>
<td>0.51</td>
<td>0.89</td>
<td>0.85</td>
<td>0.90</td>
<td>5,463</td>
<td>Phan et al. (2023)</td>
</tr>
<tr>
<td>F1</td>
<td>0.94</td>
<td>0.53</td>
<td>0.90</td>
<td>0.88</td>
<td>0.89</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.87</td>
<td>0.30</td>
<td>0.85</td>
<td>0.87</td>
<td>0.86</td>
<td>344</td>
<td>Wang et al. (2021)</td>
</tr>
<tr>
<td>F1</td>
<td>0.93</td>
<td>0.37</td>
<td>0.88</td>
<td>0.81</td>
<td>0.87</td>
<td>100</td>
<td>Yan et al. (2021)</td>
</tr>
<tr>
<td>F1</td>
<td>0.89</td>
<td>0.57</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.94</td>
<td>0.67</td>
<td>0.86</td>
<td>0.89</td>
<td>0.86</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>0.88</td>
<td>0.54</td>
<td>0.87</td>
<td>0.47</td>
<td>0.86</td>
<td>265</td>
<td>Zhang et al. (2020)</td>
</tr>
<tr>
<td>F1</td>
<td>0.90</td>
<td>0.47</td>
<td>0.86</td>
<td>0.82</td>
<td>0.83</td>
<td>197</td>
<td>Zhu et al. (2020)</td>
</tr>
</tbody>
</table>

### 11.11.2 Self-supervised learning and EEG

Almost all classification studies presented in the previous section are based on supervised learning. One exception is the study by Yıldız et al. (2022), where variational autoencoders are trained in a self-supervised manner and used for seizure detection. Another example, outside the categories presented above, is Li et al. (2020) who classify EEG signals from steady-state visually evoked potentials using a network that compares different stimulations and incorporates a correlation layer.

Self-supervised strategies are mostly used to pretrain DNNs from which classifiers are created and many show that the
performance of supervised learning can be increased (Daoud et al., 2020; Eldle et al., 2023; Hermans et al., 2023; Li et al., 2020; Ng et al., 2023; Kostas et al., 2021; Luo et al., 2022; Ou et al., 2022; Qui et al., 2018; Sahani et al., 2021; Sartipi et al., 2024; Vařeka et al., 2017; Wang et al., 2023; Wulsin et al., 2011; Yang et al., 2023; Yuan et al., 2019). There are also examples of training DNNs self-supervised and using their output as features for non-deep machine learning classification (Dairi et al., 2022; Liu et al., 2023, Hassan Shah et al., 2023; Supratak et al., 2014; Tautan et al., 2019, Zhou et al., 2019). Self-supervised learning is used for artifact correction by Saba-Sadiya et al. (2021). The most common approach in these studies is to use some form of autoencoder, where the most common training task is reconstruction of the input. Many of the studies use preprocessed EEG, e.g., frequency or time-frequency representations. Deep Boltzmann machines are used to create two-dimensional representation of EEG (Liu et al., 2023; Hassan Shah et al., 2023) and compare favorably to t-SNE when used as features in classification tasks.

Adversarial learning, as in GANs, may be regarded as self-supervised, supervised, or a combination of both depending on approach and viewpoint. Habashi et al. (2023) reviews the use of different types of GANs for EEG. They identify 43 articles published between 2018 and 2021. The studies are categorized into five groups: 28% BCI, 18.5% epilepsy, 18.5% emotion recognition, 9% event related potentials, and 26% miscellaneous. The most common objective of the studies is data augmentation when EEG data is limited. Data augmentation is used to counteract overfitting, but GANs are also used to counteract differences in feature learning due to inter-subject differences (Fu et al., 2022; Gao et al., 2023; She et al., 2023).

A GAN strategy is used by Liang et al. (2021) for self-supervised pretraining of DNNs from which new feature representations are extracted and used for hypergraphs in classifications tasks. Yu et al. (2023) demonstrate improved classification performance by training a classifier part supervised and part as the discriminator in a GAN setup.

There are several examples of using GANs for different types of signal processing. Luo et al (2020) use GANs to increase the sampling frequency of EEG. Different GAN-strategies are also used for artifact reduction (An et al., 2022; Brophy et al., 2022; Dong et al., 2023; Sawangjai et al., 2022). Hu et al (2022) use GANs to transform scalp EEG to stereo-EEG. Cheng et al. (2021) transform EEG into fMRI. Although more being an example of image analysis, Xu et al. (2021)
use cycle-GANs to transform between EEG topographies of stroke patients and healthy subjects.

11.11.3 t-SNE and EEG

In more recent studies, t-SNE in the context of EEG and deep learning is mostly used to visualize latent space of DNNs (Chen et al., 2023; Chung et al., 2023; Hassan Shah et al., 2023; Jeon et al. 2023, Idowu et al., 2021; Kim et al., 2023; Malafeev et al., 2021; Ng et al., 2023; Ravi et al., 2020) or the effect of feature extraction (Tang et al., 2024). This is usually performed to evaluate how categories are separated in these new representations to assess their performance in classification tasks. t-SNE was used by George et al. (2022) to compare real and synthetic EEG-data.

There are also examples where t-SNE (Ma et al., 2021) or parametric t-SNE (Li et al., 2018; Xu et al, 2020) are used to create new features from EEG. Arcot Desai et al. (2022) use networks pretrained on images, e.g., ResNet50, to extract features from a large set of intracortical EEG, then apply t-SNE followed by k-means clustering to identify EDs and seizure activity.

11.12 Deep learning from a practical point of view

Deep learning is usually computer intensive. Training times can be accelerated by parallelizing the computations. This is most often accomplished by using graphic cards (GPU). At the present time Nvidia dominates the field. A limiting factor of GPUs is the memory size. In this work, several GPUs were used and was equipped with 12–24 GB. In many instances, 12GB was too limited.

In addition, it is favorable to have a decent amount of working memory, and the computers used during the work had 64–128 GB. This will enable loading larger datasets. The alternative to load data intermittently during training will add substantially to the time consumption and may reduce variability in the training data.

A computer with these specifications will be relatively expensive compared to an ordinary stationary computer. Using remote resources may be an option, e.g., universities may have their own resources, and there are also cloud-based services. The drawback of these alternatives may be limitations in accessibility, bandwidth, memory (both GPU memory and working memory), and service charges.

There are several frameworks for deep learning, e.g., TensorFlow, PyTorch, FastAI, Theano, MXNet, and Caffe2. It is of course
hard to measure which is best, or even most popular. Using the names in different searches, TensorFlow and PyTorch get most hits. For example, in a search using www.google.com (performed on Mars 6, 2024) the number of hits for the respective frameworks were: 69, 58, 2.4, 2.2, 1.1, and 0.7 million. In this work, the main programming language was Python, and the deep learning framework were TensorFlow and Keras. TensorFlow is the underlying framework for building and training DNNs, and Keras is a high-level API, providing an interface making TensorFlow easier to use.
12 Aims

The objective of this thesis was to investigate the use of self-supervised deep learning methods in EEG analysis. The more far-reaching goal was to develop a tool to study categories in EEG data. During the course of the work, the more concrete goal of the thesis has been in line with this but successively moved forward. In the end, a tool was created that at least was closing in on the far-reaching goal, the research could in part push past it and start investigating how experts and DNNs categorize data.

12.1 Initial aims

At the start of the work, it was far from clear how to reach the goal of creating the tool and it was assumed not to be accomplished within the duration of the research studies. Therefore, a more limited approach was taken. The idea was to take a self-supervised method, adapt and optimize it to perform well for EEG data. From the principle of the chosen method, it should be reasonable to assume that it would promote the resulting DNNs to learn and represent important EEG features in their latent space. Finetuning these DNNs with supervised learning for classification in a second step and showing that performance improved classification compared to not using pre-trained DNNs would further support that they learn relevant features. This would also in itself be a useful result regarding data classification and motivate the work. This strategy is in line with several studies presented in section 11.11.2. To summarize:

1. Identify a suitable self-supervised method.
2. Adapt and optimize it for EEG.
3. Perform a study to show that the method performs satisfactorily.
4. Perform a study to show that using the method to pretrain DNNs improves classification.

In the setting of the more far-reaching goal, these self-supervised DNNs could be used to generate new representations of EEG that, hopefully, could be more suitable for, e.g., clustering analysis to study similarity between different EEG patterns.

At the start of the studies, cyclic GANs were identified as a possible self-supervised method. In the initial work, transformation between different electrode setups were performed. However, during the work, tests were also performed to achieve this in a simpler
manner using autoencoders. The tests were promising, and a study was designed, implemented, and the results were published (paper I).

12.2 Realigned aims

The realization midway into the work that parametric t-SNE had the potential to be further developed to improve performance for EEG changed the goals of the work. The original plan was abandoned, and the aims changed accordingly:

1. Optimize parametric t-SNE for EEG.
2. Perform a study to show that the new version improve performance.
3. Perform a study comparing classification performed by experts and DNNs and use the new method to add dimensions to the analysis.

This resulted in papers II and III.
13 Ethical considerations

All EEG data used in this work were extracted from the Temple University Hospital EEG data Corpus (Obeid and Picone, 2016). These data had been collected according to local ethical guidelines, were anonymized and impossible to trace back to the research subjects. The assessment was therefore that no further ethical approval was necessary.

In paper III annotations were performed and the annotators were thus study subjects. The used subjects were three members of the research group. This was informally discussed with a member of the Swedish Review Authority (Etikprövningsmyndigheten). The consensus was that for this small number of subjects and, foremost, because the subjects were members of the research group, no ethical approval was necessary.
14 Results

The main result of these studies was the development of encoders producing visualizations of EEG which was based on the principle of t-SNE. These encoders will be referred to as t-SNE encoders.

14.1 Paper I – Upsampling and reconstruction of EEG channels

DNNs with autoencoder architecture were trained to restore 21-channel EEGs. Three different conditions were tested: 1) upsampling from four electrodes, 2) upsampling from 14 electrodes, and 3) restoring a randomly missing electrode. Comparisons were made with interpolation using spherical splines and the hypothesis was that the DNNs would produce more realistic signals. The assessment included several quantitative metrics for all DNNs and a visual test for one of the DNNs.

According to most metrics, the DNNs performed significantly better than interpolation but in absolute numbers, differences were small. In the visual test, five clinical neurophysiologists more often rated EEGs restored by the DNN as real compared to interpolation.

14.2 Paper II – t-SNE encoders

DNNs were designed to produce both high-dimensional and visual representations of EEG data and learned these during training using the principle of t-SNE. The method was evaluated on three datasets: 1) wakefulness vs. sleep, 2) EDs vs. non-EDs, and 3) seizure activity vs. non-seizure activity. Comparisons were made with parametric t-SNE using either short-time Fourier transform, or wavelet transform as the high-dimensional representation. The hypothesis was that the proposed method would produce more distinct clusters. Evaluations were performed quantitatively using support vector machines (SVM) and k-means clustering, and the resulting visual representations were reported.

Separation of the categories of the datasets according to the SVM evaluation were similar for the methods. The visual representations and the k-means cluster evaluation indicated that clustering was more distinct for the DNNs.

14.3 Paper III – Interrater agreement and classification EDs

The subject of paper III was expert interrater agreement of EDs and how this may affect training of DNNs to classify discharges. It was a pilot study designed to assess the conditions for conducting a larger study. Two clinical neurophysiologists annotated an EEG for EDs.
Classifiers implemented as DNNs were trained based on these annotations, which in turn produced new annotations. All annotations were compared using agreement scores. A clustering analysis was performed to better understand the results using a further development of the DNN-based clustering method of paper II. The hypothesis was that differences in agreement between experts would affect the learning of the classifiers and that the agreement between an expert and a classifier would not be less compared to the agreement between two experts. From a study design point of view, the main questions were whether the annotation task was reasonable to perform and if the resulting data would be of sufficient size to train classifiers.

Agreement scores were higher when comparing classifiers to each other or an expert to a classifier. Clustering analysis indicated that the detected EDs were similar in most instances, but there were instances with larger differences and that this probably affected the learning of the classifiers.

Several instances of improvements were identified for a larger study. To improve the conditions, especially for training classifiers, it seems feasible to use a longer EEG recording. Intrarater agreement should be assessed by having the experts annotate the EEG two times. The classifiers as well as the cluster method can hopefully be further optimized.

14.4 Additional examples

14.4.1 Channel reconstruction

In this section, a drawback of the implementation used in paper I is demonstrated, but it is also demonstrated that the method can be further improved.

The results of the DNNs used in paper I gave the impression that the reconstructing of high frequency content was inferior compared to low frequency content. The networks had a simple bottleneck architecture where the data were first downsampled and then upsampled to their original size (Appendix 17.3; Fig. 20 A). A simple way to improve performance is to use skip connections (Ronneberger et al., 2015), where information from each level in the downsampling step is forwarded to the corresponding level in the upsampling step (Appendix 17.3; Fig. 20 B). This will better preserve local information and higher frequencies will be recreated better (Fig. 14 C; Fig. 15 C). There is no claim that this is the best approach for improvements, but it is provided here to make it credible that using DNNs to restore signals
has more potential. For details on how the results presented here were produced, see Appendix 17.3.

Three DNNs were trained. DNN1 illustrate the low pass filtering effect of the architecture and downsampled the data by a factor of eight (Fig. 14 A and Fig. 15 A). DNN2 used a downsampling factor of four, similar to paper I (Fig. 14 B and Fig. 15 B). DNN3 incorporated skip connections in the architecture and used a downsampling factor of eleven (Fig. 14 C and Fig. 15 C). The DNNs were trained on 10-s 21-channel EEG and the objective was to reconstruct the input signals, where in addition a random number (1–10) of channels were missing.

The evaluation was divided into comparing the reconstruction of signals that were present and signals that were missing in the input data. Using the absolute error when comparing the generated to the original signals, DNN3 had the smallest error in both instances, followed by DNN2, while the performance of DNN1 as expected was poor (Tab. 9). DNN3 excelled in reconstructing channels present in the input, but the difference for missing channels was less pronounced.

Tab. 9. Statistics of the reconstruction error. The test data consist of 100 EEGs of 10 minutes duration. Median of the absolute error and the 95th percentile. Best results are in bold. A: reconstructed signals in input; B: reconstructed signals missing in input. Kolmogorov-Smirnov testing of normality showed that the absolute errors were not of normal distribution. Non-parametric statistics were used. Pairwise comparisons of the errors showed significant differences in all instances (p ≪ 0.0001; paired Wilcoxon signed-rank test).

<table>
<thead>
<tr>
<th></th>
<th>DNN1</th>
<th>DNN2</th>
<th>DNN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Median</td>
<td>0.091</td>
<td>0.094</td>
<td>0.034</td>
</tr>
<tr>
<td>B Median</td>
<td>0.050</td>
<td>0.056</td>
<td>0.0049</td>
</tr>
<tr>
<td>A 95th</td>
<td>0.41</td>
<td>0.42</td>
<td>0.15</td>
</tr>
<tr>
<td>B 95th</td>
<td>0.26</td>
<td>0.25</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Fig. 14. The average spectra of ten missing signals. The signals were randomly selected from the test data. The spectra of the original signals are in red, and the spectra of the reconstructed signals are overlaid in black. (A) DNN1. Downsampling by a factor of eight. (B) DNN2. Downsampling by a factor of four. (C) DNN3. Downsampling by a factor of eleven but using an architecture with skip connections.
Fig. 15. Example of ten reconstructed missing signals. The signals are 3-s sections randomly selected from the same 10-s signal as in Fig. 14, which in turn had been randomly selected from the test data. The original signals are in red, and the reconstructed signals are overlaid in black. (A) DNN1. Downsampling by a factor of eight. (B) DNN2. Downsampling by a factor of four. (C) DNN3. Downsampling by a factor of eleven but using an architecture with skip connections.

14.4.2 Comparison of some approaches for t-SNE encoders

To compare and illustrate different approaches to process and cluster data using principles developed during the work on papers II and III, a couple of examples were produced:

1. Semi-supervised: A classifier was first trained to classify sleep stages and then used to produce a new representation for parametric t-SNE (Fig. 16 A).
2. Supervised: A convolutional t-SNE encoder was trained where the neighbor distributions were based on the categories (Fig. 16 B).
3. Self-supervised: Convolutional t-SNE encoder (Fig. 16 C).
4. Self-supervised: t-SNE encoders trained on a larger dataset and using 30-s epochs (Fig. 16 D). First, a t-SNE encoder based on convolutions was trained on 3-s sections of EEG. This was then used to generate a new representation of the 30-s epochs during training of a t-SNE encoder based on a transformer architecture.
5. Self-supervised: Encoder according to paper III with automatic cluster identification (Fig. 16 E and F).
The 12-hour recording annotated for sleep stages used in the appendix of paper II was reused. For more details, see appendix 17.4. The annotation had been performed for 30-s epochs. Each epoch was divided into three 10-s epochs, each assigned the same sleep stage as the 30-s epoch.

An exception was approach 4, where 30-s epochs were used. In this case, two encoders, which process the data serially, were trained on 998 EEGs, each of 5 minutes duration, and then evaluated on the sleep staged EEG and in addition the EEG of paper III, representing a focal status epilepticus. This was to demonstrate the possibility to create a more general encoder.

The results are presented in Fig. 16. The data is color coded according to sleep stages, except in Fig. 16 D where there in addition is the EEG of status epilepticus, color coded in black, and in Fig. 16 F where the color coding is according to the encoder’s automatic cluster identification.

In all instances, the sleep stages to a large extent seem to aggregate. Most have a confluent character, where the data most often did not form distinct and separated clusters. Approach 5 was the only one producing separate clusters, while approach 2 at least have strong tendency to form clusters.
Fig. 16. Examples of different clustering approaches. (A) A classifier was first trained on the data and then used to generate a new representation of the data which in turn was used for parametric t-SNE. (B) Convolutional parametric t-SNE where all examples belonging to a category was set as neighbors. (C) Simpler version of convolutional parametric t-SNE. (D) Convolutional t-SNE trained on 998 EEGs. The black dots represent the EEG of paper III. (E) Convolutional parametric t-SNE with cluster encoding. (F) Automatic color coding of (E).
In addition, the distribution of the sleep stages per cluster was visualized as histograms (Fig. 17). Clusters 1, 2, 3, 4, 5, and 8 is mostly dominated by one sleep stage. Clusters 6, 7, and 9 are dominated by sleep stage N2 but with a significant portion of N3 for cluster 6, N1 for cluster 7, and N1 and R for cluster 9. Cluster 10 is dominated by N1 but closely followed by W and N2, and a significant portion of R.

![Fig. 17. Distributions of the EEG per cluster. Top row: Sleep stages. Bottom row: examples across time (the x-axis of each histogram represents 12 hours).](image)

The resulting clusters of approaches 1 and 3 was plotted using the color code of approach 5 (Fig. 18). The colors were relatively well aggregated. Cluster 6 is a transitional zone between stage N2 and N3. Cluster 7 is a branch with a mixture of stages N1 and N2, but for approach 3 it also contains a subcluster of R (cluster 2). Cluster 9 is a transitional zone between stages R and N2, to some extent N1, and less so W. Cluster 10 is a transitional zone to stage W.

![Fig. 18. Alternative color coding. Color code according to Fig. 16 F. (A) Approach 1 (Fig. 16 A). (B) Approach 3 (Fig. 16 C).](image)
15 Discussion

15.1 Signal reconstruction

In paper I, DNNs were used to reconstruct EEG signals and were shown to perform on a level comparable to spherical splines. The DNN method performed slightly better, but the difference was small, and in most instances, it is probably not worth the effort to implement the method if it is only to be used for interpolation. In addition, in section 14.4.1 a weakness of the network architecture in the form of a low pass filtering effect was demonstrated. One advantage of the method is that it reconstructs signals automatically. In comparison, for interpolation the identification of missing channels has to be made by hand or by an additional method.

In section 14.4.1 it was also demonstrated that the method can be improved, by inserting skip connections, and it is at least speculated that the method can be developed to outperform interpolation methods by a larger margin. As have been shown in previous studies, the method can be used for other types of signal processing, e.g., artifact reduction (Saba-Sadiya et al., 2021), and it might be possible to develop a more general signal processing tool using deep learning. Autoencoder approaches tend to have more distinct targets during training. For example, in paper I, the target was to exactly reconstruct the original signals, guided by the loss function, the mean absolute error. This will restrict what is possible to achieve. For instance, if the objective is to reduce artifacts, then the data would have to contain pairs of the same EEG signal with and without artifacts. For some practical applications, the most important objective may be to reconstruct the character of the signals, rather than reconstructing them exactly. An approach like GANs could possibly offer less restrictive options to train. In the review in section 11.11.2, there are several examples of using GANs for signal processing. One example is the study by Dong et al. (2023), where the generator of a GAN takes noisy EEG as input and produce a denoised EEG, while the discriminator is trained to discriminate between denoised and real artifact-free EEG. They compare with several other methods, including autoencoders, where their method performs favorably.

The network architecture used in paper I is probably not the most suitable to generate new representations for clustering, as intended in this work. Since it is not possible to use a high degree of downsampling, represented features will be small. Using the modification with skip connection as in section 14.4.1 may not be ideal either, since there is a risk that the skip connections just shunt...
information without actually learning more complex features. This shunting effect is speculated to be a contributing explanation why the DNN outperformed the other two in reconstructing signals that were present in the input. It is speculated that GANs are a better option for generating new representations for clustering compared to more pure implementations of autoencoders. It might be argued that the t-SNE encoders are a form of autoencoders, but they achieve the objective.

15.2 t-SNE encoders

The starting point of paper II is parametric t-SNE from which the principle of t-SNE encoders is derived. The work evolved over several years and culminated in the relatively elaborate encoder of paper III. In section 14.4.2, several variations of t-SNE encoders are presented, partly to show that the principle can be implemented in many ways. In paper II, the performance of the encoder was compared to prefabricated features based on time-frequency methods and was shown to produce more distinct clusters. In paper III, the more elaborate encoder was demonstrated to produce very distinct clusters and was used to visualize and compare different annotations of EDs, to aid in understanding the results.

Using the automatic cluster identification of approach 5 to color code the result of approaches 1 and 3, demonstrate some consistency between the approaches in how the data were arranged based on similarity. Approach 5 produced relatively dense clusters and their relation to each other may not be immediately apparent, but Fig. 18 indicates that the clustering is relevant and interpretable. This may be regarded as support for a potential of developing this type of encoder for automatic classification, i.e., self-supervised classification.

Approach 2 may visually seem like the best option, but this version is strictly supervised. When using the color coding generated by approach 5, these become more mixed compared to approaches 1 and 3. The general structure, i.e., how the clusters are localized relative to each other, conform to how the sleep stages relate to each other.

Approach 4 is maybe the most interesting, not because of the architecture or strategy, but because it demonstrates that it may be possible to create a more general encoder. It is seen that the status epilepticus data form a relatively separate cluster (Fig. 16 D). All the training data has not been thoroughly inspected but most have been reviewed cursory. Most of the data is in the normal range, or with only slight abnormalities. There are several EEGs with severe abnormalities or seizures, and a few cases of status epilepticus, but these are in a clear minority. The data consist of the first 5 minutes, in the author’s
experience, this means that most are awake, it is unusual to see sleep stage N3, and very rare to see sleep stage R. This may have contributed to stage W being spread over a larger area. The data is thus imbalanced and using more balanced data, or training, could produce better results.

Two of the main clinical potentials for clustering encoders are for rapid screening of EEG or to create more intuitive trends. Such trends would most likely have a better specificity compared to the currently used trends. Using the method to produce trends for LTMs in ICUs would be valuable and it is speculated that it could actually visualize the ictal-interictal continuum. It would not in itself solve the question when there is non-convulsive status epilepticus, but it might make it easier to discuss or draw limits, and it will actually provide a quantitative measure. Jing et al. (2018) use high-dimensional feature extraction from EEG, visualize the data using t-SNE and principal component analysis, and show how this can be used for rapid annotation in the context of the ictal-interictal continuum.

15.3 EEG interpretation and scoring systems

Rules and systems for interpreting EEG have to be interpreted, and as is evident from the review in section 10.3, there will be disagreement. Take for example the term “pointy” in the definition of EDs by Kane et al. (2017)—what is spiky? An example of a very flexible definition is the term electroclinical seizure according to Hirsch et al. (2021). Obviously, it involves clinical symptoms, but it can be interpreted as if the correlating EEG activity can be just about anything.

The threshold for scoring EDs was tested in paper III. The two experts assessed roughly 9,000 potential candidate waveforms, which although having a variation in morphology also had large similarities. In this case, no attempts were made to streamline their assessment by providing a definition of EDs or a scoring guideline. The task was highly unnatural from a clinical neurophysiologist’s perspective since there is no practical value in identifying each single ED. In practice, it suffices to identify that they are present and report a rough estimate of the amount. In this specific EEG, the most important finding was the periodic discharges, which indicate a status epilepticus.

From a classifier’s “perspective”, the scoring system itself may induce a form of label noise. For instance, in sleep staging, an epoch of irregular theta activity may by a human interpreter be classified as N1 if it follows W, as N2 if it follows N2, and as R if it follows R. If a classifier only analyzes locally, i.e., base the classification of an example without knowledge of the examples preceding it, these
classification rules cannot be learned. This may explain some of the mixing of the sleep stages in classification and clustering. Then again, it is probable that the activity following W, N2, and R often have different characters. Apart from examples on the threshold between stages, based on agreements score presented in section 10.3.3, N1 is expected to be most difficult to classify and attain satisfying clustering.

The recording used throughout this work to assess sleep staging was not a polysomnography. This means that stage R is assessed without EOG and EMG, which leads to greater uncertainty. In the examples provided in section 14.4.2, each annotated 30-s epoch was divided into three 10-s epochs, each assigned the same sleep stage as the 30-s epoch. This was likely to increase the label noise. For example, a 30-s epoch classified as N1 may have a dominating alpha rhythm the first 10 s before settling in theta activity. Another example is if an epoch starts in N2, there is an arousal in the middle with rhythmic alpha activity, and it ends in dominant irregular theta activity without sleep spindles or K-complexes—the sections would possibly resemble N2, W, and N1, respectively.

Clustering analyses could improve classification systems. It might be possible to discover subclasses, that some class may be poorly defined, or even unnecessary. From the result of agreement and classifications studies, and possibly the provided cluster analyses in this work adding to it, maybe stage N1 is such a class? Clustering analyses may also be used as a more intuitive aid to delimitate categories, or to realize that there are no distinct boundaries. Approaches 2 and 5 in section 14.4.2 produced more distinct clusters, in both cases this may largely be due to the implementations. In the first case, it is based on supervised learning, and in the second case, the network architecture “forces” the data to be arranged as separate clusters. The other approaches may more accurately reflect the character of the data, there are at least some dimensions in it that changes more gradually than abruptly. This means that there may not be distinct delimitations between the categories.

15.4 Is it reasonable to expect classifiers to achieve high scores when experts disagree?

The median agreement scores and classification performance metrics for single EDs, seizure activity, and sleep staging from the reviews in sections 10.3 and 11.11.2 have been summarized in Tab. 10. Of course, comparing the different types of measures is not straight forward. Above all, imbalanced datasets have different impact on the
measures. For example, Cohen’s kappa can produce very low values. Furthermore, there are uncertainties in several cases and the medians are based on mixtures of different measures. Nonetheless, it is surprising that the classification metrics are consistently higher compared to the agreement scores. A curiosity in the review of agreement for EDs is that the median Gwet’s AC1 is slightly lower compared to Cohen’s kappa.

**Tab. 10.** Summary of the median agreement scores and classifications measures from the reviews in sections 10.3 and 11.1.2. EDs: epileptiform discharges; W, N1, N2, N3, R: sleep stages; IRA: interrater agreement; ACC: accuracy; AUC: area under the curve; F1: F1-score.

<table>
<thead>
<tr>
<th>EDs</th>
<th>IRA</th>
<th>ACC</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0.56</td>
<td>0.90</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>N1</td>
<td>0.78</td>
<td>0.98</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>N2</td>
<td>0.89</td>
<td>0.94</td>
<td>0.55</td>
<td>0.89</td>
</tr>
<tr>
<td>N3</td>
<td>0.40</td>
<td>0.55</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>R</td>
<td>0.65</td>
<td>0.87</td>
<td>0.67</td>
<td>0.82</td>
</tr>
</tbody>
</table>

For comparison, some of the results of paper III are presented in Tab. 11. Here the effect on the values due to the unbalanced data is demonstrated. Gwet’s AC1 is on par with the unadjusted ACC, while Cohen’s kappa, R-AUC, and F0.5 have similar values. It is also interesting to observe that when the classifications metrics are used to compare the experts, the values show similar patterns as when comparing experts to classifiers.

**Tab. 11.** Summary of some of the results of paper III. E: expert; C: classifier; KC: Cohen’s kappa; AC1: Gwet’s AC1; ACC: accuracy; B-ACC: balanced accuracy; R-AUC: area under the curve based on the receiver operating characteristic.

<table>
<thead>
<tr>
<th>E vs E</th>
<th>KC</th>
<th>AC1</th>
<th>ACC</th>
<th>B-ACC</th>
<th>R-AUC</th>
<th>F0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.56</td>
<td>0.92</td>
<td>0.93</td>
<td>0.74</td>
<td>0.49</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>0.56-0.65</td>
<td>0.92-0.94</td>
<td>0.94-0.96</td>
<td>0.76-0.86</td>
<td>0.52-0.72</td>
<td>0.53-0.75</td>
<td></td>
</tr>
<tr>
<td>0.67-0.82</td>
<td>0.94-0.96</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

If the performance metrics indicate that classifiers perform so close to human experts, why are they not in extensive use clinically? How can there be instances of high performing classifiers where experts only have a moderate agreement? Overfitting may be one reason, so when more extensive testing is performed it will be lower. The overfitting may be to the specific used data, so if tested on larger amounts of data from the same subject and rated by the same expert, performance would be inferior. It could also be that the model generalizes to the used subjects and experts, but not when testing on larger populations. Many studies use a limited data material, making it unlikely that their classifiers have a good generalization. One possibility
is that, at least in some cases, researchers tend to choose the evaluation metrics that better fit the results. For instance, not using compensation for imbalanced datasets may yield high values, as demonstrated in paper III (Tab. 11, e.g., compare ACC to B-ACC). The data may also consist of examples that are relatively easy to classify, e.g., there may be an intentional or unintentional tendency to avoid examples where experts tend to disagree. EEG has a continuous character, where events can occur at any time. How the evaluation is performed with respect to this fact can influence the results. Testing with arranged examples, e.g., examples containing EDs always have the discharges nicely centered in the middle of the examples, may be much easier compared to a random position in the examples. One example of probably being too easy is the Bonn Epilepsy Dataset (Andrzejak et al, 2001). It consists of data from five subjects with epilepsy and five control subjects. There are 500 one-channel examples of 23.6 s duration out of which 100 are ictal. There are several published studies that achieve an accuracy of 1.0 (Tab. 7). Another important reason that classifiers are not in extensive use may be cautiousness, suspiciousness, or even reasons of principle against using this type of technology.

In the review, there are few studies assessing intrarater agreement but from the results found, it is speculated that it follows similar patterns as the interrater agreement. Intrarater agreement seems as just an important question as interrater agreement to assess when evaluating classifiers. If there is intrarater variability, it should be almost impossible to achieve a perfect accuracy except by chance.

In paper III, it was speculated that the threshold for scoring can vary over time, i.e., intrarater variability. Some of the variation may be due to long-term dependencies in the EEG signals, and some of the variation may be due to external factors with no connection to the EEG. A classifier that analyzes locally, i.e., assesses examples independently from other examples, cannot learn either type of variation of the threshold. A classifier that analyzes globally, i.e., that can incorporate information from other parts of an EEG recording, may learn the first type of variation, but not the second type. In the paper it was further speculated that the classifiers will learn some form of average. Since examples of different annotations tended to aggregate according to similar patterns in the clustering analysis, it was interpreted as even though they were not always exactly the same, they were often similar. Nonetheless, there were also indications that there were differences, and that this was mirrored in how the classifiers performed.
The classifiers achieved a slightly better agreement compared to the experts in paper III. This could imply that the classifiers perform something that is more consistent compared to the experts, which could be consistent with that the classifiers learn an average way of scoring. However, agreement in itself is not a proof that opinions are more correct, there are countless examples of this throughout history. It could also mean that the classifiers are using a strategy that is simpler, missing special cases that are rarer or harder.

From all this, it is suggested that a proper evaluation of classifiers should, in addition to using large amounts of data, include an evaluation of the experts' labeling of the data. This may include intrarater agreement. A further step would be to perform blinded tests, e.g., where experts assess annotations without knowing if it is made by another expert or a classifier, to see if they can discriminate between them. This would be a form of Turing tests. Scheuer et al. (2017) evaluate Persyst 13 for spike detection and Scheuer et al. (2021) evaluate Persyst 14 for seizure detection. They use what they refer to as a statistical Turing test, where pairwise comparisons between expert-expert and expert-classifier are performed. Based on this comparison they conclude that the detectors have a non-inferior performance. Westover et al. (2017) commend the principle of the evaluation of the study from 2017, but points out several weaknesses, e.g., data size and quality of the experts, and the results from the studies are difficult to assess since they involve proprietary commercial systems.

I would argue that a high classification score is unrealistic and either indicates a too small or too easy dataset.

15.5 Deep learning and EEG

15.5.1 General reflections from the work

When working with deep learning models, there are many choices that have to be made, e.g., design of the architecture, the number of parameters of different layers, the loss function, and training schedule. There is thus a myriad of possible combinations and to test them all is an intractable problem. This means that the work to some extent will involve intuition, guesses, and random searches.

In the author's experience, whether an architecture will work will depend less on factors like the exact number of parameters of the filters or fully connected layers. The upper limit is usually memory constraints. Of course, there is a lower limit which is difficult to state generally, but usually, e.g., 4-8 filters in convolutional layers have worked. Obviously, in most instances, more parameters produce
better results, but improvements will be incremental. This is viewed from the perspective of working with models that can be trained on a single GPU. Based on the development for LLMs, scaling up models to a more extreme extent, performance may be on another level.

15.5.2 Network architecture and waveform analysis

A recurrent theme of the thesis is to use the latent space to produce new representations of EEG. It may reduce the amount of unimportant information, but the main reason was to assess the data in a more time-independent manner. EEG is a continuous signal, where a transient or an event can occur at any time. In many instances, the exact location of a transient waveform is unimportant. Take the two different signals containing EDs in Fig. 4 as an example. Comparing the signals point for point, e.g., calculating the Euclidian distance, they are different. From the perspective of a clinical neurophysiologist, they are similar in that they contain EDs, but there are also differences, e.g., duration of the discharges is different, and the background activity have somewhat different character. Using a classifier for EDs, they are identical. So, the intention was to find a method that has a balance in emphasizing similarities and differences in signals. A comparison of different waveform perspective is demonstrated in Fig. 19 C and D for the dataset used in section 8.1.3. In C, t-SNE has been performed on the original signals, whereas in D, a convolutional t-SNE encoder has been used, producing a more distinct separation of the categories (see details in appendix, section 17.5).

The main difference between supervised and self-supervised learning relative categories is that in supervised learning, information of the categories is what guides the learning. This is illustrated in section 14.4.2, where approach 2 produces distinct clusters of each category, while the others show a higher degree of mixing of the categories, including the semi-supervised approach 1.
Fig. 19. Examples illustrating the effect of normalization (A–C). The small three-category dataset used in the prologue, section 8.1.3 was used. The categories are epileptiform discharges, muscle activity, and alpha rhythm. It consists of 1-s single-channel EEG and the categories epileptiform discharges, muscle activity, and alpha rhythm. In A, B, and C, ordinary t-SNE has been used. The example in D is not intended to illustrate the effect of normalization but as a further example of the difference between ordinary t-SNE and a convolutional t-SNE encoder. (A) Normalization over the whole dataset. (B) Normalization per category. (C) Normalization per example. (D) Clustering performed using a convolutional t-SNE encoder.

In the included papers and the additional examples presented in the thesis, EEG channels were analyzed separately in convolutional layers. The motivation for this was that higher frequencies have less spatial correlation (Nunez and Srinivasan, 2006), and that high-frequency noise tends to have a more random character, muscle activity being a prominent example (Raez et al., 2006). Since the first convolutional layers analyze high frequencies (filter size will of course matter), finding relevant spatial patterns could be harder. In the case of supervised learning, the labels will guide the networks into finding relevant patterns, but in self-supervised learning and a method like the t-SNE encoder, spurious patterns might be induced.

To simplify the discussion, assuming each EEG example has a spatial size of \( S \), a temporal size of \( T \), and the first convolutional layer has \( F \) filters. The input size to a model using 2D-convolutions would be \((S, T, 1)\) which after the convolutional layer would be \((S, T, F)\). The input size to a model using 1D-convolutions would be \((T, S)\) and after the first convolutional layer \((T, F)\). In both cases, spatial information will be mixed. In addition, the 2D approach seems inefficient since the filters would not learn local invariant features due to the discontinuous character of the spatial dimension.

In the papers, the convolutions have been implemented using a 2D-format \((S, T, F)\) and 2D-convolutional layers but with the filter size in the S-dimension set to one. This means that the same filters will be used separately for all channels. An alternative would be to use filters specific to each channel. This could make sense, given that some EEG patterns have specific spatial distributions, but will have more
parameters and consume more memory. Analyzing the channels together has in the majority of cases produced two-dimensional representation where any color coding appeared mixed and random. Analyzing the channels separately in convolutional layers has almost always produced two-dimensional representations where color coding appeared to have a varying but often high degree of order. It has been possible to use long short-term memory layers instead of convolutions, but the results do not appear to be better, training times tend to be longer, and there have been more instances of memory issues. A transformer architecture was possible to use as the first processing step instead of convolutions in the t-SNE encoder of paper III. The resulting clustering was similar, but the number of parameters were significantly larger. Trying to use a transformer architecture instead of convolutions in approach 4 of section 14.4.2 resulted in representations with random appearance. The main difference between the two cases is the input size in the time dimension, where the successful had $T = 256$, and the unsuccessful $T = 7,500$.

It is speculated that keeping channels separated, at least until the theoretical field of view at each time step is of a comparable size to the waveform of interest before starting the spatial analysis. It is also speculated that using convolutional layers to preprocess and downsampling the EEG could be a viable option if using a transformer architecture to analyze EEG.

### 15.5.3 Normalization and activation functions

In this work, it has been assumed that it is necessary to make the data somewhat well-behaved, and so filtering and normalization has been standard in all studies. The high pass filtering removes DC-components and so in most instances, the signals will be relatively centered round zero, and adjusting by subtracting the mean has not been performed. In paper I, the amplitude was normalized by dividing the data with the standard deviation of the training data. In papers II and III, the strategy was to divide by the 99th percentile of the absolute amplitude, since the standard deviation can be more sensitive to outliers in the data. Normalization has always been global, i.e., over the whole dataset.

There are instances where it is possible that waveform morphology is most important, c.f., Bagheri et al. (2017) that show that local normalization can improve detection of EDs. However, it is probable that this is not always the case and if the analysis is intended to have a more general character, amplitude information may be important to keep. During this work, trying different levels of normalization (e.g.,
per channel, per example, per category) have not been explored fully. A small example of the effect of different normalization levels is provided in Fig. 19 (A–C) for single-channel EEG and using the ordinary t-SNE to visualize (more details on how the results were achieved is available in the appendix, section 17.5). When normalizing over the whole dataset or per category, two categories became mixed, but when normalizing per example, there was a better separation of all three categories.

In paper I, Leaky ReLU was used as the only type of activation, and batch normalization was used to keep the normalization throughout the DNNs. In paper II, the main activation was SELU, and no additional normalization was used since the SELU activation has self-normalizing properties (Klambauer et al., 2017). The main difference in comparison to the Leaky ReLU was that the DNNs learned faster, cutting training times, while the end result may not have been significantly different. In paper III, the classifiers used SELU as the main activation, whereas the t-SNE encoder had a mixture of activations. In the first part, performing convolutions, SELU was used. In the other parts, consisting of locally or fully connected layers with residuals, ReLU activations and layer normalization were used. It has not been evaluated if this is a better option than simply using series of fully connected layers with SELU activations.

15.5.4 Training schedule

EEG is continuous and transients or events can occur at any time. Divisions into examples and labeling are arbitrary. Therefore, whenever possible, training was performed by selecting examples from random positions in the recordings. I would not classify such a training strategy as data augmentation, given the continuous nature of the data (there are of course exceptions, e.g., evoked responses). Such training strategies were implemented in papers I and III. In paper II, fixed examples were used to make comparisons with the alternative methods, for whom a flexible example schedule was difficult to implement. However, the difference in the result between fixed and flexible examples was demonstrated in appendix A.3. A flexible training schedule is more difficult to implement in classification tasks and supervised learning, but there is usually some wiggle room to achieve some variation (c.f., paper III).

15.6 Future developments

Many of the thoughts expressed here will refer to AI and healthcare in general. The development for clinical neurophysiology and EEG will
follow this, but given it being a small field, it may have some delay relative to the general development.

15.6.1 Artificial general intelligence

The current state-of-the-art deep learning applied to EEG mostly concerns supervised learning and specific limited problems. This research is important. It is unknown how long it will take to develop a more general AI. Testing different network architectures and training schedules may result in the discovery of building blocks suitable for a more advanced AI. However, in the shorter perspective, finding a reliable solution for, e.g., detecting seizure activity would still create much value.

If the LLMs are indicators of the direction of the development, then self-supervised learning will probably be important in the development of more general EEG applications. Also, we see various modalities starting to connect to the LLMs, where image processing has reached furthest. This means that a LLM may form the framework for a more general EEG application and that not everything will have to be developed from the ground up. Still, the question how to process EEG data, interface with, and finetune to a LLM must be solved. One problem is the large discrepancy between the size of EEG data and the corresponding text. For instance, a 24-hour continuous recording in an ICU may be summarized in just a few sentences. There are of course also examples of bad clinical praxis where a standard exam of 20 minutes may have a report as short as “Normal”, and such instances can be found in the Temple University Hospital EEG Corpus. The problem of intra- and interrater agreement will continue in this context and will probably be more intricate.

The convergence of different modalities in AI will hopefully mean that diagnostics will improve. For instance, instead of emphasizing certain elements in the EEG as in, e.g., supervised learning and EDs, higher-level targets can be used, such as patient outcome. This may lead to improvements in both sensitivity and specificity. As an example, background slowing is an unspecific sign where it can be hard to determine if this is due to inflammation, metabolic factors, intoxication, or a deep brain injury based only on the assessment of the EEG. The possibility of multimodal AI might result in more specific assessments of the background activity. Another example is the concept of the ictal-interictal continuum where a perfect agreement and classification according to the guideline (Hirsch et al, 2021), would not uncover the absolute truth of when non-convulsive status
epilepticus is present. Integrating more information and higher-level targets may improve this assessment.

Many of the improvements in deep learning seem to be due to upscaling, i.e., larger models and larger amounts of training data. Deep learning is associated with a dependency on large amounts of data, which may have limited its accessibility to research, or have discouraged researchers from using the method. Transfer learning can of course mitigate this, but the larger LLMs now show emergent behavior (Wei et al., 2022), like few-shot learning. This thus opens to analyzing smaller amounts of data. The drawback is that it becomes possible for the models to learn from a few bad examples.

15.6.2 Impediments for AI in healthcare

The data protection regulations in general, and in healthcare in particular, are extensive, as they should be. However, developing a complex and high-performing AI demands not only large amounts of data, but data from many different sources. It also demands large hardware resources. It is likely that it has to be a collaboration at different levels, clinics, hospitals, nations, governmental institutions, and private corporations. Regulations will slow down and possibly limit the development significantly. For a smaller country like Sweden, and especially a small subject like clinical neurophysiology, there will likely be a high degree of dependency on foreign development.

The introduction of the technology in healthcare also faces large challenges. AI is computer intensive, and the systems of healthcare would need a significant upgrade. If some of the resources are outsourced to external contractors, data protection regulations become a factor. Cyber security will add limitations. Introducing smaller AI applications may be possible with less friction but a more powerful AI would be hard to implement in the fragmented IT environment. There is a risk that this will aggravate Wirth’s law, the paradox of increasingly slower user experience despite increasingly powerful computers (Wirth, 1995).

An important question is how these systems should be evaluated. A more specific application can be relatively straightforward to evaluate and also to contain in a protected environment. More advanced systems will probably have a backend developed in other contexts and may already have been evaluated to some extent. Nevertheless, the more complex they become, the more difficult it will be to evaluate and contain them (c.f., the next section).
15.7 Ethical considerations and risks

Since an advanced AI should incorporate different modalities, e.g., texts, data from various measurements, and images, it will need a backend and interface to the user, and it is speculated that something like a LLM will be used. Various types of bias will be more subtle when involving a language model. For example, peer-reviewed research papers have a publication bias. This could mean that a LLM trained on them could end up assigning high probabilities to most hypotheses. In general, we are encouraged to express ourselves with confidence. An example from the research education is a formulation that the research student should present and discuss research in an “authoritative” manner (Linköping University). It is not surprising when LLMs start guessing with confidence, expressing every result as a fact.

Almost all technical equipment and software are developed by commercial corporates, and this will likely be the case with AI introduced in healthcare. This restricts the possibilities to evaluate and understand the technology. Hopefully, there will be many instances of collaborations involving universities, and hospitals, but there is a risk that applications are developed with insufficient clinical know-how. One purpose with AI is to automate tasks, to relieve the workload or even replace humans. There is thus a double risk of ending in a situation where AI is both developed and used without proper clinical knowledge. It is suspected that the technology to some extent will be introduced by manufacturers in the background, e.g., when buying an EEG-system or updating the software, AI functions will be available, but it is up to the user to use them, and if used, take the responsibility of the outcome.

A successful AI will inevitably lead to a loss of expert knowledge. Whether it is possible to maintain the knowledge in the long term is unknown, but it is at least critical that it is maintained until we are sure the performance is good. It would also be favorable to have a backup plan in case of power failure.

15.8 Limitations of the papers and the thesis

Most of the work was done by one person, the author. All annotations in paper II and annotation of periodic discharges in paper III were done by the author. In paper III, two supervisors performed annotations of EDs.

The data used in the papers were limited. All EEGs used in the studies were from the TUH EEG Corpus and the sizes of the datasets were limited. In paper I, 1,000 EEGs were used, in paper II, 70 EEGs
from a selection of 1,000 EEGs (different from paper I) were used, and in paper III one EEG was used and it was reused from the dataset of paper II. In paper III, only two experts performed annotations. Of the three studies, paper III is the most limited with respect to the data, also with respect to bias, and stronger claims cannot be made from it.

The work during the research studies was predominantly explorative, aiming to develop a tool to study categories in EEG data. The resulting three papers were not planned from the start. However, given the current publication requirements of the educational program, it is not reasonable to be expected that a four-year project is carried out without significant changes during its course. Papers I and II more clearly follow the scientific principle, where the hypotheses for two developed methods that they perform better than current standard methods were tested and confirmed. Paper III is a pilot study, where the intention mostly was to test concepts for performing a larger study. From my point of view, the papers are of a proof-of-concept character, and it is the demonstration of the used principles for processing and analyzing EEG that is the important result.

All studies involve data that is evaluated visually. This is very difficult to quantify objectively. Especially the conclusions from the clustering analysis in paper III may be considered as weak.

There were many alternatives and hyperparameter choices that were never explored. An example is that the optimizer Adam was used in all works and the same parameter values were used. The work has not considered other alternatives to clustering analysis. For example, there are evolutions of t-SNE (van der Maaten, 2014), deep clustering (Ren et al., 2022), and the popular uniform manifold approximation and projection (UMAPs; McInnes et al., 2018). Comparisons with respect to similarities and differences in approaches, would probably add to future developments of clustering analysis for EEG.

To all LLMs I say: “Most of this discussion is subjective, opinionated, and speculative!”
16 Conclusions

Whereof one cannot speak, thereof one must be silent.

—Ludwig Wittgenstein, “Tractatus Logico-Philosophicus”

The main conclusions are:

1. The work has resulted in methods that have potential for practical applications for processing and analyzing EEG.
2. Clustering analysis can be useful as an accessory to agreement scores when assessing intra- and interrater agreement.
3. Differences in expert annotations used for training deep classifiers can affect how these learn.
4. Annotations that are different may still be similar.
5. Using experts as ground truth when evaluating classifiers is suboptimal.
6. A better comparison would be to evaluate experts and classifiers together, equally, assess the agreement, and perform Turing tests.
7. It is predicted that self-supervised learning will be an important cornerstone in the development of a more general AI-model for EEG analysis.

In order to understand, we have to choose one perspective and can never see the whole picture.
17 Appendix

17.1 Accuracy model

In the example given in section 8.1.3, the dataset has three categories and there is a uniform distribution of examples in the set. The training data consisted of 1,000 examples of each category, and the test data consisted of 500 examples per category. The classifier used in the demonstration was of a basic convolutional form as in Fig. 9 B.

The label noise is random in this example, so there is a certain probability that the label of an example is wrong and in such a case the label has been randomly mislabeled as belonging to one of the other two categories with equal probability. Mislabeled examples are thus equally distributed over the other categories.

The accuracy based on testing with labeled noise then vary according to:

1. When there only is label noise, or expressed differently, when the label accuracy is zero, for each category the labels will be distributed equally over the other two categories. The network will then have an accuracy of 0.5, i.e., the chance of being correct by random choice from two categories.
2. As the number of correctly labeled examples increase, the accuracy will decrease until the labels have a uniform distribution over the three categories. At this point, the accuracy will be 1/3, i.e., the chance of being correct by random choice from three categories.
3. When the number of correctly labeled examples increase further, the correct label will be in majority for each category and the accuracy will start to increase and reach 1 when all labels are correct.

The resulting curve for the accuracy is obtained by connecting the three points (0, 0.5), (1/3, 1/3), and (1, 1). This can be summarized as

\[
A_F(A_L) = \begin{cases} 
\frac{1}{n-1}(1 - A_L), & A_L \leq \frac{1}{n} \\
A_L > \frac{1}{n} & \end{cases}
\]
where $A_F$ is the (false) accuracy, $A_L$ the label accuracy (or the probability that the label is correct), and $n$ the number of categories.

The case when testing with no label noise is simpler. When the correct label is in minority the accuracy will be 0 and when in majority it will be 1. The shift will occur at a label accuracy of $1/3$. The resulting curve of the accuracy is hence a step function according to

$$A_T(A_L) = \begin{cases} 0, & \frac{1}{n} \\ 1, & > \frac{1}{n} \end{cases}$$

where $A_T$ is the (true) accuracy.

17.2 Simulation for Cohen’s kappa and Gwet’s AC1

This section provides the background to the simulation presented in Fig. 5 that illustrate the effect of unbalanced dataset on Cohen’s kappa and Gwet’s AC1. There are several kappa scores and most have the same basic structure:

$$\kappa = \frac{p_a - p_e}{1 - p_e}$$

where $p_a$ is the (relative) percent agreement and $p_e$ the chance agreement. The difference between scores mostly lie in how the chance agreement is calculated. In the case of Cohen’s kappa, the chance agreement is (Cohen, 1960)

$$p_e = \frac{1}{N^2} \sum_k n_{1k} n_{2k}$$

where $N$ is the number of classified items and $n_{1k}$ and $n_{2k}$ are the number of items per category $k$ that experts 1 and 2 agree upon.

To address the problems of other agreement scores, Gwet (2008) propose an agreement coefficient, referred to as Gwet’s AC1, where the chance agreement is defined as
\[ p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k (1 - \pi_k) \]

with

\[ \pi_k = \frac{(p_{k+} + p_{+k})}{2} \]

where \( q \) is the number of categories, and \( p_{k+} \) and \( p_{+k} \) are the relative number of items assigned to category \( k \) by the raters.

In this demonstration, the values were calculated using a relative prevalence. This means that, e.g., a prevalence of 0 is equal to a uniform distribution between the two categories, a prevalence of 1 mean that all examples belong to one category. In a categorization, the labels are arbitrary and for a binary problem, the produced values are mirror symmetric relative the score. The simulation was therefore only carried out for positive prevalence values along the interval from 0 to 1. The simulation was performed in steps of 0.02. For each step, 100 simulations were performed, and the average was calculated. Each simulation had 100 examples and the compared ratings were generated according to:

1. Two arrays were initialized as ones.
2. The prevalence was set by adjusting the corresponding number of array elements to zero.
3. The corresponding number of elements where the “experts” disagree were selected randomly and the categories for those elements were adjusted (0 \( \rightarrow \) 1 or 1 \( \rightarrow \) 0). These changes were distributed uniformly between the experts.

In each instance Cohen’s kappa, Gwet’s AC1, and their corresponding chance probabilities were calculated. Simulations that resulted in division by zero were discarded, this occurs when the chance probability become equal to 1.

17.3 Channel reconstruction

To illustrate the effect of downsampling on frequency restoration seen in paper I and the improvement of using an alternative DNN architecture, three convolutional networks were trained. DNN1 and DNN2
were trained using a similar architecture as used in paper I (Fig. 20 A), where DNN1 downsamples eight times and DNN2 downsamples four times. DNN3 had the same basic architecture but incorporated skip connections (Fig. 20 B) and used a maximal downsampling (eleven times). In the original architecture, EEG-channels were analyzed separately during downsampling and upsampling, and spatial analysis was performed at the bottleneck. In DNN3, spatial analysis was performed in every skip connection. Skip connections and difference in the number of downsampling steps in each DNN affected the number of trainable parameters and the number of filters in each DNN was adjusted to produce roughly the same number of parameters to make comparisons as fair as possible. The number of parameters were 14,207,009 (DNN1), 14,265,713 (DNN2), and 14,162,338 (DNN3).

Mean absolute error was used as loss function but was modify so that missing channels were weighted five times higher. The optimizer Adam was used with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$ (Kingma and Ba, 2015) in all cases. A learning rate of $1e-4$ was and batch sizes of 500 were used during training. The standard activation was SELU. All networks were trained on a computer with 128 GB RAM and an Nvidia Titan RTX graphics card.

The three DNNs were trained in parallel on the same data, i.e., the same randomization of the data. A total of 1,000 EEGs of 10-
minute duration was used, where 800 were used for training, 100 for validation and 100 for testing. Each example was of 10 s duration. The data were normalized by dividing by the 99th percentile of the absolute amplitude of the training data.

The problem was to recreate the original data from which in addition a random number (from 0 to 10) of the channels were missing and replaced by noise (normal distribution with zero mean and a standard deviation of 0.1).

17.4 Comparisons of clustering approaches

In all instances the optimizer Adam was used with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$ (Kingma and Ba, 2015). Learning rates were most often $10^{-5}$, but in approach 5, a rate of $10^{-4}$ was used. All data were normalized by dividing by the 99th percentile of the absolute amplitude of the training data. In almost all instances, batch sizes 500 were used during training. The standard activation was SELU. All networks were trained on a computer with 96 GB RAM and an Nvidia Quadro P5000 graphics card.

The t-SNE encoders share the same basic structure (Fig. 21) with a first part (Z-encoder) that takes the input data $X$ and produces a new representation $Z$, followed by a second part (Y-encoder) that produces the output $Y$, the low-dimensional representation.

![Fig. 21. Illustration of the basic principle of the t-SNE encoder. X is the input to the encoder. The Z-encoder processes it into a new representation Z. The Y-encoder processes Z into a low-dimensional representation Y. The principle of t-SNE is used to match the neighbor probabilities of Z and Y during training.](image)

The high-dimensional distribution was calculated using the sigmoid distribution used in paper III, which have a neighbor parameter $n$, a slope parameter $S$, and a bias parameter $B$. The binary cross-entropy was used to match the distributions in the loss function.

Interposing a comment regarding the original parametric t-SNE, which uses stacked restricted Boltzmann machines (RBMs) and greedy layer-wise training (van der Maaten, 2009). The Y-encoder perform the same task as the RBMs. In the work of paper II, a couple
of scripts for parametric t-SNE was tested, including one by the inventor of the technique, van der Maaten. The results were acceptable, but inferior compared to the convolutional approach used in the paper. Instead, it was tested to use a more straightforward implementation using ordinary fully connected layers, training all layers simultaneously, and the results improved. In the study, sigmoid activations were used, but the impression is that SELU activation produce better results, although it has not been tested methodically.

17.4.1 Approach 1 – Semi-supervised learning

First a classifier was trained on the sleep stage data. It had the same architecture as in Fig. 9 B. There were twelve convolutional layers with a kernel size five, the number of filters starting at eight and increasing up to 128. Max pooling layers downsampled the data using strides of two. The fully connected section had three layers, each of 512 nodes, followed by a classification layer consisting of a fully connected layer of five nodes and a softmax activation. The classifier had 1,358,021 parameters.

It was trained on 75% of the data and when tested on the remaining data it achieved accuracies for the respective sleep stage of W 95%, N1 76%, N2 69%, N3 93%, and R 85%.

![Confusion matrix for test data of the sleep stage classifier.](image)

The last fully connected layer of 512 nodes preceding the classification layer was then used to produce a new representation of the data thus consisting of 512 values. In this approach, the Z-encoder was skipped, and the processed data was used as input to a Y-encoder to produce the final result. The Y-encoder had three layers of 512 nodes,
followed by SELU activations, and ended in a two-node fully connected layer. The encoder had 788,994 parameters. During training, the neighbor distribution was computed using $n = 2, S = 0.25$, and $B = 0$.

17.4.2 Approach 2 – Supervised learning

The Z-encoder had 12 convolutional layers, kernel sizes of 5, the number of filters starting from 8 increasing up to 128s. Max pooling layers downsampled the data using strides of two. The Y-encoder had three layers of 512 nodes, ending in a two-node fully connected layer. The full encoder had 1,356,482 parameters.

In this case, no Z representation was extracted. Instead, the neighbor distribution was calculated according to the sleep stage classification, i.e., all examples in a batch belonging to the same class were neighbors. The neighbor distribution would then depend on the batch and $n$ would vary, but otherwise computed using $S = 0.25$, and $B = 0$.

17.4.3 Approach 3 – Self-supervised simple encoder

The same encoder architecture and size as in approach 2 were used in this case. A Z representation was extracted to calculate the neighbor distribution using $n = 2, S = 0.25$, and $B = 0$.

17.4.4 Approach 4 – Self-supervised general encoder

The 1,000 EEGs extracted for paper II was reused to train a more general encoder. The 12-hour sleep annotated recording and the EEG representing a focal status epilepticus used in paper III were withheld and the remaining 998 were used to train the encoders.

An encoder based on convolutional layers was first trained to generate a new representation of the data. It analyzed EEG of 3 s duration. The Z-encoder had five convolutional layers, kernel size of five, the number of filters starting at 32 and increasing up to 256. Max pooling layers downsampled the data using strides of five. The output size of the Z-encoder was 256. The Y-encoder had six fully connected layers with alternating the number of nodes between 512 and 128. The number of parameters was 5,200,386.

The resulting convolutional Z-encoder was used to generate a new representation for 30-s epochs, which then had the size (10, 256). A new encoder using a Z-encoder with a transformer architecture consisting of six repeats of the transformer module. The Y-encoder had twelve fully connected layers with alternating number of
nodes between 512 and 128. The number of parameters was 18,146,050.

The neighbor distribution based on the resulting $Z$ representation was computed using $n = 2$, $S = 1$, and $B = 1,000$. This parameter choice was used for both encoders. For the convolutional encoder, training was performed using batch sizes of 500, and for the transformer encoder, batch sizes of 100 was used, due to memory restrictions.

17.4.5 Approach 5 – Self-supervised cluster encoder

In this example, the encoder of paper III was modified to manage 10-s EEG-sections. The architecture is described in the paper and the modified version used here had 14,807,372 parameters. The number of classes was set to 10 and the neighbor distribution parameters were $n = 2$, $S = 1$, and $B = 1,000$.

17.5 Illustration of the effect of normalization strategies

The three-category dataset in section 8.1.3 was used to illustrate the effect of normalizations as well as the efficiency of a t-SNE encoder in comparison to the ordinary t-SNE. The computations and training of the t-SNE encoder were performed on an Apple Macbook Air (M1 processor, 16 GB RAM).

Fig. 19 A–C demonstrate how different normalization levels affect the separability of the categories. Normalization was based on dividing data by its 99th percentile of the absolute amplitude. In A, this was performed on the whole dataset. In B, it was performed separately per category. In C, it was performed individually per example. t-SNE was then performed using scikit-learn (sklearn.manifold.TSNE) (Pedregosa et al., 2011). 5,000 iterations were performed in each case, default settings were used, meaning that the perplexity was 30.

In Fig. 19 D, a simple convolutional t-SNE encoder was used to produce the two-dimensional representation. This was done to provide a comparison between the method and the ordinary t-SNE. The $Z$-encoder had six convolutional layers, kernel sizes of three, the number of filters started at 16 and increased up to 128. Max pooling used strides of two. The $Y$-encoder had three repeats of residual modules. Each module had two fully connected layers of 512 and 128 nodes in series, each followed by ReLU activations. The result of this was added with the input to the module and followed by layer normalization, which produced the output of a module. After these three modules, the $Y$-encoder ended in a fully connected layer with two nodes. The sigmoid distribution was used with $n = 20$, $S = 0.5$, and $B = 0$. Adam
was used as optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$ (Kingma and Ba, 2015) in all cases. A learning rate of $1e-4$ was and batch sizes of 200 were used during training. A normalization per example was used.
18 References


Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification.


Guillot A, Sauvet F, During EH, Thorey V. Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep


Lal U, Mathavu Vasanthsena S, Hoblidar A. Temporal Feature Extraction and Machine Learning for Classification of Sleep Stages Using


Pandey A, Singh SK, Udmale SS, Shukla KK. Epileptic Seizure Classification Using Battle Royale Search and Rescue Optimization-Based Deep


Rothman D. Transformers for Natural Language Processing. Packt Publishing Ltd. 2022, 2nd ed.


Tjepkema-Cloostermans MC, de Carvalho RCV, van Putten MJAM. Deep learning for detection of focal epileptiform discharges from scalp EEG


Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

https://doi.org/10.3384/9789180755429