1 GENERALIZATION TO OTHER TASKS

TODO: textbox

After our initial work designing and evaluating convolutional neural networks for movement decoding from EEG, we evaluated the resulting networks on a wide variety of other EEG decoding tasks found that they generalize well to a large number of settings such as error-related decoding, online BCI control or auditory evoked potentials and also work on intracranial EEG.

Text and content condensed from a number of publications, namely Schirrmeister et al. [5], Völker et al. [7], Burget et al. [3], Volker et al. [6], Behncke et al. [2], Wang et al. [8] and Heilmeyer et al. [4]. In all of these works except Schirrmeister et al. [5], I was not the main contributor, I assisted in adapting the code and training for the various settings and helped in the writing process.

1.1 DECODING DIFFERENT MENTAL IMAGERIES

FBCSP	DEEP CONVNET	SHALLOW CONVNET
71.2	+1.0	-3.5

Table 1.1: **Accuracies on the Mixed-Imagery dataset.** ConvNet accuracies show the difference to the FBCSP accuracy. Results from Schirrmeister et al. [5].

The Mixed Imagery Dataset (MID) was obtained from 4 healthy subjects (3 female, all right-handed, age 26.75±5.9 (mean±std)) with a varying number of trials (S1: 675, S2: 2172, S3: 698, S4: 464) of imagined movements (right hand and feet), mental rotation and mental word generation. All details were the same as for the High Gamma Dataset, except: a 64-electrode subset of electrodes was used for recording, recordings were not performed in the electromagnetically shielded cabin, thus possibly better approximating conditions of real-world BCI usage, and trials varied in duration between 1 to 7 seconds. The dataset was analyzed by cutting out time windows of 2 seconds with 1.5 second overlap from all trials longer than 2 seconds (S1: 6074 windows, S2: 21339, S3: 6197, S4: 4220), and both methods were evaluated using the accuracy of the predictions for all the 2-second windows for the last two runs of roughly 130 trials (S1: 129, S2: 160, S3: 124, S4: 123).

For the mixed imagery dataset, we find the deep ConvNet to perform slightly better and the shallow ConvNet to perform slightly worse than the FBCSP algorithm, as can be seen in Table 1.1.

1.2 DECODING ERROR-RELATED SIGNALS

1.2.1 *Decoding Observation of Robots Making Errors*

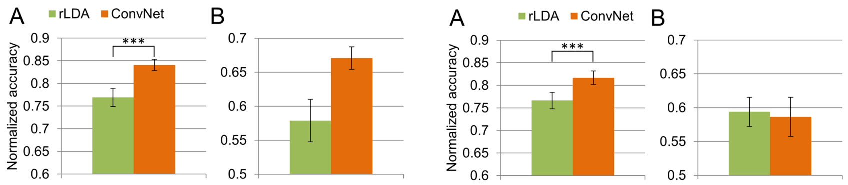
ROBOT TASK	TIME INTERVAL	DEEP CONVNET	RLDA	FBCSP
Pouring Liquid	2-5s	78.2 ± 8.4	67.5 ± 8.5	60.1 ± 3.7
Pouring Liquid	3.3-7.5s	71.9 ± 7.6	63.0 ± 9.3	66.5 ± 5.7
Lifting Ball	4.8-6.3s	59.6 ± 6.4	58.1 ± 6.6	52.4 ± 2.8
Lifting Ball	4-7s	64.6 ± 6.1	58.5 ± 8.2	53.1 ± 2.5

Table 1.2: **Accuracies for robot error observation.** Task was to decode whether a person watches a successful or unsuccessful robot-liquid pouring or ball-lifting. Results from Behncke et al. [1].

In this study, we aimed to classify whether a person had watched a video of a successful or an unsuccessful attempt of a robot performing one of two tasks (lifting a ball or pouring liquid) based on EEG recorded during the video observation. We compared the performance of our deep ConvNet to that of regularized linear discriminant analysis (rLDA) and FBCSP on this task. Our results, presented in Table 1.2, demonstrate that the deep ConvNet outperformed the other methods for both tasks and both decoding intervals.

1.2.2 *Decoding of Eriksen Flanker Task Errors and Errors during Online GUI Control*

In two additional error-related decoding experiments, we evaluated an Eriksen flanker task and errors during an the online control of a graphical user interface through a brain-computer-interface. In the Eriksen flanker task, the subjects were asked to press the left or right button on a gamepad depending on whether an 'L' or an 'R' was the middle character of a 5-letter string displayed on the screen. For the online graphical user interface (GUI) control, the subjects were given an aim to reach using the GUI, also see **??**. They had to think of one of the classes of the aforementioned Mixed Imagery Dataset to choose one of four possible GUI actions. The correct GUI action was always determined by the specificed aim given to the subject, hence an erroneous action could be detected. The decoding task in this paper was to distinguish whether the BCI-selected action was correct or erroneous. Results in Figure 1.1a and Figure 1.1b show that deep



(a) **Comparison of within-subject decoding by rLDA and deep ConvNets.** Error bars show the SEM. A) Eriksen flanker task (mean of 31 subjects), last 20% of subject data as test set. Deep ConvNets were 7.12% better than rLDA, pval = 6.24 *10-20 (paired t-test). B) Online GUI control (mean of 4 subjects), last session of each subject as test data. Figure from [7].

(b) **Mean normalized decoding accuracy on unknown subjects.** Error bars show the SEM. A) Eriksen flanker task, trained on 30 subjects, tested on 1 subject. Deep ConvNets were 5.05% better than rLDA, p = 3.16 *10-4 (paired t-test). B) Online GUI control. Trained on 3 subjects, tested on the respective remaining subject. Figure from cite:t‘volker2018deep‘.

Figure 1.1: **Error decoding accuracy on Eriksen flanker task and online GUI control.**

ConvNets outperform rLDA in all settings except cross-subject error-decoding for online GUI control, where the low number of subjects (4) may prevent the ConvNets to learn enough to outperform rLDA.

1.3 PROOF-OF-CONCEPT ASSISTIVE SYSTEM

We also evaluated the use of our deep ConvNet as part of an assistive robot system where the brain-computer interface was sending high-level commands to a robotic arm. In this proof-of-concept system, the robotic arm could be instructed by the user via the BCI to fetch a cup and directly move the cup to the persons mouth to drink from it. An overview can be seen in Figure 1.2. Results from Table 1.4 show that 3 out of 4 subjects had a command accuracy of more than 75% and were able to reach the target using less than twice the steps of the minimal path through the GUI (path optimality > 50%).

1.4 INTRACRANIAL EEG DECODING

1.4.1 *Intracranial EEG Decoding of Eriksen Flanker Task*

We further evaluated whether the same networks developed for non-invasive EEG decoding can successfully learn to decode intracranial EEG. Therefore, in one work we used the same Eriksen flanker task as described in Section 1.2.2, but recorded intracranial EEG from 23 patients who had pharmacoresistant epilepsy [6]. Results for single-channel decoding **??** show the deep and shallow ConvNet to clearly outperform rLDA (59.3%/58.4% vs. 53.8%) , whereas the residual

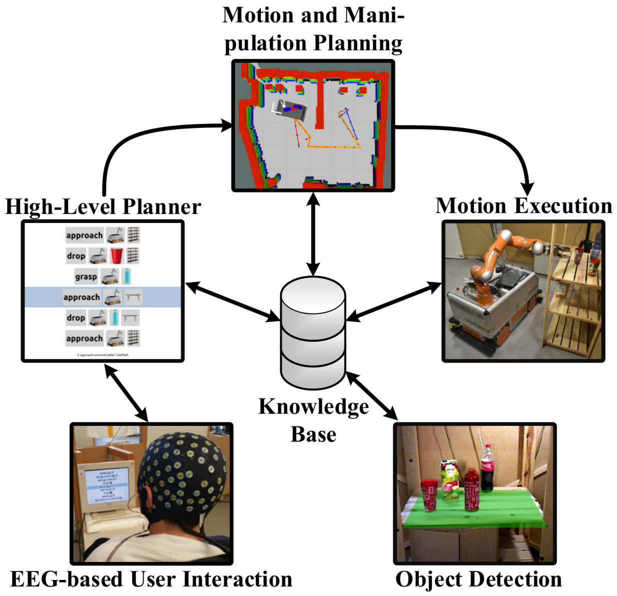


Figure 1.2: **Overview of the proof-of-concept assistive system. The system uses the deep ConvNet in the BCI component.** Robotic arm could be given high-level commands via the BCI, high-level commands were extracted from a knowledge base. The commands were then autonomously planned and executed by the robotic arm. Figure from Burget et al. [3]

ConvNet has low accuracy (52.5%). In contrast, results for all-channel decoding Figure 1.3 show the residual ConvNet to perform well with the residual ConvNet and the deep ConvNet outperforming the shallow ConvNet (72.1% and 73.7% vs. 60.3% class-normalized accuracies (average over per-class accuracies)).

1.4.2 *Transfer Learning for Intracranial Error Decoding*

We further tested the potential of ConvNets to transfer knowledge learned from decoding intracranial signals in error-decoding paradigm to decoding signals in another a different error-decoding paradigm [2]. The two error-decoding paradigms were the aforementioned Eriksen flanker task (EFT) and a car driving task (CDT), where subjects had to use a steering wheel to steer a car in a computer game and avoid hitting obstacles, where hitting an obstacle was considered an error event (see Figure 1.4). Results in Figure 1.5 show that pretraining on CDT helps EFT decoding when few EDT data is available.

1.4.3 *Microelectrocorticography Decoding of Auditory Evoked Responses in Sheep*

In this study, we evaluated the ConvNets for decoding auditory evoked responses played to a sheep that was chronically implanted with a

SUBJECT	RUNS	ACCURACY [%]	TIME [S]	STEPS	PATH OPTIMALITY [%]	TIME / STEP [S]
S1	18	84.1 ± 6.1	125 ± 84	13.0 ± 7.8	70.1 ± 22.3	9 ± 2
S2	14	76.8 ± 14.1	150 ± 32	10.1 ± 2.8	91.3 ± 12.0	9 ± 3
S3	17	82.0 ± 7.4	200 ± 159	17.6 ± 11.4	65.7 ± 28.9	11 ± 4
S4	3	63.8 ± 15.6	176 ± 102	26.3 ± 11.2	34.5 ± 1.2	6 ± 2
Average	13	76.7 ± 9.1	148 ± 50	16.7 ± 7.1	65.4 ± 23.4	9 ± 2

Table 1.3: **Results for BCI control of the GUI.** Accuracy is fraction of correct commands, time is time per command, steps is steps needed to reach the aim, path optimality is ratio of miniminally needed nubmer of steps to actually used number of steps when every step is optimal, and time/step is time per step. Results from Burget et al. [3].

*μ*ECoG-based neural interfacing device [8].

3-seconds-long sounds were presented to the sheep and two decoding tasks were defined from those 3 seconds as well as the second immediately before and after the playing of the sound. The first decoding task was to distinguish the 3 seconds when the sound was playing from the second immediately before and the second immediately after the sound. The second task was distinguishing the first, second and third second of the playing of the sound to discriminate early, intermediate and late auditory evoked response (see Figure 1.6). Results in Figure 1.7 show that the deep ConvNet can perform as good as FBSCP and rLDA, and perform well on both tasks, whereas rLDA performs competitively only on the first and FBSCP only on the second task.

CLASSIFIER	BALANCED ACCURACY	ACCURACY CORRECT CLASS	ACCURACY ERROR CLASS
Deep4Net	59.28 ± 0.50	69.37 ± 0.44	49.19 ± 0.56
ShallowNet	58.42 ± 0.32	74.83 ± 0.25	42.01 ± 0.40
EEGNet	57.73 ± 0.52	57.78 ± 0.48	57.68 ± 0.56
rLDA	53.76 ± 0.32	76.12 ± 0.26	31.40 ± 0.38
ResNet	52.45 ± 0.21	95.47 ± 0.14	09.43 ± 0.28

Table 1.4: **Results for single-channel intracranial decoding of errors during an Eriksen flanker task.** Balanced Accuracy is the mean of the accuracies for correct class ground truth labels and error class ground truth labels. Results from **intracranial-error-results-table**.

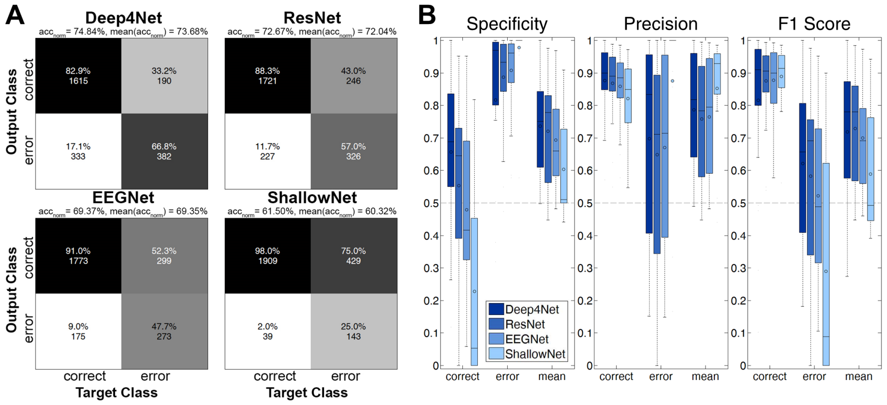


Figure 1.3: **Results for all-channel intracranial decoding of errors during an Eriksen flanker task.** Here, the classifiers were trained on all available channels per patient. A) Confusion matrices of the four models used for decoding. The matrices display the sum of all trials over the 24 recordings. On top of the matrices, the class-normalized accuracy (average over per-class accuracies) over all trials, i.e., acc_{norm}, and the mean of the single recordings' normalized accuracy, i.e., $\text{mean}(\text{acc}_{\text{norm}})$ is displayed; please note that these two measures differ slightly, as the patients had a varying number of total trials and trials per class. B) Box plots for specificity, precision and F1 score. The box represents the interquartile range (IQR) of the data, the circle within the mean, the horizontal line depicts the median. The lower whiskers include all data points that have the minimal value of 25^{th}percentile $- 1.5 \cdot \text{IQR}$, the upper whiskers include all points that are maximally 75^{th}percentile $+ 1.5 \cdot \text{IQR}$. Figure from Volker et al. [6].

1.5 EVALUATION ON LARGE-SCALE TASK-DIVERSE DATASET

"'

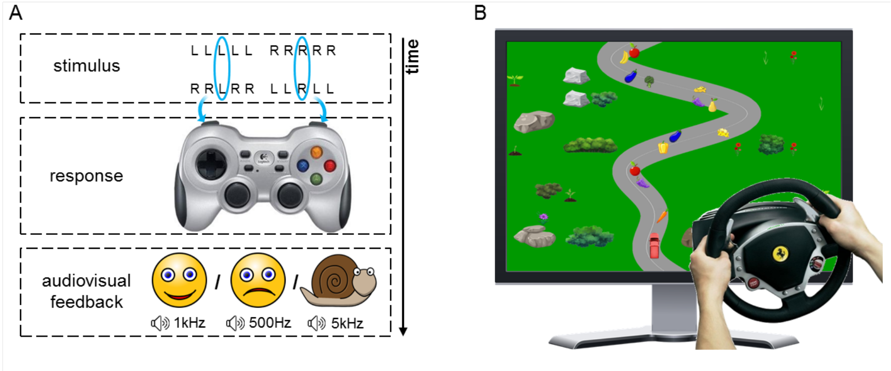


Figure 1.4: **Sketch of the Eriksen flanker task (A) and screenshot of the car driving task (B).** Figure from Behncke et al. [2].

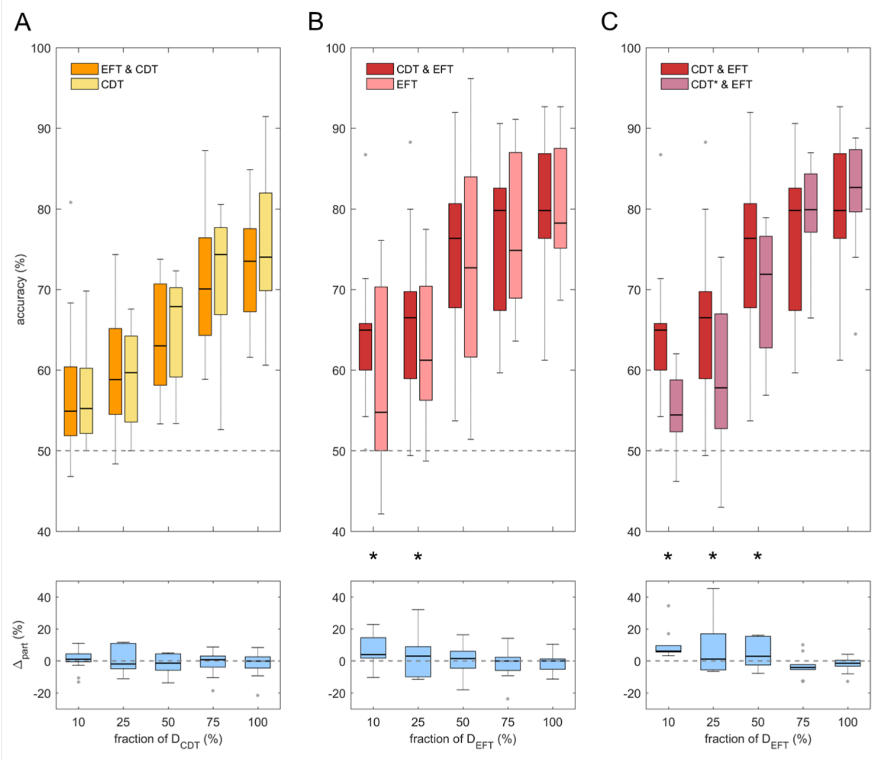


Figure 1.5: **Results for transfer learning on the Eriksen flanker task (EFT) and the car driving task (CDT).** All results are computed for a varying fraction of available data for the target decoding task (bottom row). **A** compares CDT accuracies after training only on CDT or pretraining on EFT and finetuning on CDT. **B** compares EFT accuracies after only training on EFT or after pretraining on CDT and finetuning on EFT. As a sanity check for the results in **B**, **C** compares EFT accuracies when pretraining on original CDT data and finetuning on EFT to pretraining on CDT data with shuffled labels (CDT*) and finetuning on EFT. Results show that pretraining on CDT helps EFT decoding when little EFT data is available. Figure from Behncke et al. [2].

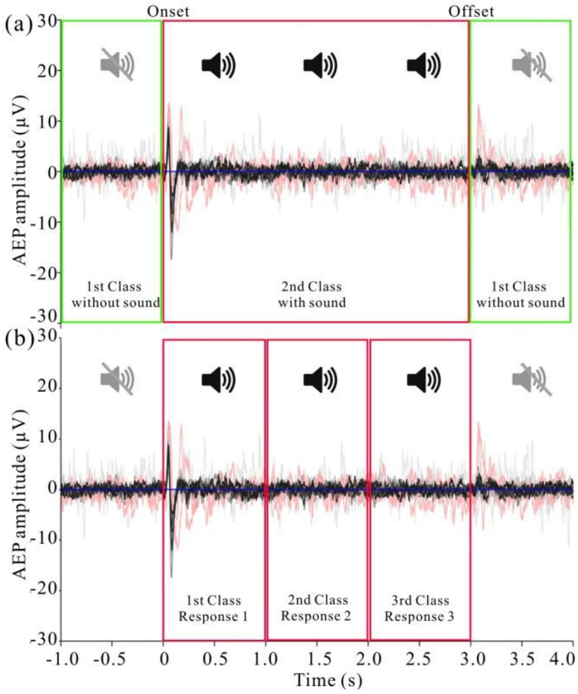


Figure 1.6: **Overview over decoding tasks for auditory evoked responses in a sheep.** First task (top) was to distingish 3 seconds when the sound was playing from the second before and the second after. Second task (bottom) was to distinguish the first, second and third second during theplaying of the sound. Signals are averaged responses from one electrode during different days, with black and grey being signals while the sheep was awake and red ones while the sheep was under general anesthesia. Figure from Wang et al. [8].

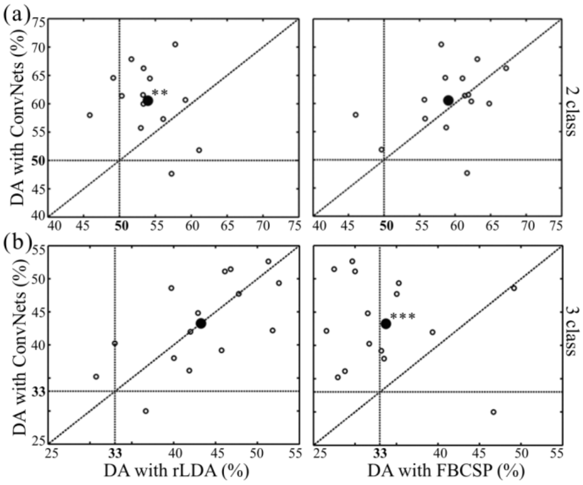


Figure 1.7: **Results of decoding auditory evoked responses from sheep.** rlDA, FBSCP and the deep ConvNet were the decoding models. Open circles represent accuracies for individual experiment days and closed circles represent the average over these accuracies. Figure from Wang et al. [8].

NAME (ACRONYM)	#CLASSES	TASK TYPE	#SUBJECTS	TRIALS PER SUBJECT	CLASS BALANCE
High-Gamma Dataset (Motor)	4	Motor task	20	1000	balanced
KUKA Pouring Observation (KPO)	2	Error observation	5	720-800	balanced
Robot-Grasping Observation (RGO)	2	Error observation	12	720-800	balanced
Error-Related Negativity (ERN)	2	Eriksen flanker task	31	1000	1/2 up to 1/15
Semantic Categories	3	Speech imagery	16	750	balanced
Real vs. Pseudo Words	2	Speech imagery	16	1000	3/1

Table 1.5: **Datasets for the large-scale evaluation framework.** Used in Heilmeyer et al. [4].