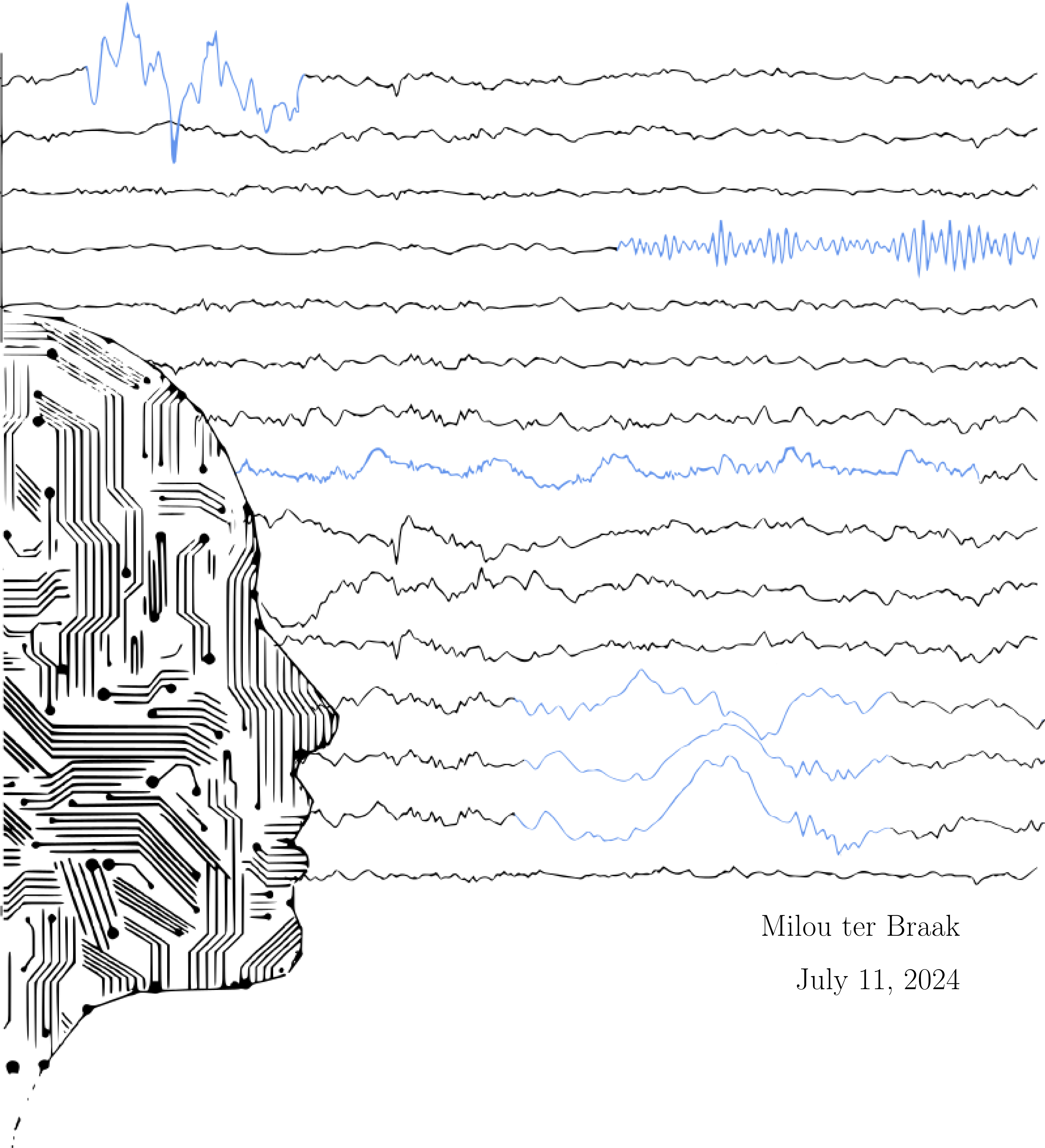
Transfer learning for multi-class artefact classification in dry EEG using convolutional neural networks



Milou ter Braak

July 11, 2024

Transfer learning for multi-class artefact classification in dry EEG using convolutional neural networks

by

Milou ter Braak

Student number: 4560132

July 11, 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)
Dept. of Biomechanical Engineering, TU Delft

Project duration:	January 29, 2024 – July 25, 2024
Supervisor(s):	Dr. Maritta van Stigt, medical/daily supervisor Eva Groenendijk MSc, daily supervisor Dr. ir. Mario Negrello, technical supervisor
Thesis committee:	Dr. ir. Christos Strydis, TU Delft, Erasmus MC (chair) Dr. ir. Mario Negrello, TU Delft, Erasmus MC Dr. Maritta van Stigt, Amsterdam UMC

Abstract

Background Electroencephalography (EEG) using a dry electrode cap is currently being investigated as a pre-hospital stroke triage instrument. Developing an algorithm for automatic interpretation of the EEG signals is challenging, considering the amount of artefacts often in the signal. Ideally, an algorithm should be capable of distinguishing between different artefact types to determine the appropriate action: whether to correct them, reject them, or consider their potential predictive value. Neural networks have demonstrated their value for these types of classifications in wet EEG data. However, this approach requires enormous amount of data and dry EEG data is sparse.

Objective This study aims to develop a multi-class artefact classification model for dry EEG data using transfer learning.

Methods First, a convolutional neural network (CNN) for multi-class (*clean, eye movement, muscle activity* and *electrode artefact*) classification was developed. Wet electrode EEG recordings from a publicly available dataset were used, containing data of 213 patients (Part I). Second, this model was implemented and transfer learned for multi-class (*clean, pulse artefact, muscle activity* and *artefact*) classification of dry EEG recordings using data of 13 subjects (Part II). The models were trained using annotated multi-channel input. Model performances were evaluated on unseen test data using accuracy, area under the receiver operating characteristic curve, F1-score, precision, and recall.

Results The pre-trained multi-class model achieved an overall accuracy of 74.8%. The fine-tuned model was able to correctly differentiate between the classes with an accuracy of 71.2%, with the best performance for the classes *muscle activity* (AUC 0.92, F1-score 0.80) and *artefact* (AUC 0.94, F1-score 0.80).

Conclusion Transfer learning enabled the development of a good performing multi-class artefact classification model specifically tailored for dry EEG data, even though available data was limited. The developed model could assist in assessing appropriate action for different artefact types in dry EEG data interpretation algorithms.

Contents

1 Introduction

Acute ischemic stroke is a leading cause of death and disability, with in 2020 an incidence of approximately 7.6 million worldwide [1]. A contribution of 10-20% is assigned to large vessel occlusion (LVO) stroke, an proximal obstruction of large, cerebral arteries [2]. Acute LVO strokes are associated with a more than twofold increased risk of death and permanent disability compared to non-LVO ischemic strokes. This stroke type contributes to 95% of post-ischemic stroke mortality [2]. Standard treatment for LVO stroke is intravenous thrombolysis (IVT) and endovascular thrombectomy (EVT) [3]. Immediate treatment is of the utmost importance to improve patient outcome since the efficacy of EVT is highly time-dependent [4]. The largest postponement of treatment in acute stroke care is attributed to pre-hospital delays [2].

In the Netherlands, triage is done according to a drip and ship model. Patients with a suspected stroke are brought to the nearest hospital by the ambulance for diagnostics. Following confirmation of LVO stroke, patients may require transfer to a comprehensive stroke centre capable of performing EVT. A study showed transfer is needed in 54% of the patients [5]. As a consequence, time from onset to treatment was delayed by 40 minutes which was associated with a worse functional outcome [5].

Electroencephalography (EEG) has shown its capabilities of detecting LVO stroke through various in-hospital studies [6, 7, 8]. These studies overcome the limitation of traditionally wet electrode EEG, which require extensive preparation and individual placement on the skin, making them impractical for a pre-hospital setting. Sergot et al. implemented a portable wet electrode EEG combined with somatosensory-evoked potentials device which made use of a 13-electrode cap [6]. Electro conductive gel only needed to be applied in the openings during application. Their device achieved a sensitivity and specificity of 80% for predicting LVO strokes, outperforming traditional clinical scales. Although the wet electrode cap had a short application time of 4.6 minutes, implementation of dry electrodes can decrease application time even more. Utilising dry electrodes, a study has identified the theta/alpha ratio, representing distinct frequency characteristics of the EEG, as the most effective EEG feature for LVO prediction, achieving an area under the receiver operating characteristic curve (AUC) of 0.80 [7]. Furthermore, Erani et al. demonstrated moderate accuracy (ACC) in LVO detection using ipsilesional relative theta and alpha power, with an AUC of 0.69 [8]. These results highlight the diagnostic potential of dry EEG electrodes, which are not only quick to apply but also user-friendly and cost-effective, making them well-suited for a pre-hospital setting [9].

In an ideal scenario, automatic interpretation of an EEG recording would indicate the probability of LVO stroke. A current challenge of automatic interpretation algorithms is the presence of artefacts in the signal. Artefacts could have a masking negative or positive effect on the performance of algorithms, not representing the EEG data underneath it [10]. Especially, a pre-hospital setting in combination with the dry EEG electrodes has shown difficulties in obtaining adequate data quality, necessitating exclusion of 32% of the data from analysis in a previous study [7]. A binary (*clean vs artefact*) convolutional neural network (CNN) algorithm was used to detected and rejected artefacts. However, ideally, an algorithm should be capable of distinguishing between different types of artefacts to determine the appropriate action: whether to correct them, reject them, or consider their potential predictive value.

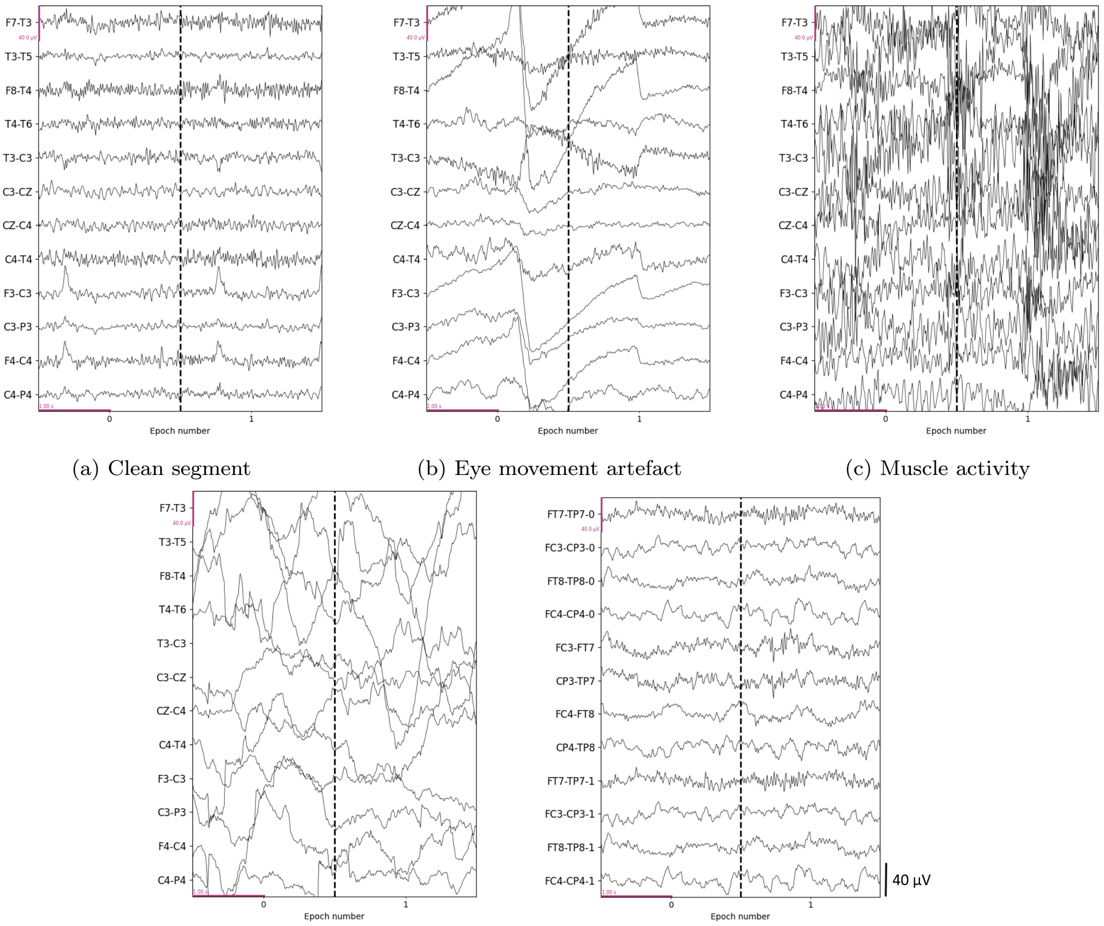
Many techniques have been developed to detect artefacts in EEG data [11, 12]. Recent years, neural networks (NNs) have shown promising results in capturing different artefacts types [13, 14]. These networks are able to learn from raw data, without the need for pre-processed features. The raw data is used to find non-linear relations, explaining the differences between specific artefacts. However, in order to exhibit the profit of a NN, it needs an enormous amount of data to learn from. Dry EEG recordings are sparse and training such a network solely on the available data will likely not result in an optimal classification model.

Therefore, this study aims to develop a multi-class artefact classification model for dry EEG data. Initially, by building a NN based on wet EEG recordings. Followed by implementing the model and translating this to dry EEG recordings.

2 Background

2.1 Artefacts

An artefact can be defined as a signal distortion in the data, not representing the measured brain activity. Artefacts can be divided into two types, technological and physiological artefacts. Technical artefacts arise primarily due to electrical and electro-magnetic noise coming from the power lines or electric lights [16]. Physiological artefacts arise from a subject's own activities, including muscle movements, heart activity, eye blinks, or eye movements [16, 17, 18]. Technological artefacts can be minimised during the processing of the data. Their distinctive frequency characteristics allow for effective suppression using simple filters [10]. Physiological artefacts can be minimised by giving instructions to the patient. However, total elimination is never possible.



(a) Clean segment

(b) Eye movement artefact

(c) Muscle activity

(d) Electrode artefact

(e) Pulse artefact in FC4-FC8

Figure 1: Artefact types. Examples from the Temple University Hospital database [15] (a, b, c, d) and Amsterdam University Medical Centre (e) database. On the x-axis; 2 epochs of 2 seconds, so a total time of 4 seconds is visualised. On the y-axis the according channels are depicted.

Artefacts caused by eye movement do not completely disrupt the brain signals, rather, they add linearly to the data (e.g. Fig. 1b) [19]. Muscle activity can be seen as bursts with high amplitudes (e.g. Fig. 1c). These artefacts can arise due to chewing, talking, clenching etc. They are usually present in the 20-40 Hz range [19]. Depending on electrode position, a pulsation artefact can be found in the signal. The latter artefact distinguishes itself by its repetitive occurring waveform, representing heart pulses (Fig. 1e). Lastly, instrumental artefacts can be found in EEG data. They differ in their appearance depending on the specific source. An example can be seen in Fig. 1d. Movement of an electrode can cause a change in contact leading to the EEG baseline of the signal to start wandering.

Because each artefact type exerts its influence in a different frequency range, complete removal of the artefacts without removing signal as well remains challenging. In normal practice, artefacts are visually ignored during interpretation. However, this is not feasible for its use in automated classification models.

2.2 Related work

To minimise the influence of artefacts on EEG signal processing and automated classification models, algorithms have been developed to removed or correct artefacts. Approaches vary from regression to decomposition methods to deep learning (DL) methods, with the most common method being independent component analysis (ICA). ICA is a form of blind source separation, which tries to break the EEG signal down into different components. These components can subsequently be identified as artefact by an expert or another model [20]. The emphasis is on a specific artefact, particularly eye blinks and muscle movement.

Most recent developments have been in the field of DL. In contrast to other methods, DL can automatically learn pre-processing, feature extraction and classification details without expert knowledge [21]. DL is based on the principle of NNs [22]. The basic structure of a NN consists of an input layer, multiple hidden layers and an output layer. Each layer consists of its own set of neurons to which nonlinear transformations are applied to [22]. For EEG signal analysis, models often utilise extracted features as input [23]. Alternatively, the raw signal or its frequency-based image transformation can be used as input options [23].

Common DL methods used for multi-class artefact classification for wet EEG recordings include CNNs and recurrent neural networks (RNNs). Kim et al. [14] explored different architectures (shallow CNN, deep CNN, RNN, and an ensemble method) to obtain the highest classification ACC for a 5-class (*clean, eye movement, muscle, electrode* and *chewing* artefact) classification model. The ensemble method (ACC 67.59%) was demonstrated as the optimal choice, closely followed by the shallow CNN (ACC 65.15%). Webb et al. [13] used a deep residual CNN for 6-class (*clean, device interference, EMG, movement, electrode* artefacts and *biological rhythms*) artefact classification in neonatal EEG data. The performance resulted in a high ACC of 84.8%. A common limitation was found in accurately classifying all classes. Kim et al. [14] only classified 28% of the muscle artefacts correctly with their ensemble method and Webb et al. [13] only classified the biological rhythms correctly in 4.3% of the cases on the validation set.

DL has also been applied to artefact classification of dry EEG recordings. A binary classifier (*clean* vs *artefact*) was developed by van Stigt et al. [24]. Their model yielded a performance of 90.7% ACC, F1-score of 90.2% and a recall and precision of 91.2% and 89.1%, respectively. This model was not trained to differentiate between different artefact types. The development of a multi-class artefact classification algorithm specifically for dry EEG data would represent a novel contribution to the field of EEG research.

3 Methods

This study consists of two distinct parts. First, a multi-class artefact classification model based on a publicly available dataset was developed. Second, the pre-trained model was implemented and transfer learned to dry EEG data.

3.1 Part I: Multi-class classification model

3.1.1 Dataset description

The model was trained on a publicly available EEG Artifact Corpus (version 3.0.1) provided by the Temple University Hospital (TUH) of Philadelphia [15]. Recordings were performed with electrodes positioned according to the 10-20 system. Data was re-referenced to 20 or 22 channels, depended on the presence of the two earlobe electrodes, with a temporal central parasagittal montage [26]. Three different configurations were available; averaged reference, linked ear reference and a modified version of the averaged reference without using the auricular channels (electrodes A1 and A2). Recordings were available with varying sampling frequencies between 250 - 500 Hz. A total of 310 files of 213 subjects were available, with an average file duration of 19 minutes.

The dataset contained EEG data with artefacts annotated per channel and sample; eye movement (*eyem*), chewing, shivers, muscle artefact (*musc*) and combination of electrode pop, electrostatic and lead artefacts (*elec*). The remaining signal was considered as clean background. Annotations were carried out by a trained team of undergraduate students. At least two individual annotators reviewed a file [27]. In case of uncertainty, a third annotator was added to serve as a tiebreaker.

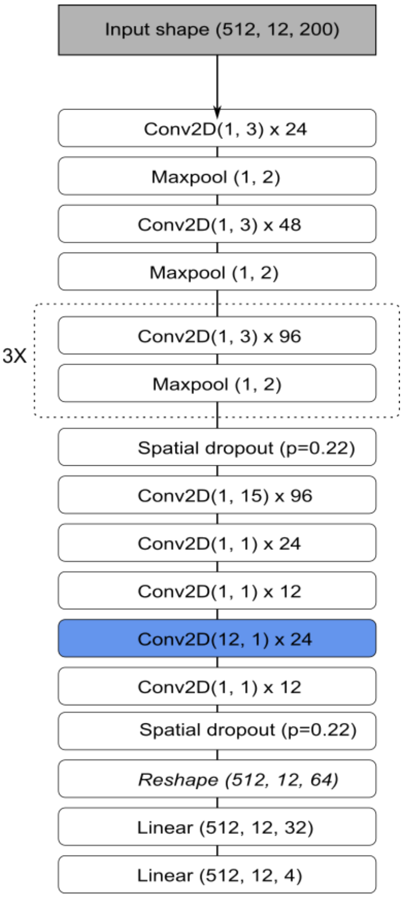


Figure 2: Model architecture based on Hermans et al.[25]

3.1.2 Pre-processing

Data were band-pass filtered (0.5 - 35 Hz) and resampled to 100 Hz. The recordings were further divided into 2-second segments with 50% (1 second) overlap. Inclusion of artefact segments was limited to *eyem*, *musc*, *elec* and *clean* EEG segments based on the availability of annotations, relevance for the model to be fine-tuned and complexity of the pre-trained model. To match the channels as much as possible with the available Amsterdam UMC data (Part II), 12 channels from the 20-22 available channels were selected from the frontal, temporal and parietal regions (Table. 2, Fig. 3).

Labels were given to individual channels as one-hot encodings (*clean*: [1, 0, 0, 0], *eyem*: [0, 1, 0, 0], *musc*: [0, 0, 1, 0], *elec*: [0, 0, 0, 1]). If the annotation of the artefact started/stopped within the epoch, a soft label was generated to provide the model with a more accurate label representation. The soft label was calculated as a percentage based on the presence of the clean data and the corresponding artefact.

E.g. if 80% was annotated as *eyem*, the label of the corresponding segments was [0.2, 0.8. 0, 0]. The labels were saved per segments in a 12-channel epoch instead of single channel segments, to include dependencies between channels as well. *Clean* epochs were downsampled to obtain a more balanced dataset.

Lastly, data were split into a train-test set (80% - 20%). Stratification was done on subject, ensuring no overlap of subjects between different groups. The train set was further divided into 5-folds for cross validation. It was aimed to maintain the class balance within the splits as much as possible (Appendix Table. A).

3.1.3 Model structure, training and evaluation

The input of the model were 12-channel filtered EEG epochs, with as output a class prediction (*clean, eyem, musc, elec*) per channel. The final implemented model was based on the CNN developed by Hermans et al. for artefact detection in neonatal multi-channel EEG data [25]. Their model is a combination of an auto-encoder for unsupervised learning plus a classifier used for supervised learning. A model, based on the encoder part of the autoencoder and classifier, was implemented in this study.

The optimal width of the kernels, batch size, learning rate and dropout probability were determined during hyperparameter tuning (Table. E.3). Hyperparameters were chosen using hyperopt with a tree-structured parzen estimator (TPE), with a maximum of 20 iterations [28, 29]. Per iteration a selection was made of the available parameters (Table. E.3). Values per parameter were carefully chosen, based on previous research. TPE iteratively tries to optimise the relationship between the hyperparameter choice and the average loss of the 5 folds, rather than evaluating all possible combinations.

Table 1: Hyperparameter search space and optimal value implemented in the model.

Hyperparameter	Value(s)	Optimal
Batch size	128, 256, 512, 1024	512
Convolutional layers	9, 10, 11	9
Kernel size width	3, 5, 7	3
Kernel size height	12	12
Dropout probability	0.0 - 0.5	0.22
Learning rate	0.00001, 0.0001	0.0001

The final model consisted of 10 (9 temporal, 1 spatial) convolutional layers (Fig. 2). The first five convolutional layers, were followed by a max pooling layer. A dropout layer was added before the spatial convolutional layer and after all convolutions. Lastly, in contrast to the 1 x 1 convolutional layers of Hermans et al.[25], two fully-connected layers were connected to the network to combine all information from previous layers. The output of the model is a tensor with the probabilities for each class with the shape (512, 12, 4), representing the batch size, the number of channels, and the four classes, respectively. After each convolution, batch normalisation was applied for regularisation and AdamW was used as an optimizer. Two other explored architectures are described in Appendix E.

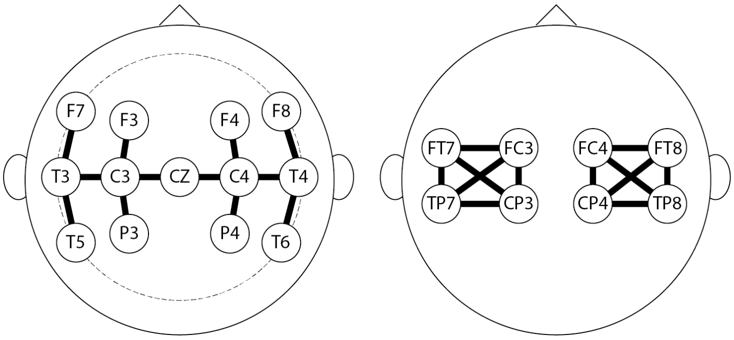


Figure 3: Bipolar montages used. Visualises channels chosen from the Temple University Hospital (left) and Amsterdam University Medical Centre (right).

Table 2: Bipolar montages used. Selected channels from TUH dataset and all available channels from the Amsterdam UMC dataset.

TUH dataset	Amsterdam UMC dataset
F7 - T3	FT7 - TP7
T3 - T5	FC3 - CP3
F8 - F4	FT8 - TP8
T4 - T6	FC4 - CP4
T3 - C3	FC3 - FT7
C3 - CZ	CP3 - TP7
CZ - C4	FC4 - FT8
C4 - T4	CP4 - TP8
F3 - C3	FT7 - CP3*
C3 - P3	FC3 - TP7*
F4 - C4	FT8 - CP4*
C4 - P4	FC4 - TP8*

Part I: Temple University Hospital (TUH), Part II: Amsterdam University Medical Centre (UMC). *Input replaced by available annotated channels.

The model was trained by minimising the cross entropy loss and the learning rate was scheduled with a cosine annealing scheduler. Training was done for a maximum of 100 epochs. Early stopping was implemented when no further minimisation of the validation loss could be achieved for 10 consecutive epochs. During training the class imbalance was corrected for by adjusting the class weights. Weights calculation was based on inverse frequency weighting. Performance was computed for each fold. Finally, the five models were combined using a majority vote to determine the final classification per channel: (*clean, eyem, musc* or *elec*).

3.2 Part II: Fine-tuning model

3.2.1 Dataset description

The dataset was provided by the Amsterdam UMC. The dataset contained a total of 13 participants, 10 healthy subjects and 3 outpatient clinic patients. Recordings were acquired using a dry EEG cap with 8 Ag/AgCl coated electrodes. Electrodes were positioned at FC3, FC4, CP3, CP4, FT7, FT8, TP7 and TP8 (Fig. 3). During recordings, physiological and technological artefacts were induced according to specific protocols. These included eye movement, muscle activities (jaw clenching, talking, frowning) and electrode artefact (cable movement and high electrode-skin impedances for the reference, ground and cap electrodes).

The dataset contained annotations of the labels *clean*, pulsation artefact (*pulse*), (*musc*) and artefact (*art*). The latter label was given to segments not belonging to any of the other groups or when a combination of *pulse* and *musc* was present. Labelling was done sample and channel-wise at a sampling frequency of 100 Hz by 3 to 4 trained reviewers.

3.2.2 Pre-processing

Similar to the TUH dataset, data was band-pass (0.5 - 30 Hz) filtered and resampled to 100 Hz. Data was re-referenced to a 12-channel bipolar montage. The order of the channels was maintained to be as similar as possible to the TUH dataset. Annotations were only available for 8 out of the 12 bipolar channels. Since the model is dependent on 12-channel input, it was chosen to fill the remaining four channels with annotated data. Therefore, for the development of the model, FT7-TP7, FC3-CP3, FT8-TP8 and FC4-CP4 were duplicated to replace FT7-CP3, FC3-TP7, FT8-CP4, and FC4-TP8 at this stage (indicate with * in Tab. 2). Lastly, recordings were split in 2-second segments with 1.9 second overlap.

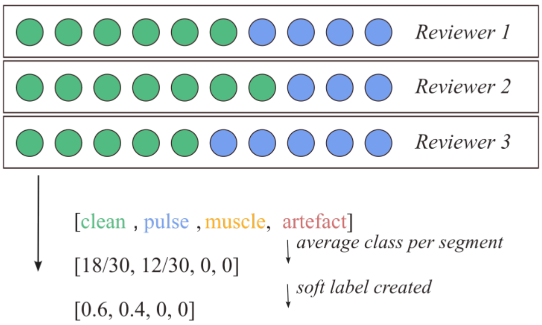


Figure 4: Example of soft label generation for one segment. Per 2-second segment annotated data points of different reviewers are averaged.

The *musc* label was further specified as 'clean with small high frequency component', and thus labelled *clean*, or *musc* to isolate strong muscle activity in the *musc* label. This categorization was achieved by calculating the power ratio per segment between the 5-15 Hz and 25-40 Hz frequency bands. A power ratio lower than 0.5 was classified as *musc*.

Labels were translated to one-hot encodings ([*clean, pulse, musc, art*]) for each 2-second segment. Soft labels were created based on the annotations of the different reviewers. The class labels of all data points within a 2-second segment were averaged and given to the specific class accordingly (Fig. 4).

Moreover, in order to obtain a more balanced dataset, segments with $\geq$ 10 out of 12 channels (one epoch) labelled *clean* or *art* were downsampled by a factor 2. Lastly, the data was split into train (80%) and test (20%) set. The train set was further divided into three folds for cross validation. During all splits effort was made to keep the class balance as equal as possible. (Appendix. A). Data of the same subject was maintained within one group during splitting.

3.2.3 Model fine-tuning

The model of Part I was implemented as pre-trained CNN and fine-tuned for the classes *clean, pulse, musc* and *art*. Similar as in Part I, four classes were of interest. Therefore, output structure could remain the same. The optimal balance between layers to freeze and layers to update was optimised on the train data, as well as batch size, learning rate and dropout probability. This resulted in updating all layers, including training the last layer from scratch, a learning rate of 0.001, a batch size of 128 and a dropout probability of 0.25. Similarly to Part I, the CNNs were trained with an AdamW optimizer by minimizing the cross entropy loss.

$$\text{4-Class Accuracy} = \left(\sum_i True\right)/Total, \quad (1)$$

where i=[four classes]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

3.3 Performance assessment

A similar performance assessment was executed for the pre-trained model (Part I) as well as for the fine-tuned model (Part II). Performance metrics were calculated based on the unseen test set (20% of the original TUH/Amsterdam UMC data). The predictions generated by the models during cross-validation were combined through averaging. Subsequently, the argmax operation was applied to the averaged predictions to determine the majority vote. Similarly, the true soft labels were converted to hard labels based on the largest portion present.

The performance of the classification model was assessed by a confusion matrix (four by four). Moreover, ACC was computed as percentage of accurately classified segments (Eq. 1) and AUC was computed to provide more inside in the possibility to discriminate between classes. Per class, a one-versus-rest ROC was calculated resulting in one overall AUC. Lastly, recall and precision were calculated and combined in a F1-score (Eq. 4). Class weights were incorporated in the performance assessment.

4 Results

4.1 Data availability

4.1.1 TUH dataset

After resampling and segmentation, almost 4-million 2-second single channel segments were obtained, which can be translated to 326.781 12-channel epochs. Downsampling of *clean* segments resulted in a final number of 969,744 single channel segments (Table. 3), 80,812 12-channel epochs. Of these, 1,030 are completely clean epochs (12 *clean* 2-second segments).

4.1.2 Amsterdam UMC dataset

Pre-processing of the data results in over 1 million 2-second single channel segments. Downsampling of *clean* and *art* segments led to a final 518,868 single channel segments, which are 43,239 2-second epochs (Table. 3).

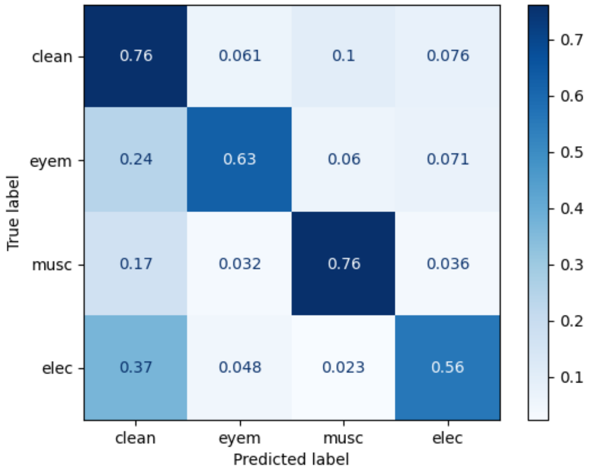


Figure 5: Confusion matrix using the classification model on the test set. The majority vote is used of the models created during 5-fold cross validation. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (clean, eye movement (eyem), muscle activity (musc), electrode artefact (elec)).

Table 3: Description of labelled segments. Total number of 2-second segments per class (*clean*, eye movement (*eyem*), muscle activity (*musc*), electrode artefact (*elec*), pulsation artefact (*pulse*), artefact (*art*)) before and after downsampling. One epoch consists of 12 segments. Temple University Hospital (TUH), Amsterdam University Medical Centre (UMC)

		Number of 2s segments (%)	
		Unbalanced data distribution	Corrected data distribution
TUH dataset	*Clean*	3,417,993 (87.9)	466,365 (48.1)
	Eyem	53,313 (1.4)	53,313 (5.5)
	Musc	325,330 (8.3)	325.330 (33.6)
	Elec	124,736 (3.1)	124,736 (12.8)
	Total	3,921,372	969,744
Amsterdam UMC dataset	*Clean*	524,810 (50.3)	257,135 (49.6)
	Pulse	28,257 (2.8)	28,257 (5.4)
	Musc	25,814 (2.5)	25,814 (5.0)
	Art	455,091 (44.4)	207,662 (40.0)
	Total	1,023,972	518,868

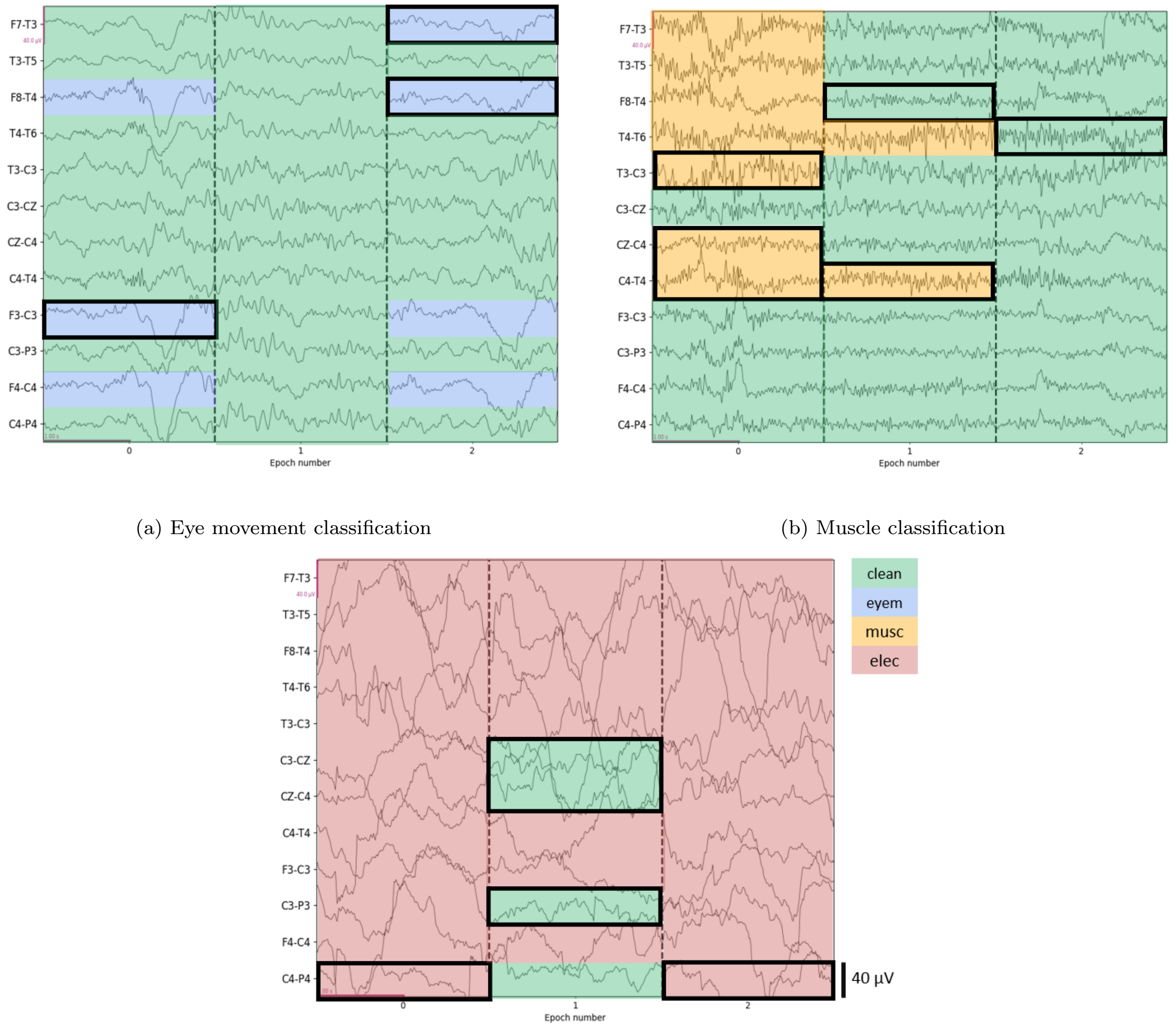
Table 4: Performance metrics of classification model using the majority vote on the test set. Showing the overall accuracy (ACC) and per artefact class (clean, eye movement (*eyem*), muscle activity (*musc*), electrode artefact (*elec*)) the recall, precision, area under the curve (AUC) and F1-score.

	ACC (%)	Recall (%)	Precision (%)	AUC	F1-score
Overall	74.8	68.4	71.7	0.87	0.68
Clean	-	76.2	49.8	0.84	0.60
Eyem	-	62.3	80.3	0.85	0.70
Musc	-	76.2	81.0	0.91	0.79
Elec	-	56.0	76.1	0.89	0.64

4.2 Part I: Multi-class classification performance

The 4-class classification model achieved an ACC of 74.8%, with a recall, precision, AUC and F1-score of 68.4%, 71.7%, 0.87 and 0.68 respectively using the majority vote of the five models on the test set (Tab. 4). Performance per individual model of the train, validation and test set can be found in Appendix B. Highest performance was found for the *musc* label (recall 76%, precision 81%, AUC 0.91, F1-score 0.79). When the accurate artefact class (*eyem, musc, elec*) was not predicted, the model often classified the data as clean (Fig. 5). This resulted in a high precision for the artefact classes (*eyem* 80%, *musc* 81%, *elec* 76%) at the expense of the precision of the clean label (precision 50%).

In Fig. 6 segments are visualised along with their corresponding predictions. In general, it can be seen that clean segments are classified correctly. Most of the epochs contain clean segments plus one artefact class (*eyem, musc, elec*). A combination of artefacts classes is not often seen in one epoch.



(a) Eye movement classification

(b) Muscle classification

(c) Electrode artefact classification

Figure 6: Examples of correctly classified and misclassified segments, predictions are visualised. Black boxes indicate a disagreement between the annotation and prediction. (a) Classification of eye movement (*eyem*). Three segments were wrongly classified as *eyem*, whereas they were annotated clean. (b) Muscle (*musc*) classification. Four clean segments were predicted *musc* but annnotated clean, and two *musc* segments were predicted clean. (c) Electrode artefact classification (*elec*). Two clean segments are wrongly classified as *elec*, and three are wrongly classified clean while they were annotated *elec*.

4.3 Part II: Fine-tuning performance

Fine tuning the model resulted in an ACC of 71.2% with a recall and precision of 62.3% and 65.4% respectively, using the majority vote. High classification performances were found for the *musc* artefact and the *art* class. Most instances from the *musc* class were accurately classified (recall 86%), whereas the *art* class obtained the highest precision (85%). An overall low performance was seen for the *pulse* artefacts (recall 17%, precision 59%). The performance of the *pulse* artefacts on the test set was considerably lower compared to the validation groups during cross validation (Appendix Fig. C.3). Most of the *pulse* artefacts were classified as clean. This negatively impacted the model's overall performance as well as the precision of the *clean* segments.

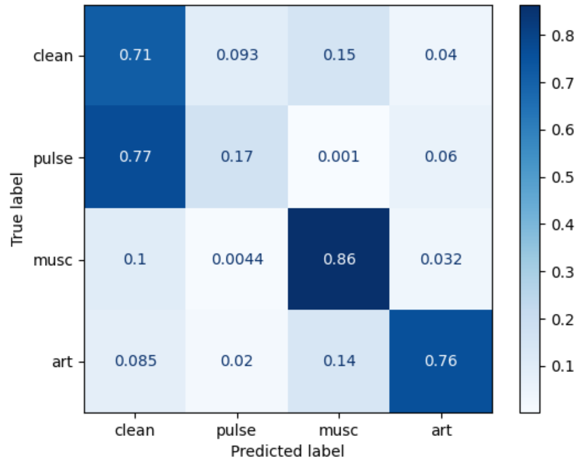
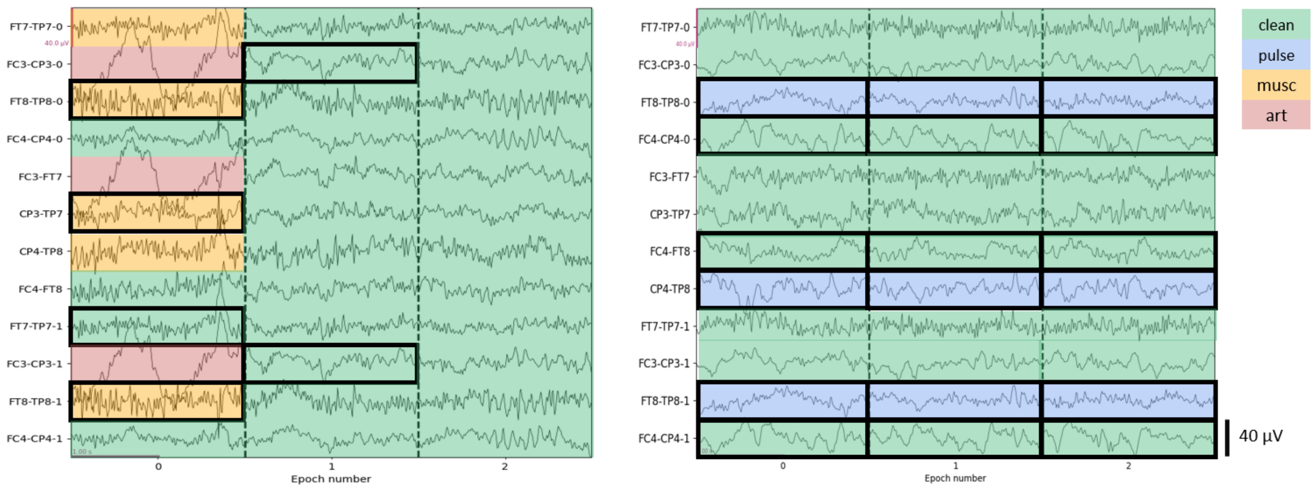


Figure 7: Confusion matrix using the fine-tuned model on the test set. The majority vote is used of the models created during 3-fold cross validation. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (clean, pulsation artefact (pulse), muscle activity (musc), artefact (art)).

Table 5: Performance metrics of the fine-tuned model using the majority vote on the test set. Showing the overall accuracy (ACC) and per artefact class (*clean*, pulsation artefact (*pulse*), muscle activity (*musc*), artefact (*art*)) the recall, precision, area under the curve (AUC) and F1-score.

	ACC	Recall	Precision	AUC	F1-score
	(%)	(%)	(%)		
Overall	71.2	62.3	65.4	0.82	0.60
Clean	-	71.4	42.8	0.88	0.54
Pulse	-	16.9	59.0	0.65	0.26
Musc	-	86.4	74.8	0.92	0.80
Art	-	75.7	85.0	0.94	0.80



(a) Muscle and artefact classification

(b) Pulsation artefact classification

Figure 8: Examples of correctly classified and misclassified segments, predictions are visualised. Black boxes indicate a disagreement between the annotation and prediction. The first four channels are duplicated for the purpose of training of the model. (a) Epoch 0 contains three predicted muscle artefacts (*musc*), which are annotated clean. The clean predicted segment is annotated *musc*. Epoch 1, contains two artefact (*art*) segments, which are predicted clean. (b) Channel 'FC4-CP4' is predicted clean, but annotated as *art*. 'CP4-TP8' is annotated clean, while predicted pulsation artefact (*pulse*). 'FC4-TP8' is predicted clean, while annotated *pulse*.

In Fig. 8 examples of segments with their predicted labels are visualised. On the left plot, the first epoch shows difficulties in differentiating between *musc* artefacts and *clean* segments. In the second epoch, the model predicted all segments to be clean, however 'FC3-CP3' was actually labelled as containing *art.* The last epoch is consistent with the true labels, containing all *clean* segments. On the right plot, difficulties in capturing *pulse* artefacts are found. Segments from 'CP4-TP8' are predicted to be *pulse*, but are annotated clean. The other way around is seen in 'FC4-TP8', where all *pulse* segments are predicted clean.

5 Discussion

In this study, a multi-class artefact CNN-based algorithm for dry EEG data was developed using transfer learning. To our knowledge, this is the first algorithm for multi-class artefact classification in dry EEG data. A pre-trained CNN was developed leveraging the large public TUH dataset. Subsequently, this model was fine-tuned for its application in dry EEG data. The pre-trained multi-class (*clean, eyem, musc* and *elec*) model achieved an overall ACC of 74.8%. The fine-tuned model was able to correctly differentiate between the classes (*clean, pulse, musc* and *art*) with an ACC of 71.2%, with the best performance for artefact classes' *musc* (AUC 0.92, F1-score 0.80) and *art* (AUC 0.94, F1-score 0.80).

Other multi-class artefact classification models exist, but are only applied for EEG recordings obtained with wet electrodes [13, 14]. In general, wet and dry electrode EEG data have similar characteristics, but dry electrode EEG has a slightly increased power in the lower frequency bands (< 8 Hz) and is more prone to artefacts [7, 30]. Therefore, transfer learning was implemented in this study. Findings have been reported per model; Part I - development of pre-trained model and Part II - fine-tuning the model on dry EEG data.

For the pre-trained model (Part I) it was observed that for all artefact classes (*eyem, musc* and *elec*) the most falsely predicted segments were predicted *clean.* The false prediction of *clean* segments can be attributed to: I) the high prevalence of the *clean* class (48.1%) in the overall dataset used for development of the model, II) the presence of periods of uncontaminated EEG data within false predicted *clean* segments, and III) the difficulty of capturing artefact classes represented by distinct phenomena. Regarding the latter point, *elec* is a combination of various electrode artefacts, such as electrode pop, electrostatic, and lead artefacts, which are characterised differently within an EEG signal [15]. In contrast, *musc* artefacts have a more typical appearance, characterised by an increased frequency and higher amplitudes, which may be learned more easily by the model. To decrease the amount of falsely classified *clean* segments, annotations should be reviewed to improve time resolution. Moreover, alternative methods to balance classes could be explored. Currently, these falsely predicted *clean* segments could influence further analysis of this class and result in a lower precision (50%) for the *clean* class. Conversely, the developed CNN achieved a high precision for the artefact classes (*eyem* 80%, *musc* 81% and *elec* 76%). The model rarely falsely predicted artefact classes as the wrong type of artefact, indicating the ability of the model to accurately distinguish between different artefact types.

Similar to the pre-trained model, Kim et al. used the TUH dataset to develop a multi-class (*clean, eyem, musc, elec* and *chewing*) classification model [14]. A higher recall was obtained for *musc* (76% vs 28%, respectively) and *elec* (56% vs 41%, respectively) artefacts by our model in comparison to the ensemble method of Kim et al. Kim et al. employed a single-channel approach, whereas the present study utilised a multi-channel approach. The observed improvement can be attributed to spatial learning across all channels within each 2-second epoch, as opposed to learning from individual channels. Contrary to the increased recall for *musc* and *elec* artefacts, our study demonstrated a decreased recall for *eyem* (63% vs 72%). In contrast to this study, Kim et al. utilised all available channels, resulting in a different class distribution to train on. Specifically, their study incorporated a greater proportion of the *eyem* class compared to the present study (26% vs. 6%, respectively), which was a result of excluding the most frontal channels in our methodology. This factor potentially contributed to their higher recall. Lastly, both Kim et al. and our study experienced similar difficulty in accurately classifying the *elec* artefacts. In this study, as well as in Kim et al., 37% of the *elec* artefacts were classified as clean on the test set [14]. This suggests that clean segments were hard to differentiate from electrode artefacts, despite visual discernibility. In addition to the challenge of capturing different phenomena within a single class, as discussed in the previous paragraph, model characteristics, such as epoch length, might not have been optimal to effectively differentiate and capture electrode artefact features. This study focused on 2 second segments and Kim et al. implemented 1 second segments. Peh et al. researched classification performances using different segments lengths (1s, 3s, 5s) [31]. They showed that in a multi-channel configuration, *elec* artefacts were captured best with a segment length of 5 seconds (ACC 88%, AUC 0.82) compared to using segments of 1 second (ACC 80%, AUC 0.79) or 3 seconds (ACC 86%, AUC 0.82). This suggests that increasing epoch length to 5 seconds, could improve *elec* classification.

The fine-tuned model (Part II) showed a good performance for each individual class, except for *pulse artefact* (*clean* [AUC 0.88, F1-score 0.54], *musc* [AUC 0.92, F1-score 0.80], *art* [AUC 0.94, F1-score 0.80], *pulse* [AUC 0.65, F1-score 0.26]). Performance for the detection of *pulse* artefacts differed per fold (Appendix C.3). Also, a difference in validation and test set was seen: 30-75% is accurately captured in the validation sets, whereas this is only 7-46% in the test set (Appendix C). During all splits (train, validation and test), it was made sure that the class balance was maintained as much as possible (Appendix A). However, it was seen that especially for *pulse* artefacts, the class distribution per channel differed between the train and test set (Appendix D). In the train set, *pulse* artefacts were mainly present in the right more temporal located channels (CP4-TP8 [37%] and FT8-TP8 [17%]). Similarly, *pulse* artefact predictions, based on the test set, were also mainly found in these channels (CP4-TP8 [61%] and FT8-TP8 [11%]). However, the original distribution of the test set contains *pulse* artefacts predominately in the right more frontal channels (FC4-FT8 [27%] and FC4-CP4 [25%]). This type of event is more clearly visualised in Fig.8b. A *pulse* artefact is detected in the analysed epoch, but in a wrong channel (Fig. 8), contributing to the lower performance of *pulse artefacts*. This suggests the problem of overfitting and more focus might be on the location of the *pulse* artefact rather than the specific shape. In this study, it was not possible to balance each individual class per channel due to limited data availability. To address this limitation in the future, data augmentation can be used to maximise the generalisation capability of deep learning models [32]. Data augmentation can be implemented through DL or feature transformation techniques [32].

It is suggested by Roy et al. that artefact removal might be redundant for DL EEG analyses, as 47% of the studies included in the review did not use any artefact handling methods while accurate task performance was preserved [21]. However, due to the high prevalence of artefacts in dry EEG recordings alternative handling methods, as opposed to removal or ignorance, are needed to prevent exclusion due to insufficient data quality [7]. The algorithm developed in this study could facilitate the research on the impact of artefact correction, exclusion, or assessment of predictive value per artefact type on LVO prediction models. Translating this to the Amsterdam UMC dataset, the following considerations are made:

1. The *art* class contains data that is unusable due to the severity of the artefacts. For example, if electrode contact is inadequate, underlying brain activity will not be captured. Hence, correction is not possible and it is advisable to omit these segments from analysis.
2. *Pulse* artefacts are identified. A study by Paxton et al. demonstrated that cranial accelerometry, combined with clinical data, can identify LVO strokes with a good performance (AUC 0.91, sensitivity 84.6%, specificity 82.6%) [33]. This technology utilises subtle head oscillations in combination with cardiac contractions to assess pathological changes in cranial blood flow. Therefore, it would be of interest to explore the predictive promise *pulse* artefacts may hold for LVO classification algorithms.
3. *Musc* artefacts are identified, and their predictive value could be researched. In case they do not add value, multiple studies have shown that *musc* artefacts can be corrected rather than eliminated [34, 35, 36]. Artefact containing segments are down-sampled and clean EEG data were generated by implementing an autoencoder [36] or an UNet [34].

This approach could ensure that data quality is optimised and the potential predictive value of different artefact types is thoroughly evaluated, while simultaneously maintaining more data for the LVO classification algorithms.

Improvement of the performances of the models may be achieved by expanding the dataset and reviewing current labels rather than by researching different model architectures as different architectures have proven to result in similar performances (Appendix E). Differences were seen between performances of the validation folds and the losses of the validation sets only decreased for the first few epochs (Fig. C.3). This may indicate the model learns limited information and might benefit from an increase in data. Moreover, it could be debated whether the model was wrong in its prediction or that the annotated label was not optimal. Visual inspection of the predictions highlighted difficulties in classifying *musc* and *clean* segments (Fig. 8). The segments predicted as *musc* in channel 'FT8-TP8' in epoch 1, could also have been classified as *musc* visually instead of *clean*. This distinction is challenging because there are no objective criteria, such as a detection threshold, for determining when to classify something as clean or an artefact. Semi-supervised learning could be incorporated to let the model learn features in an unsupervised manner before it is applied to a supervised dataset [25]. In this way, more data could be incorporated and the feature learning is not solely dependent on the annotated data. However, the performance should still be evaluated on a supervised dataset, for which accurate labelling is required.

In contrast to other multi-class models, a multi-channel input was used in this study [13, 14]. This has been shown to outperform single-channel input, but has one main limitation [25, 31]. Whereas single channel models can process all channels individually, a multi-channel model becomes dependent on the number of channels it is trained on. A 12-channel configuration was chosen to match the number of bipolar channels available in the Amsterdam UMC data. This allowed us to optimally make use of spatial characteristics between the channels. However, when 12 channels are not available, the model cannot be applied. Since the model is specifically developed to be applied for the current dry EEG set-up of the Amsterdam UMC, this is not seen as a problem. Instead, now it can incorporate spatial features as well.

For further research, a first step would be to implement the annotations of the four remaining bipolar channels (FC3-TP7, FT7-CP3, FC4-TP8, FC8-CP4) to investigate the performance with all 12-channels incorporated. At this stage, four channels were duplicated, since the model was dependent on 12-channel input. Implementing the additional channels could allow the model to learn the most optimal spatial features and hence potentially improve performance. Furthermore, it would be of interest to implement a post-processing step to increase performance per artefact class. This was implemented by Webb et al. in the form of temporal smoothing [13]. They averaged the predictions over a time period that was specific for each artefact type. Based on a balanced dataset, this increased their ACC from 82.0% to 84.8% [13]. It is thought that especially for *pulse* artefact temporal smoothing could improve performance, since this type of artefact is often seen during the entire length of the recording in one channel. If surrounding segments contain *pulse* artefacts, it is more likely that the corresponding segment contains a *pulse* artefact as well.

6 Conclusion

In this study, a pre-trained model was developed using a publicly available dataset and subsequently fine-tuned on a smaller dataset with dry EEG recordings. This approach demonstrated a good capability to differentiate between artefact classes in the pre-trained model (*eyem, musc* and *elec*) as well as in the fine-tuned model (*pulse, musc* and *art*), although dry EEG data were sparse for training the latter model. The fine-tuned model holds potential to facilitate further research aimed at determining the predictive value of each artefact class on LVO stroke prediction models using dry EEG data. Additionally, the impact of excluding and correcting various artefact types could be investigated.

References

[1] Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

[2] Vasu Saini, Luis Guada, and Dileep R Yavagal. Global epidemiology of stroke and access to acute ischemic stroke interventions. *Neurology*, 97(20 Supplement 2):S6–S16, 2021.

[3] William J Powers, Alejandro A Rabinstein, Teri Ackerson, Opeolu M Adeoye, Nicholas C Bambakidis, Kyra Becker, José Biller, Michael Brown, Bart M Demaerschalk, Brian Hoh, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 50(12):e344–e418, 2019.

[4] Jeffrey L Saver, Mayank Goyal, AAD Van der Lugt, Bijoy K Menon, Charles BLM Majoie, Diederik W Dippel, Bruce C Campbell, Raul G Nogueira, Andrew M Demchuk, Alejandro Tomasello, et al. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. *Jama*, 316(12):1279–1289, 2016.

[5] Esmee Venema, Adrien E Groot, Hester F Lingsma, Wouter Hinsenveld, Kilian M Treurniet, Vicky Chalos, Sanne M Zinkstok, Maxim JHL Mulder, Inger R De Ridder, Henk A Marquering, et al. Effect of interhospital transfer on endovascular treatment for acute ischemic stroke. *Stroke*, 50(4):923–930, 2019.

[6] Paulina B Sergot, Andrew J Maza, Bruce J Derrick, Lane M Smith, Liam T Berti, Madeleine R Wilcox, Matthew R Kesinger, and W Frank Peacock. Portable neuromonitoring device detects large vessel occlusion in suspected acute ischemic stroke. *Stroke*, 52(4):1437–1440, 2021.

[7] Maritta N van Stigt, Eva A Groenendijk, Laura CC van Meenen, Anita AGA van de Munckhof, Monique Theunissen, Gaby Franschman, Martin D Smeekes, Joffry AF van Grondelle, Geertje Geuzebroek, Arjen Siegers, et al. Prehospital detection of large vessel occlusion stroke with eeg: Results of the electra-stroke study. *Neurology*, 101(24):e2522–e2532, 2023.

[8] Fareshte Erani, Nadezhda Zolotova, Benjamin Vanderschelden, Nima Khoshab, Hagop Sarian, Laila Nazarzai, Jennifer Wu, Bharath Chakravarthy, Wirachin Hoonpongsimanont, Wengui Yu, et al. Electroencephalography might improve diagnosis of acute stroke and large vessel occlusion. *Stroke*, 51(11):3361–3365, 2020.

[9] Laura CC van Meenen, Maritta N van Stigt, Arjen Siegers, Martin D Smeekes, Joffry AF van Grondelle, Geertje Geuzebroek, Henk A Marquering, Charles BLM Majoie, Yvo BWEM Roos, Johannes HTM Koelman, et al. Detection of large vessel occlusion stroke in the prehospital setting: Electroencephalography as a potential triage instrument. *Stroke*, 52(7):e347–e355, 2021.

[10] Hangyu Zhu, Yonglin Wu, Ning Shen, Jiahao Fan, Linkai Tao, Cong Fu, Huan Yu, Feng Wan, Sio Hang Pun, Chen Chen, et al. The masking impact of intra-artifacts in eeg on deep learning-based sleep staging systems: A comparative study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1452–1463, 2022.

[11] Sari Sadiya, Tuka Alhanai, and Mohammad M Ghassemi. Artifact detection and correction in eeg data: a review. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 495–498. IEEE, 2021.

[12] Xiao Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from eeg signals: a review. *Sensors*, 19(5): 987, 2019.

[13] Lachlan Webb, Minna Kauppila, James A Roberts, Sampsa Vanhatalo, and Nathan J Stevenson. Automated detection of artefacts in neonatal eeg with residual neural networks. *Computer Methods and Programs in Biomedicine*, 208:106194, 2021.

[14] Dong Kyu Kim and Sam Keene. Fast automatic artifact annotator for eeg signals using deep learning. *Biomedical Signal Processing: Innovation and Applications*, pages 195–221, 2021.

[15] Ahmed Hamid, Katherine Gagliano, Safwanur Rahman, Nikita Tulin, Vincent Tchiong, Iyad Obeid, and Joseph Picone. The temple university artifact corpus: An annotated corpus of eeg artifacts. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE, 2020.

[16] Gernot R Müller-Putz. Electroencephalography. *Handbook of Clinical Neurology*, 168:249–262, 2020.

[17] C Rashmi and C Shantala. Eeg artifacts detection and removal techniques for brain computer interface applications: A systematic review. *Int. J. Adv. Technol. Eng. Explor*, 9:354, 2022.

[18] Rakesh Ranjan, Bikash Chandra Sahana, and Ashish Kumar Bhandari. Cardiac artifact noise removal from sleep eeg signals using hybrid denoising model. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–10, 2022.

[19] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014.

[20] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197, 2019.

[21] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16 (5):051001, 2019.

[22] Tao Qin and Tao Qin. Deep learning basics. *Dual Learning*, pages 25–46, 2020.

[23] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[24] MN van Stigt, EA Groenendijk, HA Marquering, JM Coutinho, and WV Potters. High performance clean versus artifact dry electrode eeg data classification using convolutional neural network transfer learning. *Clinical Neurophysiology Practice*, 8:88–91, 2023.

[25] Tim Hermans, Laura Smets, Katrien Lemmens, Anneleen Dereymaeker, Katrien Jansen, Gunnar Naulaers, Filippo Zappasodi, Sabine Van Huffel, Silvia Comani, and Maarten De Vos. A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection. *Journal of Neural Engineering*, 20(2):026013, 2023.

[26] Sean Ferrell, Vishnu Mathew, Mark Refford, Vincent Tchiong, Tariq Ahsan, Iyad Obeid, and Joseph Picone. The temple university hospital eeg corpus: Electrode location and channel labels. https://www.isip.piconepress.com/publications/reports/2020/tuh_eeg/electrodes, 2020.

[27] Daniel Ochal, Saifur Rahman, Sean Ferrell, Tamer Elseify, Iyad Obeid, and Joseph Picone. The temple university hospital eeg corpus: Annotation guidelines. https://www.isip.piconepress.com/publications/reports/2020/tuh_eeg/annotations/, 2020.

[28] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[29] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[30] Patrique Fiedler, Paulo Pedrosa, Stefan Griebel, Carlos Fonseca, Filipe Vaz, Eko Supriyanto, F Zanow, and J Haueisen. Novel multipin electrode cap system for dry electroencephalography. *Brain topography*, 28:647–656, 2015.

[31] Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3599–3602. IEEE, 2022.

[32] Chao He, Jialu Liu, Yuesheng Zhu, and Wencai Du. Data augmentation for deep neural networks model in eeg classification task: a review. *Frontiers in Human Neuroscience*, 15:765525, 2021.

[33] James H Paxton, Kevin J Keenan, John M Wilburn, Stefanie L Wise, Howard A Klausner, Matthew T Ball, Robert B Dunne, K Derek Kreitel, Larry F Morgan, William D Fales, et al. Headpulse measurement can reliably identify large-vessel occlusion stroke in prehospital suspected stroke patients: Results from the episode-ps-covid study. *Academic Emergency Medicine*, 2024.

[34] Md Shafayet Hossain, Sakib Mahmud, Amith Khandakar, Nasser Al-Emadi, Farhana Ahmed Chowdhury, Zaid Bin Mahbub, Mamun Bin Ibne Reaz, and Muhammad EH Chowdhury. Multiresunet3+: A full-scale connected multi-residual unet model to denoise electrooculogram and electromyogram artifacts from corrupted electroencephalogram signals. *Bioengineering*, 10(5):579, 2023.

[35] Haoming Zhang, Mingqi Zhao, Chen Wei, Dante Mantini, Zherui Li, and Quanying Liu. Eegdenoisenet: a benchmark dataset for deep learning solutions of eeg denoising. *Journal of Neural Engineering*, 18(5):056057, 2021.

[36] Le Xing and Alex Casson. Deep autoencoder for real-time single-channel eeg cleaning and its smartphone implementation using tensorflow lite with hardware/software acceleration. *Authorea Preprints*, 2023.

[37] Yuqi Wang, Lijun Zhang, Pan Xia, Peng Wang, Xianxiang Chen, Lidong Du, Zhen Fang, and Mingyan Du. Eeg-based emotion recognition using a 2d cnn with different kernels. *Bioengineering*, 9(6):231, 2022.

A Cross validation splits

Both models were trained or fine-tuned using cross validation. Class balance was strived for in each fold. Each individual fold was used as validation group, while data from the other folds was used to train on. Tab.A.1 shows the data distribution of the dataset used in Part I. Tab. A.2 shows the data distribution of the dataset used in Part II.

Table A.1: Artefact class (*clean*, eye movement (*eyem*), muscle activity (*musc*), electrode artefact (*elec*)) segments per cross-folds: Part I - TUH dataset. Train dataset is divided into 5 cross folds

	Train dataset					Test dataset
Fold - Validation	1	2	3	4	5	
Clean (%)	49301 (53.3)	103891 (52.8)	74659 (48.3)	55447 (49.3)	67694 (49.0)	115373 (47.7)
Eyem (%)	26441 (24.1)	7875 (3.8)	6981 (4.4)	4165 (3.5)	4143 (2.9)	3708 (1.5)
Musc (%)	18890 (17.2)	58316 (28.8)	51433 (32.3)	46340 (40.2)	48.387 (34.4)	101964 (42.1)
Elec (%)	15216 (13.6)	32202 (15.9)	26515 (16.7)	9380 (8.1)	20500 (14.5)	20923 (8.6)
2 sec segments	109.948	202.284	159.588	115.332	140.724	242.068
12-channel epochs	9.154	16.857	13.299	9.611	11.727	20.164

Table A.2: Artefact class (*clean*, pulsation artefact (*pulse*), muscle activity (*musc*), artefact (*art*)) segments per folds: Part II - Amsterdam UMC dataset. Train dataset is divided into 3 cross folds.

	Train dataset			Test dataset
Fold - Validation	1	2	3	
Clean (%)	116,492 (65.3)	32,557 (34.5)	58,465 (47.3)	49,621 (40.5)
Pulse (%)	4,592 (2.6)	9,361 (9.9)	7,437 (6.0)	6,867 (5.6)
Musc (%)	6,326 (3.5)	8,792 (9.3)	4,873 (3.9)	5,823 (4.8)
Art (%)	50,970 (28.6)	43,694 (46.2)	52,741 (42.7)	60,257 (49.2)
2 sec segments	178,380	94,404	123,516	122,568
12-channel epochs	14,865	7,867	10,293	10,214

B Cross validation performance - Part I: TUH dataset

The model of Part I was trained using five fold cross-validation. The results per fold can be seen in Fig.B.3 for the train and validation sets and Fig. B.2 for the test set. An overall stable performance was seen for *clean*, *musc* and *elec* classification in all five model created (Fig. B.1) . More variation was seen for *eyem* classification. The performances of the validations set were in a similar range as the test set (Table B.1).

Table B.1: Accuracy (ACC) and area under the curve (AUC) metrics for the models created using 5-fold cross validation

Model	Dataset	ACC (%)	AUC
1	Train	82	0.97
	Validation	65	0.83
	Test	72	0.87
2	Train	77	0.96
	Validation	70	0.93
	Test	73	0.88
3	Train	80	0.97
	Validation	70	0.91
	Test	71	0.86
4	Train	83	0.98
	Validation	69	0.84
	Test	74	0.86
5	Train	83	0.98
	Validation	73	0.91
	Test	72	0.86

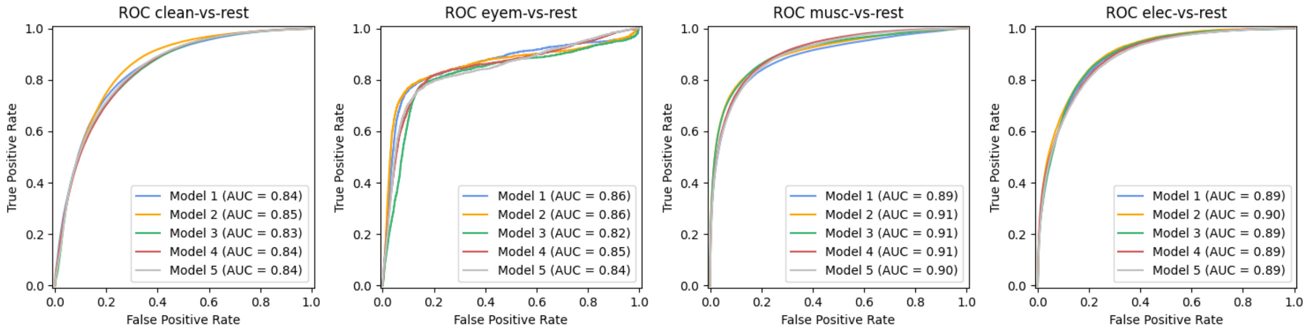
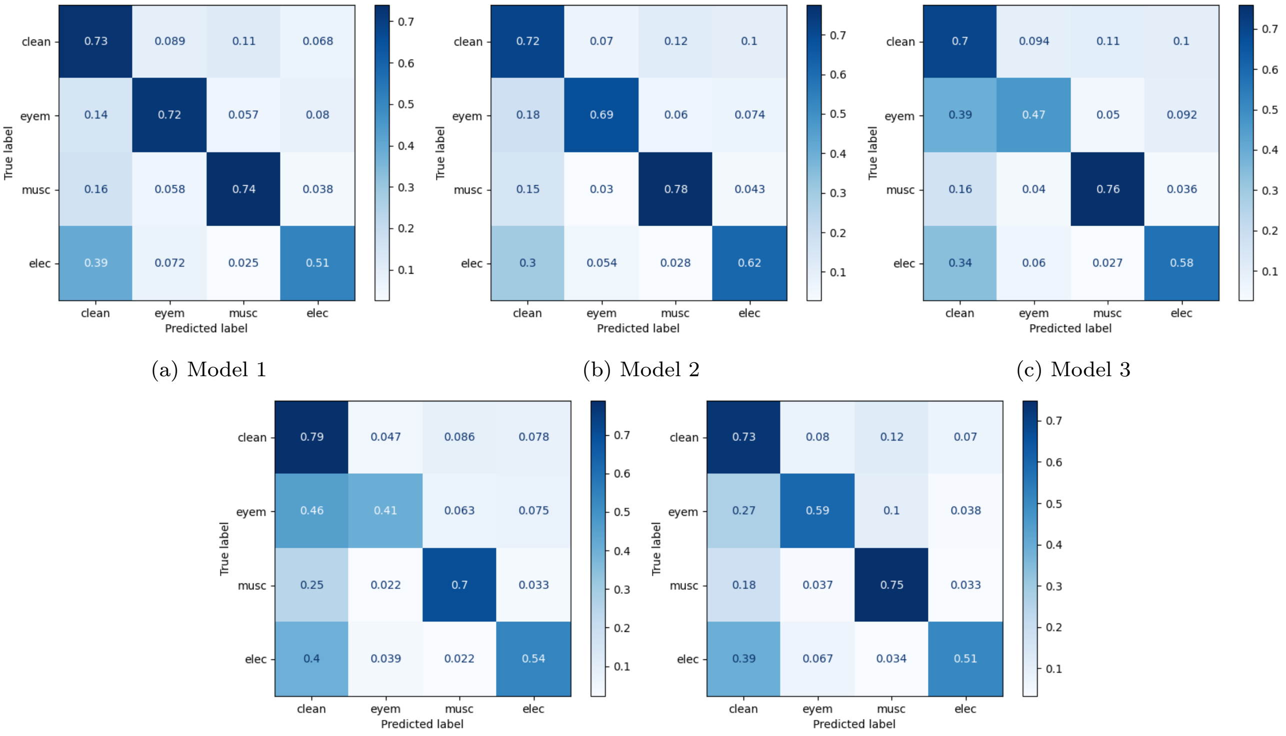


Figure B.1: Receiver operating characteristic (ROC) curve per class (clean, eye movement (eyem), muscle activity (musc), electrode artefact (elec)) one-vs-rest. Model 1-5 are created during five-fold cross validation. Performances are shown on the test set. Area under the ROC curve (AUC).



(a) Model 1

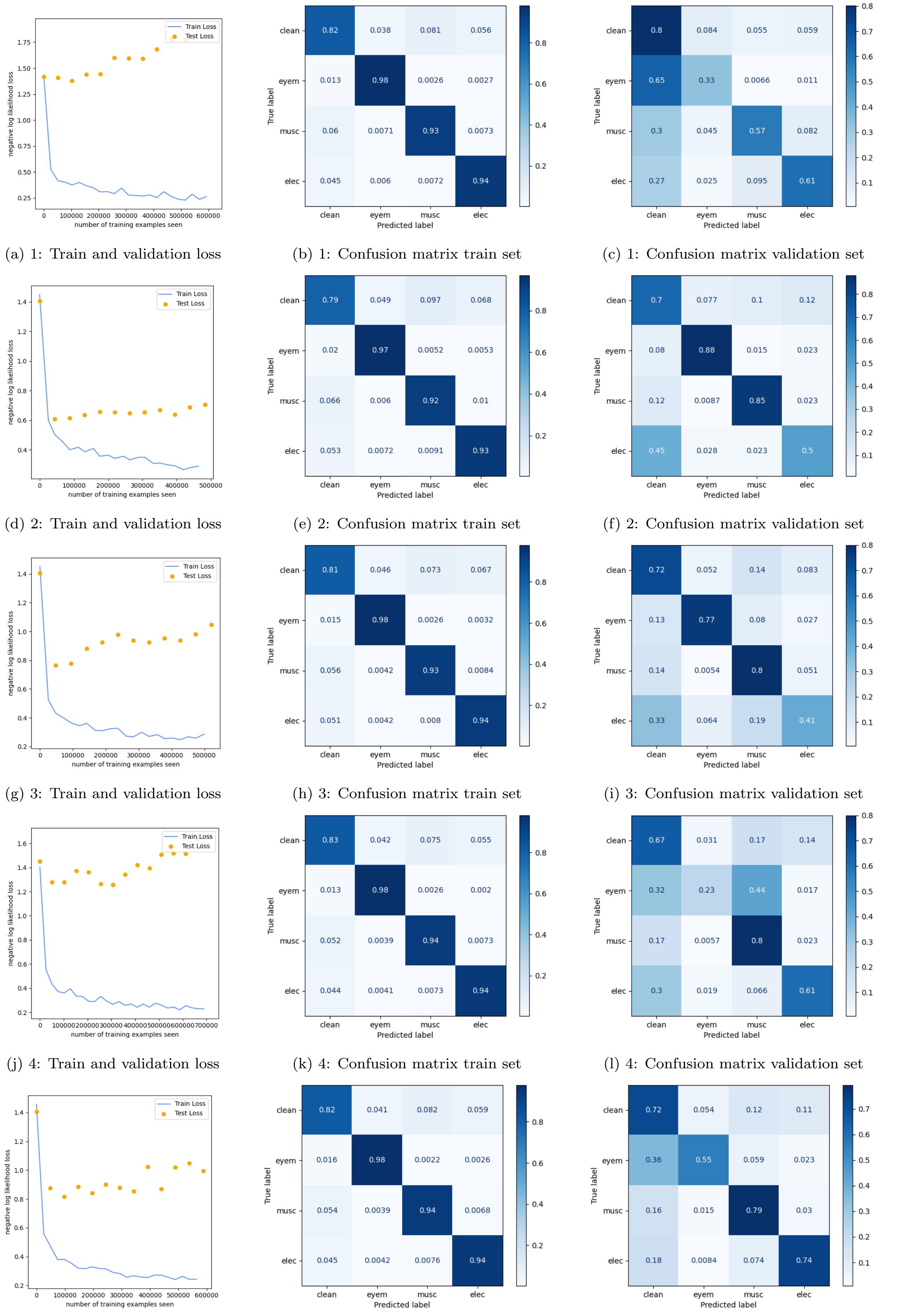
(b) Model 2

(c) Model 3

(d) Model 4

(e) Model 5

Figure B.2: Performance of the test set. Confusion matrices of models created during training with 5-fold cross validation. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (*clean*, eye movement (*eyem*), muscle activity (*musc*), electrode artefact (*elec*)).



(a) 1: Train and validation loss

(b) 1: Confusion matrix train set

(c) 1: Confusion matrix validation set

(d) 2: Train and validation loss

(e) 2: Confusion matrix train set

(f) 2: Confusion matrix validation set

(g) 3: Train and validation loss

(h) 3: Confusion matrix train set

(i) 3: Confusion matrix validation set

(j) 4: Train and validation loss

(k) 4: Confusion matrix train set

(l) 4: Confusion matrix validation set

(m) 5: Train and validation loss

(n) 5: Confusion matrix train set

(o) 5: Confusion matrix validation set

Figure B.3: Performance of the train and validation sets per model created using 5-fold cross validation. From top to bottom: model 1 - 5. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (*clean*, eye movement (*eyem*), muscle activity (*musc*), electrode artefact (*elec*)).

C Cross validation performance - Part II: Amsterdam UMC dataset

Fine-tuning was performed using three fold cross validation. Performance during training can be seen in Fig.C.3 and during testing in Fig.C.2. Comparison between the validation groups and the final test sets, shows the largest performance difference in accurately identifying *pulse* artefacts. During training it accurately classifies the *pulse* artefact in 30-75% of the cases, which is dropped to 7-46% during testing.

The three remaining classes (*clean, musc* and *art*) have an overall good stable performance, which is also seen in the ROC curves of the test set (Fig. C.1). In this figure, it can also be seen that the classification of *pulse* artefacts by the third model is almost as low as chance level (0.50).

Table C.1: Accuracy (ACC) and area under the curve (AUC) metrics for the models created using 3-fold cross validation

Model	Dataset	ACC	AUC
		(%)	
1	Train	81	0.98
	Validation	85	0.95
	Test	70	0.86
2	Train	80	0.92
	Validation	67	0.89
	Test	67	0.81
3	Train	84	0.98
	Validation	77	0.82
	Test	71	0.80

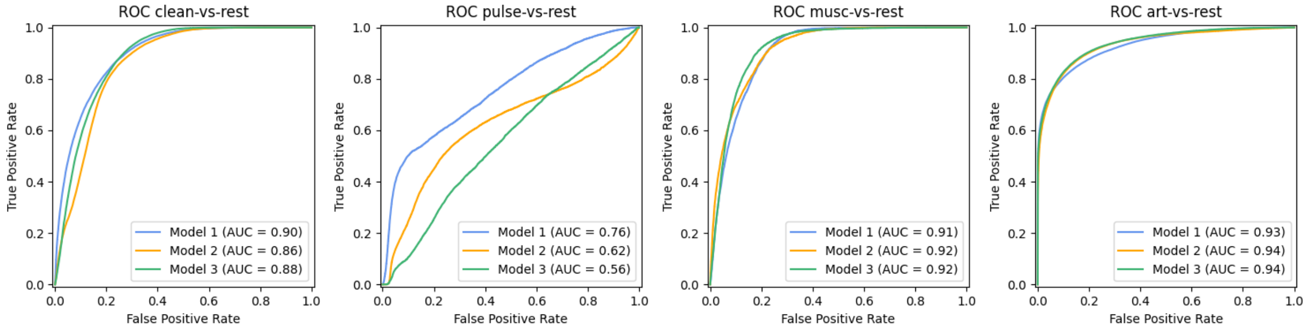
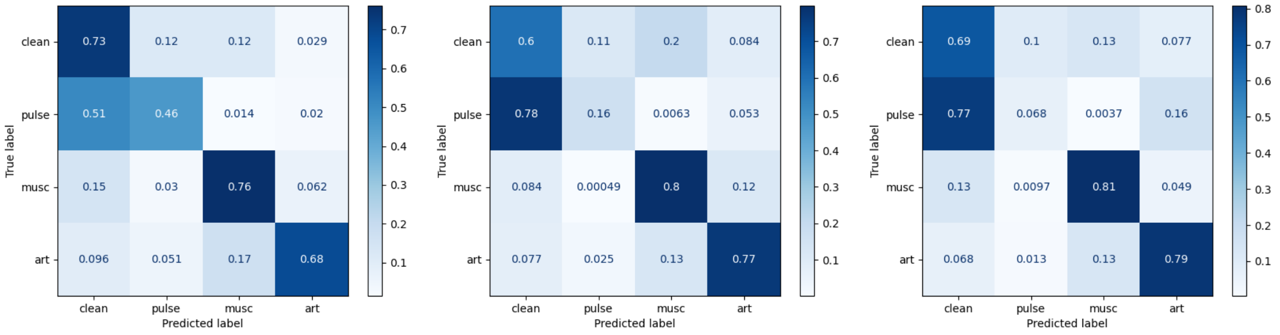


Figure C.1: Receiver operating characteristic (ROC) curve per class (clean, pulsation artefact (pulse), muscle activity (musc), artefact (art)) one-vs-rest. Model 1-3 are created during three-fold cross validation. Performances are shown on the test set. Area under the ROC curve (AUC).

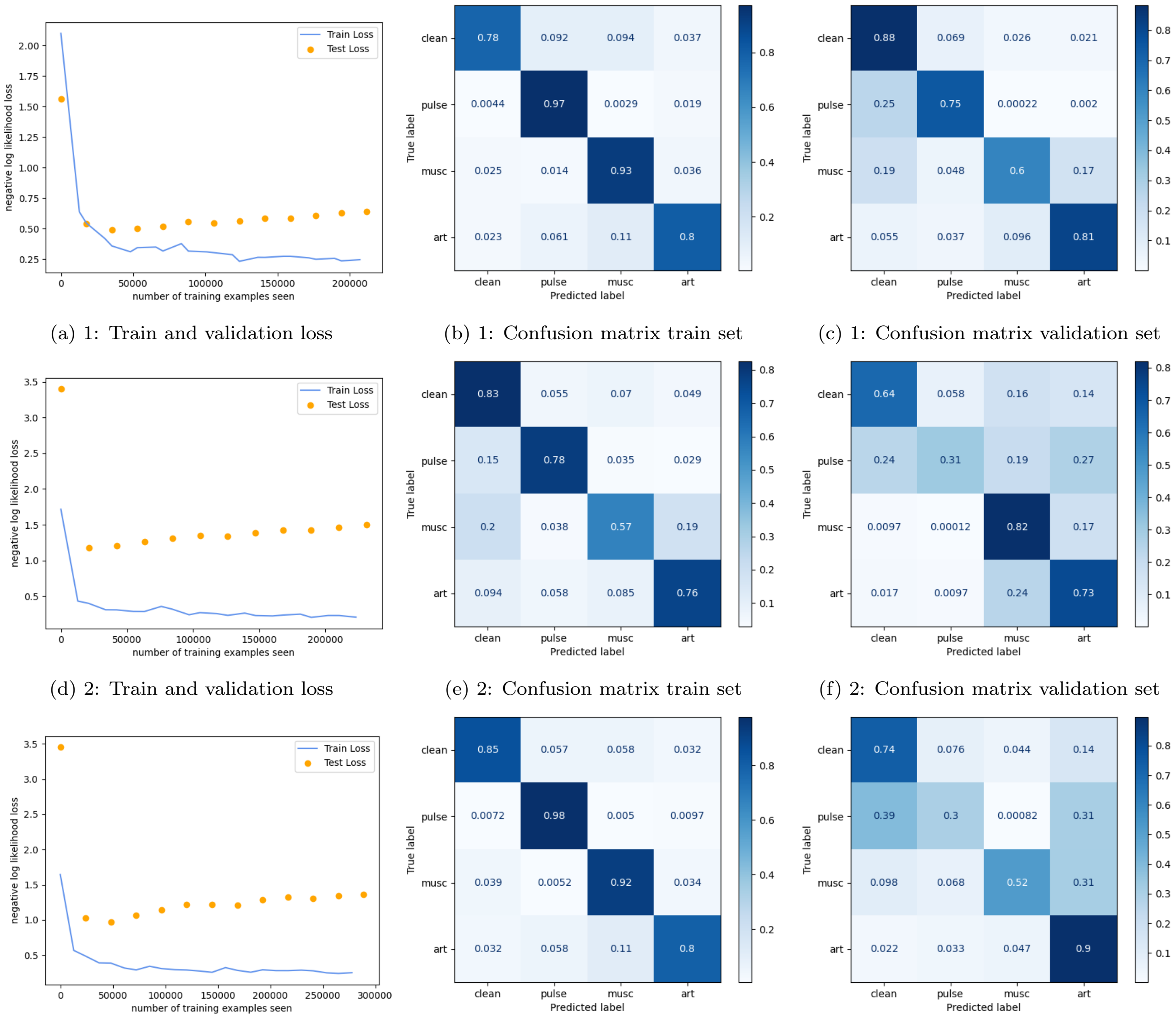


(a) Model 1

(b) Model 2

(c) Model 3

Figure C.2: Performance of the test set. Confusion matrices of models created during training with 3-fold cross validation. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (clean, pulsation artefact *pulse*), muscle activity (*musc*), artefact (*art*)).



(a) 1: Train and validation loss

(b) 1: Confusion matrix train set

(c) 1: Confusion matrix validation set

(d) 2: Train and validation loss

(e) 2: Confusion matrix train set

(f) 2: Confusion matrix validation set

(g) 3: Train and validation loss

(h) 3: Confusion matrix train set

(i) 3: Confusion matrix validation set

Figure C.3: Performance of the train and validation sets per model created using 3-fold cross validation. From top to bottom: model 1 - 3. The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class (clean, pulsation artefact *pulse*), muscle activity (*musc*), artefact (*art*)).

D Class distribution per channel

Class distribution per bipolar channel are displayed. For the model of part I - TUH dataset (Tab. D.1), the included classes are: *clean* segments, eye movement (*eyem*), muscle artefacts (*musc*) and electrode artefacts (*elec*). Data includes the complete dataset, and pred is the prediction based on the test set. It can be seen that for all channels, the class distribution for the data is similar to that of the prediction.

Table D.1: Artefact class distribution per channel of train and test set - TUH dataset. Eye movement (*eyem*), muscle artefact (*musc*), electrode artefact (*elec*).

Channel	Clean		Eyem		Musc		Elec	
	Data	Pred	Data	Pred	Data	Pred	Data	Pred
F7-T3	0.08	0.05	0.17	0.15	0.11	0.11	0.08	0.09
T3-T5	0.08	0.07	0.04	0.04	0.10	0.10	0.08	0.08
F8-T4	0.08	0.05	0.18	0.19	0.10	0.10	0.08	0.09
T4-T6	0.08	0.09	0.04	0.05	0.09	0.09	0.09	0.08
T3-C3	0.08	0.07	0.04	0.05	0.11	0.11	0.09	0.08
C3-CZ	0.08	0.09	0.04	0.04	0.08	0.08	0.10	0.09
CZ-C4	0.08	0.10	0.04	0.05	0.08	0.08	0.10	0.09
C4-T4	0.08	0.08	0.05	0.05	0.09	0.09	0.08	0.09
F3-C3	0.08	0.07	0.18	0.17	0.08	0.08	0.09	0.09
C3-P3	0.09	0.10	0.04	0.04	0.07	0.08	0.08	0.08
F4-C4	0.08	0.08	0.15	0.16	0.07	0.07	0.10	0.10
C4-P4	0.09	0.17	0.00	0.02	0.02	0.01	0.02	0.04

In the Table D.2, the class distribution per channel are displayed for part II - Amsterdam UMC dataset. The included classes are: *clean* segments, pulsation artefacts (*pulse*), muscle artefacts (*musc*) and an artefact class (*art*). The latter label is given to segments, not belonging to any of the other groups.

For the classes *clean*, *musc* and *art* it can be seen that the predictions are in a similar range to the class distribution of the overall dataset. A similar phenomenon can be seen for the *pulse* artefacts, however differences can be seen when data is further split into the train and test set (Fig. D.2 grey highlighted channels). The prediction tends to align more closely with the training class distribution than with the distribution of the test set on which the prediction is based. This is most evidently seen in channels 'FC4-FT8' and 'CP4-TP8'.

Table D.2: Artefact class distribution per channel of train and test set - Amsterdam UMC dataset. Pulsation artefact (*pulse*), muscle artefact (*musc*), artefact (*art*).

Channel	Clean		Pulse		Musc		Art	
	Data *train - test*	Pred	Data *train - test*	Pred	Data *train - test*	Pred	Data *train - test*	Pred
FT7-TP7	0.09 *0.09 - 0.09*	0.08	0.02 *0.02 - 0.00*	0.00	0.20 *0.22 - 0.19*	0.18	0.08 *0.07 - 0.08*	0.06
FC3-CP3	0.08 *0.10 - 0.09*	0.11	0.02 *0.00 - 0.03*	0.00	0.03 *0.04 - 0.02*	0.03	0.09 *0.08 - 0.09*	0.10
FT8-TP8	0.08 *0.07 - 0.09*	0.06	0.12 *0.17 - 0.01*	0.11	0.08 *0.07 - 0.12*	0.14	0.08 *0.10 - 0.08*	0.08
FC4-CP4	0.08 *0.08 - 0.06*	0.09	0.12 *0.07 - 0.25*	0.10	0.03 *0.03 - 0.01*	0.01	0.09 *0.09 - 0.10*	0.10
FC3-FT7	0.08 *0.09 - 0.09*	0.09	0.02 *0.02 - 0.00*	0.00	0.12 *0.13 - 0.09*	0.11	0.09 *0.08 - 0.09*	0.08
CP3-TP7	0.09 *0.10 - 0.11*	0.10	0.04 *0.04 - 0.01*	0.00	0.08 *0.09 - 0.08*	0.10	0.08 *0.07 - 0.07*	0.07
FC4-FT8	0.08 *0.08 - 0.04*	0.09	0.11 *0.05 - 0.27*	0.01	0.06 *0.03 - 0.13*	0.05	0.08 *0.10 - 0.09*	0.10
CP4-TP8	0.09 *0.07 - 0.11*	0.03	0.28 *0.37 - 0.12*	0.61	0.05 *0.05 - 0.03*	0.06	0.07 *0.07 - 0.06*	0.07
FT7-TP7*	0.09 *0.09 - 0.09*	0.08	0.02 *0.02 - 0.00*	0.00	0.20 *0.22 - 0.19*	0.17	0.08 *0.07 - 0.08*	0.06
FC3-CP3*	0.08 *0.10 - 0.09*	0.11	0.02 *0.00 - 0.03*	0.00	0.03 *0.04 - 0.02*	0.02	0.09 *0.08 - 0.09*	0.10
FT8-TP8*	0.08 *0.07 - 0.09*	0.07	0.12 *0.17 - 0.01*	0.07	0.08 *0.07 - 0.12*	0.12	0.08 *0.10 - 0.08*	0.09
FC4-CP4*	0.08 *0.08 - 0.06*	0.10	0.12 *0.07 - 0.25*	0.10	0.03 *0.03 - 0.01*	0.01	0.09 *0.09 - 0.10*	0.10

Channels are grey highlighted that have a significant difference between train and test set in class distribution for pulsation artefacts.

E Explored model architectures

During the development of the final model, multiple options were considered. Two models, which were not used in the end, will be briefly discussed here. Both models were depended on the same input shape (batch size, 12 channels, 200 data points) as the implemented model. Similarly, output was produced like (batch size, 12 channels, 4 classes).

1: 1D CNN

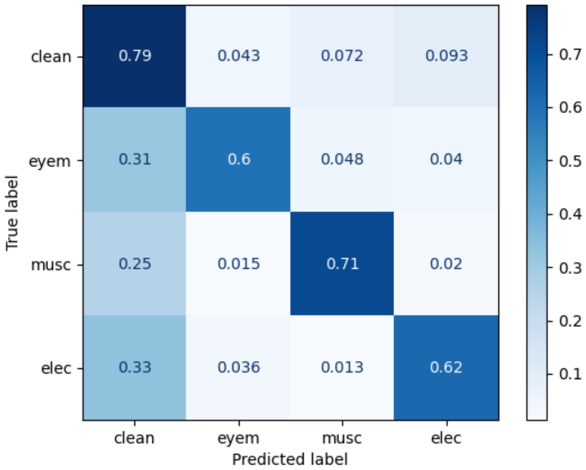


Figure E.1: Confusion matrix of test set using a 1D-CNN for the classes clean, eye movement (eyem), muscle activity (musc), and electrode artefact (elec). The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class.

Table E.1: Optimal parameters implemented in model using 1D-convolutional layers

Hyperparameter	Value(s)
Convolutional layers	6
Batch size	128
Kernel size	5
Dropout probability	0.20
Learning rate	0.001

A 1-dimensional CNN was developed consisting of 6 convolutional layers, followed by two fully connected layers. Each convolutional layers was followed by a batch normalisation and pooling layer. Only incorporating 1D layers allowed to capture temporal dimensions, as well as the place of the channel. However, no spatial dimensions were learned from. Optimal parameters can be seen in Tab. E.1.

2: 2D CNN - with temporal and spatial blocks

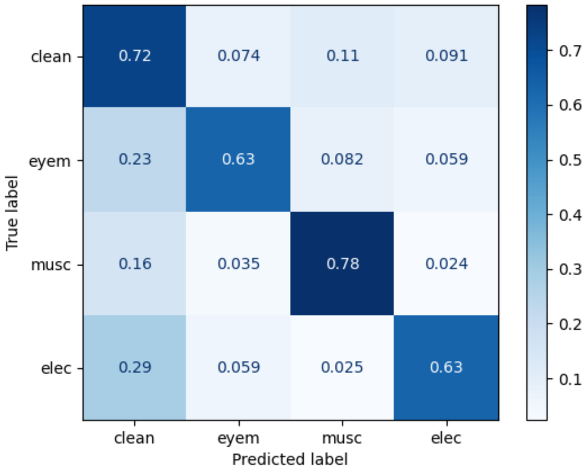


Figure E.2: Confusion matrix of test set using a 2D-CNN with blocks of spatial and temporal layers based on Wang et al. [37] for classes clean, eye movement (eyem), muscle activity (musc), and electrode artefact (elec). The diagonal line represented correct classification. The numbers were calculated as row percentages, indicating the proportion of true labels for each predicted class.

Table E.2: Optimal parameters implemented in model using blocks of temporal and spatial convolutional layers.

Parameter	Value(s)
Convolutional layers	8
Batch size	256
Kernel size width	20
Kernel size height	2
Dropout probability	0.49
Learning rate	0.0001

A second 2D-CNN was developed based on the model created by Wang et al. [37]. Their model used multi-channel input for emotion classification based on EEG recordings. Blocks of 2 convolutional layers, one temporal plus one spatial layer, were added to the network. For the considered model a total of four blocks were used, resulting in a total of 8 convolutional layers. Optimal parameters can be seen in Tab.E.2.

Comparing the performances of the models, similar scores were obtained. Main differences were found in the precision of the *clean* class. The 1D-CNN had a false positives in 25-33% of the cases for the other classes (Fig. E.1. The 2D-CNN with the temporal and spatial convolutional layers performed better, wrongly classifying *clean* in 16-29% of the cases while actually an artefact was present. The implemented model exhibited moderate performance (17-37%). The final choice for the implemented model was based on the incorporation of spatial kernels, the highest recall for the majority classes and overall performance.

Table E.3: Performance comparison of the explored model architectures. Included measures are accuracy (ACC), recall, precision, area under the receiver operating characteristic curve (AUC) and F1-score computed for the overall dataset and per class (*clean*, eye movement (*eyem*), muscle artefact (*musc*), electrode artefact (*elec*).

Model	ACC	Recall	Precision	AUC	F1-score
Class	(%)	(%)	(%)		
1D-CNN	75.0	68.1	74.4	0.86	0.69
Clean	-	79.3	47.0	-	0.59
Eyem	-	59.7	86.2	-	0.71
Musc	-	71.2	84.3	-	0.77
Elec	-	62.0	80.3	-	0.70
2D-CNN temporal + spatial blocks[1]	73.9	69.3	71.2	0.86	0.70
Clean	-	72.5	52.3	-	0.61
Eyem	-	63.0	77.5	-	0.70
Musc	-	78.3	78.9	-	0.79
Elec	-	63.1	79.0	-	0.70
2D-CNN 1 spatial layer[2]	74.8	68.4	71.7	0.87	0.68
Clean	-	76.2	49.8	0.84	0.60
Eyem	-	62.3	80.3	0.85	0.70
Musc	-	76.2	81.0	0.91	0.79
Elec	-	56.0	76.1	0.89	0.64

[1] based on Wang et al. [37], [2] based on Hermans et al. [25]