Online Continual Decoding of Streaming EEG Signal with a Balanced and Informative Memory Buffer

Tiehang Duan[a], Zhenyi Wang[b], Fang Li[d], Gianfranco Doretto[c], Donald Adjeroh[c], Yiyi Yin[a], Cui Tao[d]

[a]*Meta AI, Seattle, WA, 98005, United States*
[b]*University of Maryland, College Park, MD, 20742, United States*
[c]*West Virginia University, Morgantown, WV, 26506, United States*
[d]*University of Texas Health Science Center at Houston, Houston, TX, 77030, United States*

Abstract

Electroencephalography (EEG) based Brain Computer Interface (BCI) systems play a significant role in facilitating how individuals with neurological impairments effectively interact with their environment. In real world applications of BCI system for clinical assistance and rehabilitation training, the EEG classifier often needs to learn on sequentially arriving subjects in an online manner. As patterns of EEG signals can be significantly different for different subjects, the EEG classifier can easily erase knowledge of learnt subjects after learning on later ones as it performs decoding in online streaming scenario, namely catastrophic forgetting. In this work, we tackle this problem with a memory-based approach, which considers the following conditions: 1) subjects arrive sequentially in an online manner, with no large scale dataset available for joint training beforehand, 2) data volume from the different subjects could be imbalanced, 3) decoding difficulty of the sequential streaming signal vary, 4) continual classification for a long time is required. This online sequential EEG decoding problem is more challenging than classic cross subject EEG decoding as there is no large-scale training data from the different subjects available beforehand. The proposed model keeps a small balanced memory buffer during sequential learning, with memory data dynamically selected based on joint consideration of data volume and informativeness. Furthermore, for the more general scenarios where subject identity is unknown to the EEG decoder, aka. subject agnostic scenario, we propose a kernel based subject shift detection method that identifies underlying subject changes on the fly in a computationally efficient manner. We develop challenging benchmarks of streaming EEG data from sequentially arriving subjects with both balanced and imbalanced data volumes, and performed extensive experiments with a detailed ablation study on the proposed model. The results show the effectiveness of our proposed approach, enabling the decoder to maintain performance on all previously seen subjects over a long period of sequential decoding. The model demonstrates the potential for real-world applications.

1. Introduction

Emerging machine learning techniques enable the decoding of brain activities based on electroencephalographic (EEG) recordings, and serve an important role in current BCI systems [1]. The wide applications include different forms of clinical assistance such as autonomous robotic navigation [2], digital interface control assistance of phones and tablets [3], clinical event detection of seizures etc. [4, 5]. It has also been used for entertainment purposes such as gaming control [6] and live interaction between game players [7].

In real world applications of BCI system for clinical assistance and rehabilitation training etc., the EEG decoder often needs to learn on sequentially arriving subjects in an online manner, e.g. 1) the robotic wheelchair at reception desk sequentially hosts different patients during the day for clinical assistance purpose; 2) the gait training and arm training BCI system at rehabilitation center utilized by different patients sequentially in multiple sessions of the day. Given the non-stationarity of the signal and the significant variance in signal patterns across the subjects, the model can easily forget the knowledge of previous subjects after adaptation to new subjects, namely the catastrophic forgetting phenomenon. This makes it difficult for the model to retain knowledge and maintain performance of learnt subjects during online sequential decoding.

An effective approach for solving this issue is to keep samples of previous subjects in a memory buffer and jointly train with current subject. The recently emerged memory-based models in the field of continual learning show promising result in tackling the catastrophic forgetting problem [8, 9]. However the current widely used memory selection approaches such as reservoir sampling may not achieve the desired performance for sequential EEG decoding with imbalanced data volumes and varying decoding difficulty, e.g. some subjects are producing data in significantly higher volumes, for which the data in memory can easily be skewed and only emphasize on certain subjects, or the selected signal segments are not informative enough. We need to always keep data in memory buffer *balanced and representative* for each subject, and this need to be done on the fly with streaming EEG signal. The memory update and replay mechanism should incorporate such requirements and effectively retain knowledge on all previous subjects over a long period of sequential decoding. Additionally, the subject identity and shift boundary may be unavailable during sequential decoding in real world scenarios. The model should work well in this more generalized subject-agnostic scenario.

In this work, we tackle this challenging setting of online sequential EEG decoding with a balanced memory selection and sampling approach. We need to consider the following two aspects for the proposed approach: 1) how to determine the segments of data be moved into memory and the data to be replaced, for which we propose a dynamic memory allocation mechanism based on importance estimation at the cluster level; 2) how to determine which segments to be sampled from memory buffer for replay, for which

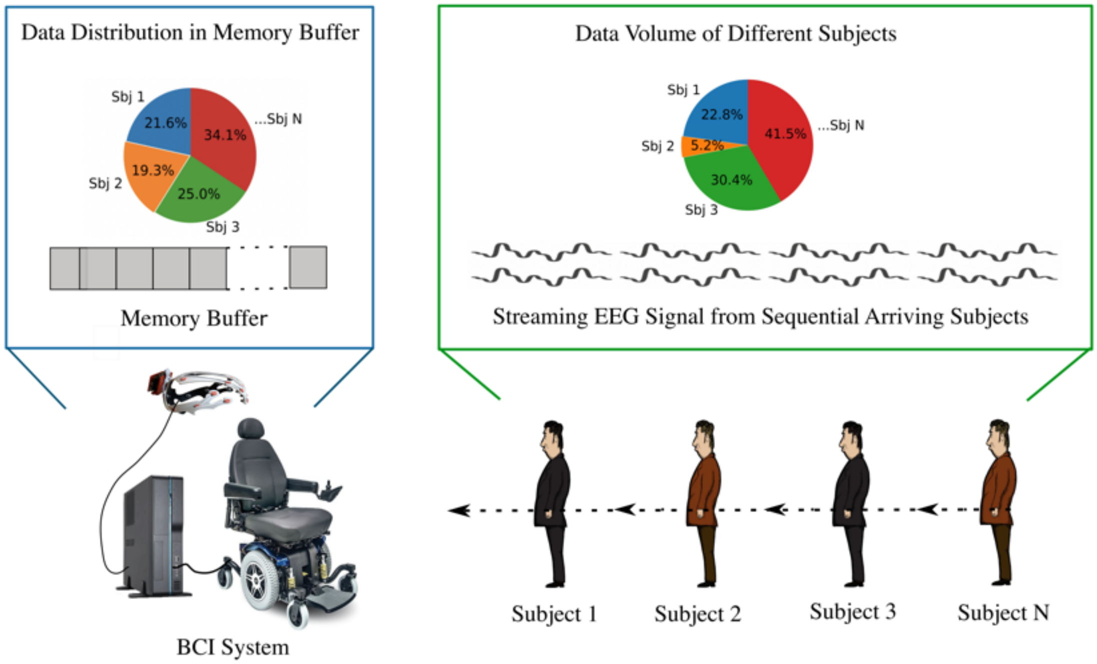


Figure 1: Illustration on the continual EEG decoding scenario, with different subjects arriving in sequential order, producing streaming EEG signal with imbalanced data volume and varying decoding difficulty. The EEG decoder needs to retain knowledge and maintain performance on learnt subjects after adaptation to a new subject. MUDVI keeps a balanced memory buffer and selects informative samples for replay purpose.

we derive an effective sampling approach that reduces gradient estimation variance and increases model convergence speed, this helps to reduce the memory size needed. Additionally, we consider the realistic subject-agnostic setting where subject identity and shift occurrence are unknown to the decoder. We propose a kernel based method on lower dimensional projected space for subject shift detection. The detection method constructs a distance metric that encodes the subject context information over a long period of time and is robust to variance in the EEG signal. Then, the distribution shift is estimated based on the reproducing kernel Hilbert space (RKHS) constructed on adjacent distance metrics.

We develop several different benchmarks to mimic the real world scenario of imbalanced EEG signals from different subjects being sequentially fed into the model. We evaluate model accuracy after sequential learning ends, and also measure the information on backward transfer (BWT). We performed a detailed ablation study on the proposed benchmarks including different imbalance ratios, different subject ordering, and varying number of sequential subjects, etc. This offers an in-depth understanding on model performance, and demonstrates the effectiveness of the proposed approach.

The contributions of this work are as following:

- We propose an effective memory based approach for online sequential EEG decoding over long periods of streaming EEG signal from various sequentially arriving subjects. The model preserves knowledge of previous subjects after learning on later ones.
- We jointly consider the imbalance of data volume and informativeness of recorded signal from different subjects, and design a memory update mechanism that tackles such imbalance issues. An effective memory sampling approach for replay is proposed to increase convergence speed. In addition, we developed a kernel based subject shift detection algorithm, which enables the model to work in both subject-aware and subject-agnostic settings.
- We introduce new benchmarks for model evaluation, with imbalanced data volumes and varying decoding difficulty for the sequential arriving subjects. The benchmarks mimic real world scenarios of BCI system usage.
- We conducted extensive experiments and demonstrated the effectiveness of the proposed method. Our approach achieved significant margin on top of strong baselines. The proposed approach is ready to be integrated into current widely used BCI systems.

2. Related Work

2.1. EEG Classification

Machine learning techniques have been the central part of BCI systems, which have seen rapid progress in the past few years. A continued trend is on adopting deep learning techniques for extracting and decoding EEG signals [10]. Recent works in the field have achieved promising results in terms of accuracy, interpretability and usage in online streaming settings [11][12][13].

For performance improvement, current approaches have explored novel model architectures such as EEGNet[10], CTCNN[14] and CRAM[15], and also proposed domain adaptation and transfer learning based models to cope with differing patterns across subjects, and hence improve the applicability of the models. For example, [16] proposed an inter-subject transfer learning framework built on top of CNN model;[17] tackle the problem of variability across subjects by transferring non-stationary information in the data; [18] and [19] explored performance of multiple domain adaptation methods including transfer component analysis (TCA-EEG), maximum independence domain adaptation (MIDA-EEG) and information theoretical learning (ITL-EEG). [20] proposes a new hard parameter sharing mechanism which enables the model to transfer knowledge across datasets and effectively tackles the cross-dataset EEG classification problem. [21] utilizes Riemannian geometry-based adaptive boosting and voting ensemble algorithm, with the cross-session and cross-subject variations being efficiently represented as Riemannian transformations of the covariance matrices. Different from these models which focus on adapting to future subjects, our approach aims to preserve learned knowledge during sequential EEG decoding and maintain performance on previous subjects to mitigate catastrophic forgetting. It is worth noting that the performance of sequential EEG decoding could be further improved by incorporating neuro-feedback techniques during signal recording stage. Specifically, the provided neuro-feedback assists subjects to control their brain waves consciously [22] and enables the subject to make real time adjustments, therefore improves the quality of EEG recordings [23, 24, 25]. It is a promising direction to explore in future work.

2.2. Continual Learning

Continual learning approaches [26, 27, 28, 29, 30, 31] have been applied to tackle the problem of catastrophic forgetting when sequentially learning across different tasks. Different approaches have been recently proposed to achieve this. These can be classsified into three broad groups: 1) those that utilize a memory buffer to store samples of previous data [32, 33, 34, 8, 35, 36], and then perform sampling in memory buffer for replay and jointly train with current data. To improve the data selection process, Herding dynamic algorithm [37] is applied in ICARL [38] to emphasize on data closer to class means. GDumb [39] proposed a greedy approach on selection of samples from different classes; 2) those that use regularization terms to guide the parameter update process [30, 40, 41, 42]. The regularization approaches can further be divided into data-focused [43] and prior-focused methods [30][44]; and 3) those that explore on expandable network structures for the evolving data [45, 46, 29, 47], with size of the network dynamically increasing as the training progresses.

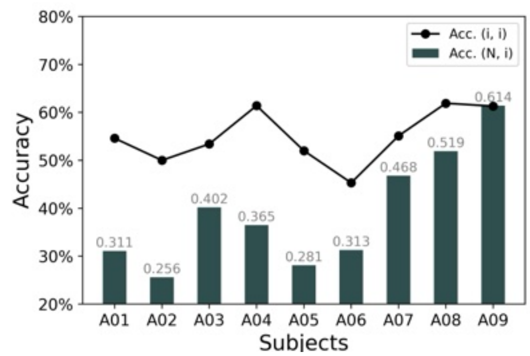


Figure 2: An example sequential EEG decoding trial run on BCI-IV 2a dataset. Acc. (i, i) denotes the accuracy evaluated immediately after learning on each subject, and Acc. (N, i) is the accuracy evaluated on each subject after finished sequential learning on all N subjects. The decoder quickly forgets knowledge on subjects before A07 after sequentially learning on all subjects.

Another related field of research is continual domain adaptation [48][49], which tackles domain adaptation with continuous domain shift. IADA [50] performs continuous adaptation by aligning source and target features through adversarial learning. Volpi et al [51] continuously adapts to sequential visual domains while mitigating the problem of forgetting on previous domains with a meta-learning inspired regularization strategy.

3. Method

This work proposes to preserve learned information on previous subjects during sequential EEG decoding. As illustrated in Fig. 2, the EEG decoder can quickly forget information on subjects before A07 after sequential learning ends with A09, aka catastrophic forgetting. Memory based approaches that keep a small memory buffer to train together with current data have been shown to be relatively effective in solving this problem. Yet, the current widely used memory update methods such as reservoir sampling do not meet the need of challenging scenarios in sequential EEG decoding with imbalanced data volumes and varying decoding difficulty from the different subjects. In this section, we first provide an introduction on the problem setting and conventional reservoir sampling approach. Then, we present details of our proposed memory update and sampling methods.

Definition 1. ***Continual EEG decoding***.
EEG decoder receives streaming EEG signal input from sequentially arriving subjects $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_J$. The model does not have information on subject identity or when the subject shift happens. The EEG decoder needs to preserve knowledge and maintain performance on learnt subjects after adaptation to later subjects during online sequential decoding.

We adopt the memory based approach for tackling this problem with a memory small in size for replay of previous data samples, with the assumption that 1) previous data is not available to revisit unless it is stored in memory buffer, 2) the sequential decoding of EEG signal lasts for a long period of time and the total data volume is much larger than memory buffer capacity. The conventional way to update memory buffer is with reservoir sampling

Reservoir Sampling. For a memory buffer of size M, it will store the first M data points until full. For later batch $\boldsymbol{x}^k$ arriving, it will generate random number i which is in [1, k]. And the sample will be selected and replace the ith data in memory if $i < M$.

With imbalanced data volumes from different subjects, the samples in memory would also be skewed. Furthermore, it is important to select the most informative samples into memory that could bring incremental knowledge to the model. To tackle these challenges, we propose to actively detect subject changes during sequential learning, and perform memory updates jointly considering data volume and informativeness of different subjects accordingly. As illustrated in Fig. 1, with sequentially arriving subjects producing imbalanced data volumes, memory buffer with reservoir sampling would similarly be skewed on $\mathcal{S}_1$ and $\mathcal{S}_3$, and will lack representation of $\mathcal{S}_2$. Our proposed update mechanism jointly considers data volume and informativeness, and creates a balanced representation on all subjects.

3.1. Memory Update Based on Volume and Informativeness

The proposed **M**emory **U**pdate on **D**ata **V**olume and **I**nformativeness model (MUDVI) estimates on the probability of current data $\boldsymbol{x}^t$ moving into memory and also determines the data to be replaced in memory. The update process considers both data volume imbalance and varying decoding difficulty. As illustrated in Fig. 3, for efficiency improvement, we keep a hierarchical architecture in memory where data points are organized into clusters, with each cluster corresponding to a subject. The subject identity is detected with proposed kernel based subject shift detection algorithm for subject-agnostic scenarios, as detailed in Section 3.3. The sample to be replaced in memory is determined first at cluster level based on the following:

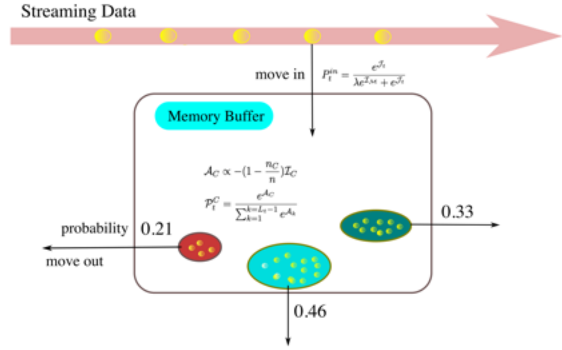


Figure 3: Illustration on the memory update process. The colored circles are clusters formed on different subjects. Data to be moved out of memory is determined first at the cluster level and then sampled in the cluster.

$$\mathcal{A}_C \propto -(1 - \frac{n_C}{n})\mathcal{I}_C \tag{1}$$

$$\mathcal{P}_t^C = \frac{e^{\mathcal{A}_C}}{\sum_{k=1}^{k=L_t-1} e^{\mathcal{A}_k}} \tag{2}$$

where $\mathcal{I}_C$ is the importance score of cluster C defined in Section 3.2. n_C is the number of data samples in cluster C. The data to be replaced in cluster C is sampled based on $-\mathcal{I}_i, i \in C$, with $\mathcal{I}_i$ the gradient norm of x_i. The intuition for Eq. 1 is that the data is more likely to be removed in memory if there exists large number of similar data from same subject in memory, or if the data is less informative.

For current streaming EEG data $\boldsymbol{x}^t$, we compute the probability of it moving in memory. The intuition is to select more informative ones with incremental knowledge to the memory buffer. The average importance of samples in $\mathcal{M}$ is $\mathcal{I}_{\mathcal{M}} = \frac{1}{L_t}\sum_{c=1}^{L_t} \frac{n_C}{n}\mathcal{I}_C$, where L_t is the number of subjects at time step t, n_C is the number of data in cluster C and n is the total number of data in memory. With the gradient norm of $\boldsymbol{x}^t$ being $\mathcal{I}_t$, denote $\mathcal{J}_t = (1 - \frac{n_{L_t}}{n})\mathcal{I}_t$ which jointly considers the prevalence of current subject in memory and the informative level of current data, the probability of moving $\boldsymbol{x}^t$ into memory is

$$P_t^{in} = \frac{e^{\mathcal{J}_t}}{\lambda e^{\mathcal{I}_{\mathcal{M}}} + e^{\mathcal{J}_t}} \tag{3}$$

where λ is a hyperparameter, We provide the complete algorithm on this memory update mechanism in Algorithm 1.

Algorithm 1 Memory Update based on Volume and Informativeness

---

Require: streaming EEG data $\mathcal{D}_t$; maintained memory buffer $\mathcal{M}$; the number of subjects L_t detected at time t.

if subject shift detected $L_t = L_{t-1} + 1$ **then**
  create new cluster in memory $\mathcal{M}_{L_t} = \{\}$
end if
if memory $\mathcal{M}$ has free space **then**
  $\mathcal{M}_{L_t} \leftarrow \mathcal{M}_{L_t} \cup \mathcal{D}_t$
else
  if $\mathcal{D}_t$ is selected into memory based on Eq. 3 **then**
    compute moving-out probability for the clusters
    determine cluster C based on Eq. 1 and 2.
    sample data from $\mathcal{M}_C$ and replace it with $\mathcal{D}_t$
  end if
end if
return memory buffer $\mathcal{M}$

3.2. Adaptive Sampling for Joint Training

During sequential learning, we perform sampling on memory to joint train with each batch of current data. The commonly adopted uniform sampling neglects the varying importance among the samples, and previous work has shown it also introduces unwanted variance which makes training less stable [52]. Here we propose a sampling approach based on the informative level of the data. The approach renders more effective mitigation of catastrophic forgetting, and also reduces gradient estimation variance.

To improve efficiency, the sampling is also at cluster level. A small number of R data is randomly obtained from each cluster for computing cluster importance $\mathcal{I}_C$, which is the average gradient norm of these R representatives. Then the probability to sample on cluster is

$$\mathbf{q}_t^C = \frac{n_C \mathcal{I}_C}{\sum_{k=1}^{k=L_t} n_k \mathcal{I}_k} \tag{4}$$

with L_t being the number of detected subjects at time step t. The proposed approach increases convergence speed and reduces the memory size needed. With the model convergence speed can be effectively represented as

$$V = -\mathbb{E}_{\mathbf{q}_t}[||\boldsymbol{\theta}_t - \boldsymbol{\theta}^*||_2^2 - ||\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*||_2^2] \tag{5}$$

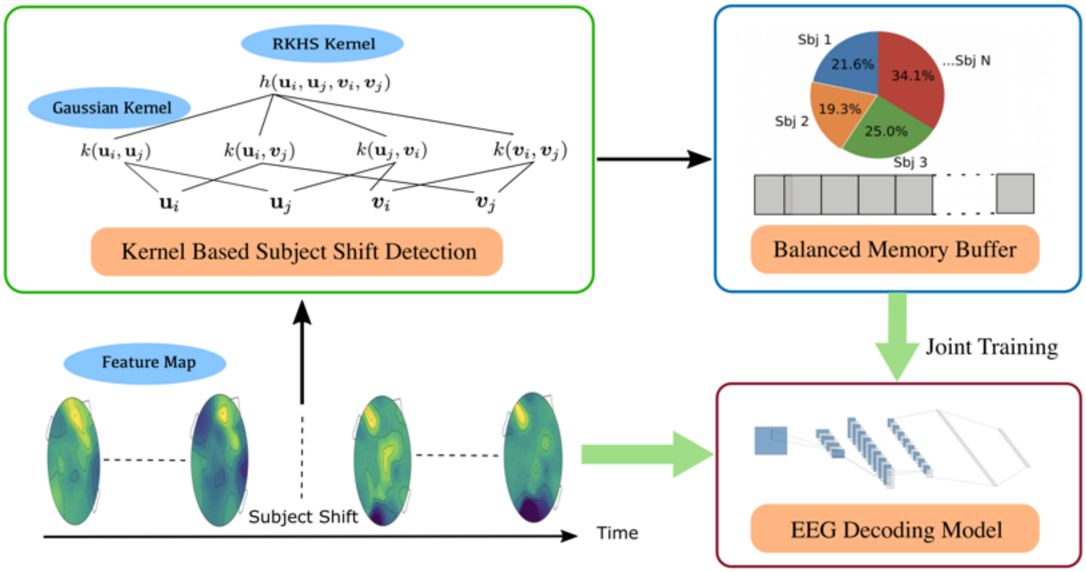


Figure 4: Illustration on the workflow of proposed approach. Kernel based subject shift detection functions on the streaming EEG signal and guides the memory update process. Samples from memory buffer jointly train with current data for forgetting mitigation.

$\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-1}$ are the model parameters of two consecutive time steps, $\boldsymbol{\theta}^*$ is the optimal parameter. $\mathbf{q}_t$ is the data distribution at current time step.

Eq. 5 can be expanded as (proof is available in Appendix A)

$$V = 2\eta(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)\Omega - \eta^2\Omega^T\Omega - \eta^2\mathrm{Tr}(\mathbb{V}_{\mathbf{q}_t}[\Omega]) \tag{6}$$

where η is the learning rate, Ω is the expected gradient $\Omega = \mathbb{E}_{\mathbf{p}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})$. $\mathbb{V}_{\mathbf{q}_t}[\Omega]$ is the covariance matrix on Ω and $\mathrm{Tr}(\mathbb{V}_{\mathbf{q}_t}[\Omega])$ is the trace of the matrix.

From Eq. 6, the convergence speed can be improved by minimizing on $\mathrm{Tr}(\mathbb{V}_{\mathbf{q}_t}[\Omega])$. This is achieved with (proof is available in Appendix B)

$$\mathbf{q}_t^*(\mathcal{D}) = \frac{\mathbf{p}_t(\mathcal{D})||\nabla_{\boldsymbol{\theta}_t}\mathcal{L}_{\boldsymbol{\theta}_t}(\mathcal{D})||_2}{\int_{\mathcal{D}}\mathbf{p}_t(\mathcal{D})||\nabla_{\boldsymbol{\theta}_t}\mathcal{L}_{\boldsymbol{\theta}_t}(\mathcal{D})||_2\, d\,\mathbf{p}_t(\mathcal{D})}. \tag{7}$$

with $\mathbf{p}_t(\mathcal{D})$ being the data distribution in $\mathcal{M}$ and $\mathbf{q}_t^*(\mathcal{D})$ the optimal sampling distribution. We can see our sampling approach in Eq. 4 is an approximation of this distribution.

3.3. Reliable Subject Identity Detection

We propose a reliable subject identity detection algorithm for constructing the clusters in memory. Based on our initial experiment, the naive way of setting a threshold on loss function does not work well in this case given the high variance and pattern difference in the signal. The proposed approach utilizes kernel methods to perform detection in a projected lower dimensional space.

We compute the moving average on the extracted features with $\boldsymbol{e}_t = \alpha \boldsymbol{f}_t + (1-\alpha)\boldsymbol{e}_{t-1}$. And a distance metric $\boldsymbol{d}_t$ with dimensionality being m is computed between $\boldsymbol{f}_t$ and $\boldsymbol{e}$ of previous m steps $d(\boldsymbol{f}_t, \boldsymbol{e}_{t-i})_{i=0:m-1}$. $\boldsymbol{d}_t$ captures context over longer period of time and is robust to variance in EEG signal.

For distributions of two adjacent batches of distance metric $\mathcal{U}$ and $\mathcal{V}$ with $\{\boldsymbol{d}_{t-2B}, \boldsymbol{d}_{t-2B+1}, \cdots, \boldsymbol{d}_{t-B-1}\} \sim \mathcal{U}$, $\{\boldsymbol{d}_{t-B}, \boldsymbol{d}_{t-B+1}, \cdots, \boldsymbol{d}_t\} \sim \mathcal{V}$, here B is the batch size, we utilize maximum mean discrepancy (MMD) to measure the distance between these two distributions.

$$\text{MMD}[\mathcal{U}, \mathcal{V}] := \sup_{f \in \mathcal{F}} \{\mathbb{E}_{\mathbf{u} \sim \mathcal{U}}[f(\mathbf{u})] - \mathbb{E}_{\boldsymbol{v} \sim \mathcal{V}}[f(\boldsymbol{v})]\} \tag{8}$$

Similar to [53], we utilize U-statistics for its estimation

$$\begin{aligned} \delta =& \text{MMD}^2[\mathcal{U}, \mathcal{V}] = \frac{1}{B(B-1)} \sum_{i \neq j}^{B} h(\mathbf{u}_i, \mathbf{u}_j, \boldsymbol{v}_i, \boldsymbol{v}_j) \\ & \text{with } \mathbf{u}_i, \mathbf{u}_j \in \mathcal{U} \text{ and } \boldsymbol{v}_i, \boldsymbol{v}_j \in \mathcal{V} \end{aligned} \tag{9}$$

$h(\cdot)$ is formed with the RKHS kernel as follows

$$h(\mathbf{u}_i, \mathbf{u}_j, \boldsymbol{v}_i, \boldsymbol{v}_j) = k(\mathbf{u}_i, \mathbf{u}_j) + k(\boldsymbol{v}_i, \boldsymbol{v}_j) - k(\mathbf{u}_i, \boldsymbol{v}_j) - k(\mathbf{u}_j, \boldsymbol{v}_i) \tag{10}$$

here $k(\cdot)$ follows a normal distribution

$$k(x, x') = exp(-||x - x'||^2 / 2\sigma^2)$$

Intuitively, the data distribution shift is more significant and more abrupt when subject change happens, comparing to distribution drift within the same subject. Therefore, $\mathcal{U}$ and $\mathcal{V}$ deviate from each other and δ significantly increases when subject identity changes. Hence, a threshold h can be placed on δ to detect a change in subject identity. Each processed data $\boldsymbol{x}^t$ is associated with subject identity L_t, starting from $L_0 = 1$, and the following rules apply

$$\begin{cases} L_t = L_{t-1} \text{ if } \delta_t \leq h \\ L_t = L_{t-1} + 1 \text{ if } \delta_t > h \end{cases} \tag{11}$$

With the signal pattern and variance changing drastically across subjects, choosing a fixed threshold doesn't perform well in our exploration. Here we propose a dynamic scaling mechanism on the threshold. It is reasonable to deem δ_t to have a normal distribution based on [54] and the fact that δ_t converges to a linear combination of normal distributions. (See Appendix C for details). We perform adaptive estimation of its μ_t and σ_t, viz:

$$\begin{cases} \mu_t = (1-\rho)\mu_{t-1} + \rho(\delta_t) \\ \mu_t^{(2)} = (1-\rho)\mu_{t-1}^{(2)} + \rho(\delta_t)^2 \\ \sigma_t = \sqrt{\mu_t^{(2)} - \mu_t^2} \end{cases} \tag{12}$$

And $h = \mu_t + a\sigma_t$ with a being the desired quantile on the distribution.

4. Experiments

In this section, we construct various benchmarks for sequential EEG decoding with imbalanced data on three publicly available datasets, namely, BCI IV-2a [55] [1], SEED dataset [56] [2] and DEAP dataset [57] [3]. We first describe benchmark construction and evaluation metrics. Then, we discuss the detailed experimental setting, followed by in-depth analysis and ablation study on model performance.

4.1. Benchmark

We mimic the real-world scenario of subjects arriving in sequential order with data volumes varying for each subject. We created benchmark datasets based on sequence ordering of subjects, for both data balanced and imbalanced settings. For the imbalanced setting, we randomly select half of the subjects and downsample to 20% of the original volume. Detailed ablation study is conducted on the relationship between model performance and imbalance ratio, different subject ordering etc. The streaming EEG data is processed into batches of $t \times n$, with t the temporal span and n the number of channels. Details on data processing for each dataset are provided in Section 4.3.

4.2. Evaluation Metrics and Baselines

We evaluate the accuracy for all subjects after sequential learning ends to measure the performance in forgetting mitigation. We also measure BWT for information on backward transfer. The definition of BWT is given as follow: BWT $= \frac{1}{N-1}\sum_{i=1}^{N-1} a_{N,i} - a_{i,i}$, with

[1] `http://bnci-horizon-2020.eu`
[2] http://bcmi.sjtu.edu.cn/ seed/downloads.html
[3] `https://www.eecs.qmul.ac.uk/mmv/datasets/deap/download.html`

N the number of subjects, and $a_{j,i}$ the accuracy evaluated on subject i after the model finished sequential learning on subject j. Negative value of BWT reveals the occurrence of catastrophic forgetting after learning the new subject, while positive value shows learning on new subject improves performance of previous subjects.

Table 1: Accuracy evaluated on all subjects after sequential learning ends, with both data imbalance and balanced scenarios. We performed Dunn's post hoc test between MUDVI and baselines, with p-value provided in brackets following the result. Overall, the proposed model shows less deterioration when data becomes imbalanced. This is achieved with the more balanced data kept in memory for the proposed approach.

Dataset	BCI IV-2a Mean + SD (p-value)		DEAP Mean + SD (p-value)		SEED Mean + SD (p-value)	
Method	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
sequential	$32.51_{\pm 1.08}$ (1.08e−11)	$36.40_{\pm 0.51}$ (1.01e−14)	$28.46_{\pm 2.28}$ (6.51e−09)	$32.57_{\pm 0.84}$ (1.28e−12)	$36.21_{\pm 1.45}$ (1.03e−12)	$38.72_{\pm 0.80}$ (4.76e−15)
joint training	$77.54_{\pm 0.61}$ (<1e−20)	$79.63_{\pm 0.47}$ (<1e−20)	$67.82_{\pm 0.79}$ (2.08e−15)	$71.36_{\pm 0.63}$ (2.67e−16)	$82.54_{\pm 0.63}$ (7.81e−16)	$84.15_{\pm 0.42}$ (<1e−20)
EWC	$37.95_{\pm 1.83}$ (1.11e−07)	$42.69_{\pm 1.27}$ (7.87e−09)	$36.24_{\pm 1.51}$ (2.42e−07)	$41.38_{\pm 2.13}$ (4.80e−05)	$45.62_{\pm 1.47}$ (1.14e−10)	$47.49_{\pm 0.81}$ (3.47e−13)
UCB	$36.37_{\pm 0.92}$ (5.30e−11)	$41.43_{\pm 0.62}$ (3.37e−12)	$35.80_{\pm 2.16}$ (2.73e−06)	$39.14_{\pm 1.21}$ (1.45e−08)	$46.40_{\pm 1.18}$ (2.67e−11)	$49.51_{\pm 2.06}$ (5.33e−09)
ER	$39.24_{\pm 1.35}$ (3.59e−08)	$45.96_{\pm 0.83}$ (2.74e−08)	$37.47_{\pm 1.32}$ (5.37e−07)	$42.59_{\pm 1.78}$ (1.31e−04)	$51.75_{\pm 0.43}$ (3.35e−13)	$56.24_{\pm 1.57}$ (3.30e−07)
ER+GMED	$40.39_{\pm 2.17}$ (9.63e−06)	$47.53_{\pm 1.49}$ (1.38e−04)	$39.61_{\pm 0.70}$ (3.59e−07)	$44.34_{\pm 2.13}$ (0.026)	$52.39_{\pm 1.21}$ (6.81e−09)	$57.06_{\pm 2.49}$ (4.63e−05)
MIR	$42.53_{\pm 1.15}$ (2.76e−06)	$48.57_{\pm 0.74}$ (2.74e−05)	$38.34_{\pm 1.49}$ (7.62e−06)	$43.20_{\pm 0.96}$ (5.47e−06)	$55.26_{\pm 0.82}$ (1.80e−08)	$59.87_{\pm 1.64}$ (5.60e−04)
MIR+GMED	$42.80_{\pm 1.71}$ (1.17e−04)	$49.13_{\pm 1.25}$ (0.010)	$40.15_{\pm 2.02}$ (4.26e−03)	$44.71_{\pm 1.82}$ (0.039)	$56.53_{\pm 1.05}$ (2.62e−06)	$61.18_{\pm 1.96}$ (0.051)
MUDVI	**$45.98_{\pm 1.83}$ (−)**	**$50.24_{\pm 1.67}$ (−)**	**$42.29_{\pm 1.81}$ (−)**	**$45.85_{\pm 2.23}$ (−)**	**$59.70_{\pm 1.34}$ (−)**	**$62.31_{\pm 0.72}$ (−)**

Table 2: Model performance evaluated with both subject aware and subject agnostic settings. In general, there is a modest deterioration on model performance when subject identity becomes unknown and needs detection.

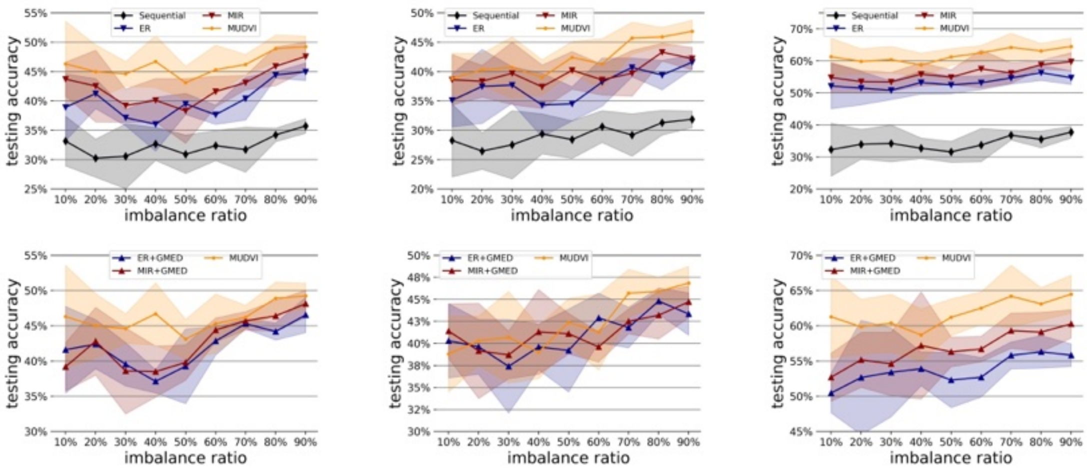
Dataset	BCI IV-2a Mean + SD (p-value)		DEAP Mean + SD (p-value)		SEED Mean + SD (p-value)	
Scenario	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
Subj. Agnostic	$45.98_{\pm 1.83}$ (0.023)	$50.24_{\pm 1.67}$ (0.135)	$42.29_{\pm 1.81}$ (0.005)	$45.85_{\pm 2.23}$ (0.078)	$59.70_{\pm 1.34}$ (0.034)	$62.31_{\pm 0.72}$ (0.305)
Subj. Aware	$47.31_{\pm 1.37}$ (−)	$50.86_{\pm 0.92}$ (−)	$44.15_{\pm 1.66}$ (−)	$46.94_{\pm 1.80}$ (−)	$60.58_{\pm 0.96}$ (−)	$62.43_{\pm 1.25}$ (−)

We compared to widely used strong baselines in the experiment, with details as follows:

1) Top and bottom bound of model performance. The bottom bound is named **classic sequential learning**, where the subjects data arrive sequentially on base model. The top bound is **joint training** with data of all subjects jointly available.

2) The continual learning models that are used for comparison includes regularization-based models, such as **EWC** [30], **UCB** [58], and memory-based models, such as **ER** [8], **MIR** [9] and **GMED** [59]. Elastic weight consolidation (**EWC**) [30] performs dynamic adjustment on learning progress for important parameters. Uncertainty guided continual

Bayesian neural networks (**UCB**) [58] adapt the parameter update speed based on uncertainty in the weights. Experience replay (**ER**) [8] stores a small subset of data from previous tasks with reservoir sampling to train together with current data. Gradient editing in memory (**GMED**) [59] makes stored samples hard to remember and mitigates overfitting. Maximally interfered retrieval (**MIR**) [9] replays examples with larger estimated interference.



(a) BCI IV-2a (b) DEAP dataset (c) SEED dataset

Figure 5: Model performance with different levels of volume imbalance. In this ablation study, we randomly select half of the subjects to perform downsample on data volume in each run, thus incurring higher variance in the result. Each imbalance setting is repeatedly run for 10 times. The sampling ratio is in the range of [10%, 90%]. (a) BCI IV-2a dataset, (b) DEAP dataset, (c) SEED dataset.

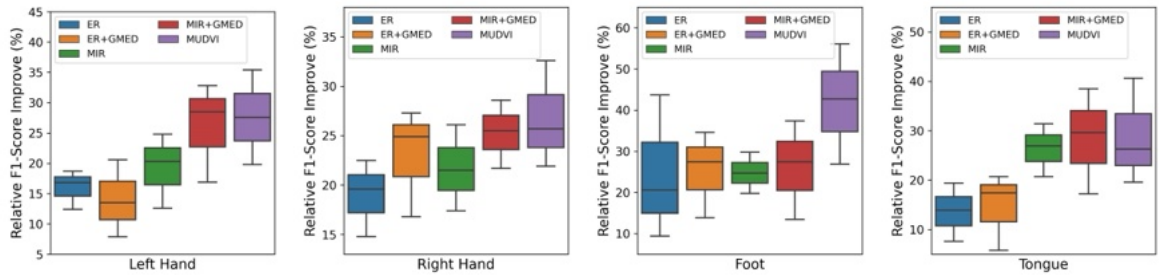
4.3. Settings

Data Processing The processing on each dataset is as follows:

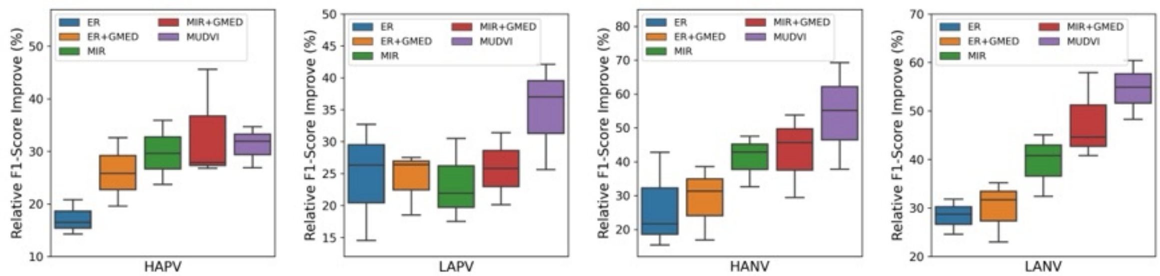
1) BCI IV-2a: Each trial is divided into segments of size 400×22, with 22 channels and temporal span of 400. The step size is 50 for adjacent segments. In each trial, the period between $t = 3s$ and $t = 6s$ is extracted for decoding. This produces 8 data samples per trial with 250Hz sampling rate.

2) DEAP dataset: Trials are processed into 768×32 segments, with adjacent segments having a step size of 128. We used the last 50 seconds of each trial. With downsampling rate of 128Hz this produces 45 segments per trial.

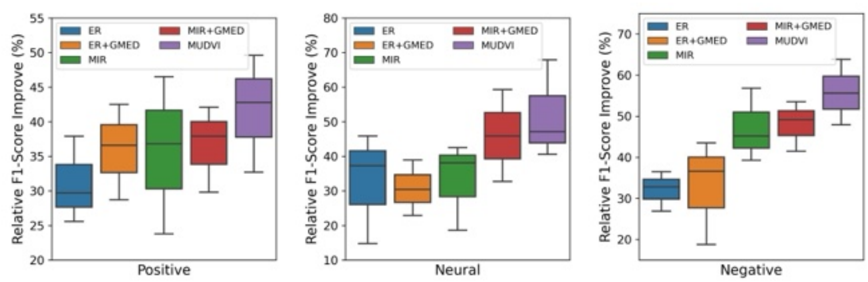
3) SEED dataset: Trials are separated into segments of 800×62 with a step size of 100. The data is downsampled to 200Hz, producing 472 segments per trial.



(a) BCI IV-2a dataset



(b) DEAP dataset



(c) SEED dataset

Figure 6: Relative improvement of F1-score compared to the baseline of sequential learning on base decoder. The rows from top to bottom shows performance on individual tasks of the three datasets respectively.

Model Settings

We used a single Titan-V GPU for model training. The base model is a three layer convolutional neural network. The first layer performs temporal convolution for frequency information extraction, with filter size being $(1, C)$, C is the number of channels. The second layer performs depthwise convolutions with temporal specific spatial filters, with filter size being $(2, 32)$ across all three datasets. Third layer performs pointwise convolution operations for improvement on computational cost. Zero padding is added between neighboring layers to keep the data dimensionality. We set the hyperparameter λ to be 1.

Further performance improvement is possible with additional fine tune on λ. We set $R = 6$ for estimation of cluster importance in memory. The context window size of distance metric is $m = 8$ by default, with detailed analysis in Appendix G. In terms of subject shift detection, the averaging factor ρ is set to 0.2 by default, with significant level $a = 1.96$ and confidence interval of 95%. For our problem setting, training data from the subjects is not available beforehand. We do not use pre-trained feature extractor and the model learns sequential arriving subjects on the fly.

4.4. Analysis on Balanced vs. Imbalanced Data

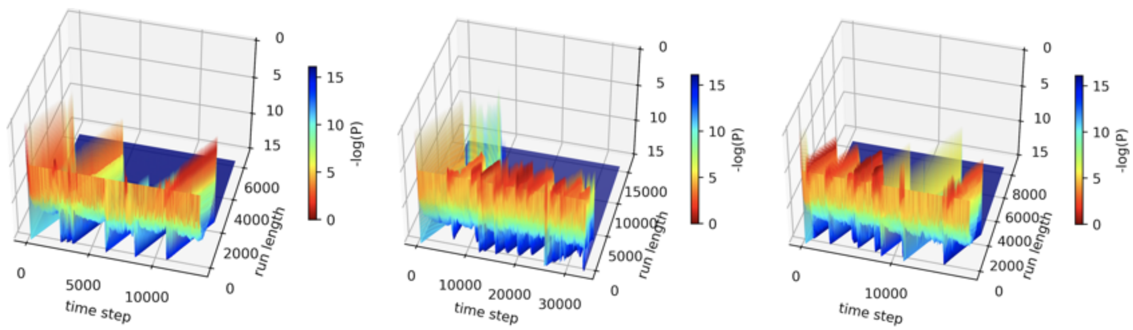
Result on performance comparison is shown in Table 1. We performed comparision for both data balanced and imbalanced settings. Overall, the proposed model shows less deterioration when data becomes imbalanced. For BCI IV-2a dataset, the accuracy of proposed approach reduced by 4.26% after data becomes imbalanced, while memory based comparison models have a reduction of at least 6.04%. Similar improvements are observed for DEAP and SEED datasets. For data imbalanced scenario, the proposed model has a margin of at least 3.18% on BCI IV-2a dataset, 2.14% on DEAP dataset and 3.68% on SEED dataset over baselines in terms of testing accuracy. This is achieved with the more balanced data kept in memory for the proposed approach. Currently the performance gap between online sequential decoding and the top bound of joint training is still large, given the fact that online sequential decoding setting is significantly more challenging in the following two aspects: 1) the inter-subject variability causing the EEG decoder to lose information of previous subjects after learning on later subjects; 2) with no data jointly available for pre-training, the knowledge learned from previous subjects doesn't readily fit to future subjects. Further performance improvement is needed in future work.

4.5. Analysis on Subject Aware vs. Subject Agnostic

We performed comparison on model performance with the subjects identity either known or unknown to the decoding model, with the former referred to as subject-aware and latter referred to as subject-agnostic. The subject-agnostic setting is more challenging for balanced memory data selection and we performed detection on underlying subject distribution shift. The performance comparison between the two scenarios is presented in Table 2. In general, there is a modest deterioration in terms of model performance when subject identity becomes unknown and distribution shift needs to be detected. For example, BCI IV-2a sees the performance reduced by 1.33% and 0.62% for imbalanced and balanced data respectively. DEAP dataset shows higher sensitivity towards subject identity, with performance deteriorating by 1.86% for imbalanced setting and 1.09% for balanced setting. On SEED dataset, the model performance has a slight decrease of 0.67% and 0.29% for imbalanced and balanced settings respectively. We performed detailed analysis of the proposed subject shift detection algorithm in ablation study.

Table 3: Influence of memory size on model performance. The performance of forgetting mitigation is improved with larger memory size. MUDVI needs a smaller memory to achieve comparable performance to other memory-based baselines.

Dataset	BCI IV-2a			DEAP			SEED		
Memory Size	100	200	500	100	200	500	100	200	500
ER	37.47±0.69	39.24±1.35	47.60±0.81	34.65±0.83	37.47±1.32	45.23±1.60	46.02±0.76	51.49±1.31	56.74±0.63
ER+GMED	38.82±1.61	40.39±2.17	47.52±1.29	35.92±1.59	39.61±0.70	46.49±0.98	47.85±1.34	52.66±2.18	58.17±1.72
MIR	38.33±1.34	42.53±1.15	48.87±2.08	36.17±2.26	38.34±1.49	47.04±1.32	48.30±1.78	55.51±0.74	60.54±1.29
MIR+GMED	39.58±0.70	42.80±1.71	51.13±1.54	37.94±1.30	40.15±2.02	47.96±2.19	50.43±2.49	56.17±0.40	61.15±2.16
MUDVI	43.14±1.49	45.98±1.83	52.05±2.36	39.53±1.35	42.29±1.81	49.80±1.43	52.71±2.64	59.85±1.63	63.29±0.98



(a) BCI IV-2a (b) Deap Dataset (c) Seed Dataset

Figure 7: visualization of -log(P) on run length with respect to the time steps in subject shift detection trial runs. The majority of subject shift occurrences match with a new peak along the run length axis

4.6. Ablation Study

Evaluation on Memory Size

The influence of memory size on the proposed method is evaluated in Table 3. The performance on forgetting mitigation is improved with larger memory size. We observed the proposed method is able to achieve comparable performance as baselines with less data, e.g. MUDVI with memory size of 100 achieved similar performance as baselines using 200 data points. We used memory size of 200 as default setting in our experiments.

Subject Shift Detection

We performed analysis on the performance of the subject shift detection algorithm. For each trial of experiment, the number of shifts is the number of subjects minus 1. Running the subject shift detection algorithm for 10 times on each dataset, it successfully detected 58 of the 80 shift occurrences for BCI IV-2a dataset, 196 of the 310 shift occurrences for DEAP dataset and 117 out of the 140 shift occurrences for SEED dataset. This implies accuracy of 72.5% on BCI IV-2a, 63.2% on DEAP and 83.6% on SEED. The task of shift detection is in general easier for SEED dataset, probably due to its data being less noisy.

With the assumption of normal distribution on δ, we can plot out the probability of run length at each time step based on the estimated adaptive mean and variance. Example plot of a trial run for each dataset is shown in Fig. 7. We observed the majority of subject shift occurrences match with a new peak along the run length axis, indicating that the algorithm successfully detected the underlying subject shift.

Performance with Different Memory Selection Functions

Alternative memory selection functions exist for balanced data selection in addition to the proposed approach. In this section, we perform ablation study on the different memory selection approaches and their influence on model performance, including $\mathcal{A}_C^{(1)} = -\frac{\mathcal{I}_C}{(n_C+1)}$ and $\mathcal{A}_C^{(2)} = -\frac{\mathcal{I}_C}{exp(n_C)}$. These two alternatives offer different levels of emphasis between volume balance and informativeness of data. The comparison result is shown in Table 5. We observed $\mathcal{A}_C^{(2)}$ suffers modest performance deterioration. $\mathcal{A}_C$ achieves the best performance among the three options, with proper balance between data volume and its informativeness.

Table 4: The effect of different subject ordering on model performance. We study three different ordering scenarios, 1) sequential order based on subject id, 2) ascending order based on decoding accuracy, 3) descending order based on decoding accuracy.

Dataset	Scenario	EWC	UCB	ER	ER+GMED	MIR	MIR+GMED	MUDVI
BCI IV-2a	Sequential	37.95±1.83	36.37±0.92	39.24±1.35	40.39±2.17	42.53±1.15	42.80±1.71	45.98±1.83
	Ascending	39.46±2.54	37.53±1.32	39.58±2.61	41.40±1.23	44.72±1.96	43.49±1.05	47.20±1.26
	Descending	37.63±1.27	35.81±2.45	38.26±0.79	40.71±2.08	41.47±0.84	42.56±0.90	45.63±1.62
DEAP	Sequential	36.24±1.51	35.80±2.16	37.47±1.32	39.61±0.70	38.34±1.49	40.15±2.02	42.29±1.81
	Ascending	36.81±1.72	35.98±1.41	38.13±1.13	38.92±1.45	39.15±0.68	42.31±1.67	42.87±1.04
	Descending	35.18±0.96	34.23±1.15	36.10±1.84	39.46±1.39	38.26±1.72	40.74±1.43	41.64±1.49
SEED	Sequential	44.83±1.08	46.32±1.54	51.49±1.31	52.66±2.18	55.51±0.74	56.17±0.40	59.85±1.63
	Ascending	44.61±2.14	46.67±1.89	52.63±0.75	53.82±1.29	56.14±1.60	56.52±1.29	60.56±2.17
	Descending	45.59±1.35	45.54±0.93	51.06±1.82	52.31±1.74	55.27±1.13	55.74±0.95	59.52±1.21

Performance with Different Levels of Volume Imbalance

We explored model performance with different levels of volume imbalance. We randomly select half of the sequential subjects to perform downsampling on data volume, with downsampling ratio ranges between [10%, 90%]. The result is shown in Fig. 5. We observed the model performance gradually increases with the increase of downsampling ratio for DEAP and SEED datasets. For BCI IV-2a dataset, the performance first decreases and then increases with the increase of sampling ratio. A possible reason is that lower sampling ratio makes the sequence shorter and easier to remember, while higher sampling ratio offers more data for the model to learn and also keeps data more balanced.

Effect of Subject Ordering on Model Performance

Different subjects have varying decoding difficulty and also shares different levels of similarity with each other. The arriving order of sequential subjects thus could influence model performance. In this section we perform ablation study on three different subject ordering scenarios: 1) sequential order based on subject id, 2) ascending order based on decoding accuracy of individual subjects and 3) descending order on decoding accuracy of the individual subjects. The result is summarized in Table 4. With variances exist in the performance, in general, ascending ordering of subjects based on decoding accuracy achieved better performance than the other two scenarios.

Table 5: Model performance with different memory selection functions. The two alternatives $\mathcal{A}_C^{(1)}$ and $\mathcal{A}_C^{(2)}$ offers different levels of emphasis on volume balance.

Dataset	Func.	ER	ER+GMED	MIR	MIR+GMED	MUDVI
BCI IV-2a	$\mathcal{A}_C$	$39.24_{\pm 1.35}$	$40.39_{\pm 2.17}$	$42.53_{\pm 1.15}$	$42.80_{\pm 1.71}$	$45.98_{\pm 1.83}$
	$\mathcal{A}_C^{(1)}$	$37.88_{\pm 2.37}$	$39.75_{\pm 1.24}$	$41.68_{\pm 1.52}$	$42.26_{\pm 0.94}$	$43.39_{\pm 1.20}$
	$\mathcal{A}_C^{(2)}$	$37.04_{\pm 0.98}$	$38.91_{\pm 0.71}$	$40.34_{\pm 2.29}$	$41.15_{\pm 2.33}$	$42.52_{\pm 2.14}$
DEAP	$\mathcal{A}_C$	$37.47_{\pm 1.32}$	$39.61_{\pm 0.70}$	$38.34_{\pm 1.49}$	$40.15_{\pm 2.02}$	$42.29_{\pm 1.81}$
	$\mathcal{A}_C^{(1)}$	$36.65_{\pm 2.58}$	$38.17_{\pm 1.73}$	$37.42_{\pm 2.48}$	$39.34_{\pm 3.50}$	$41.63_{\pm 2.44}$
	$\mathcal{A}_C^{(2)}$	$35.23_{\pm 1.66}$	$38.68_{\pm 0.49}$	$37.10_{\pm 1.14}$	$38.25_{\pm 1.43}$	$39.92_{\pm 2.19}$
SEED	$\mathcal{A}_C$	$51.49_{\pm 1.31}$	$52.66_{\pm 2.18}$	$55.51_{\pm 0.74}$	$56.17_{\pm 0.40}$	$59.85_{\pm 1.63}$
	$\mathcal{A}_C^{(1)}$	$50.28_{\pm 3.05}$	$52.15_{\pm 1.74}$	$53.72_{\pm 1.39}$	$54.40_{\pm 2.62}$	$56.52_{\pm 3.24}$
	$\mathcal{A}_C^{(2)}$	$50.01_{\pm 1.76}$	$50.89_{\pm 3.32}$	$53.26_{\pm 2.17}$	$53.68_{\pm 1.31}$	$57.41_{\pm 1.95}$

Performance on Individual Decoding Tasks

We performed analysis on model performance towards the individual tasks, as illustrated in fig. 6. The model shows different levels of effectiveness in terms of relative improvement on F1 score for the individual recognition tasks. For example, the model is more effective on decoding foot movements compared to other approaches in terms of the BCI IV-2a motor imagery tasks, and it performs better in detecting negative valence (HANV and LANV) emotions for the emotion recognition tasks of DEAP dataset.

Effect of Offline Pre-training on Model Performance

Pre-training offline on a disjoint set of subjects before online sequential decoding offers a nice tradeoff between online sequential decoding from a cold start and joint training with all data available. We explore the effectiveness of offline pre-training on performance improvement. We allocate 1/3 of the subjects in each dataset for pre-training purpose, and test the model performance of sequentially decoding through the rest 2/3 of subjects. Specifically, first 3 subjects in BCI IV-2a, first 10 subjects in DEAP and first 5 subjects in

Table 6: Comparison on running time of MUDVI and other replay-based baselines.

Methods	BCI IV-2a (sec)	DEAP (min)	SEED (min)
ER	120.2	13.4	96.7
ER+GMED	167.4	18.6	158.4
MIR	143.5	17.2	124.9
MIR+GMED	206.9	23.8	225.2
MUDVI (Sbj. Aware)	139.7	18.2	145.6
MUDVI (Sbj. Agnostic)	243.8	26.3	211.0

SEED are utilized for pre-training. For a fair comparison, the cold start scenario in this setting also perform sequential decoding on the rest 2/3 of subjects. The result is shown in Table 7.

Table 7: Evaluation of the effect of pre-training on model performance. We compare the two different scenarios of online sequential decoding either from a cold start or pre-trained on another disjoint set of subjects.

Scenarios	Dataset	BCI IV-2a		DEAP		SEED	
	Method	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
Cold Start	ER	$43.57_{\pm 1.72}$	$48.24_{\pm 0.95}$	$39.73_{\pm 1.14}$	$43.82_{\pm 2.09}$	$54.15_{\pm 0.81}$	$57.62_{\pm 1.26}$
	MIR	$45.41_{\pm 1.20}$	$49.93_{\pm 1.78}$	$40.85_{\pm 2.38}$	$45.15_{\pm 0.71}$	$56.87_{\pm 1.43}$	$62.21_{\pm 0.38}$
	MUDVI	$48.03_{\pm 1.54}$	$52.46_{\pm 1.37}$	$43.79_{\pm 1.63}$	$46.34_{\pm 1.86}$	$59.65_{\pm 1.08}$	$64.36_{\pm 0.89}$
Pre-trained	ER	$57.31_{\pm 0.96}$	$61.72_{\pm 1.40}$	$50.48_{\pm 1.14}$	$52.90_{\pm 1.67}$	$70.12_{\pm 0.65}$	$71.97_{\pm 1.23}$
	MIR	$59.39_{\pm 0.51}$	$61.97_{\pm 0.83}$	$52.16_{\pm 0.40}$	$53.35_{\pm 1.72}$	$72.08_{\pm 0.84}$	$73.20_{\pm 0.51}$
	MUDVI	$61.22_{\pm 1.14}$	$63.15_{\pm 0.62}$	$54.21_{\pm 2.39}$	$55.04_{\pm 1.25}$	$73.33_{\pm 1.27}$	$73.86_{\pm 0.79}$

Computation Cost

We evaluated the computational cost of the proposed approach with detailed comparison to baselines in Table 6. Compared to other replay-based methods, MUDVI adds a small overhead with the update of cluster metrics such as cluster level importance, at the same time it reduces the sampling cost from $O(N)$ to $O(L_t + N_C^{avg})$, with N_C^{avg} the average number of data samples per cluster. Please note with each time step, cluster metric update only takes $O(1)$ in terms of time complexity as we only need to update the changed clusters. It takes an additional $O(L_t)$ space to store the cluster information, with $L_t \ll n$ in most cases. The model doesn't involve additional forward and backward propagation in its functionality. Its overall running time is comparable to GMED and MIR for subject aware settings. For subject agnostic settings, we also take computational efficiency into consideration with the subject shift detection algorithm. Specifically, the detection is performed with the projected low dimensional distance metric d_t and its computational overhead is significantly less than operations directly on feature space. Further improvement is possible if detection is performed with stride size larger than 1 along the time axis. We also expect more speed up on running time with more optimized implementation.

5. Conclusion

In this work, we proposed an effective memory based method for continual decoding of streaming EEG signal from sequential arriving subjects with consideration on data imbalance issues. The proposed memory update and sampling approach jointly consider the volume from different subjects and data informativeness, keeping a balanced memory for replay purposes. We design the memory sampling approach following the distribution that maximizes convergence speed and reduces memory size. We constructed challenging benchmarks for sequential EEG decoding with imbalanced data on top of numerous widely used datasets covering different BCI paradigms, and conducted extensive analysis compared to related strong baselines. The model achieved significant improvement in numerous different ablation scenarios. The directions of future work include 1) exploration on sequential EEG decoding of heterogeneous classes, 2) methods for sequential EEG classification without usage of memory buffer.

Appendix A. Derivation of Convergence Speed

Denote the convergence speed as $S = -\mathbb{E}_{P_t}\left[\|\theta_{t+1} - \theta^*\|_2^2 - \|\theta_t - \theta^*\|_2^2\right]$ and with Ω the expected gradient of one time step, we have

$$\begin{aligned}
S =& -\mathbb{E}_{P_t}\left[\left(\theta_{t+1} - \theta^*\right)^T\left(\theta_{t+1} - \theta^*\right) - \left(\theta_t - \theta^*\right)^T\left(\theta_t - \theta^*\right)\right] \\
=& -\mathbb{E}_{P_t}\left[\theta_{t+1}^T\theta_{t+1} - 2\theta_{t+1}\theta^* - \theta_t^T\theta_t + 2\theta_t\theta^*\right] \\
=& -\mathbb{E}_{P_t}\left[\left(\theta_t - \eta\Omega\right)^T\left(\theta_t - \eta\Omega\right) + 2\eta\Omega^T\theta^* - \theta_t^T\theta_t\right] \\
=& -\mathbb{E}_{P_t}\left[-2\eta\left(\theta_t - \theta^*\right)\Omega + \eta^2\Omega^T\Omega\right] \\
=& 2\eta\left(\theta_t - \theta^*\right)\mathbb{E}_{P_t}\left[\Omega\right] - \eta^2\mathbb{E}_{P_t}\left[\Omega\right]^T\mathbb{E}_{P_t}\left[\Omega\right] - \\
& \eta^2\,\mathrm{Tr}\left(\mathbb{V}_{P_t}\left[\Omega\right]\right)
\end{aligned} \tag{A.1}$$

Appendix B. Theoretical Proof on Sampling Distribution

$$\begin{aligned} g &= \mathbb{E}_{\mathbf{p}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) \\ Tr(\mathbb{V}_{\mathbf{q}}[\Omega]) &= \mathbb{E}_{\mathbf{q}(\mathcal{D})}[(\frac{\mathbf{p}(\mathcal{D})}{\mathbf{q}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) - g)(\frac{\mathbf{p}(\mathcal{D})}{\mathbf{q}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D}) - g)^T] \\ &= \mathbb{E}_{\mathbf{q}(\mathcal{D})}[||\frac{\mathbf{p}(\mathcal{D})}{\mathbf{q}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2^2] - ||g||_2^2 \end{aligned} \tag{B.1}$$

By Jensen's inequality:

$$\mathbb{E}_{\mathbf{q}(\mathcal{D})}[||\frac{\mathbf{p}(\mathcal{D})}{\mathbf{q}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2^2] \geq \mathbb{E}_{\mathbf{q}(\mathcal{D})}[||\frac{\mathbf{p}(\mathcal{D})}{\mathbf{q}(\mathcal{D})}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2]^2 = (\mathbb{E}_{\mathbf{p}(\mathcal{D})}[||\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2])^2 \tag{B.2}$$

with the equality achieved at $\mathbf{q}^*(\mathcal{D}) = \frac{\mathbf{p}(\mathcal{D})||\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2}{\int_{\mathcal{D}}\mathbf{p}(\mathcal{D})||\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}}(\mathcal{D})||_2\, d\,\mathbf{p}(\mathcal{D})}$.

Appendix C. Convergence of δ^B to linear combination of normal distribution

Suppose $\mathbf{z}_i$ drawn i.i.d. from distribution Q, with the assumption that $\mathbb{E}_Q||k(\mathbf{z},\cdot)||^4 < \infty$. Define $\mu \overset{def}{=} \mathbb{E}_Q k(\mathbf{z},\cdot)$ and $K(\mathbf{z},\mathbf{z}') \overset{def}{=} \langle k(\mathbf{z},\cdot)-\mu, k(\mathbf{z}',\cdot)-\mu\rangle$. With the following conditions satisfied for eigenvalue ξ_l and eigenvectors ϕ_l^2 of K: $\xi_l \geq 0$ and $\mathbb{E}_Q\phi_l^2 < \infty$ such that $K(\mathbf{z},\mathbf{z}') = \sum_{l\geq 1}\xi_l\phi_l(\mathbf{z})\phi_l(\mathbf{z}')$ and $\langle\phi_l,\phi_{l'}\rangle = \mathbf{1}_{l=l'}$. Then,

$$\delta \xrightarrow{d} \beta\sum_{l\geq 1}\xi_l Z_l^2 \tag{C.1}$$

the symbol $\xrightarrow{d}$ means converge in distribution. $(Z_l)_{l\geq 1}$ is a collection of independent normal random variables and β is a constant. The theorem and proof follow from [60, 61].

Appendix D. Performance Evaluation on SSVEP and P300 Paradigms

We performed detailed evaluation of model performance on SSVEP and P300 tasks to gain deeper insights on model's versatility towards the different BCI paradigms. We utilized the SSVEP-benchmark dataset [62] for performance evaluation on SSVEP paradigm, and also analyzed in detail of model performance on P300 speller dataset [63]. The SSVEP-benchmark includes 35 subjects, with subject underwent 6 sessions each containing 40 trials. The EEG signal is recorded with 64 channels at sampling rate of 1000Hz with a Synamps2 system. We adopted the same pre-processing steps (downsample and filtering) as introduced in [62]. The P300 speller dataset has a total of 55 subjects. EEG signal is recorded with 32 electrodes at sampling rate of 512Hz using the Biosemi ActiveTwo system. During test the subjects spells four words including ”SUBJECT”, ”NEURONS”, "IMAGINE" and "QUALITY". Please refer to [63] for detailed preprocessing and feature extraction procedures. In accordance with the online sequential decoding setting explored in this work, we set the number of training repetitions to be 1 for the SSVEP-benchmark task and also the number of letter repetitions to be 1 for P300 speller task (the reason that there exists a large performance gap between online sequential decoding and joint training). We set the number of training repetitions and letter repetitions both to be 10 for the joint training setting following the settings in [63] and [64].

Table D.8: Evaluation of model performance on SSVEP and P300 paradigms, for both data imbalance and balanced scenarios. We performed Dunn's post hoc test between MUDVI and baselines, with p-value provided in brackets following the result.

Dataset	SSVEP-Benchmark Mean + SD (p-value)		P300 Speller Mean + SD (p-value)	
Method	Imbalanced	Balanced	Imbalanced	Balanced
sequential	$46.16_{\pm 1.80}$ (2.08e−10)	$50.48_{\pm 0.94}$ (2.36e−12)	$31.35_{\pm 4.73}$ (5.53e−07)	$35.72_{\pm 2.09}$ (7.45e−10)
joint training	$87.39_{\pm 2.12}$ (1.71e−11)	$90.17_{\pm 1.75}$ (2.40e−12)	$79.41_{\pm 2.66}$ (2.06e−11)	$85.83_{\pm 3.20}$ (4.64e−11)
EWC	$50.83_{\pm 2.95}$ (3.18e−07)	$53.10_{\pm 1.31}$ (2.97e−10)	$34.62_{\pm 3.04}$ (7.57e−08)	$36.57_{\pm 5.85}$ (8.24e−06)
UCB	$50.27_{\pm 0.64}$ (2.76e−13)	$55.54_{\pm 0.89}$ (8.34e−11)	$36.27_{\pm 2.18}$ (1.25e−08)	$38.91_{\pm 4.36}$ (2.88e−06)
ER	$54.91_{\pm 1.58}$ (6.53e−08)	$57.32_{\pm 2.40}$ (3.29e−06)	$43.13_{\pm 1.50}$ (5.77e−07)	$46.04_{\pm 2.23}$ (7.85e−06)
ER+GMED	$55.63_{\pm 1.29}$ (2.74e−08)	$57.86_{\pm 1.73}$ (4.16e−07)	$45.95_{\pm 3.61}$ (0.024)	$49.82_{\pm 2.78}$ (0.020)
MIR	$57.02_{\pm 2.46}$ (4.14e−05)	$60.21_{\pm 1.05}$ (2.68e−07)	$44.42_{\pm 2.37}$ (1.85e−04)	$50.20_{\pm 3.19}$ (0.061)
MIR+GMED	$58.44_{\pm 1.35}$ (4.31e−06)	$61.93_{\pm 2.26}$ (3.85e−03)	$46.80_{\pm 5.43}$ (0.167)	$50.31_{\pm 3.54}$ (0.092)
MUDVI	$62.28_{\pm 2.71}$ (–)	$64.37_{\pm 3.14}$ (–)	$48.56_{\pm 2.82}$ (–)	$51.93_{\pm 1.75}$ (–)

The result for online sequential decoding performance on these two datasets is summarized in Table D.8. It reveals the MUDVI model is significantly more effective for forgetting mitigation compared to other baselines. For SSVEP-benchmark, it outperforms other comparison models by 3.84% for data imbalanced settings and achieves a margin of 2.44% for data balanced settings. For P300 speller task, the MUDVI model achieved improvement of 1.62% on accuracy for data balanced setting compared to the best comparison model, while having margin of 1.76% for data imbalanced settings. We further explored the differences between subject-agnostic and subject-aware settings, with the result provided in Table D.9. The model performance shows modest deterioration when the subject identity becomes unknown. Specifically, the accuracy reduced by 1.34% and 1.86% for data imbalanced and balanced settings respectively on SSVEP-benchmark. The subject identity information has more influence for P300 speller task, with performance deteriorating by 2.68% for data imbalanced setting and 0.78% for data balanced setting respectively.

Table D.9: Performance with subject agnostic and subject aware settings for SSVEP-benchmark and P300 speller tasks.

Dataset	SSVEP-Benchmark Mean + SD (p-value)		P300 Speller Mean + SD (p-value)	
Scenario	Imbalanced	Balanced	Imbalanced	Balanced
Subj. Agnostic	62.28±2.71 (0.076)	64.37±3.14 (0.047)	48.56±2.82 (0.007)	51.93±1.75 (0.096)
Subj. Aware	63.62±1.49 (–)	66.23±2.45 (–)	51.24±1.68 (–)	52.71±2.26 (–)

Appendix E. Detailed Training Setting for the Joint Training Baseline

For the joint training baseline serving as upper bound of model performance, its training setting is as follows:

Data from each subject is split into training/testing sets, and the EEG decoder are jointly trained on training sets of all subjects. For BCI IV-2a dataset, we adopt the original dataset's train/test setting with A0XT files jointly used for training and A0XE files for testing, the train/test split is approximately 1:1. For DEAP dataset, we generate the train/test splits utilizing the train_test_split function of sklearn package, with test_size set to 20% of processed data for each subject. For SEED dataset, we combined the files of id_1.mat and id_2.mat to be training data and id_3.mat is used for testing (the id in file name refers to specific subject ids), this makes the train/test split approximately 2:1.

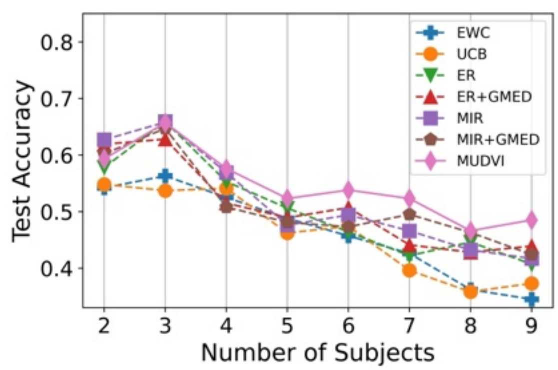


Figure E.8: Testing accuracy of sequential training with varying number of subjects for BCI IV-2a dataset.

Appendix F. Performance on Backward Transfer

The model performance and its comparison with baseline in terms of BWT is reported in Table F.10. We observed MUDVI outperforms baselines in data imbalanced settings. MIR-based approach achieved on-par or slightly better performance for balanced settings. Please note that result on BWT is known to be noisy with relatively higher variance given it is influenced by both foward adaptation and forget mitigation.

Table F.10: Performance on backward transfer (BWT), with both imbalanced and balanced scenarios.

Dataset	BCI IV-2a		DEAP		SEED	
Method	Imbalanced	Balanced	Imbalanced	Balanced	Imbalanced	Balanced
sequential	12.19±3.12	12.27±2.04	10.84±1.58	12.19±2.43	12.14±2.06	13.76±3.27
joint training	-	-	-	-	-	-
EWC	7.95±2.47	11.41±1.60	8.06±2.71	7.52±1.62	17.07±1.25	16.81±2.42
UCB	10.43±4.25	13.07±2.61	9.90±1.95	9.16±3.17	15.28±3.32	15.64±1.85
ER	10.56±1.68	10.34±3.44	8.83±3.36	6.61±2.38	9.91±1.69	9.95±2.61
ER+GMED	12.51±3.39	9.67±1.85	9.89±2.25	7.26±1.59	10.64±2.47	9.37±3.04
MIR	11.17±2.51	10.03±3.43	9.76±1.32	**6.30±2.45**	7.59±1.53	**7.76±1.79**
MIR+GMED	10.40±4.12	**7.87±2.17**	9.25±2.43	7.99±4.31	7.53±3.28	7.84±2.10
MUDVI	**6.52±1.83**	8.26±2.56	**7.91±2.26**	7.05±2.74	**6.45±1.41**	8.11±3.27

Appendix G. Window Size of Distance Metric m

Here we explore the influence of window size m for distance metric d_t on model performance. The window size denotes the past context that is put into consideration for distribution shift detection. The result is shown in Table G.11. For m increasing from 2 to 10, the model performance first increases and then peaks before slight decrease on larger size. We set $m = 8$ by default in our experiments.

Table G.11: Influence of context size on model performance for all three datasets.

Value of m	2	4	6	8	10
Dataset	Acc.				
BCI IV-2a	44.87±1.39	45.06±2.40	**45.98±1.83**	45.72±2.25	45.31±1.46
DEAP	41.32±2.14	41.74±1.59	42.17±2.72	**42.29±1.81**	41.83±1.05
SEED	58.51±1.93	59.18±0.85	59.42±1.16	**59.85±1.63**	59.60±2.17

Appendix H. Performance with Varying Number of Sequential Subjects

We performed study on the model performance with varying numbers of sequential subjects. Fig. E.8 shows the change in averaged testing accuracy with different numbers of sequential subjects on BCI IV-2a dataset. Here the accuracy is evaluated after sequential learning ends. For each setting with number of subjects being N_s, we perform 20% downsample on randomly chosen $int(N_s/2)$ subjects. The proposed model outperforms comparison models in most cases, especially for larger number of subjects.

Table I.12: Comparison on corruption error (averaged error rate across the different types of corruption operations) of online sequential EEG decoding.

Dataset	BCI IV-2a (↓)	DEAP (↓)	SEED (↓)
MUDVI	$63.25_{\pm 0.83}$	$67.82_{\pm 1.04}$	$52.41_{\pm 1.49}$
MUDVI+FT Surrogate	$61.87_{\pm 1.31}$	$66.16_{\pm 2.25}$	$50.03_{\pm 0.90}$
MUDVI+BandStopFilter	$63.63_{\pm 1.52}$	$67.44_{\pm 0.71}$	$51.62_{\pm 0.36}$
MUDVI+DADA	$60.10_{\pm 2.48}$	$\mathbf{65.23_{\pm 1.67}}$	$\mathbf{48.35_{\pm 1.14}}$
MUDVI+AugMix	$\mathbf{59.46_{\pm 1.23}}$	$65.58_{\pm 2.10}$	$49.76_{\pm 0.61}$

Table I.13: List of corruptions for robustness evaluation with detailed settings. The corruption operations are applied to normalized source signal to test model performance.

Corruptions	Description
Shot noise	Discrete electronic noise, voltage $V = 0.2$, frequency $f = 10\%$
Gaussian noise	Gaussian signal noise, mean $\mu = 0$, variance $\sigma = 0.1$
Intensity	The magnitude variation in EEG signal, reduction by 0.2
Zoom blur	Disruption in scaling of signal with scaling factor 110%

Appendix I. Robustness for Data Corruption Scenarios

The unknown variances and noise are important factors influencing decoding performance for both classic decoding settings (e.g. cross subject classification) and proposed sequential decoding setting. Denoising and data augmentation techniques are promising directions to tackle such corruptions in the signal. We performed detailed evaluation on integration of denoising and data augmentation methods with proposed approach and tested their performance on different types of corruptions including shot noise, Gaussian noise, zoom blur etc. (the detailed setting is provided in Table I.13). The result is provided in Table I.12. It shows the model has moderate performance deterioration in the occurrence of data corruptions in general. The different denoising and data augmentation approaches integrated in the study shows varying levels of effectiveness in improving model robustness.

We also performed ablation study on the influence of bad sensors and occurrence of channel blackout towards sequential decoding performance. The result is provided in Table I.14. It shows integration with denoising and data augmentation methods is useful to mitigate the performance deterioration in the occurrence of bad sensors. Another observation is that the model could achieve performance much better than random guess even with only 1-2 channels available for decoding.

Table I.14: Model performance in terms of sequential decoding accuracy with different number of channels being blackout on BCI IV-2a dataset. We tested the integration of proposed approach with a number of different denoising and data augmentation approaches for robustness improvement.

No. of Blackout Channels	0	4	8	12	16	20	22
MUDVI	$45.98_{\pm 1.83}$	$44.51_{\pm 2.36}$	$43.23_{\pm 2.52}$	$43.11_{\pm 1.97}$	$41.80_{\pm 2.79}$	$40.25_{\pm 2.23}$	$25.14_{\pm 0.86}$
MUDVI+FT Surrogate	$47.20_{\pm 2.17}$	$45.19_{\pm 0.72}$	$45.45_{\pm 1.94}$	$43.76_{\pm 3.28}$	$43.32_{\pm 1.43}$	$43.58_{\pm 2.71}$	$24.93_{\pm 0.42}$
MUDVI+BandStopFilter	$43.71_{\pm 1.05}$	$44.28_{\pm 2.60}$	$42.26_{\pm 0.49}$	$41.99_{\pm 2.25}$	$41.67_{\pm 1.84}$	$40.92_{\pm 0.60}$	$25.35_{\pm 0.57}$
MUDVI+DADA	$47.83_{\pm 1.96}$	$46.41_{\pm 1.29}$	$44.72_{\pm 3.10}$	$44.27_{\pm 1.22}$	$42.19_{\pm 0.78}$	$42.83_{\pm 1.49}$	$25.67_{\pm 0.81}$
MUDVI+AugMix	$47.32_{\pm 2.48}$	$45.56_{\pm 0.95}$	$45.13_{\pm 2.38}$	$44.94_{\pm 1.76}$	$42.73_{\pm 2.15}$	$41.30_{\pm 0.74}$	$24.71_{\pm 1.25}$

References

[1] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Human Brain Mapping 38 (11) (2017) 5391–5420. `arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23730`, `doi:https://doi.org/10.1002/hbm.23730`.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730`

[2] I. Iturrate, J. Antelis, J. Minguez, Synchronous eeg brain-actuated wheelchair with automated navigation, in: 2009 IEEE International Conference on Robotics and Automation, 2009, pp. 2318–2325. `doi:10.1109/ROBOT.2009.5152580`.

[3] A. Campbell, T. Choudhury, S. Hu, H. Lu, M. K. Mukerjee, M. Rabbi, R. D. Raizada, Neurophone: Brain-mobile phone interface using a wireless eeg headset, in: Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds, MobiHeld '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 3–8. `doi:10.1145/1851322.1851326`.
URL `https://doi.org/10.1145/1851322.1851326`

[4] S. Moghimi, A. Kushki, A. Marie Guerguerian, T. Chau, A review of eeg-based brain-computer interfaces as access pathways for individuals with severe disabilities, Assistive Technology 25 (2) (2013) 99–110.

[5] K. Gadhoumi, J.-M. Lina, F. Mormann, J. Gotman, Seizure prediction for therapeutic devices: A review, Journal of neuroscience methods 260 (2016) 270–282.

[6] T. Hafeez, S. M. Umar Saeed, A. Arsalan, S. M. Anwar, M. U. Ashraf, K. Alsubhi, Eeg in game user analysis: A framework for expertise classification during gameplay, PLOS ONE 16 (6) (2021) 1–21. doi:10.1371/journal.pone.0246913.
URL https://doi.org/10.1371/journal.pone.0246913

[7] T. Thompson, T. Steffert, T. Ros, J. Leach, J. Gruzelier, Eeg applications for sport and performance, Methods 45 (4) (2008) 279–288, neuroimaging in the sports sciences. doi:https://doi.org/10.1016/j.ymeth.2008.07.006.
URL https://www.sciencedirect.com/science/article/pii/S1046202308001163

[8] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, M. Ranzato, Continual learning with tiny episodic memories, CoRR abs/1902.10486 (2019). arXiv:1902.10486.
URL http://arxiv.org/abs/1902.10486

[9] R. Aljundi, L. Caccia, E. Belilovsky, M. Caccia, M. Lin, L. Charlin, T. Tuytelaars, Online Continual Learning with Maximally Interfered Retrieval, Curran Associates Inc., Red Hook, NY, USA, 2019.

[10] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, Journal of Neural Engineering 15 (5) (2018) 056013. doi:10.1088/1741-2552/aace8c.

[11] D. Zhang, L. Yao, K. Chen, J. Monaghan, A convolutional recurrent attention model for subject-independent eeg signal analysis, IEEE Signal Processing Letters 26 (5) (2019) 715–719. doi:10.1109/LSP.2019.2906824.

[12] B. Mirkovic, S. Debener, M. Jaeger, M. de Vos, Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications, Journal of Neural Engineering 12 (2015).

[13] D. Borra, S. Fantozzi, E. Magosso, Interpretable and lightweight convolutional neural network for eeg decoding: Application to movement execution and imagination, Neural Networks 129 (2020) 55–74. doi:https://doi.org/10.1016/j.neunet.2020.05.032.
URL https://www.sciencedirect.com/science/article/pii/S0893608020302021

[14] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Human Brain Mapping 38 (11) (2017) 5391–5420. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23730, doi:10.1002/hbm.23730.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730

[15] D. Zhang, L. Yao, K. Chen, J. Monaghan, A convolutional recurrent attention model for subject-independent eeg signal analysis, IEEE Signal Processing Letters 26 (5) (2019) 715–719. doi:10.1109/LSP.2019.2906824.

[16] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, C. Guan, Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI, Journal of Neural Engineering 16 (2) (2019) 026007. doi:10.1088/1741-2552/aaf3f6.
URL https://doi.org/10.1088%2F1741-2552%2Faaf3f6

[17] W. Samek, F. C. Meinecke, K. Müller, Transferring subspaces between subjects in brain–computer interfacing, IEEE Transactions on Biomedical Engineering 60 (8) (2013) 2289–2298. doi:10.1109/TBME.2013.2253608.

[18] W.-L. Zheng, B.-L. Lu, Personalizing eeg-based affective models with transfer learning, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, 2016, pp. 2732–2738.
URL http://dl.acm.org/citation.cfm?id=3060832.3061003

[19] Z. Lan, O. Sourina, L. Wang, R. Scherer, G. R. Müller-Putz, Domain adaptation techniques for eeg-based emotion recognition: A comparative study on two public datasets, IEEE Transactions on Cognitive and Developmental Systems 11 (1) (2019) 85–94. doi:10.1109/TCDS.2018.2826840.

[20] Y. Xie, K. Wang, M. Jiayuan, J. Yue, L. Meng, W. Yi, T.-P. Jung, M. Xu, D. Ming, Cross-dataset transfer learning for motor imagery signal classification via multi-task learning and pre-training, Journal of Neural Engineering 20 (09 2023). doi:10.1088/1741-2552/acfe9c.

[21] P. Lincong, K. Wang, L. Xu, X. Sun, W. Yi, M. Xu, D. Ming, Riemannian geometric and ensemble learning for decoding cross-session motor imagery elec-

troencephalography signals, Journal of neural engineering 20 (11 2023). doi:10.1088/1741-2552/ad0a01.

[22] H. Marzbani, H. Marateb, M. Mansourian, Methodological note: Neurofeedback: A comprehensive review on system design, methodology and clinical applications, Basic and Clinical Neuroscience Journal 7 (2016) 143–158. doi:10.15412/J.BCN.03070208.

[23] Y. Wang, J. Luo, Y. Guo, Q. Du, Q. Cheng, H. Wang, Changes in eeg brain connectivity caused by short-term bci neurofeedback-rehabilitation training: A case study, Frontiers in Human Neuroscience 15 (2021). doi:10.3389/fnhum.2021.627100.
URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.627100

[24] A. Haugg, F. M. Renz, A. A. Nicholson, C. Lor, S. J. Götzendorfer, R. Sladky, S. Skouras, A. McDonald, C. Craddock, L. Hellrung, M. Kirschner, M. Herdener, Y. Koush, M. Papoutsi, J. Keynan, T. Hendler, K. Cohen Kadosh, C. Zich, S. H. Kohl, M. Hallschmid, J. MacInnes, R. A. Adcock, K. C. Dickerson, N.-K. Chen, K. Young, J. Bodurka, M. Marxen, S. Yao, B. Becker, T. Auer, R. Schweizer, G. Pamplona, R. A. Lanius, K. Emmert, S. Haller, D. Van De Ville, D.-Y. Kim, J.-H. Lee, T. Marins, F. Megumi, B. Sorger, T. Kamp, S.-L. Liew, R. Veit, M. Spetter, N. Weiskopf, F. Scharnowski, D. Steyrl, Predictors of real-time fmri neurofeedback performance and improvement – a machine learning mega-analysis, NeuroImage 237 (2021) 118207. doi:https://doi.org/10.1016/j.neuroimage.2021.118207.
URL https://www.sciencedirect.com/science/article/pii/S1053811921004845

[25] S. Bhattacharyya, S. Das, A. Das, R. Dey, R. S. Dhar, Neuro-feedback system for real-time bci decision prediction, Microsystem Technologies 27 (10 2021). doi:10.1007/s00542-020-05146-4.

[26] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, Advances in Neural Information Processing Systems (2017).

[27] A. Chaudhry, M. Ranzato, M. Rohrbach, M. Elhoseiny, Efficient lifelong learning with a-gem, Proceedings of the International Conference on Learning Representations (2019).

[28] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, G. Tesauro, Learning to learn without forgetting by maximizing transfer and minimizing interference, International Conference on Learning Representations (2019).

[29] J. Yoon, E. Yang, J. Lee, S. J. Hwang, Lifelong learning with dynamically expandable networks, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=Sk7KsfW0-

[30] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences (2017).

[31] Z. Wang, L. Shen, T. Duan, D. Zhan, L. Fang, M. Gao, Learning to learn and remember super long multi-domain task sequence, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 7972–7982.

[32] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., USA, 2017, pp. 6470–6479.
URL http://dl.acm.org/citation.cfm?id=3295222.3295393

[33] H. Shin, J. K. Lee, J. Kim, J. Kim, Continual learning with deep generative replay, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
URL https://proceedings.neurips.cc/paper/2017/file/0efbe98067c6c73dba1250d2beaa81f9-Paper.pdf

[34] M. PourKeshavarzi, G. Zhao, M. Sabokrou, Looking back on learned experiences for class/task incremental learning, in: International Conference on Learning Representations, 2022.
URL https://openreview.net/forum?id=RxplU3vmBx

[35] E. Arani, F. Sarfraz, B. Zonooz, Learning fast, learning slow: A general continual learning method based on complementary learning system, in: International Conference on Learning Representations, 2022.
URL https://openreview.net/forum?id=uxxFrDwrE7Y

[36] Z. Wang, T. Duan, L. Fang, Q. Suo, M. Gao, Meta learning on a sequence of imbalanced domains with difficulty awareness, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 8927–8937.

[37] M. Welling, Herding dynamical weights to learn, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 1121–1128. doi:10.1145/1553374.1553517.
URL https://doi.org/10.1145/1553374.1553517

[38] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, icarl: Incremental classifier and representation learning, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 5533–5542.

[39] A. Prabhu, P. Torr, P. Dokania, Gdumb: A simple approach that questions our progress in continual learning, in: The European Conference on Computer Vision (ECCV), 2020.

[40] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3987–3995.

[41] J. von Oswald, C. Henning, B. F. Grewe, J. Sacramento, Continual learning with hypernetworks, in: International Conference on Learning Representations, 2020.
URL https://openreview.net/forum?id=SJgwNerKvB

[42] H. Liu, H. Liu, Continual learning with recursive gradient optimization, CoRR abs/2201.12522 (2022). arXiv:2201.12522.
URL https://arxiv.org/abs/2201.12522

[43] Z. Li, D. Hoiem, Learning without forgetting, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (12) (2018) 2935–2947. doi:10.1109/TPAMI.2017.2773081.

[44] Q. Wang, O. Fink, L. Van Gool, D. Dai, Continual test-time domain adaptation (2022). doi:10.48550/ARXIV.2203.13591.
URL https://arxiv.org/abs/2203.13591

[45] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, CoRR abs/1606.04671 (2016). arXiv:1606.04671.
URL http://arxiv.org/abs/1606.04671

[46] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, D. Wierstra, Pathnet: Evolution channels gradient descent in super neural networks,

CoRR abs/1701.08734 (2017). arXiv:1701.08734.
URL http://arxiv.org/abs/1701.08734

[47] Q. Qin, W. Hu, H. Peng, D. Zhao, B. Liu, Bns: Building network structures dynamically for continual learning, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 20608–20620.
URL https://proceedings.neurips.cc/paper/2021/file/ac64504cc249b070772848642cffe6ff-Paper.pdf

[48] Q. Wang, O. Fink, L. Van Gool, D. Dai, Continual test-time domain adaptation (2022). doi:10.48550/ARXIV.2203.13591.
URL https://arxiv.org/abs/2203.13591

[49] M. Rostami, Lifelong domain adaptation via consolidated internal distribution, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 11172–11183.
URL https://proceedings.neurips.cc/paper/2021/file/5caf41d62364d5b41a893adc1a9dd5d4-Paper.pdf

[50] M. Wulfmeier, A. Bewley, I. Posner, Incremental adversarial domain adaptation for continually changing environments, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE Press, 2018, p. 1–9. doi:10.1109/ICRA.2018.8460982.
URL https://doi.org/10.1109/ICRA.2018.8460982

[51] R. Volpi, D. Larlus, G. Rogez, Continual adaptation of visual representations via domain randomization and meta-learning (2020). doi:10.48550/ARXIV.2012.04324.
URL https://arxiv.org/abs/2012.04324

[52] S. M. R. Arnold, P.-A. Manzagol, R. Babanezhad, I. Mitliagkas, N. L. Roux, Reducing the Variance in Online Optimization by Transporting Past Gradients, Curran Associates Inc., Red Hook, NY, USA, 2019.

[53] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (null) (2012) 723–773.

[54] A. W. v. d. Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998. doi:10.1017/CBO9780511802256.

[55] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Mueller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, B. Blankertz, Review of the bci competition iv, Frontiers in Neuroscience 6 (2012) 55. doi:10.3389/fnins.2012.00055.
URL https://www.frontiersin.org/article/10.3389/fnins.2012.00055

[56] R.-N. Duan, J.-Y. Zhu, B.-L. Lu, Differential entropy feature for EEG-based emotion classification, in: 6th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2013, pp. 81–84.

[57] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis ;using physiological signals, IEEE Transactions on Affective Computing 3 (1) (2012) 18–31. doi:10.1109/T-AFFC.2011.15.

[58] S. Ebrahimi, M. Elhoseiny, T. Darrell, M. Rohrbach, Uncertainty-guided continual learning with bayesian neural networks, arXiv preprint arXiv:1906.02425 (2019).

[59] X. Jin, A. Sadhu, J. Du, X. Ren, Gradient based memory editing for task-free continual learning, ArXiv abs/2006.15294 (2020).

[60] R. J. Serfling, Approximation theorems of mathematical statistics, John Wiley & Sons, 2009.

[61] N. Keriven, D. Garreau, I. Poli, Newma: A new method for scalable model-free online change-point detection, IEEE Transactions on Signal Processing 68 (2020) 3515–3528. doi:10.1109/TSP.2020.2990597.

[62] Y. Wang, X. Chen, X. Gao, S. Gao, A benchmark dataset for ssvep-based brain–computer interfaces, IEEE Transactions on Neural Systems and Rehabilitation Engineering 25 (10) (2017) 1746–1752. doi:10.1109/TNSRE.2016.2627556.

[63] K. Won, M. Kwon, M. Ahn, S. Jun, Eeg dataset for rsvp and p300 speller brain-computer interfaces, Scientific Data 9 (2022) 388. doi:10.1038/s41597-022-01509-w.

[64] Ssvep-dan: A data alignment network for ssvep-based brain computer interfaces, arxiv (2023).