# Resource-Efficient Continual Learning for Personalized Online Seizure Detection

Amirhossein Shahbazinia*, Flavio Ponzina†, Jose A. Miranda*, Jonathan Dan*, Giovanni Ansaloni*, David Atienza*
*École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
†University of California San Diego (UCSD), La Jolla, California
*{amirhossein.shahbazinia, jose.mirandacalero, jonathan.dan, giovanni.ansaloni, david.atienza}@epfl.ch, †fponzina@ucsd.edu

***Abstract*—Epilepsy, a major neurological disease, requires careful diagnosis and treatment. However, the detection of epileptic seizures remains a significant challenge. Current clinical practice relies on expert analysis of EEG signals, a process that is time-consuming and requires specialized knowledge. This paper explores the potential for automated epileptic seizure detection using deep learning techniques, with a particular focus on personalized models based on continual learning. We highlight the importance of adapting these models to each patient's unique EEG signal features, which evolve over time. Our approach addresses the fundamental challenge of integrating new data into existing models without losing previously acquired information, a common issue in static deep learning models when applied in dynamic environments. In this study, we propose a novel continual learning algorithm for seizure detection, which integrates a replay buffer mechanism. This mechanism is key to retaining relevant information on past data while acquiring new one, thus effectively enhancing the model's performance over time. Our methodology is designed to be resource-efficient, making it suitable for implementation in embedded systems. We demonstrate the effectiveness of our approach using the CHB-MIT dataset, achieving an improvement of 35.34% in the F1 score with respect to a fine-tuning approach that does not consider catastrophic forgetting. Furthermore, we show that a small 1-hour data replay buffer suffices to achieve F1 scores comparable to that of a resource-unlimited scenario, while also decreasing the False Alarm Rate in 24 hours by 33% compared to a resource-unconstrained method.**

***Index Terms*—seizure detection, continual learning, incremental learning, deep learning, personalized models, wearable devices**

## I. Introduction

Epilepsy is a neurological disease that affects millions of people around the world [1]. Seizures are characterized by sudden and unexpected discharges of large groups of neurons, and their detection is crucial for timely intervention and treatment [2]. The detection of epileptic seizures commonly relies on analyzing electroencephalography (EEG) signals, which capture the electrical activity of the brain. However, manual analysis of long-term EEG signals can be time consuming and requires expert knowledge. Therefore, there has been growing interest in developing automated methods for the detection of epileptic seizures, with deep learning and machine learning techniques showing promising results [3, 4].

Deep learning techniques, such as convolutional neural networks (CNN), have shown great promise in seizure detection [5]. Unlike traditional machine learning, deep learning models can automatically learn relevant features from raw EEG data, eliminating the need for application-specific feature extraction [6]. This ability to learn from raw data makes deep learning particularly suitable for seizure detection, where the relevant features may not be easily identifiable [7, 8].

Due to their flexibility, deep learning approaches are also well suited for the development of personalized models [9]. Such scenarios are particularly attractive in the automated EEG analysis domain because each patient presents unique EEG characteristics [10], and considering them can improve performance in the detection of seizures. Training personalized models requires a substantial amount of data [11]. However, these scenarios are increasingly practical, thanks to advances in wearable devices, which allow long-term data collection [12] and the possibility of fine-tuning existing models, rather than training them from scratch.

A further challenge related to the personalization of deep learning models stems from the non-stationary acquisitions over time [13]. This is especially true in healthcare, where physiological parameters can fluctuate, leading to changes such as electrode impedance, variations in environmental noise, and alterations in exercise routines of patients [10]. Deep learning models, even after personalized fine-tuning, are inherently static and, hence, ill-equipped for dynamic environments where data distributions change over time.

Continual learning strategies address this challenge by proposing methodologies that enable deep learning models to continuously adapt to new information and environments while avoiding the need to re-train from scratch [14]. The application of continual learning to seizure detection presents specific challenges that require tailored solutions. Seizures are rare events that lead to a highly imbalanced dataset, which complicates the training process. It is therefore key to update models according to newly acquired data, while at the same time still retaining information on past events of high significance (such as seizures), hence avoiding Catastrophic

This research was supported by InnoHK funding, Hong Kong SAR: ACCESS – AI Chip Center for Emerging Smart Systems; the Swiss National Science Foundation Sinergia grant 193813: PEDESITE - Personalized Detection of Epileptic Seizure in the Internet of Things (IoT) Era; and the Wyss Center for Bio and Neuro Engineering: Lighthouse Non-invasive Neuromodulation of Subcortical Structures.

Amirhossein Shahbazinia is the corresponding author. The code is available at https://gitlab.epfl.ch/ashahbaz/personalized-online-seizure-detection.

Forgetting [14]. Moreover, solutions intended for personalized healthcare should achieve these goals within the constrained resources typical of wearable devices.

To the best of our knowledge, no study has yet applied continual learning for personalized epileptic seizure detection. We aim at filling this gap by proposing a methodology to train adaptable seizure detection models. These can adjust to the evolving data patterns of individual patients without losing knowledge of previous patterns, including, but not limited to, seizure episodes. Indeed, our experiments indicate that fine-tuning models based solely on new data makes them prone to forgetting, which can even diminish performance. To counteract this effect, our methodology employs a replay buffer [15], a mechanism that stores and reuses past experiences to mitigate the problem of forgetting. In more detail, important samples are judiciously selected and stored in memory. The neural network then retains the previous knowledge by regularly replaying these samples, thereby improving its performance.

The main contributions of this paper are as follows:

1) We introduce a methodology for personalized epileptic seizure detection that can adapt over time while avoiding forgetting.
2) We propose the utilization of a replay buffer mechanism for the strategic storage of critical data over time, which enhances the learning efficiency of the model by integrating new data and the information retained in the replay buffer.
3) We show that this methodology has high performance with notable memory and computational efficiency, making it ideally suited for real-world implementation, particularly in resource-constrained wearable devices.
4) We validated our framework in the CHB-MIT dataset, achieving an improvement in the F1 score of 35.34% compared to a fine-tuning approach that does not consider catastrophic forgetting. Moreover, we showcase that a small 1-hour data replay buffer suffices to achieve F1 scores comparable to that of a resource-unconstrained scenario.

The rest of this paper is organized as follows. In Section II, we review related work on seizure detection and continual learning. In Section III, we present our proposed algorithm to tackle the challenges of continual learning and the update of the deep learning model for personalized seizure detection. We evaluate the performance of the proposed method in Section IV. Finally, we highlight the main outcomes of the work in Section V.

## II. Related Works

### A. Machine Learning for Epileptic Seizure Detection

Conventional techniques for epileptic seizure detection extract features from the EEG. These features correspond both to linear and non-linear descriptors of the signal. They relate either to the time domain, the frequency domain, the spatial domain (across channels), or on their combinations. This process is aimed at extracting significant features, which are manually crafted and then used to train a classifier in order to differentiate between seizure and non-seizure events [16, 17].

This approach has two limitations. First, it relies on expert knowledge and trial and error when identifying features of interest, leading to a lack of generalizability. Moreover, it is vulnerable to changes in seizure patterns, caused by the inherently non-stationary nature of the EEG, which makes its statistical components change across subjects and time [18].

Deep learning has been extensively employed for the automated processing of EEG signals in various contexts, such as sleep analysis [19, 20], brain-computer interfacing [21, 22], epileptic seizure prediction [23] and detection [5, 24], showcasing its ability to learn from raw data and potentially surpass traditional feature extraction-based methods [25].

While traditional deep learning shows remarkable performance, it often assumes that data are independent and identically distributed (i.i.d.). Such models, developed based on static snapshots of data, may not effectively adapt to dynamic, non-stationary environments where this i.i.d. assumption is violated.

### B. Continual Learning

Continual learning (CL, also known as incremental learning or life-long learning), is a machine learning approach where a model is sequentially trained on a series of tasks, addressing the challenge of catastrophic forgetting — a phenomenon where learning new information causes the model to forget previously acquired knowledge — and thereby retaining knowledge from previous tasks even when their data is no longer available. This approach is categorized into methods including neural architecture modification (introducing new neurons for new tasks), regularization strategies (controlling the modification of model parameters), and replay buffers [14]. The latter technique relies on retaining and utilizing past data points through a buffer in the learning phase and is often seen as the most effective [26]. Research in continual learning has predominantly focused on computer vision [26]. In healthcare, the emphasis has largely been on medical imaging, as seen in [27, 28, 29, 30, 31]. For instance, [31] demonstrates the potential of continual learning in general-purpose, shareable AI for medical imaging.

The application of continual learning in healthcare time-series data is also gaining attention [32, 33, 26, 34]. For example, in [34], authors propose a multi-disease detection framework using wearable medical sensors and continual learning, which overcomes the limitations of traditional methods by employing a multi-headed neural network and an exemplar-replay-style algorithm for efficient, adaptable disease detection with a single model. Moreover, a continual learning approach with a replay buffer and novel trainable task-specific parameters is designed in [26] to address performance degradation in deep learning systems due to non-i.i.d. clinical data.

One important aspect of deploying continual algorithms for personalized monitoring is their computational efficiency. Under limited computation resources, simpler approaches such as uniform sampling from memory, might outperform most CL

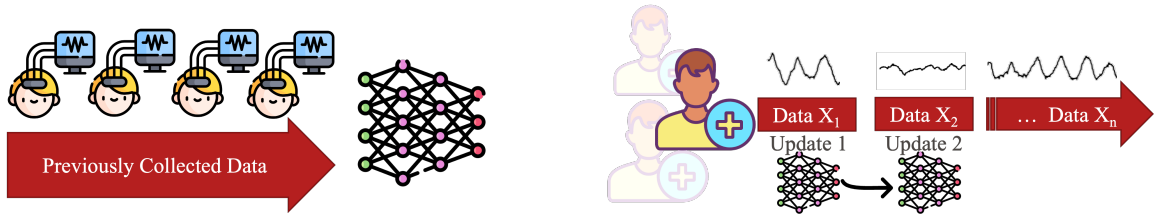


Fig. 1. Considered personalized healthcare monitoring scenario. Initially, a subject-independent model is trained using previously collected data (left). The model is then fine-tuned and personalized to an individual patient by incorporating information from his/her own recordings as they become available over time (right).

methods [35]. Our methodology is inspired by these works, but puts them in the context of our considered scenario, where data is highly skewed towards one of the two classes (as seizure episodes are rare).

Summing up, despite the growing interest in continual learning in the health domain, its application to personalized epileptic seizure detection remains, to the best of our knowledge, unexplored. To this end, our novel strategy tackles three main challenges in this scenario: non-stationarity of data, imbalance between acquisitions of seizure/non-seizure data, and resource constraints in personalized health monitors.

## III. Methodology

### A. Overview and Problem Description

As depicted in Fig. 1, our framework for continual learning in seizure detection considers as input a subject-independent model. This is usually derived by training a model on a database containing signals from a large number of subjects. This initial model is the starting point for personalized updates, which are made progressively as new data is acquired from a patient. Therefore, our methodology enables the patient-specific refinement of the initial model to suit the unique and changing physiological signals of each individual. It achieves this goal while a) coping with a memory (or storage) constraint (which must be accounted for in personal health monitors), and b) retrieving information on relevant past data in order to avoid catastrophic forgetting.

Our framework aims to improve the performance of a model, indicated as $M$, while learning from the acquired data. To this end, we divide acquisitions in chunks $\mathcal{X} = \{X_1; X_2; \dots; X_N\}$, where each chunk $X_i \in \mathbb{R}^{n \ T}$ consists of raw EEG samples captured across channels $n$ over a period of time $T$. At the end of each time period, a fine-tuning of the model is triggered.

In this scenario, the key challenge that our proposed methodology tackles lies in developing an updating strategy, where new EEG recordings are regularly introduced and older recordings are periodically phased out. In fact, due to storage restrictions, only a subset of previous data chunks is stored and utilized at any given time.

### B. Proposed Method

Our approach is depicted in Fig. 2. As illustrated in the figure, at each time period $T$, two sources of data are considered as inputs to fine-tune the model. The first is the data that are acquired as part of the current chunk, while

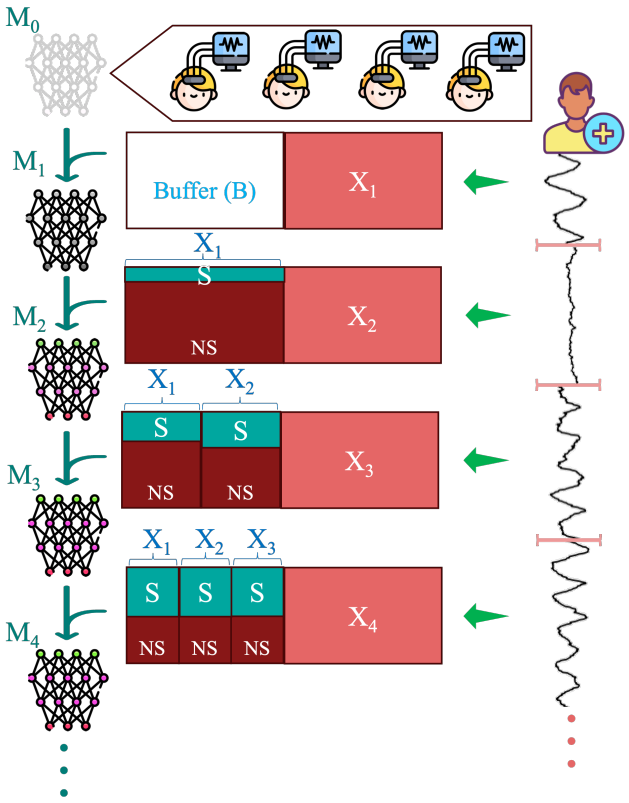


Fig. 2. Overview of the proposed fine-tuning methodology for seizure detection, where 'S' denotes Seizure and 'NS' denotes Non-Seizure data. Models $M_i$ are refined at the boundaries of each period. Inputs for model updates are the data acquired in the period (indicated as $X_i$), as well as a buffer storing representative samples from past periods.

the second (stored in a dedicated buffer $B$) is a subset of previous representative data. Hence, the first data source allows accounting for changes in data distributions over time, while the second counters catastrophic forgetting of past events.

When first deployed, both memory regions are empty. During the first period ($T_1$), the first data chunk is stored, and at the end of this period, it is used to fine-tune/personalize the model. These data are then transferred to the buffer $B$, freeing space to store the second data chunk. Then, at the end of the second period, both $X_1$ and $X_2$ are employed for the second fine-tuning step, which results in the $M_2$ model.

Note that, at this point, there is no more free space in $B$, so it is not possible to retain the entire data in the $X_1$ and $X_2$ fragments while acquiring $X_3$. Although it would be possible to overwrite the buffer with $X_2$ data (hence discarding $X_1$ data), such a solution would not retain information about

important but rare events such as seizures. Instead, when refreshing the content of $B$ before the start of a period, we judiciously selected samples from all the data collected so far.

To this end, the buffer is equally divided between seizure and non-seizure data, with the goal of preserving a representative amount of samples from the two classes. Such partitioning is only enforced on a best-effort basis. In practice, at the beginning of the recording, there will not be enough samples of the underrepresented class (seizure). In this case, all available space in the seizure partition (after storing seizures) will also be filled by non-seizure data. Note that the presence of non-seizure data in the seizure partition indicates that all seizures up to the present time period have been entirely preserved. This is usually the case even for the long health monitoring sessions considered in this study[1]. This is beneficial for model updates because seizures tend to have different patterns in different subjects, hence their inclusion during fine-tuning is key for effective personalization.

In-between time periods, buffer data is refreshed according to the following policy:

1) An equal amount of samples in $B$ originating from each chunk is discarded, in order to retain an unbiased view of past data. If a class is over-represented (e.g., non-seizure) in the buffer, samples are selected from it. Otherwise, if the two classes have the same amount of data, the same number of samples is discarded from both classes.
2) New data, selected from the current chunk, is written in the newly freed space in the buffer. The data is selected on the same best-effort basis outlined above, trying to select as much as possible the same amount of samples from the two classes.

At the end of a buffer refresh, its content contains an equal amount of data from each chunk, with (as much as possible) the same amount of seizure and non-seizure samples. Samples belonging to the same class and originating from the same chunk are selected for inclusion in the buffer, or discarded from it, according to a simple random strategy. We comparatively assess the performance of random sampling with respect to a more computationally complex alternative in Section IV-D.

In order to fine-tune the model, data in the buffer and in the newly acquired chunk are processed in the same way. The imbalance between the number of seizure and non-seizure samples is resolved by data augmentation via time-shifting and repetition, as discussed in Section IV-A1, in order to equalize the size of the two classes.

## IV. Experiments

### A. Experimental Setup

*1) Dataset Overview and Data Preparation:* The database used in this study is the CHB-MIT dataset [36, 37]. This public dataset comprises 982 hours of EEG recordings from 24 pediatric patients suffering from intractable epilepsy. Out of the 982 hours of recordings, only three hours contain seizure events, representing a total of 198 seizures. This dataset is one of the largest public datasets on the hours of recording per patient, with an average of approximately 40 hours per patient. The EEG signals are sampled at 256 Hz with a 16-bit resolution. Most, but not all, records feature 23 bipolar channels with electrodes placed according to the International 10–20 system. Here, for the sake of uniformity and comparability, we have considered only the 18 channels present in all patients.

The EEG data for each patient are divided into 1-hour chunks, with each chunk generating approximately 31.6 MB of data. Three filters were applied to preprocess the data: a 0.5 Hz highpass, a 60 Hz lowpass, and a 50 Hz notch. All filters are 4th-order Butterworth filters.

We considered a replay buffer of the same size as a data fragment, which resulted in a storage requirement of 63.2MB for implementing our methodology. Each data chunk is randomly divided into training and validation sets in an 80% to 20% ratio. We arranged the EEG recording data so that the datastream of each patient starts and finishes with a seizure event, so that personalization on seizure data can be performed from the first data chunk, and F1 scores can be computed on the last chunk.

We used three-fold cross-validation to evaluate the methodologies, partitioning the 24 patients into three groups, each comprising eight patients similar to [5]. In each experimental run, one group is designated for training an initial subject-independent model. Subsequently, this model is personalized for the patients in the remaining two groups. Each experiment is repeated three times using different seeds to ensure robustness, and the mean of these results is considered for analysis.

Data augmentation was employed on seizure windows in training sets, both when deriving the initial patient-independent model and for patient-dependent fine-tuning steps, in order to equalize the number of seizure and non-seizure data. We used a combination of time-shifting (shifting consecutive windows by 1/8 of a second to generate new data windows) and data repetition.

*2) Implementation:* Our methodology is agnostic with respect to the structure of the deep neural network employed. Without loss of generality, in all experiments, we employed a domain-specific fully convolutional network (FCN) tailored for EEG analysis introduced in [5]. The FCN takes input windows of 4 seconds and predicts a binary label corresponding to a seizure or non-seizure window. It comprises three blocks, each consisting of a convolution layer, batch normalization, a Rectified Linear Unit (ReLU) activation function, and a pooling layer. This is followed by two fully convolutional layers and a concluding SoftMax layer. The network has low resource requirements, employing approximately 300K parameters, and hence is a representative choice in the context of personalized health monitoring.

Training is performed for 100 epochs in all cases. Note that, as opposed to when training the initial model, when

[1]Indeed, only for one patient among the ones considered in Section IV the seizure partition filled up completely.

performing fine-tuning each epoch only processes a small amount of data, i.e. that of the current chunk and that in the replay buffer. The training process uses cross-entropy loss and Adam optimizer, with an initial learning rate set at $10^{-4}$. To optimize performance, a learning rate scheduler is incorporated to reduce the learning rate by a factor of 5 at specified epochs, notably at epochs 50 and 75.

*3) Post Processing:* The model is designed to predict a label for each 4-second window. These windows overlap by 3 seconds, providing continuity in data analysis. However, seizure events exhibit temporal dependencies, which means that seizures are not random, but can be correlated over time. Post-processing can enhance detection performance by enforcing these temporal dependencies, taking into account the context of seizure events over time. Therefore, first, a simple moving average filter is applied to smooth out the predictions. Subsequently, for a more accurate evaluation, event-based scoring is adopted [38]. This method assesses the performance of the model at the level of seizure episodes instead of using sample-by-sample based metrics.

*4) Metrics:* In our experiments, we utilize event-based F1 Score and False Alarm Rate (FAR) to assess the dataset's seizure detection performance. The F1 Score, a harmonic mean of precision and recall, evaluates the balance between accurately identifying seizures and correctly ruling out non-seizure events. The FAR measures the frequency of incorrect seizure predictions, an important metric for minimizing ineffective seizure detection. To evaluate our methods, we examine Remembering [39] and Future Scores:

**Remembering Score:** This metric is calculated using Backward Transfer (BWT), which measures the impact of learning new data on previously learned data. It is defined by the change in a metric on previously-seen data after a model update, thus quantifying the extent of catastrophic forgetting. BWT is calculated as:

$$BWT = \frac{\sum_{i=2}^{N}\sum_{j=1}^{i-1}(R_{i;j} - R_{j;j})}{\frac{N(N-1)}{2}}; \quad (1)$$

where $R_{i;j}$ denotes the score of the model trained at time step $i$ and evaluated on the data observed at time step $j$, and $N$ is the total number of the time steps. Using BWT, we define the Remembering metrics for both F1 score and FAR. The Rememebering F1 Score is defined as $1 - |min(BWT; 0)|$. It ranges from 0% to 100%, with 100% meaning the model remembers everything after learning new other data. The Remembering FAR is defined as $max(BWT; 0)$. This metric quantifies the increase in FAR after learning new data.

**Future Score:** Future Score measures how well the most recently updated model performs on unseen testing data. This metric is crucial for assessing the model's adaptability and current relevance in continual learning scenarios. It is calculated as the average performance of a model on the next future data chunk after a fine-tuning step. The equation for the Future Score is given by:

$$\text{Future Score} = \frac{\sum_{i=1}^{N-1} R_{i;i+1}}{N-1}; \quad (2)$$

where $R_{i;i+1}$ denotes the performance of the model trained at time step $i$ and evaluated in the subsequent data chunk $i+1$, and $N$ is the total number of the time steps. The Future F1 Score and Future FAR are derived in a similar way.

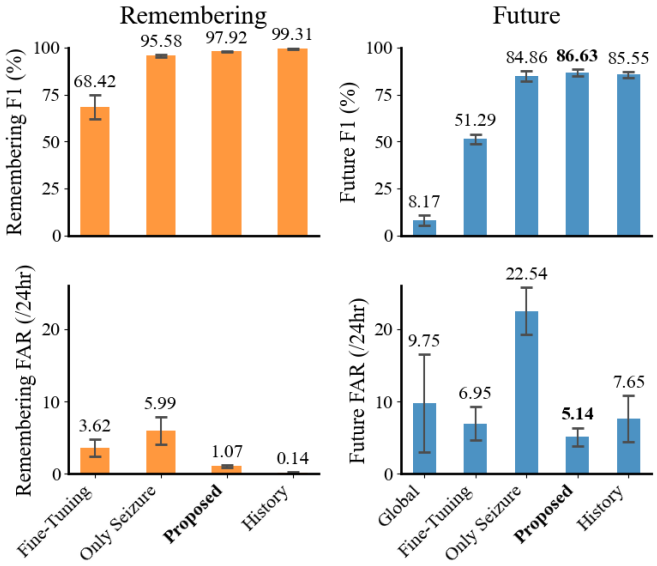


Fig. 3. Remembering (left) and Future (right) Scores for different seizure detection methods. In both cases, F1 Scores (higher is better) and False Alarm Rates per 24 hours (lower is better) are presented.

*5) Proposed and Baseline Methods:* In our study, several baseline methods are evaluated alongside our proposed approach, each offering distinct strategies:

1) **Global**: This method employs a subject-independent pretrained model based on data from other patients, without considering any personalized fine-tuning.
2) **Fine-tuning**: Updates the model for each new chunk without employing a replay buffer to reduce forgetting.
3) **Only-Seizure**: Updates the model only on data chunks that include a seizure, also without using a replay buffer.
4) **Proposed Method**: Besides the storage required for the current chunk, it also uses a replay buffer of the same size, following the strategy detailed in Section III.
5) **History**: Employs an unconstrained replay buffer, using all past data of a patient at each fine-tuning step.

### B. Experimental Results

In this section, the experimental results of our proposed method are presented. We report the performance of the proposed method with the evaluation metrics introduced in Section IV-A4. F1 scores (Remembering and Future) are only reported for data chunks with seizures, where false negatives can occur. On the other hand, all chunks are considered in the reported FAR data.

*1) Backward Transfer:* Fig. 3-left presents the results for the Remembering metrics. The Global baseline method, in which the model does not update over time, is excluded from this experiment. In the top-left subfigure, the Remembering F1 Score is shown; a higher score signifies that models are less prone to forgetting. Similarly, the Remembering FAR is shown in the bottom-left subfigure, with a lower score indicating an increase in only that amount of false alarms.

The Remembering Score for the Fine-Tuning method is lower than that of other strategies, at $68{:}42 \pm 6{:}35$. This is attributed to two main factors. First, the highly imbalanced nature of seizure events often results in chunks lacking seizure data, and fine-tuning the model predominantly with such non-seizure data can reduce its seizure detection capabilities. More critically, this method is susceptible to catastrophic forgetting.

The Remembering Score of the Only Seizure method is higher, at $95{:}58 \pm 0{:}58$. In this baseline, the issue of training on data chunks without seizures is avoided; therefore, it updates less frequently compared to other methods. Still, this method only remembers 95:58% (forgets 4:42%), likely due to catastrophic forgetting, along with a $5{:}99 \pm 1{:}92$ increase in FAR.

The History method represents an upper bound for these experiments, as all available data are used for training. This baseline has a high Remembering Score of $99{:}31 \pm 0{:}42$ and a minimal increase of $0{:}14 \pm 0{:}10$ in FAR. Our proposed method approximates these results well, presenting a Remembering F1 Score of $97{:}92 \pm 0{:}23$ and only a $1{:}07 \pm 0{:}19$ increase in FAR. This indicates that the proposed method appropriately retains past information when adapting to new data.

*2) Forward Transfer:* The Future Score metrics are presented in Fig. 3-right.

The Future F1 score and FAR for the Global method are $8{:}17 \pm 2{:}6$ and $9{:}75 \pm 6{:}80$, respectively. The Global method's poor performance underscores the limitations of using a subject-independent model in real-world scenarios, emphasizing the need for personalized models.

The Fine-Tuning method achieves a Future F1 Score of $51{:}29 \pm 2{:}64$ and a FAR of $6{:}95 \pm 2{:}33$. As mentioned earlier, its performance is hindered by including chunks without seizure data and proneness to catastrophic forgetting.

The Only Seizure method achieves an impressive Future F1 score of $84{:}86 \pm 2{:}83$, indicating its efficacy in cases where seizure data is present. This can be attributed to infrequent updates, leading to less forgetting. However, it records a high FAR of $22{:}54 \pm 3{:}25$. This is primarily due to the infrequent updates (as only chunks with seizures trigger a model fine-tuning), which result in extended periods without model adaptation and neglect the broader context of the patient's EEG data, thereby increasing the false alarm rate.

The History method shows a Future F1 score of $85{:}55 \pm 1{:}49$ and a FAR of $7{:}65 \pm 3{:}18$. Interestingly, our method outperforms this baseline, even if it employs a constrained replay buffer, with a Future F1 score of $86{:}63 \pm 1{:}67$ and a FAR of $5{:}14 \pm 1{:}25$. This improved performance is due to our method's ability to prioritize recent data, as only the most recent past chunk is used entirely for fine-tuning.

TABLE I
COMPARATIVE ANALYSIS OF REMEMBERING AND FORWARD SCORES FOR THE PROPOSED METHOD USING EITHER RANDOM OR BOUNDARY SAMPLES IN THE REPLAY BUFFER.

| Sampling | Remembering | | Forward | |
|---|---|---|---|---|
| | F1-Score ↑ | FAR(/24hr) ↓ | F1-Score ↑ | FAR(/24hr) ↓ |
| Random | $97{:}92 \pm 0{:}23$ | $1{:}07 \pm 0{:}19$ | $86{:}63 \pm 1{:}67$ | $5{:}14 \pm 1{:}25$ |
| Boundary | $98{:}89 \pm 0{:}69$ | $0{:}44 \pm 0{:}16$ | $85{:}16 \pm 1{:}53$ | $5{:}07 \pm 1{:}52$ |

### C. Memory requirements and Computational Complexity

The Fine-Tuning, Only Seizure baselines, as well as the proposed method, have a fixed-size memory footprint. In this regard, our method is at a slight disadvantage, as it requires additional storage to host the replay buffer. Nevertheless, as shown above, a reasonably sized replay buffer hosting one chunk of data suffices to counter forgetting. Conversely, the History baseline stores all acquired data, hence its memory requirements linearly increase with the length of the acquisition.

Similar considerations apply to the computational complexity of the strategies considered. The training process run-time is proportional to the number of training samples. Therefore, the complexity of our methods is almost four times that of the Fine-tuning baselines. A doubling factor arises from the use of the replay buffer. A second (almost) 2X increase in run time is instead due to the performed data augmentation, which results in rare seizure windows (and their time-shifted variations) being employed multiple times at each epoch.

As for the History method, its computational cost increases linearly with the number of chunks. Therefore, it is impractical in the long-term personalized monitoring scenario.

### D. Ablation Study

This section explores the efficacy of retaining boundary samples in the replay buffer for continual learning. Retaining boundary samples in the replay buffer has been shown to be beneficial [40]. Our study aims to demonstrate the effects of retaining boundary samples instead of random samples for non-seizure data in the replay buffer.

To conduct this study, at the end of each learning phase at a given time step, we used the recently trained model to select new samples and updated the replay buffer by choosing the most uncertain samples. Uncertainty was measured using the loss function's value. Table I presents the results. The Remembering F1 score and FAR for the boundary sampling are $98{:}89 \pm 0{:}69$ and $0{:}44 \pm 0{:}16$ respectively. These results suggest that the model performs better in terms of remembering previous data compared to the scenario where random samples are stored, likely due to the maintained accuracy of the classifier by replaying boundary samples. However, the Forward F1 Score degrades by 1:47% with a similar FAR when using boundary samples. This degradation is likely caused by the model starting to overfit previous data by focusing on outliers from previous data.

Fig. 4 illustrates this phenomenon. The four 4-second EEG excerpts (all originating from the same recording: patient 13,