An Ultra-Low Power Wearable BMI System with Continual Learning Capabilities

Lan Mei, *Graduate Student Member, IEEE*, Thorir Mar Ingolfsson, *Graduate Student Member, IEEE*, Cristian Cioflan, *Graduate Student Member, IEEE*, Victor Kartsch, Andrea Cossettini, *Member, IEEE*, Xiaying Wang, *Member, IEEE*, and Luca Benini, *Fellow, IEEE*

arXiv:2409.10654v1 [eess.SP] 16 Sep 2024

***Abstract*— Driven by the progress in efficient embedded processing, there is an accelerating trend toward running machine learning models directly on wearable Brain-Machine Interfaces (BMIs) to improve portability and privacy and maximize battery life. However, achieving low latency and high classification performance remains challenging due to the inherent variability of electroencephalographic (EEG) signals across sessions and the limited on-board resources. This work proposes a comprehensive BMI workflow based on a CNN-based Continual Learning (CL) framework, allowing the system to adapt to inter-session changes. The workflow is deployed on a wearable, parallel ultra-low power BMI platform (BioGAP). Our results based on two in-house datasets, Dataset A and Dataset B, show that the CL workflow improves average accuracy by up to 30.36% and 10.17%, respectively. Furthermore, when implementing the continual learning on a Parallel Ultra-Low Power (PULP) microcontroller (GAP9), it achieves an energy consumption as low as 0.45 mJ per inference and an adaptation time of only 21.5 ms, yielding around 25 h of battery life with a small 100 mAh, 3.7 V battery on BioGAP. Our setup, coupled with the compact CNN model and on-device CL capabilities, meets users' needs for improved privacy, reduced latency, and enhanced inter-session performance, offering good promise for smart embedded real-world BMIs.**

***Index Terms*— brain-computer interface, EEG, wearable healthcare, wearable EEG, deep learning, continual learning, on-device learning**

I. Introduction

BRAIN-MACHINE interfaces (BMIs) harness brain electrical activity to control external devices, such as prosthetic limbs and computers, providing substantial benefits to individuals with disabilities or in rehabilitation [1]. Beyond clinical applications, BMIs are becoming integral to mainstream neurotechnology, allowing consumers to explore them for activities such as gaming, virtual reality, and smart home control, presenting novel interaction methods and enhancing user experiences [2], [3]. Notably, a standard BMI paradigm uses electroencephalography (EEG) signals to discern whether a subject is performing or imagining specific motor tasks—Motor Movement (MM) or Motor Imagery (MI)—emphasizing their diverse uses [4].

Traditional BMIs based on scalp EEG are often cumbersome, stigmatizing, and require lengthy setup times [5], limiting their use primarily to laboratory and clinical environments. The advent of consumer-grade BMIs has led to the development of market-available devices that feature dry electrode setups, which are less stigmatizing and have quicker setup [6]–[8]. However, despite these advancements, EEG signals suffer from intrinsic inter-subject and inter-session variability, requiring subject-specific models [9] and continuous adaptation to new data sessions [10].

Transfer Learning (TL) has shown promise in addressing inter-session variability by continuously updating the model with data from new sessions [10]. However, TL does not allow the model to retain previously learned knowledge, which leads to catastrophic forgetting [11]. This phenomenon has been observed in many domains, including computer vision [12], [13] and seizure detection [14], resulting in progressive performance degradation over time or across domains.

To overcome this limitation, Continual Learning (CL) has been proposed, leveraging the retention of information from previous experience, and has been effective in mitigating catastrophic forgetting [12]–[14]. Few works have investigated CL for BMIs [14], [15] and they rely on remote computing platforms, leaving its application on edge BMI devices unexplored.

A crucial element for the success of next-generation BMIs is their ability to process streaming sensor data in real-time directly on edge devices, which preserves user privacy, reduces latency, and ensures sustained operation [16]. Advances in smart edge computing and TinyML offer promising solutions to enhance battery efficiency while balancing accuracy and resource demands [17], [18]. Prior work has predominantly focused on performing inference on edge devices, with the learning phase delegated to remote servers. This approach leads to delayed model updates and reliance on additional devices, limiting the system's ability to swiftly and autonomously adapt

This project was supported by the Swiss National Science Foundation under grant agreement 193813 (Project PEDESITE) and grant agreement 207913 (Project TinyTrainer), by the ETH-Domain Joint Initiative program (project UrbanTwin), and by the ETH Future Computing Laboratory (EFCL). The authors thank Xiaoling Xia for her expert advice on tailoring the headband.

Lan Mei, Thorir Mar Ingolfsson, Cristian Cioflan, Victor Kartsch, Andrea Cossettini, Xiaying Wang, and Luca Benini are with the Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland. Xiaying Wang is also with the Swiss University of Traditional Chinese Medicine, Zurzach, Switzerland. Luca Benini is also with the DEI, University of Bologna, Bologna, Italy.

Corresponding and co-last author: Xiaying Wang, xiaywang@iis.ee.ethz.ch

©2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published paper can be found here: DOI: 10.1109/TBCAS.2024.3457522

to new data or evolving conditions. Therefore, to develop adaptive BMI systems capable of real-time updates, CL methods must be implemented on the onboard processor, necessitating careful design to utilize limited resources optimally.

Within this context, we present a comprehensive approach to meet these criteria, aiming for robust and energy-efficient BMIs deployed on edge devices with minimal calibration time. This paper extends our previous work [10] and proposes the following contributions:

- We integrate BioGAP, a next-generation miniaturized BMI system with Parallel Ultra-Low Power (PULP) onboard computation capabilities for EEG processing in a user-friendly, non-stigmatizing, wearable BMI headband, and collect an EEG dataset based on motor tasks from five subjects over four different sessions with a total duration of ~22.2 hours;
- We propose a CL framework on inter-session EEG data that mitigates catastrophic forgetting and achieves average classification accuracy over all sessions of 89.93% on Dataset A and 90.88% on Dataset B, outperforming the conventional TL workflow by up to 30.36% and 10.17%, respectively;
- We demonstrate On-Device Learning (ODL) with TL and CL on the GAP9 processor, yielding a low latency of $21.6\,\mathrm{ms}$ and extended battery life, i.e., $25\,\mathrm{h}$ of operation with a $100\,\mathrm{mA\,h}$, $3.7\,\mathrm{V}$ battery, paving the way for future online, on-device BMI applications.

Finally, we open-source release the EEG dataset and the codes. [1]

II. Related Works

A. EEG Acquisition Platforms

While non-invasive medical-grade EEG acquisition platforms [19], [20] have been widely applied in clinical assessments with the advantage of good signal quality, their bulky setup causes discomfort and hinders daily-life usage [21]. Comparatively, consumer-grade EEG devices, such as Versus headset [6], Emotiv EPOC+ [7], and Muse headband [8], provide improved usability and user experience with non-stigmatizing appearance, shortened setup time, and lower costs [5].

Most of the existing platforms lack capabilities to perform on-device processing directly at the edge. They typically stream data to external computers or servers where the processing is carried out, raising concerns about user privacy and battery lifetime [5]. Very few commercial BMI devices and research platforms, such as OpenBCI Cython board [22] and Biowolf [17], provide onboard computing capabilities. Another notable and more recent platform is BioGAP [23], which represents the State-of-the-Art (SoA) in miniaturization, energy efficiency, and onboard PULP processing capabilities.

In this work, we incorporate BioGAP into a custom-tailored headband to create a user-friendly, non-stigmatizing, wearable BMI device equipped with comfortable, soft, dry electrodes. Using this in-house setup, we collect a multi-session dataset from five subjects and leverage the onboard processing capabilities to demonstrate for the first time ODL for BMIs.

[1] https://github.com/pulp-bio/bmi-odcl.git

B. Continual and Transfer Learning for EEG Data

TL has emerged as a prominent method to mitigate inter-session variability in EEG classification [24]–[26]. In [25], various methods such as Euclidean Alignment (EA) and Common Spatial Pattern (CSP) combined with TL strategies are compared on the BCI Competition IV-2a dataset [27], [28], demonstrating that TL can enhance classification accuracy in both cross-subject and cross-session contexts. Similarly, Lee et al. [26] applied a Transformer-Based Spatial-Temporal Network (TSTN) on an in-house dataset to address inter-session variability, improving $2-7\%$ per session.

Despite its benefits, TL has a notable drawback: the catastrophic forgetting phenomenon, where the model's performance degrades across different tasks as it learns new ones [11]. CL overcomes this limitation, and several CL methods have been proposed, including Learning Without Forgetting (LWF) [29], Elastic Weight Consolidation (EWC) [30], and Experience Replay (ER) [31]. These methods have improved average accuracy over TL in scenarios involving multiple tasks, particularly in image classification [29], [30].

However, applying CL methods to EEG datasets remains relatively underexplored.

Recent work by Rajpura et al. [15] evaluated different CL methods on the PhysioNet dataset [32], [33], reporting an improvement of up to 14.89% compared to naive TL. Notably, their analysis involved randomly shuffled data, ignoring chronological order, and the benefits of CL methods were not significant when averaged across all subjects. Another study by Shahbazinia et al. [14] introduced a resource-efficient CL algorithm for personalized seizure detection to preserve past knowledge over time, achieving an improvement in the F1 score of 35.34% compared to a fine-tuning approach that does not consider catastrophic forgetting, and demonstrating a 33% reduction in the false alarm rates. However, despite these performance improvements, inter-session variability has not been investigated. As a comparison, our work addresses inter-session variability through TL and CL workflows while keeping time coherency, showing the advantage of CL methods when applied to real-world BMIs.

C. Online On-Device Learning for Biosignals

Online ODL with neural networks for biosignals is an emerging area of research, distinct from the extensively studied traditional offline Machine Learning (ML) setup. Current frameworks, such as TensorFlow Lite Micro [34] and uTensor [35], facilitate on-board inference based on networks trained offline on servers. However, these frameworks typically do not support on-device fine-tuning, especially due to resource constraints, including limited computational power, memory, and energy efficiency.

Several works have begun to address these challenges. For example, Kwon et al. [36] proposed an ODL framework tested on Nvidia Jetson Nano and One Plus Pro smartphone platforms. Still, these solutions require considerable resources

and are not feasible for ultra-low-power edge devices. Ren et al. [37] introduced TinyOL, a system for incremental on-device training, which adds a trainable layer to the network to enable on-device updates. As one of the first works to implement incremental online learning on microcontroller units (MCUs), this work does not focus on CL approaches and shows sub-optimal computational costs [38]. Built on the tinyML framework for on-device replay-based CL [39], PULP-TrainLib [38] has been developed as a learning library to implement ODL on PULP MCU with an autotuning tool for latency and hardware optimization. Recently, Cioflan et al. [40] implemented ODL based on PULP-TrainLib to achieve domain adaptation in keyword spotting with up to 14% accuracy gains.

Recent studies have shown the feasibility and advantages of ODL specifically for biosignals. A survey by Li et al. [41] highlights the potential of CL in processing physiological signals such as EEG, emphasizing the need for adaptive learning models to handle the dynamic and evolving nature of healthcare scenarios. Similarly, the development of wearable devices for continuous monitoring of biosignals points to the increasing need for real-time, on-device processing to enhance user compliance and data accuracy [42].

This work leverages PULP-TrainLib to implement CL strategies for biosignals, explicitly focusing on wearable BMIs. We achieve efficient ODL by employing techniques such as selective layer updates [43], which combine 8-bit frozen layers with a 32-bit fine-tunable part. In this work, we extend the application of ODL into bio-signal classification while taking on the challenge of embedding it on resource-constrained wearable edge devices. This advancement enables a broader range of real-world applications for BMI systems, offering user-friendly, resource-efficient, and privacy-preserving solutions for continuous biosignal monitoring and adaptation.

III. Materials and Methods

A. EEG Acquisition Device

We introduce an inconspicuous EEG acquisition system comprising three major components: BioGAP, a headband, and electrodes. This system is then used to acquire the dataset we analyze in this work.

1) BioGAP: BioGAP [23] is used for ultra-low power EEG signal acquisition, preprocessing, and on-device learning. It has a miniaturized form factor ($16 \times 21 \times 14\,\mathrm{mm}^3$) and is based on GAP9, a PULP microcontroller offering SoA performance on tinyML tasks. EEG signals are captured through the ADS1298 Analog Front-End (AFE), which samples data from 8 channels at medical-grade resolution. BioGAP can operate in different modes (streaming, computation on the edge, sleep, etc.) to optimally balance power consumption and performance. We use the mode of computation on the edge in our setup, where all computations are done online without transmitting raw data wirelessly, significantly minimizing power consumption. Additionally, BioGAP communicates with a Java-based Graphical User Interface (GUI) via a Bluetooth Low Energy (BLE) dongle, enabling command transmission and result display on a computer. More details

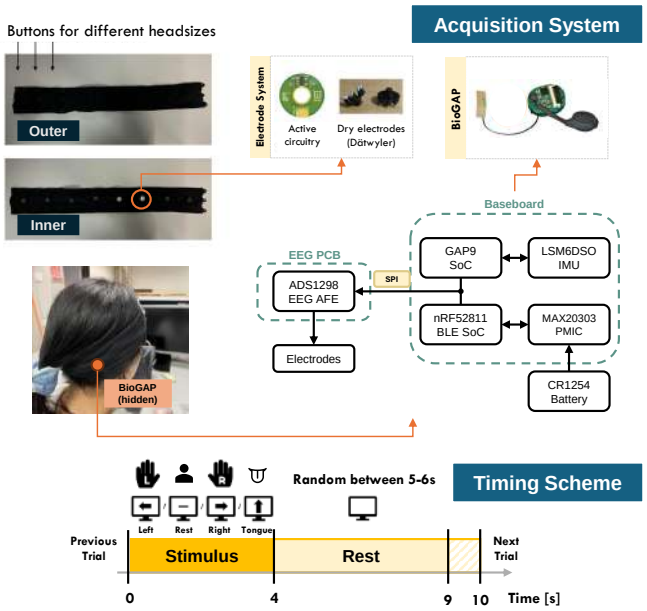


Fig. 1: Top: data acquisition setup, comprising the headband, BioGAP [23], and the dry active electrodes. Bottom: timing scheme of data trials.

can be found in [23]. This ultra-low-power, energy-efficient wearable platform is ideal for our proposed BMI system, providing functionalities for streaming EEG signals, on-device computations, and ODL with ML networks.

2) Headband: BioGAP is integrated into a custom-made, comfortable, easy-to-use headband (Fig. 1). The headband is made from elastic fabric and features a small pocket to house wearable BMI devices such as BioGAP. It includes 24 buttonholes for electrode fixation, providing flexibility in position adjustments. The headband can be customized to fit different head sizes and with differing levels of tightness, ensuring comfort during use. The headband maintains a non-stigmatizing appearance by concealing cables and BioGAP within the fabric, making it suitable for daily use.

3) Electrodes: The electrode subsystem (see Fig. 1) comprises two components: dry electrodes and active buffering circuitry. The dry electrodes, developed by Dätwyler Schweiz AG as a part of the newly developed SoftPulse Flex family [44], are made of soft, bendable, conductive elastomer materials and feature multiple silver/silver-chloride-coated pins on a circular disc. These pins pass through the hair to maintain scalp contact in a dry configuration and utilize a conductive elastomer snap connector interfacing device circuitry. Designed to flatten under pressure, the electrodes adhere to the scalp for improved user experience (for instance, participants typically reported no significant discomfort in sessions lasting 1.5-2 hours). The active buffering circuitry, embedded in a circular Printed Circuit Board (PCB) and placed on top of the electrode, minimizes cable-induced EEG artifacts and enhances Signal-to-Noise Ratio (SNR). For the experiments conducted in this work, we positioned 8 EEG channels near PO7, PO8, TP7, TP8, FC5, FC6, F1, and F2 according to the standard 10/20

system. Most electrodes are placed around the neighboring premotor cortex, with two electrodes near the occipital lobe (PO7, PO8) to capture motor-related information [18] better.

B. Dataset Description

At the beginning of each data acquisition session, a healthy volunteer (subject) sits comfortably in a chair approximately 1 meter away from a screen. The subject faces the screen with their back leaning on the chair and forearms resting on their lap. The signal quality is evaluated through an alpha-wave test before the formal acquisition. Throughout the data collection process, subjects follow instructions displayed on the screen to perform specific MM/MI tasks, while EEG signals are recorded at a sampling rate of 500 Hz. The tasks encompass four categories: MM/MI with the left hand, MM/MI with the right hand, MM/MI with the tongue, and rest.

The timing scheme for each trial is shown in Fig. 1 (bottom). Each trial begins with an instruction (one of the four above-mentioned tasks) that the subject performs for 4 seconds. This is followed by a blank screen (no task required for the subject) for a random duration between 5 and 6 seconds. A sequence of 40 trials (10 trials per category) constitutes a "run", where each category of tasks (left hand, right hand, tongue, and rest) is performed with a randomized order. Each data session consists of 10 runs, divided into 5 MM runs and 5 MI runs, i.e., 200 MM trials and 200 MI trials.

Our collected dataset comprises recordings from 5 subjects (2 females), aged between 25 and 35. Each subject participates in 4 separate data collection sessions on different days, with a minimum interval of 5 days between sessions. Once the EEG device is set up, it remains in place for the entire session to avoid variability caused by repositioning electrodes. In total, the dataset contains approximately 22.2 hours of EEG data. This paper focuses on the MM part of the dataset.

Compared to commonly used public datasets, such as the BCI Competition IV-2a dataset [27], [28] and the Physionet MM/MI dataset [32], [33], our dataset is collected using a lightweight 8-channel acquisition system as opposed to a bulky clinical setup. This represents a significant step towards evaluating the performance of consumer-grade BMI systems on real-world applications.

Procedures were in accordance with the ethical standards of the local institution and with the principles outlined in the Helsinki Declaration of 1975, as revised in 2000.

C. Preprocessing and Classification

1) Preprocessing: Several preprocessing steps are performed before feeding the raw EEG data into the classification model to enhance the signal quality and reduce noise. These steps are: (a) a fourth-order bandpass filter (0.5-100 Hz) to remove low-frequency drift and high-frequency noise; (b) a notch filter to suppress the 50 Hz power line interference; (c) a moving average filter with a sliding window of 0.25 seconds to reduce slow drifts and other low-frequency artifacts. For each 8-channel trial of 4 seconds, we take the intermediate 1900 samples (3.8 seconds with a sampling rate of 500 Hz) as the input to the network, i.e., an input size of 8×1900.

2) Classification: In this work, MI-BMInet [18] is used for classification. Among SoA ML models for EEG classification, MI-BMInet is a lightweight Convolutional Neural Network (CNN) that achieves SoA accuracy while requiring significantly fewer computational resources and power consumption than other EEG-based CNNs.

We present the network architecture in Fig. 2. The model is trained using the Adam optimizer and categorical cross-entropy loss function, with a dropout rate of 0.5 and a learning rate of 0.001. For within-session classification, we use a batch size of 4, and for TL/CL experiments, as explained below, we use a batch size of 10. All experiments are conducted on a subject-specific basis to tackle inter-session variability.

D. Inter-session Continual Learning

1) Transfer Learning Workflow: First, we analyze the conventional chain-TL workflow proposed in [10] on our collected dataset. An example of this workflow for a sequence of four data sessions is illustrated in Fig. 3a.

Based on four sessions, the TL workflow comprises one pretraining phase and three TL phases. Sessions 1, 2, 3, and 4 are numbered according to the acquisition sequence. In the pretraining phase, data from session 1 is used to obtain the pretrained model M_1. In the first TL phase, model M_2 is obtained by fine-tuning M_1 on the training data from session 2, which is divided chronologically into a training set and a test set. We guarantee time coherency (i.e., predict the future from past data) by taking the former part of the data as a training set and the latter part as a test set. Subsequent TL phases involve similar fine-tuning processes for sessions 3 and 4, resulting in models M_3 and M_4, respectively. For consistency, the three TL phases share the same number of TL epochs and the number of training/test trials in train-test splits.

The number of epochs required in the pretraining phase (training model M_1) is determined using an early-stopping strategy to five-fold Cross-Validation (CV) experiments on collected data. An average of approximately 40 epochs is found to be optimal. In each TL phase, 50 epochs are used, as this number typically allows the training set's loss and accuracy curves to converge with minimal fluctuations. For the train-test split, we use the first 60% of trials for training and the remaining 40% of trials for testing (denoted as a 60/40 split).

2) Continual Learning Methods and Workflow: Based on TL, the CL workflow additionally utilizes either previous data or previous weights to mitigate catastrophic forgetting and improve classification accuracy. An example framework is presented in Fig. 3b. The CL workflow starts with a model M_1 obtained by pretraining on session 1. Subsequent CL phases involve fine-tuning M_1 session-by-session on newer data sessions to obtain models M_2', M_3', and M_4', using specific CL updating strategies. We use 40 epochs for pretraining and 50 for CL. The train-test split is 60/40.

CL methods retain information from previous tasks or domains by utilizing old data or model parameters. As shown in Fig. 3b, two major categories of CL methods are investigated: data-based and regularization-based. As an example of data-based CL, ER maintains a replay buffer updated via reservoir

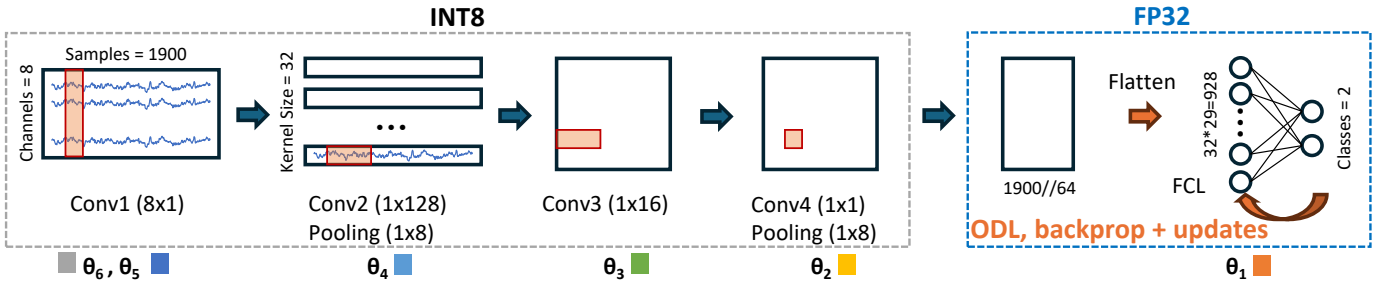
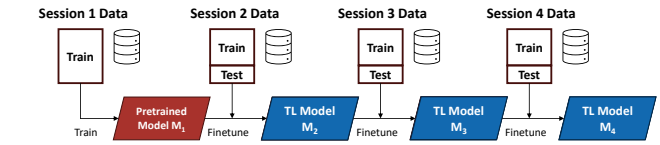
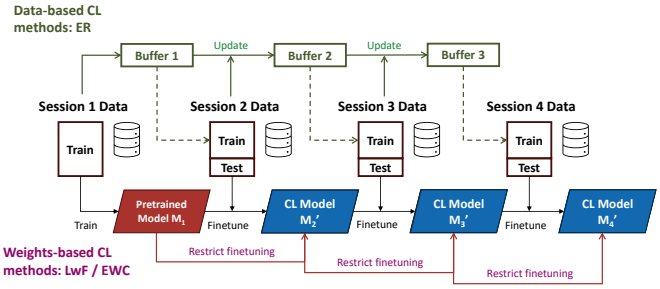


Fig. 2: The network architecture MI-BMINet [18] used in this work, selective layer update is shown for the θ_1 part of the network, this part of the network is fine-tuned in floating precision, while others are kept in `int8` format.



(a) TL workflow.



(b) CL workflow.

Fig. 3: Inter-session learning workflows for BMI systems.

sampling [45] to store training samples from previous sessions. We denote the total number of sessions as N_s and the current session number as n_s, and we have $N_s = 4$ and $1 \leq n_s \leq N_s$ in Fig. 3b. Initialized with sampled training data from session 1, the replay buffer will be updated with training data after each CL phase on session $n_s - 1$, and then used as a supplementary training set during the next CL phase on session n_s to obtain model M'_{n_s} $(n_s = 2, 3, 4)$. Using samples stored in the buffer, the model learns information from earlier experiences to mitigate the trend of forgetting.

In contrast, LWF and EWC belong to regularization-based CL and make use of only the new session's (session n_s) data while utilizing the previous model M'_{n_s-1} for regularization to preserve previous information. LWF feeds new data into the old model to obtain the output, which is used as a term of regularization in the loss function. On the other hand, EWC adds a penalty term based on the Fisher information matrix [30], which restricts the change of network weights from previous optimal weights. By using previous model weights, these methods prevent the learned model from deviating too much from the old model, thus keeping memory of the previous sessions or classification tasks.

Regarding implementing CL methods, a comparison is conducted between different methods (ER, LWF, EWC, and naive TL) using the Avalanche library based on PyTorch [46]. The buffer size of ER is set as 20 or 200 as two different options in our explorations. Additionally, the distillation hyperparameter λ_o is set to 1, temperature T to 2 in LWF as in [29], and the penalty parameter λ as 10000 in EWC, which is one of the standard options in previous works [47], [48].

E. Quantization

For on-device deployment, it is essential to quantize input data, network weights, and activations to meet the memory limitations of the GAP9 processor and exploit the ISA extensions for an efficient computation onboard. This work implements Quantization-Aware Training (QAT) to mitigate accuracy loss following quantization. The quantization is conducted on laptop and based on PyTorch and QuantLab [49], a software tool that trains and quantizes neural networks using QAT algorithms. The quantization procedure is performed once, prior to ODL, and the quantized model backbone is kept frozen and serves as the starting point for online, on-device updates on future sessions.

The following strategies and parameters are applied in our quantization workflow, converting from `fp32` to `int8`:

- QAT Workflow: The model is first trained in `fp32` for 40 epochs during the pretraining phase, followed by 50 epochs of QAT in each TL phase;
- Layer Quantization: Convolutional layers are quantized per channel, while activations are quantized per layer;
- Clipping bounds: Network weights are initialized symmetrically with their maximum values; activation bounds are initialized with constants -1 and 1;
- Learned Clipping Bounds: Clipping bounds for network weights and activations are learned through backpropagation, utilizing Trained Quantization Thresholds (TQT).

Other learning hyperparameters in QAT are consistent with those used in inter-session TL/CL with `fp32`.

F. On-device Deployment on GAP9

The computing platform used in this work is the commercial GAP9 chip by Greenwaves technologies [50]. GAP9 features 10 RISC-V cores, among which a RISC-V Fabric Controller (FC) core functions as the system controller, and the remaining nine cores, one for orchestration and eight workers, act as a

compute cluster. We use the cluster of 8 cores on GAP9 and speed up the ODL process.

Based on a hierarchical memory structure, GAP9 has a 128 kB shared L1 Tightly-Coupled Data Memory (TCDM) SRAM in the cluster, and a 1.5 MB L2 SRAM, as well as an on-chip L3 non-volatile eMRAM memory of 2 MB. Meanwhile, GAP9 exhibits up to 150.8 GOPS of processing power with a power efficiency of 0.33 mW/GOP, which facilitates deployments at ultra-low power and latency. Furthermore, GAP9 utilizes Dynamic Frequency and Voltage Scaling (DVFS) to effectively tune the energy consumption according to the requirements of the task and automatic clock gating to further reduce power consumption.

On-device deployment on the GAP9 processor is achieved using DORY, an automatic tool to deploy Deep Neural Networks (DNNs) on low-cost MCUs [51]. Specifically, DORY translates tiling into a constraint programming problem to maximize the utilization of L1 memory. Moreover, it optimizes target-specific PULP-NN heuristics, including adherence of a tile to the innermost loop, number of rows per core, and number of output channels, to maximize throughput and energy efficiency in the GAP9 target. Additionally, ODL is achieved using the combination of DORY and PULP-TrainLib, a DNN training library for multi-core RISC-V MCUs [38].

We use a selective layer update strategy [43] to comply with memory restrictions. Only the Fully-Connected Layer (FCL) in the MI-BMInet remains fine-tunable in `fp32` during on-device TL/CL, with other layers frozen and stored in `int8`. An illustration of this approach on MI-BMInet is shown in Fig. 2. The ODL workflow is described as follows with a pretrained, `int8` quantized network obtained from offline QAT on previous sessions as a starting point for online learning to a new session.

- Training Phase: Training inputs are fed into the `int8` network deployed with DORY until outputs before the FCL are obtained, producing a tensor of shape 32×29. These `int8` values are converted to `fp32` and expanded to a vector of length 928 for ODL. Forward and backward propagation during fine-tuning is conducted using PULP-TrainLib. Various methods, including TL, LWF, and ER, are implemented. Weights and biases of the FCL are updated based on gradient descent.
- Inference Phase: Inputs are fed into a combination of `int8` previous layers deployed with DORY and the `fp32` FCL deployed with PULP-TrainLib. Classification results are obtained as the outputs of the FCL.

IV. Results

A. Within-session Analyses

First, we analyze the newly collected data (before the CL analyses reported in the next section) by evaluating the within-session accuracy via 5-fold CV. Table I shows the within-session results for our 4 classes of MM. Comparatively, the performance on data sessions of subject A is overall the best, while subjects B and D perform the worst. We attribute this poor performance to the effects of BCI illiteracy [52].

TABLE I: Accuracy results [%] of 4-class MM within-session classification.

Sub.	S1	S2	S3	S4	Avg. Acc.
A	72.5 ± 8.4	70.5 ± 6.0	65.5 ± 9.9	63.5 ± 2.6	68.0 ± 3.6
B	31.3 ± 12.4	48.5 ± 7.5	45.9 ± 8.4	53.5 ± 4.1	44.8 ± 4.3
C	58.5 ± 12.5	61.0 ± 4.6	59.5 ± 8.3	51.0 ± 9.0	57.5 ± 4.5
D	35.5 ± 5.1	32.3 ± 6.0	38.0 ± 5.8	35.5 ± 5.1	35.3 ± 2.8
E	50.5 ± 7.0	68.5 ± 8.2	50.0 ± 6.5	50.5 ± 4.9	54.9 ± 3.4

TABLE II: Results of 2-class Tongue/Rest MM within-session classification.

(a) Accuracy results [%] of classification.

Sub.	S1	S2	S3	S4	Avg. Acc.
A	91.0 ± 5.8	97.0 ± 2.5	84.0 ± 7.4	88.0 ± 5.1	90.0 ± 2.7
B	71.2 ± 9.5	89.0 ± 5.8	87.5 ± 9.1	93.0 ± 5.1	85.2 ± 3.8
C	86.0 ± 6.6	86.0 ± 7.4	89.0 ± 2.0	87.0 ± 8.1	87.0 ± 3.2
D	81.0 ± 8.6	73.2 ± 8.2	71.0 ± 11.6	68.0 ± 6.8	73.3 ± 4.5
E	73.0 ± 6.8	71.0 ± 10.2	80.0 ± 6.3	91.0 ± 3.7	78.8 ± 3.6

(b) Precision results [%] of classification.

Sub.	S1	S2	S3	S4	Avg. Pre.
A	91.7 ± 7.7	95.3 ± 5.8	85.2 ± 12.3	86.5 ± 7.1	89.7 ± 4.1
B	100.0 ± 0.0	86.6 ± 9.1	84.9 ± 9.3	95.8 ± 5.3	91.8 ± 6.3
C	86.2 ± 9.8	88.1 ± 11.2	88.1 ± 9.9	90.4 ± 5.8	88.2 ± 1.5
D	79.8 ± 17.8	78.3 ± 15.3	66.6 ± 17.5	75.5 ± 16.9	75.1 ± 5.1
E	70.1 ± 5.6	71.2 ± 12.3	81.4 ± 16.6	93.2 ± 5.7	79.0 ± 9.3

(c) Recall/Sensitivity results [%] of classification.

Sub.	S1	S2	S3	S4	Avg. Rec.
A	90.7 ± 8.6	98.7 ± 2.7	83.4 ± 9.2	90.6 ± 11.0	90.9 ± 5.4
B	39.8 ± 20.1	92.2 ± 6.7	91.8 ± 7.0	89.7 ± 8.9	78.4 ± 22.3
C	87.2 ± 11.1	86.4 ± 8.1	90.5 ± 5.9	85.1 ± 9.5	87.3 ± 2.0
D	82.8 ± 11.4	68.4 ± 10.3	80.6 ± 6.9	53.2 ± 22.3	71.3 ± 11.8
E	79.7 ± 11.3	69.1 ± 10.5	77.0 ± 5.8	87.9 ± 7.5	78.4 ± 6.7

(d) Specificity results [%] of classification.

Sub.	S1	S2	S3	S4	Avg. Spe.
A	92.5 ± 7.4	96.8 ± 3.9	83.3 ± 12.1	86.3 ± 6.0	89.7 ± 5.3
B	100.0 ± 0.0	86.8 ± 8.8	81.9 ± 14.1	96.0 ± 5.0	91.2 ± 7.2
C	86.7 ± 8.1	85.4 ± 16.6	88.4 ± 7.8	88.7 ± 8.0	87.3 ± 1.3
D	82.8 ± 14.0	79.0 ± 13.3	61.1 ± 13.2	82.4 ± 11.5	76.3 ± 8.9
E	65.2 ± 9.1	73.6 ± 10.6	82.2 ± 15.4	94.5 ± 4.5	78.9 ± 10.8

Overall, the results of subject A (68.00%), C (57.50%), and E (54.88%) are comparable with the SoA. In fact, in [18], 4-class classification on 8-channel EEG data achieves 60% accuracy on PhysioNet MM/MI dataset.[2]

Considering that the regions of the motor cortex under the headband are mostly responsible for tongue movements [53], [54], we also analyze the 2-class classification between tongue and rest (see Table II). Subject A achieves overall best performance with up to 97% session-specific accuracy, optimal

[2]Note that in [18], the results are obtained from the automatically-selected 8 best channels among the 64 channels distributed over the entire scalp, whereas in our dataset the 8 channels are restricted to specific ones for the non-stigmatizing nature of the headband.

average accuracy and recall, and near-optimal precision and specificity. Different metrics mostly show similar results on the same session except for a few cases, e.g., S1 of subject B. Compared to the SoA work of [55], where the 2-class classification between tongue and rest classes using 64-channel EEG data achieved ~91-95% accuracy, our work demonstrates that the proposed non-stigmatizing dry-electrode headband and the classification model can decode useful information to perform specific BMI tasks.

B. Inter-session Continual Learning

We implement and evaluate the TL and CL methods presented in Sect. III-D on two in-house EEG datasets:

- Dataset A: a previously acquired MM dataset, with seven data sessions on one subject [10] with 2-class classification between the left- and right-hand MM. We consider the first 100 trials of each session, with the first 60 trials used as the training set and the remaining 40 trials as the test set.
- Dataset B: the newly acquired dataset presented in this work. We present the results of the 2-class tongue/rest MM classification and the 4-class classification of subject A. Each session, therefore, has a size of 100 trials and is split using the same 60/40 split.

We present the results as the average accuracy over five replications of experiments with different random seeds.

1) Performance Analysis Across Multiple Sessions: We evaluate the average performance of TL and CL methods across all acquired sessions when progressively adding more sessions (i.e., for every newly acquired session n_s, we report the average performance $P(1:n_s)$ from session 1 to session n_s in terms of accuracy $Acc(1:n_s)$, precision $Pre(1:n_s)$, recall/sensitivity $Rec(1:n_s)$, and specificity $Spe(1:n_s)$).

In naive TL, catastrophic forgetting negatively impacts model performance averaged over all sessions as the number of sessions increases. As more sessions are acquired, information from previous data sessions is lost during the fine-tuning process, resulting in a decreasing trend in average accuracy across all sessions (see Fig. 4a, black dashed line). When more novel sessions are considered, the average test accuracy decreases from 100% to 59.57%. Conversely, CL methods (ER, LWF) mitigate this degradation (Fig. 4a, red, blue, and orange lines). LWF achieves an average accuracy of 67.14%. ER with a buffer size of 20 achieves an improvement of up to 16.72% compared to naive TL. When increasing the ER buffer size to 200, the improvement increases to up to 30.36% with an average accuracy of 89.93%. Additionally, EWC sometimes performs worse than naive TL but still shows comparable performance with an average accuracy of 65.79%.

Regarding other metrics in Fig. 4, ER shows an overall good performance at each phase. ER outperforms naive TL in most cases, with all metrics above 70% for buffer size 20 and above 80% for buffer size 200. Meanwhile, ER exhibits smoother curves while naive TL, LWF, and EWC show more fluctuations. Compared to naive TL, LWF shows comparable precision and specificity and achieves better recall in most cases. On the other hand, EWC achieves significantly higher recall and lower precision and specificity than naive TL.

We repeat the same analyses for Dataset B (Fig. 5). In the 2-class MM classification, similarly to the results of Dataset A, in Dataset B we also observe a decreasing trend in average accuracy using naive TL, whereas CL methods, including ER and LWF, mitigate this trend by outperforming TL at each phase. Conversely, EWC does not always outperform TL. In this workflow, ER with a buffer size of 200 performs best among all CL methods, showing an increase in accuracy of up to 10.17% compared to naive TL. Meanwhile, ER with a smaller buffer size of 20 improves accuracy by up to 5.17%, and LWF improves accuracy by up to 5.25% compared to naive TL. We also observe from Fig. 5 that all CL methods achieve better average precision and specificity than naive TL. Regarding recall, ER shows comparable results as naive TL, while LWF performs slightly worse and EWC performs the worst among all methods.

In the 4-class classification, CL methods except EWC can also mitigate the decreasing trend of average accuracy as shown in Fig. 7a. Among all CL methods, ER with a buffer size of 200 performs overall the best with an improvement of up to 11.12% compared to naive TL.

2) Performance Analysis on the New Session: Here, when progressively considering more sessions, we evaluate the performance $P(n_s)$ of TL and CL methods on the newest session n_s in terms of accuracy $Acc(n_s)$, precision $Pre(n_s)$, recall/sensitivity $Rec(n_s)$, and specificity $Spe(n_s)$. We limit this evaluation to Dataset B.

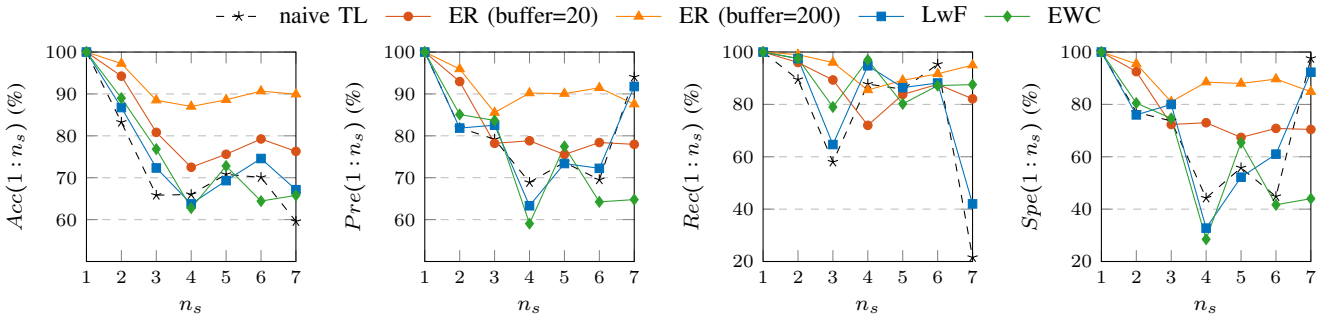
In the 2-class classification, the results in Fig. 6a show a progressive improvement with an increasing number of acquired sessions for methods including naive TL, ER (buffer size=20), and LWF. ER and LWF achieve comparable accuracy as naive TL, with a maximum difference of 4%. This demonstrates that retaining past experience using CL does not hinder the model's capability to learn from new data sessions, except for EWC. Considering other metrics in Fig. 6, ER and naive TL achieve comparable performance in terms of precision, recall, and specificity with a maximum difference of 5%. LWF exhibits significantly higher precision and specificity and lower recall in session 2, while EWC performs worse than other methods in most cases. When measuring 4-class classification on the newest session (Fig. 7b), EWC performs the worst while other CL methods show comparable or worse performance compared to naive TL.

The diminished efficacy of EWC, as also depicted in Fig. 5a, is attributable to its worse performance on the most recent session, leading to the decision against its on-device deployment.

C. On-device Deployment on GAP9

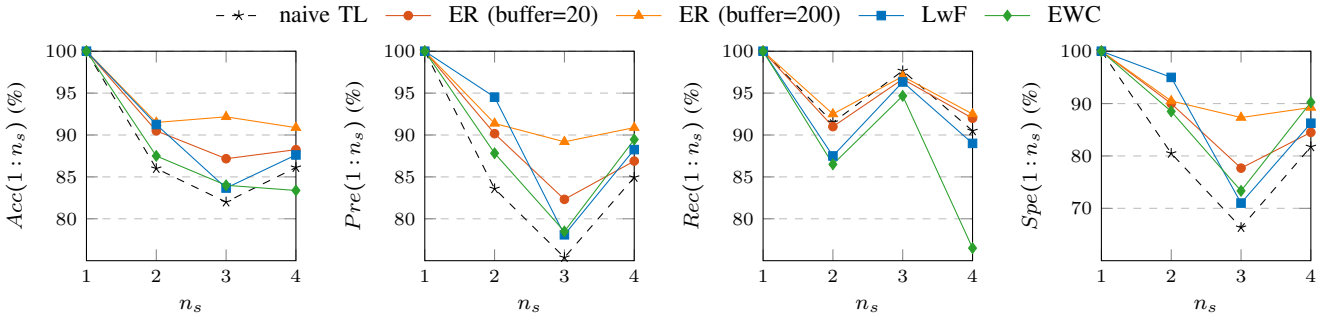
In this section, we discuss the deployment of various methodologies, including TL, ER, and LWF, on the GAP9 processor. The implementation of on-device training and inference was conducted using DORY and PULP-TrainLib. Detailed procedural information can be found in III-E.

1) Quantization: We perform quantization of the TL models fine-tuned on the data from subject A to verify the proposed



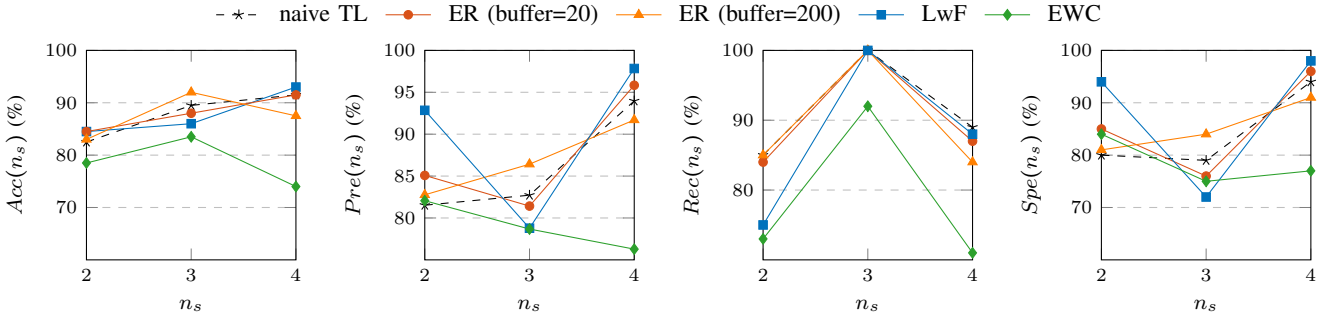
(a) Accuracy $Acc(1:n_s)$ (b) Precision $Pre(1:n_s)$ (c) Recall/Sensitivity $Rec(1:n_s)$ (d) Specificity $Spe(1:n_s)$

Fig. 4: Performance across all sessions when progressively adding more sessions, comparing TL and CL in 2-class MM classification on Dataset A.



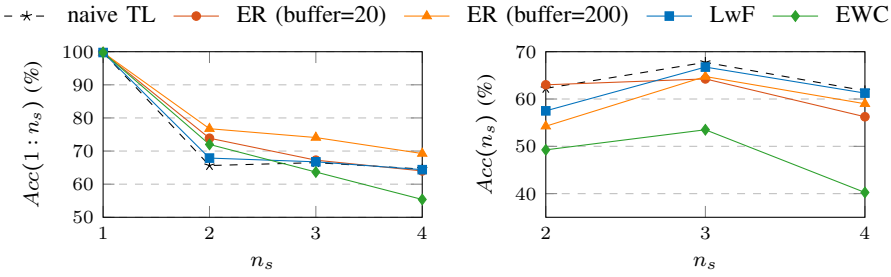
(a) Accuracy $Acc(1:n_s)$ (b) Precision $Pre(1:n_s)$ (c) Recall/Sensitivity $Rec(1:n_s)$ (d) Specificity $Spe(1:n_s)$

Fig. 5: Performance across all sessions when progressively adding more sessions, comparing TL and CL in 2-class MM classification on Dataset B.



(a) Accuracy $Acc(n_s)$ (b) Precision $Pre(n_s)$ (c) Recall/Sensitivity $Rec(n_s)$ (d) Specificity $Spe(n_s)$

Fig. 6: Performance on the newest session, comparing TL and CL in 2-class MM classification on Dataset B.



(a) Accuracy $Acc(1:n_s)$ (b) Accuracy $Acc(n_s)$

Fig. 7: Accuracy comparing TL and CL in 4-class MM classification on Dataset B. (a) Average accuracy across sessions when progressively adding more sessions; (b) Accuracy on the newest session.

TABLE III: TL accuracy results [%] before and after quantization (subject A, batchsize=10).

TL phase	`int8`	`fp32`
TL to S2	85.0 ± 3.2	84.0 ± 1.2
TL to S3	88.5 ± 3.4	91.0 ± 3.0
TL to S4	90.3 ± 1.0	88.7 ± 4.2

TABLE IV: Metrics of the on-device deployment of TL and CL methods on the GAP9 processor.

Metrics	TL&ER		LwF	
(Forward, Backward)	→	←	→	←
Time [ms]	20.1	1.5	20.1	1.4
Avg. Power [mW]	21.1	13.2	21.1	13.2
L1 [kB]	18.6		26.0	
L2 [kB]	TL: 510 ER: 814		510	

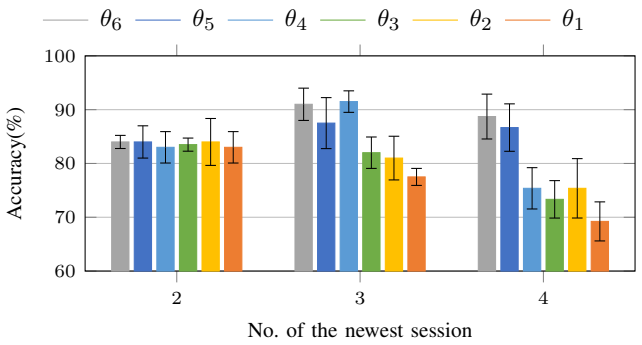


Fig. 8: Accuracy of fine-tuning network with different adaptation depths after each chain-TL phase tested on the newest session, conducted using QuantLab on one of the subjects in the MM dataset acquired in this work. As labeled in Fig. 2, θ_1 to θ_5 denotes the first fine-tunable layer with an increasing adaptation depth; θ_6 denotes the full network is fine-tunable, and further includes all BatchNorm layers.

quantization flow. We exclude 10 trials presenting outlier values due to artifacts in session 4 caused by a few electrodes losing adherence to the scalp during subject movements after visual inspection. We present the accuracy results before and after quantization using PyTorch and QuantLab in Table III. The performance differences between the `int8` and `fp32` models fall within standard deviation ranges, confirming that deploying the backbone model in `int8` on GAP9 does not compromise the accuracy.

2) On-board Measurements: Key statistics, including adaptation time, L1/L2 memory consumption, and power consumption for deploying TL and CL methods on GAP9, are summarized in Table IV. The results indicate that CL methods exhibit resource consumption comparable to TL.

Specifically, LWF requires approximately 7.4 kB more L1 memory than TL, and ER uses an additional 304 kB of L2 memory with a buffer size of 20 samples. The L1 memory constraint of GAP9 is 128kB. We are using up to 20.3% of this memory in ODL. The L2 memory constraint of GAP9 is 1.5MB, of which we are using 34% with TL and LWF, and 54.3% with ER (buffer size = 20).

All methods maintain an adaptation time of around 22 ms. Power consumption for the complete system for TL, ER, and LWF is around the same at 14.95 mW. More specifically, on-device processing and learning consume ~21 mW as shown in Table IV, and BioGAP consumes around 14.5 mW in signal acquisition, including the ADS1298 AFE (Texas Instruments) and the active channel circuitry. As all processing and learning is done on-device, the nRF52911 BLE is inactive, and it does not contribute to the total power consumption of the system. Correspondingly, ODL on BioGAP yields around 25 h of battery life with a 100 mA h, 3.7 V battery assuming fine-tuning every second, i.e., $(100 \times 3.7)/14.95 = 25$ h.

Overall, the CL methods provide an efficient BMI system, enhancing privacy, accuracy, and resource efficiency. The BMI system processes streaming EEG signals in real-time, operates under ultra-low-power conditions, and ensures low latency and extended battery life.

3) Selective Layer Updates & Future Improvements: Our approach involves fine-tuning only the FCL during ODL. While efficient, this method results in some accuracy degradation, which can be mitigated by allowing more flexibility in fine-tuning additional layers. Fig. 8 presents results from QuantLab, showing accuracy improvements when varying the number of fine-tunable layers.

The adaptation depth, indicating the number of fine-tunable layers θ_i, significantly affects accuracy. For instance, the grey bar in Fig. 8, representing full-network fine-tuning, shows that increased flexibility yields better accuracy, especially in later TL sessions, compared to fine-tuning fewer layers, e.g., only the last FCL represented in orange. Overall, accuracy improvements from 1% to 19.49% are observed across successive sessions, underscoring the benefits of fine-tuning more layers in long-term usage.

The constraint of fine-tuning only the FCL arises from the L1 memory limitations of the GAP9 processor when using the PULP-TrainLib library. Future work will focus on implementing advanced tiling strategies and dynamic memory allocation to facilitate more comprehensive layer fine-tuning.

Moreover, with an input shape of 8×1900, the output feature map of the first convolutional layer in MI-BMInet is a 32×1900 matrix, which consumes approximately 243 kB of memory in `fp32`. Optimization strategies such as downsampling [56] will be further explored to decrease memory consumption.

Future studies will also explore model architectures to enhance the model's capability to learn representative features for an optimal selective layer update with ODL.

D. Comparison with SoA Works

Table V compares our work to the SoA. Most previous works [24]–[26] focused on inter-session TL without explorations on CL methods. A few recent works [14], [15] applied CL on EEG data, yet they did not address inter-session variability. Meanwhile, none of the above-mentioned works involved on-device learning or inference. Compared to SoA

TABLE V: Comparison between SoA related works on BMIs.

	Application	Method	TL	CL	Intra-Session Adaptation	ODL	Ultra-Low-Power Deployment	Accuracy	# Params	On-board Measurements
Ma [24]	MM/MI Classification	CNN	yes	no	yes	no	no	2-class: 78.9%	291.6 k	no
Wu [25]	MM/MI Classification	Common Spatial Patterns (CSP)	yes	no	yes	no	no	2-class: 76.02%	-	no
Lee [26]	MM/MI Classification	TSTN	yes	no	yes	no	no	2-class: 77.00%	8.9 k	no
Rajpura [15]	MM/MI Classification	Long Short-Term Memory Network	yes	yes	no	no	no	2-class: 59.36%	33.2 k	no
Shahbazinia [14]	Seizure Detection	Fully Convolutional Network	yes	yes	no	no	no	F1 Score: 86.63% False Alarm Rate: 5.14%	~ 300 k	no
This work	**MM/MI Classification**	**CNN**	**yes**	**yes**	**yes**	**yes**	**yes**	**Dataset A 2-class:** 89.93% **Dataset B 2-class:** 90.88%	7.7 **k**	**Latency:** 21.5 ms **Avg. Power:** 21 mW **L1 Mem.:** 18.6 **/** 26.0 kB **L2 Mem.:** 510 **/** 814 kB

works on BMIs, this work is the first to implement on-device CL to achieve intra-session adaptation on PULP MCU. While achieving SoA classification accuracy in MM classification, our on-board deployment based on ODL pave the way to real-world applications of BMI systems on the extreme edge.

V. Conclusion

In this work, we proposed a comfortable, non-stigmatizing BMI headband based on ultra-low-power BioGAP system [23], featuring accurate classification and on-device continual learning capabilities. We collected a multi-session EEG dataset based on motor tasks from five subjects, totaling 8,000 trials of approximately ~22.2 hours of recordings. To address inter-session variability and mitigate catastrophic forgetting, we introduce CL methods, outperforming traditional transfer learning by up to 10.17% on our dataset and up to 30.36% on a previously published dataset [10].

For real-time adaptation on edge devices, we implemented ODL using TL and CL methods on the GAP9 processor within BioGAP, utilizing DORY and PULP-TrainLib. Our measurements indicate that CL methods have comparable resource consumption to TL in terms of memory and power, while allowing the model to retain knowledge from past experience. With inference and learning on the edge, our setup yields low latency and extended battery life, with $25\,\mathrm{h}$ of operation on a $100\,\mathrm{mA\,h}$, $3.7\,\mathrm{V}$ battery while performing on-device fine-tuning every second.

Apart from MM tasks, the proposed headband has the potential to be applied to visual or concentration tasks related to the visual, frontal, and parietal cortex. With a similar but more comprehensive coverage compared to the Muse headband [8], our system is capable of onboard processing and is specifically suitable for applications including concentration and stress measurements [57] and Steady State Visually Evoked Potentials (SSVEP)-based BMIs [58] that can be explored in future investigations.

References

[1] J. R. Wolpaw, J. D. R. Millan, and N. F. Ramsey, "Brain-computer interfaces: Definitions and principles," *Handbook of clinical neurology*, vol. 168, pp. 15–23, 2020.

[2] M. A. Lebedev and M. A. Nicolelis, "Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation," *Physiological reviews*, vol. 97, no. 2, pp. 767–837, 2017.

[3] J. Zhuang, K. Geng, and G. Yin, "Ensemble learning based brain–computer interface system for ground vehicle control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 9, pp. 5392–5404, 2021.

[4] G. Pfurtscheller, "Functional brain imaging based on erd/ers," *Vision research*, vol. 41, no. 10-11, pp. 1257–1260, 2001.

[5] A. Casson, "Wearable EEG and beyond," *Biomedical Engineering Letters*, vol. 9, pp. 53–71, 01 2019.

[6] Evoke Neuroscience Inc., "The versus headset," https://getversus.com/headset, Accessed: 2024-04-05.

[7] Emotiv, "Epoc+," https://www.emotiv.com/epoc/, Accessed: 2024-04-05.

[8] Muse, "EEG-powered meditation and sleep headband," https://choosemuse.com/, Accessed: 2024-04-05.

[9] T. M. Ingolfsson, U. Chakraborty *et al.*, "EpiDeNet: An energy-efficient approach to seizure detection for embedded systems," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2023, pp. 1–5.

[10] X. Wang, L. Mei *et al.*, "Enhancing performance, calibration time and efficiency in brain-machine interfaces through transfer learning and wearable EEG technology," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2023, pp. 1–5.

[11] X. Chen, S. Wang *et al.*, "Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[12] H. Qu, H. Rahmani *et al.*, "Recent advances of continual learning in computer vision: An overview," *arXiv preprint arXiv:2109.11369*, 2021.

[13] Z. Mai, R. Li *et al.*, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[14] A. Shahbazinia, F. Ponzina *et al.*, "Resource-efficient continual learning for personalized online seizure detection," in *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

[15] P. Rajpura, P. Pandey, and K. Miyapuram, "Continual learning for EEG based brain computer interfaces," in *Continual Lifelong Learning Workshop at ACML 2022*, 2022.

[16] X. Wang, L. Cavigelli *et al.*, "Sub-100 µW multispectral Riemannian classification for EEG-based brain–machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 6, pp. 1149–1160, 2021.

[17] V. Kartsch, G. Tagliavini *et al.*, "BioWolf: A sub-10-mW 8-channel advanced brain–computer interface platform with a nine-core processor and BLE connectivity," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 5, pp. 893–906, 2019.

[18] X. Wang, M. Hersche *et al.*, "MI-BMInet: An efficient convolutional neural network for motor imagery brain–machine interfaces with eeg channel selection," *IEEE Sensors Journal*, vol. 24, no. 6, pp. 8835–8847, 2024.

[19] BioSemi, "Biosemi EEG ECG EMG BSPM neuro amplifier electrodes," https://www.biosemi.com/Products_ActiveTwo.htm, Accessed: 2024-04-05.

[20] Neuroelectrics, "Enobio 32: wireless high-density EEG medical grade system for high-precision brain research," https://www.neuroelectrics.com/solutions/enobio/32, Accessed: 2024-04-05.

[21] E. Ratti, S. Waninger *et al.*, "Comparison of medical and consumer wireless EEG systems for use in clinical trials," *Frontiers in human neuroscience*, vol. 11, p. 398, 2017.

[22] OpenBCI, "OpenBCI," https://openbci.com, Accessed: 2024-05-28.

[23] S. Frey, M. Guermandi *et al.*, "BioGAP: A 10-core FP-capable ultra-low power IoT processor, with medical-grade AFE and BLE connectivity for wearable biosignal processing," in *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. IEEE, 2023, pp. 1–7.

[24] J. Ma, B. Yang *et al.*, "A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface," *Scientific Data*, vol. 9, no. 1, p. 531, 2022.

[25] D. Wu, X. Jiang, and R. Peng, "Transfer learning for motor imagery based brain–computer interfaces: A tutorial," *Neural Networks*, vol. 153, pp. 235–253, 2022.

[26] P.-L. Lee, S.-H. Chen *et al.*, "Continual learning of a transformer-based deep learning classifier using an initial model from action observation EEG data to online motor imagery classification," *Bioengineering*, vol. 10, no. 2, p. 186, 2023.

[27] C. Brunner, R. Leeb *et al.*, "BCI competition 2008–Graz data set a," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[28] M. Tangermann, K.-R. Müller *et al.*, "Review of the BCI competition IV," *Frontiers in neuroscience*, vol. 6, p. 55, 2012.

[29] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[30] J. Kirkpatrick, R. Pascanu *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[31] D. Rolnick, A. Ahuja *et al.*, "Experience replay for continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[32] G. Schalk, D. J. McFarland *et al.*, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[33] A. L. Goldberger, L. A. Amaral *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[34] R. David, J. Duke *et al.*, "Tensorflow lite micro: Embedded machine learning for tinyml systems," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 800–811, 2021.

[35] uTensor, "uTensor TinyML AI inference library." https://utensor.github.io/, Accessed: 2024-05-28.

[36] Y. D. Kwon, J. Chauhan *et al.*, "Exploring system performance of continual learning for mobile and embedded sensing applications," in *2021 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2021, pp. 319–332.

[37] H. Ren, D. Anicic, and T. A. Runkler, "Tinyol: Tinyml with online-learning on microcontrollers," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[38] D. Nadalini, M. Rusci *et al.*, "PULP-TrainLib: Enabling on-device training for RISC-V multi-core MCUs through performance-driven autotuning," in *Embedded Computer Systems: Architectures, Modeling, and Simulation*, A. Orailoglu, M. Reichenbach, and M. Jung, Eds. Cham: Springer International Publishing, 2022, pp. 200–216.

[39] L. Ravaglia, M. Rusci *et al.*, "A tinyml platform for on-device continual learning with quantized latent replays," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 4, pp. 789–802, 2021.

[40] C. Cioflan, L. Cavigelli *et al.*, "On-device domain learning for keyword spotting on low-power extreme edge embedded systems," *arXiv preprint arXiv:2403.10549*, 2024.

[41] A. Li, H. Li, and G. Yuan, "Continual learning with deep neural networks in physiological signal data: A survey," in *Healthcare*, vol. 12, no. 2. MDPI, 2024, p. 155.

[42] T. Stuart, J. Hanna, and P. Gutruf, "Wearable devices for continuous monitoring of biosignals: Challenges and opportunities," *APL bioengineering*, vol. 6, no. 2, 2022.

[43] J. Lin, L. Zhu *et al.*, "On-device training under 256KB memory," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed *et al.*, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 22 941–22 954.

[44] Datwyler Schweiz AG, "Softpulse," https://datwyler.com/company/innovation/wearable-sensors/softpulse, Accessed: 2024-04-05.

[45] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, vol. 11, no. 1, pp. 37–57, 1985.

[46] A. Carta, L. Pellegrini *et al.*, "Avalanche: A PyTorch library for deep continual learning," *Journal of Machine Learning Research*, vol. 24, no. 363, pp. 1–6, 2023.

[47] H. Liu, Y. Yang, and X. Wang, "Overcoming catastrophic forgetting in graph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 10, 2021, pp. 8653–8661.

[48] W. Zhang, D. Li *et al.*, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2864–2878, 2022.

[49] M. Spallanzani, L. Cavigelli *et al.*, "Additive noise annealing and approximation properties of quantized neural networks," *arXiv preprint arXiv:1905.10452*, 2019.

[50] GreenWaves Technologies, "GAP9 processor," https://greenwaves-technologies.com/gap9_processor/, Accessed: 2024-04-05.

[51] A. Burrello, A. Garofalo *et al.*, "DORY: Automatic end-to-end deployment of real-world dnns on low-cost IoT mcus," *IEEE Transactions on Computers*, pp. 1–1, 2021.

[52] S. Becker, K. Dhindsa *et al.*, "BCI illiteracy: It's us, not them. optimizing BCIs for individual brains," in *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, 2022, pp. 1–3.

[53] S. Kakei, D. S. Hoffman, and P. L. Strick, "Muscle and movement representations in the primary motor cortex," *Science*, vol. 285, no. 5436, pp. 2136–2139, 1999.

[54] A. Nakamura, T. Yamada *et al.*, "Somatosensory homunculus as drawn by MEG," *NeuroImage*, vol. 7, no. 4, pp. 377–386, 1998.

[55] R. L. Kæseler, T. W. Johansson *et al.*, "Feature and classification analysis for detection and classification of tongue movements from single-trial pre-movement EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 678–687, 2022.

[56] X. Wang, M. Hersche *et al.*, "An accurate EEGNet-based motor-imagery brain–computer interface for low-power edge computing," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2020, pp. 1–6.

[57] M. M. Maddox, A. Lopez *et al.*, "Electroencephalographic monitoring of brain wave activity during laparoscopic surgical simulation to measure surgeon concentration and stress: Can the student become the master?" *Journal of endourology*, vol. 29, no. 12, pp. 1329–1333, 2015.

[58] Z. Cao, C.-T. Lin *et al.*, "Extraction of ssveps-based inherent fuzzy entropy using a wearable headband EEG in migraine patients," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 1, pp. 14–27, 2019.



Lan Mei (Graduate Student Member, IEEE) received her Bachelor's degree in Electronics and Information Engineering from Beihang University in 2020. She is currently pursuing her Master's degree in Electrical Engineering and Information Technology at ETH Zürich, Switzerland. Her research interests include brain machine interface, biosignal acquisition and processing, machine learning, and edge computing on microcontrollers.



Thorir Mar Ingolfsson (Graduate Student Member, IEEE) received his Bachelor's degree in Electrical and Computer Engineering from the University of Iceland in 2018. He then completed his Master's degree in Electrical Engineering and Information Technology from ETH Zürich, Switzerland, in 2020. Currently, he is pursuing his Ph.D. at ETH Zürich, working under the supervision of Prof. Dr. Luca Benini at the Integrated Systems Laboratory. His research primarily focuses on biosignal processing and machine learning, emphasizing low-power embedded systems and energy-efficient implementation of machine learning models on microcontrollers.



Cristian Cioflan (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering and information technology from the University Politehnica of Bucharest, Bucharest, Romania, in 2018, and the M.Sc. degree from Swiss Federal Institute of Technology Zürich, Zürich, Switzerland, in 2020, where he is currently pursuing the Ph.D. degree with the Digital Circuits and Systems Group of Prof. Luca Benini. His research interests include on-device continual learning, audio processing in low-power embedded systems, and neural architecture search for energy-efficient inference.



Victor Kartsch received the Ph.D. degree in electrical engineering and computer science from the University of Bologna, in 2020 (Ph.D. Advisor Prof. Luca Benini). During the Ph.D., he worked on the hardware-software design of fully embedded human-machine interaction (HMI) systems with a full-stack perspective, targeting both EMG and EEG signals. One of the most important systems developed by Dr. Kartsch includes BioWolf, an ultra-low-power HMI for signal acquisition and real-time processing of computationally-heavy algorithms, which has been adopted by many research institutions in the field. His work experience also includes data analysis and optimization of signal processing and machine learning algorithms for embedded systems. He has published several papers in international peer-reviewed conferences and journals. Currently, he's a Research Fellow with the Integrated System Laboratory, ETH Zurich, where he also works on designing wearable systems for several applications. At ETH, he's also working on the design of hardware and software solutions for UAVs, and, by extension, on integrating such systems with HMIs for advanced control.



Andrea Cossettini (Member, IEEE) received the Ph.D. degree in Electronic Engineering from the University of Udine (Udine, Italy) in 2019, working on nanoelectrode array biosensors. Previously, he was with Acreo Swedish ICT AB (Kista, Sweden), designing waveguide-to-chip transitions at sub-mm waves, and with Infineon Technologies (Villach, Austria), working on signal integrity for high-speed serial interfaces. He joined ETH Zurich in 2019. His research interests are in biomedical circuits and systems, with a special focus on wearable/high-speed ultrasound and wearable EEG. He currently serves as Research Cooperation Manager of the ETH Future Computing Laboratory (EFCL), Project Leader at the Integrated Systems Laboratory, and Lecturer.



Xiaying Wang (Member, IEEE) received her B.Sc. and M.Sc. degrees in biomedical engineering from Politecnico di Milano, Italy, and ETH Zurich, Switzerland, in 2016 and 2018, respectively. She won the Ph.D. Fellowship funded by the Swiss Data Science Center in 2019 and obtained the Ph.D. degree in electrical engineering from ETH Zurich, Switzerland, in 2023 (Doctor of Science ETH). She is currently a postdoctoral researcher at the Integrated Systems Laboratory at ETH Zurich and at the Swiss University of Traditional Chinese Medicine in Switzerland. Her research interests include biosignal processing, brain–machine interface, smart wearable devices, edge computing, and applied machine learning.



Luca Benini (Fellow, IEEE) holds the chair of digital Circuits and systems at ETHZ and is Full Professor at the Universita di Bologna. He received a PhD from Stanford University. Dr. Benini's research interests are in energy-efficient parallel computing systems, smart sensing micro-systems and machine learning hardware. He is a Fellow of the IEEE, of the ACM and a member of the Academia Europaea. He is the recipient of the 2016 IEEE CAS Mac Van Valkenburg award, the 2020 EDAA achievement Award, the 2020 ACM/IEEE A. Richard Newton Award and the 2023 IEEE CS E.J. McCluskey Award.