Enhancing Performance, Calibration Time and Efficiency in Brain-Machine Interfaces through Transfer Learning and Wearable EEG Technology

Xiaying Wang*†, Lan Mei†, Victor Kartsch†, Andrea Cossettini†, Luca Benini†‡

†Dept. ITET, ETH Zurich, Zurich, Switzerland ‡DEI, University of Bologna, Bologna, Italy

***Abstract*—Brain-Machine Interfaces (BMIs) have emerged as a transformative force in assistive technologies, empowering individuals with motor impairments by enabling device control and facilitating functional recovery. However, the persistent challenge of inter-session variability poses a significant hurdle, requiring time-consuming calibration at every new use. Compounding this issue, the low comfort level of current devices further restricts their usage. To address these challenges, we propose a comprehensive solution that combines a tiny CNN-based Transfer Learning (TL) approach with a comfortable, wearable EEG headband. The novel wearable EEG device features soft dry electrodes placed on the headband and is capable of on-board processing. We acquire multiple sessions of motor-movement EEG data and achieve up to 96% inter-session accuracy using TL, greatly reducing the calibration time and improving usability. By executing the inference on the edge every 100 ms, the system is estimated to achieve 30 h of battery life. The comfortable BMI setup with tiny CNN and TL pave the way to future on-device continual learning, essential for tackling inter-session variability and improving usability.**

***Index Terms*—brain-computer interface, EEG, wearable healthcare, wearable EEG, deep learning, transfer learning**

I. Introduction

Electroencephalography (EEG) is the most common non-invasive technique used to measure and record the electrical activity of the brain. It is widely used in medical and research settings to study brain function, diagnose neurological disorders, and monitor the brain's status. Furthermore, EEG plays a central role in brain–machine interfaces (BMIs) bridging the human brain and external machines. One attractive BMI paradigm is based on the motor function where the BMI device can decode which body parts a subject moves—motor movement (MM)—or imagines to move—motor imagery (MI)—based on EEG signals [1].

Over recent years, many EEG devices have appeared on the consumer market. A few examples are Versus headset [2], Melomind [3], Emotiv EPOC+ [4], and Muse headband [5]. While they are much less intrusive and stigmatizing than clinically-used EEG devices, they face a major problem related to low signal quality. Compared to the gold standard EEG headsets based on gel, which are not suitable for out-of-lab daily usage mostly due to the long setup time and obtrusiveness, consumer-grade EEG devices use dry electrodes to improve the user experience and reduce the setup time but yield low signal-to-noise ratio (SNR). Authors in [6] used the Muse headband to collect MI signals reaching a two-class classification accuracy ranging between 74% and 93%. However, a significant amount of data was polluted by noise, forcing the removal of up to 65% of the acquired data. Hence, it is extremely important to guarantee an optimal signal quality while improving the wearability of the EEG devices to make them truly usable in daily life.

A second big challenge in EEG is the inter-session (or intra-subject) variability: the signals acquired from the same person can vary when measured at different times (or sessions), despite using the same setup [7]. As a result, the classification model trained on one session often yields significant accuracy degradation when applied to the next ones.

A simple solution is to train a new model for each session. However, this yields long calibration time, lowering user acceptance and engagement. Another way is to train big models on large datasets comprising multiple sessions. Even though there are various available open-source datasets, each uses different acquisition setups introducing inter-datasets variability, making the generalization problem even harder [8]. Hence, it is a big challenge to directly use the models pre-trained on open-source datasets on commercially available devices. Another downside of this solution is that the models tend to grow in complexity and are impossible to be embedded on the edge device requiring the data to be transmitted to a remote gateway or cloud, yielding high power consumption and short battery lifetime. The latest trend of smart edge computing and tinyML has proven to be a promising solution for increasing the battery life of BMI devices [9], [10]. However, it requires a careful trade-off between accuracy performance and resource usage besides the necessity of miniaturized yet high-performing microprocessors.

In summary, the requirements for a BMI device to be accepted by users and be usable in daily life scenarios are: a) Compact, comfortable, and non-stigmatizing BMI system; b) High classification accuracy with short calibration time; c) Energy-efficient design to prolong battery life and allow long-term usage. In this paper, we present a comprehensive solution to fulfill the above requirements for an accurate and robust BMI with low calibration time and high energy efficiency. The following points summarize our main contributions:

- We propose a comfortable and non-stigmatizing EEG

* Corresponding email: xiaywang@iis.ee.ethz.ch

headband featuring eight soft elastic elastomer-based dry active electrodes and BioWolf, a miniaturized acquisition device capable of onboard processing [10].
- We collect 7 sessions of EEG data from one subject while performing left- and right-hand movements and address the inter-session variability using transfer learning (TL) with a tiny convolutional neural network (CNN), i.e., MI-BMInet [9]. We achieve up to 96% inter-session accuracy, which is 30% higher than the baseline without TL.
- We also minimize the number of TL training data and reduce by 2.13× the calibration time for data acquisition at every new session down to ca. 5 min. Additionally, the proposed "chain" TL technique greatly reduces the resource requirements while performing TL.
- MI-BMInet can be deployed on the parallel ultra-low power (PULP) RISC-V microprocessor on BioWolf taking circa 6 ms and consuming up to 30 uJ to execute one inference. The system is estimated to provide more than 30 h of battery life when doing inference every 100 ms.

II. Materials and Methods

A. EEG Headband

EEG data are collected with an ultra low-power BMI platform (BioWolf [10]), which features a PULP system-on-chip (SoC) called Mr. Wolf, a 8-channel analog front-end (AFE) for biopotentials EEG channels, and a SoC with Bluetooth low energy (BLE) for wireless connectivity. BioWolf is integrated into a novel, comfortable, wearable and non-stigmatizing headband form-factor (Fig. 1). The in-house-tailored headband is made of comfortable clothing fabric, with a small pocket to hold the wearable BMI device. The headband offers 3 different levels of tightness, which can be customized to the user's head size. Cables for connecting EEG electrodes to BioWolf are hidden inside the headband (which appears as a normal sport headband that people can wear in daily life). The headband features 24 buttonholes for placing electrodes at different positions. For our experiments, we used 8 EEG channels near the positions TP8, TP7, FC6, FC2, FC1, FC5, PO7, and PO8 of the standard 10/20 system.

The electrode subsystem comprises two distinct mated components: the dry electrode and active buffering circuitry. The electrodes are fabricated by Dätwyler Schweiz AG and are a new design part of the newly developed SoftPulse Flex family [11]. Electrodes are based on a soft elastic elastomer-based material and feature multiple silver/silver-chloride-coated legs arranged in a circular disc, enabling scalp contact and hair penetration without skin preparations or gel. An elastomer-based conductive snap connector allows to interface with the acquisition device. The active circuitry, embedded in a circular printed circuit board (PCB) and placed on top of the electrode, comprises a buffering subsystem, which enhances the common-mode rejection ratio (CMRR) and reduces the artefacts caused by cable movement.

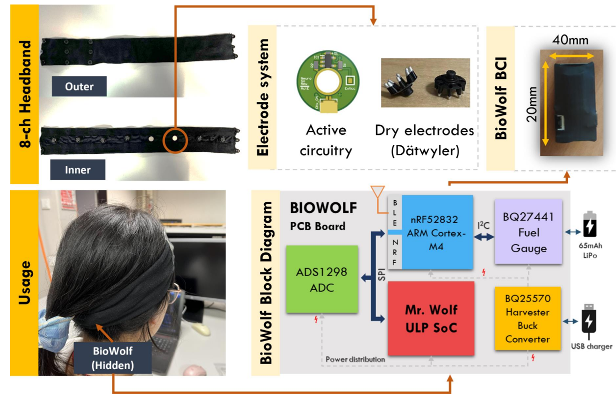

Fig. 1. Headband layout. Electrodes (Dätwyler Schweiz AG) are integrated with an active buffering circuit, plugged into dedicated buttonholes on the headband and connected to BioWolf via cables hidden in the headband textile.

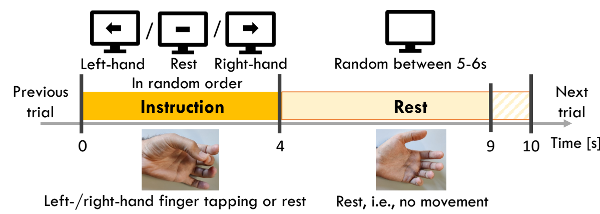

Fig. 2. Timing scheme of one data collection trial.

B. Experimental Protocol

A healthy volunteer is instructed to sit relaxed with the back against the chair and with the forearms lying on the lap. The subject follows visual instructions shown on a screen placed ca. 1 m away from the subject's eyes, and performs MM tasks of finger tapping [12], while EEG signals are recorded using the wearable device at a sampling rate of 500 Hz. Each instruction from the PC is synchronized with the EEG acquisition using BioWolf digital triggering system.

We collect seven sessions of EEG data (a total of approx. 4.375 hours) from one subject, each session containing 12–20 runs. One run consists of 15 trials, i.e., 5 with the left hand movement (left arrow), 5 with the right hand (right arrow), and 5 of resting state (dash line) in random order. As shown in Fig. 2, a trial includes a 4 s instruction and a resting period between 5 s and 6 s. In this work, we perform two-class classfication of left- and right-hand movements.

C. Classification Approach

1) Preprocessing: The 8-channel EEG data are preprocessed with a notch filter at 50 Hz and a 4^{th} order IIR band-pass filter at 0.5–100 Hz. We downsample the data in the 4 s MM window by a factor of 2 obtaining 950 samples in time.

2) Within-session Training and Validation: We use MI-BMInet [9] which is a lightweight, embedded CNN that achieves the state-of-the-art (SoA) classification accuracy while demanding orders of magnitude less resources compared to other SoA models. The reduced resource usage is an essential requirement for embedded deployment on low-power edge devices. Fig. 3 shows the network architecture and the hyperparameters obtained after a grid search on 180 trials and 5-fold cross-validation (CV) experiments.

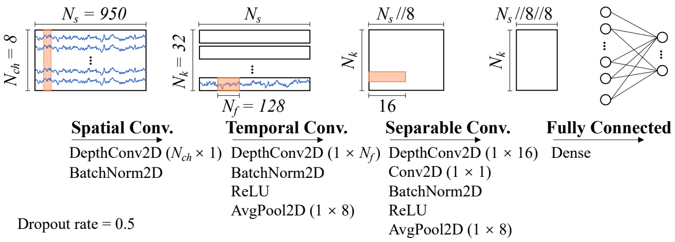

Fig. 3. MI-BMInet architecture [9].

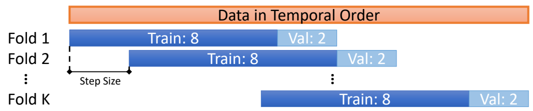

Fig. 4. K-fold rolling CV.

For the within-session baseline models, we use a rolling-window CV approach, illustrated in Fig. 4, for training and validation. For each fold, the window of the training and validation data advances in time. This is to take into account the intrinsic changes of EEG data in time. We also consider this method being the most suitable for future on-device learning as the model will need to be updated continuously with the latest data. We train the models for 500 epochs using cross-entropy loss, Adam optimizer (lr=0.001), and a batch size of 64. In each fold of rolling-window CV, eight runs, i.e., 80 trials were used as training set, and the subsequent two runs, i.e., 20 trials were used as validation set. The step size between neighboring folds is one run (Fig. 4). As in each session different number of runs were acquired, different number of folds in rolling-window CV were applied.

D. Inter-session Transfer Learning

To obtain the pre-trained model for TL, we use the rolling-window CV method as in the within-session case and choose the epoch N_{ep} where the average validation loss and accuracy curves converge. We then train a model from scratch on all the available pretraining data until N_{ep} and obtain a pre-trained model M_{pre} that is ready to be used for TL in the next session.

First, we use a one-to-one TL approach: We load the pre-trained model M_{pre} as the starting point to further train and validate on the first few runs of the new session. We select the final model M_{TL} with the best TL validation accuracy and test it on the remaining runs of this session in real time. Fig. 5 illustrates an example with eight training runs, two validation runs, and six testing runs. Additionally, we optimize the number of TL training runs by performing the same procedure and reducing the number of training runs to assess the minimum number of new data necessary for TL before the model is ready for real-time testing. This directly affects the calibration time.

Second, we explore a multi-to-one technique: Instead of using only one session of data for pre-training, we use multiple sessions to obtain the pre-trained model and perform the same TL technique as before with the minimum number of TL training runs on a new session. With more pre-training sessions, the model is expected to learn more variability between sessions and be able to generalize better.

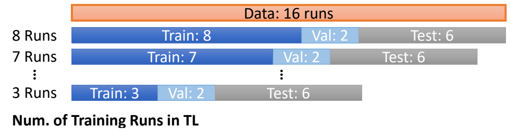

Fig. 5. Train-validation-test split of TL models with decreasing training runs.

A major problem of adding more sessions in the pre-training dataset is that every time a new session is recorded, a new pre-trained model has to be trained from scratch, requiring increasingly more memory for storing the dataset. Hence, we propose another experiment called "chain" TL scheme, where the new TL model is constantly updated based on the previous TL model. Fig. 6 shows an illustration: We obtain an initial M_{pre} using one-session data and continuously perform TL in the next sessions at every new usage of the headband by applying the same TL technique with the minimum number of training runs followed by two validation runs.

E. Drone Control

We demonstrate the proposed methodology in a real use case scenario by conducting real-time testing and inference for drone control. The drone is a Crazyflie 2.1 from BitCraze [13] that can perform tasks while keeping a stable (level) flight autonomously. During inference, the drone receives the CNN output commands (short left or right translations) using the BioWolf wireless link (BLE).

III. Results

Within-session classification. Table I reports the best within-session classification CV accuracy for each session. Apart from Session 12.09, all the sessions' CV accuracy exceeds 85% with Session 12.10 reaching 97.86%, which is comparable to SoA results for two-class classification [9]. Note that in our experiments we perform motor movement instead of imagery since our focus is on the system level. However, previous studies show that finger movements and imagery have high real-virtual congruency, especially with four or five fingers movements, as in our case [14]. If we directly compare to [6] in terms of average accuracy, ours is 6.12% higher.

One-to-one approach. We use Session 11.27 as pre-training data and Session 11.30 for TL and real-time testing. With 7-fold rolling-window CV on 11.27, we obtain N_{ep}=80 with 85% CV accuracy. We then obtain M_{pre} by training a

TABLE I
Results of within-session rolling CV. The average accuracy of all 7 sessions is 89.62%.

Session	# Runs	# Folds	Avg. Accuracy ± Std. [%]
11.27	16	7	90.00 ± 10.69
11.30	16	7	86.43 ± 10.25
12.05	12	3	85.00 ± 10.80
12.09	13	4	81.25 ± 2.17
12.10	16	7	97.86 ± 3.64
12.13	12	3	95.00 ± 4.08
12.14	20	11	91.82 ± 8.86

new model from scratch using all data from Session 11.27 for 80 epochs. Afterwards, TL is applied on Session 11.30 as explained in Sec. II-D and illustrated in Fig.5. The model with the best TL validation accuracy, M_{TL}, is finally selected and tested on the last 6 runs of the same session. To make a direct comparison without applying TL, we test M_{pre} on the same test runs and report the results in Table II. M_{TL} showed a great improvement of 21.66% in classification accuracy compared to M_{pre}, proving that the proposed TL technique can successfully mitigate the inter-session variability and achieve high accuracy on newly-acquired data.

Optimization of TL training runs. We gradually reduce the number of TL training runs from 8 to 3 to lower the calibration time in terms of acquiring new data (Fig. 5). The results are presented in Table II. The TL model using only three training runs can still achieve an excellent testing result of 96.67%. Decreasing the number of training runs does not necessarily reduce the accuracy performance on the testing set, however, we can observe that more training epochs are necessary to obtain the model with the best validation accuracy. Meanwhile, by reducing the number of training runs from 8 to 3, we can reduce the new data acquisition time from about 17 minutes to about 8 minutes, i.e., 2.13× reduction, including the two validation runs. If we fix the number of epochs and eliminate the need of validation runs, the acquisition time is reduced down to ca. 5 minutes (i.e., 10 trials×10 s) with 90% accuracy.

Multi-to-one approach. First, we take 11.27 and 11.30 sessions as pre-training data and perform the same procedure as for one-to-one approach with a step size of 2 runs (Fig. 4). We obtain the pre-trained model M_{pre2S} at N_{ep} = 60 and perform TL on Session 12.05 using the first 3 runs as TL training and the following 2 runs as validation. The model M_{TL2S} with the best validation accuracy is finally tested on the remaining 7 runs of Session 12.05 with results presented in Table III. For comparison, we do the same testing with M_{TL1S} which is obtained by applying TL on Session 12.05 using the above M_{pre}. We can see that TL offers an accuracy improvement of up to 31.43% demonstrating once again the advantage of TL techniques. When comparing M_{TL2S} and M_{TL1S}, an improvement of 5.72% is observed, inferring that more pre-training data might be beneficial for TL to new sessions. Analogously, we conduct three-to-one TL experiments on Session 12.13 using session 11.27, 11.30, and 12.05 (M_{TL3S}) and the same ablation studies by reducing the number of pre-training sessions to two sessions 11.27 and 11.20 (M'_{TL2S}) and to one session 11.27 (M'_{TL1S}). For each experiment we provide also the comparison of test accuracy without applying TL. The results are reported in Table III and same conclusions can be drawn.

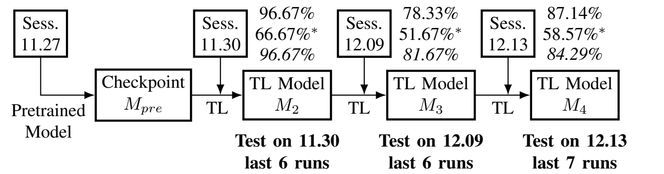

Fig. 6. Chain TL scheme with respective results and comparison without TL (marked with *) and with multi-to-one approach (italic).

Chain TL. We pre-train an initial M_{pre} model on Session 11.27. When we record the second session 11.30, we transfer M_{pre} to M_2 using 3 training and 2 validation runs and test on the remaining runs. The same is done on a next session 12.09 by transferring M_2 to M_3 and so on. We observe that chain TL performs similarly well as the multi-to-one experiments and is a more promising solution for smart edge computing in continual learning. Every time a user uses a BMI device in a new session, only a few recordings are necessary to calibrate the device and the TL model update can be potentially also executed onboard, achieving a real-time update.

Embedded implementation. An inference of MI-BMInet deployed on Mr. Wolf takes approximately 6 ms consuming up to 30 uJ measured onboard [9]. The power envelop of the whole system is estimated to be 8 mW [9], [10] while performing the inference every 100 ms, which translates to 30 h of operation with 65 mAh battery.

IV. Conclusion

We presented a novel solution for a comfortable, accurate, and energy-efficient BMI based on EEG signals and TL. We designed and implemented a wearable EEG headband with eight dry active electrodes and a miniaturized acquisition device that can perform onboard processing using a tiny CNN. We collected and analyzed EEG data from one subject performing left- and right-hand movements over seven sessions and demonstrated the effectiveness of TL in overcoming inter-session variability. We also deployed the inference of the tiny CNN on a PULP microprocessor to achieve low latency and long battery life. For future works, we will investigate the resting class to account for cases of no commands, broaden our study to more subjects, and implement the TL on the edge device for a continuous real-time adaptation of the BMI.

TABLE II
One-to-one inter-session classification accuracy w/ and w/o (marked by *) TL and with varying number of TL training runs.

# Runs	Best TL Epoch	Test Acc. [%]	Acq. Time [min]
8	121	88.33	16.7
-	-	66.67*	-
7	55	83.33	15
6	35	73.33	13.3
5	74	86.67	11.7
4	208	95.00	10
3	215	96.67	8.3
3	100 (fixed)	90	5

TABLE III
Multi-to-one inter-session test accuracy w/ and w/o TL and comparison with lower number of pre-training sessions.

Test on 12.05	Test Acc.	Test on 12.13	Test Acc.	with TL?
M_{TL2S}	81.43%	M_{TL3S}	92.86%	yes
M_{pre2S}	50.00%	M_{pre3S}	60.00%	no
M_{TL1S}	75.71%	M'_{TL2S}	84.29%	yes
M_{pre}	50.00%	M_{pre2S}	72.86%	no
-	-	M'_{TL1S}	81.43%	yes
-	-	M_{pre}	68.57%	no

ACKNOWLEDGMENT

We thank Bettina Lory for implementing the initial experimental setup, Thorir Mar Ingolfsson for fruitful discussions, and Dätwyler Schweiz AG for providing the electrodes. This project is supported by the Swiss National Science Foundation under the grant number 193813 (PEDESITE project) and the grant number 207913 (TinyTrainer project) and by the ETH Future Computing Laboratory (EFCL).

REFERENCES

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

[2] N. M. LLC, "Versus: a mobile EEG headset," https://getversus.com/, 2021, accessed: 2021-10-11.

[3] G. Spinelli, A. Odouard, M.-C. Nierat, S. Campion, M. Bensoussan, F. Grosselin, K. Pandremmenou, A. Breton, M. Raux, Y. Attal, T. Similowski, and X. Navarro-Suné, "Validation of melomind™ signal quality: a proof of concept resting-state and ERPs study," *bioRxiv:2020.02.28.969808*, 2020.

[4] Emotiv, "Epoc+," https://www.emotiv.com/epoc/, accessed: 2023-05-30.

[5] Muse, "EEG-powered meditation and sleep headband," https://choosemuse.com/, accessed: 2023-05-30.

[6] F. M. Garcia-Moreno, M. Bermudez-Edo, J. L. Garrido, and M. J. Rodríguez-Fórtiz, "Reducing response time in motor imagery using a headband and deep learning," *Sensors*, vol. 20, no. 23, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/6730

[7] S. Saha and M. Baumert, "Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review," *Frontiers in Computational Neuroscience*, vol. 13, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2019.00087

[8] T. Zaremba and A. Atyabi, "Cross-subject & cross-dataset subject transfer in motor imagery bci systems," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[9] X. Wang, M. Hersche, M. Magno, and L. Benini, "Mi-bminet: An efficient convolutional neural network for motor imagery brain–machine interfaces with eeg channel selection," *arXiv preprint arXiv:2203.14592*, 2022.

[10] V. Kartsch, G. Tagliavini, M. Guermandi, S. Benatti, D. Rossi, and L. Benini, "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.

[11] Datwyler Schweiz AG, "Softpulse," https://datwyler.com/company/innovation/wearable-sensors/softpulse.

[12] H. Cho, M. Ahn, M. Kwon, and S. Jun, "A step-by-step tutorial for a motor imagery–based BCI," in *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*, F. L. Chang S. Nam, Anton Nijholt, Ed. Taylor & Francis Group, 2018, ch. 23, pp. 445–460.

[13] BitCraze, "Crazyflie 2.1." [Online]. Available: https://www.bitcraze.io/about/bitcraze/

[14] M. Díaz, C. Llanos, S. Gonzalez, and M. Sabate, "How similar are motor imagery and movement?" *Behavioral neuroscience*, vol. 122, pp. 910–6, 09 2008.