

Caveats of deep learning for BCI diagnosis in DOC

Idorenyin Amaunam
BCI-NE, CSEE
University of Essex
0000-0002-0737-1619

Mushfika Sultana
BCI-NE, CSEE
University of Essex
0009-0001-0382-772X

Marina Lopes da Silva
ANU, Department of Clinical Neurosciences
Centre Hospitalier Universitaire Vaudois
marina.lopes-da-silva@chuv.ch

Jane Jöhr
ANU, Department of Clinical Neurosciences
Centre Hospitalier Universitaire Vaudois
jane.johr@chuv.ch

Karin Diserens
ANU, Department of Clinical Neurosciences
Centre Hospitalier Universitaire Vaudois
karin.diserens@chuv.ch

Serafeim Perdikis
BCI-NE, CSEE
University of Essex
0000-0003-2033-2486

Abstract—Deep learning models have met with remarkable success in various domains due to their ability to extract hierarchical features from high-dimensional data. However, these advantages may not directly apply to electroencephalography (EEG)-based Brain-Computer Interface (BCI) applications for diagnosing Disorders of Consciousness (DOC), where datasets are usually small, heterogeneous, and highly variable across patients and sessions. Here, we compare three EEG Deep Neural Network (DNN) models—EEGNet, DeepConvNet, EEGConformer—with a shallow Linear Discriminant Analysis (LDA) on a DOC dataset. Under within-session leave-one-trial-out (LOTO) evaluation, all DNNs show overoptimistic performance, including above chance results for the coma and all unresponsive wakefulness syndrome (UWS) subjects. Under leave-one-session-out (LOSO) evaluation, in contrast, DNN performance collapses to near random levels. Conversely, session-specific, shallow LDA models are able to avoid these caveats. To interpret these effects, we propose different hypotheses involving the subject-specificity and overall low separability of Sensorimotor Rhythms (SMR) features in DOC, as well as indirect data leakage and/or overfitting to neural correlates of behaviour not related to command following and motor imagery or attempt.

Index Terms—disorders of consciousness, deep learning, electroencephalography, brain-computer interface

I. INTRODUCTION

The appeal of Brain-Computer Interfaces (BCIs) lies in their ability to facilitate direct interaction with neural activity independent of muscular control. However, their rapid adoption—particularly in clinical contexts, has at times outpaced the establishment of methodological standards and validated protocols. Classifier performance is critical, as reliable accuracy is indispensable for practical utility.

In recent years, deep learning has transformed classification across multiple fields, including computer vision, medical imaging, and natural language processing. Within the BCI domain, deep learning architectures began to emerge roughly a decade ago, displacing traditional shallow models. Despite these promising advances, their impact in BCI applications has remained moderate, constrained largely by data limitations and the inherent variability of neuro-imaging signals.

In this paper, we extend this line of inquiry by applying state-of-the-art deep learning models with architectures

designed specifically for electroencephalography (EEG) signals, namely, EEGNet [1], DeepConvNet [2], and EEGConformer [3], to a novel, sizeable EEG Disorders of Consciousness (DOC) patient dataset. EEGNet has been shown to perform robustly in low-data EEG settings [1]. Additionally, both DeepConvNet and EEGConformer have been widely adopted in the literature, with several studies reporting strong performance even under limited data conditions [4]–[15]. The performance of Deep Neural Networks (DNNs) is systematically compared with a shallow Linear Discriminant Analysis (LDA) model, providing a direct assessment of their relative strengths and limitations in this challenging clinical setting. The performance of all models was evaluated using multiple cross-validation strategies: (i) leave-one-trial-out (LOTO), implemented session-wise with entire trials as inputs; (ii) leave-one-session-out (LOSO), also trial-based, where all trials from one session were held out; and (iii) a leave-one-sample-out scheme, conducted session-wise using 1 s windows as inputs.

Deep learning models were anticipated to struggle to achieve reliable performance on the present dataset under within-session cross-validation schemes, given the limited number of trials per class in a session (typically fewer than 100). On the contrary, we assumed that DNNs may be able to outperform LDA with LOSO given what seems to be adequate data cardinality when nearly all of the 62 sessions in our dataset are available for training. Surprisingly and very interestingly, both expectations were not ultimately upheld. On this basis, we take the opportunity to interpret these outcomes and extract useful insights on some particularities of using Deep Learning (DL) in EEG and BCI research.

II. MATERIALS AND METHODS

A. Patients and data

Twenty-eight patients (8 female; age 53.5 ± 16.6 , range 20–75) with varying levels of awareness across the DOC spectrum were recruited from the Acute Neurorehabilitation Unit (ANU) of the University Hospital of Lausanne (CHUV), Switzerland. The protocol complied with the Declaration of Helsinki and was approved by the ethical committee of the

Canton of Vaud (No. 142/09), with written informed consent obtained from participants' relatives.

A total of 179 runs across 62 recording sessions were acquired. Where multiple sessions per patient existed, these corresponded to recordings on different days, reflecting in general changes in awareness given the sample's acute clinical status. All patients were evaluated by experienced clinical neurologists and repeatedly assessed throughout their hospital stay using the Coma Recovery Scale-Revised (CRS-R) [16]. Based on these behavioral assessments, the neurology team assigned each recording session to one of seven categories corresponding to the associated CRS-R scores: Coma, unresponsive wakefulness syndrome (UWS), Minimally Conscious State (MCS)-, MCS+, emergence from Minimally Conscious State (eMCS), Locked-in Syndrome (LIS), and Healthy. Although LIS is not classified as a DOC, it is commonly included in the literature as a conscious control group, particularly in studies investigating Cognitive Motor Dissociation (CMD) and related DOC conditions. The variability across sessions motivated a session-wise analytical approach throughout, as each clinical assessment can reasonably be assumed to reflect only the EEG data acquired within the same recording session.

Patients were lying down on their beds in the hospital ward of the acute neurorehabilitation unit during each recording session. An active, 16-channel EEG electrode montage concentrated on the motor cortex according to the 10-20 EEG sensor positioning system was used for data acquisition [17]. EEG data were captured either with a g.Nautilus wireless amplifier at 500 Hz sampling rate or with a wired g.USBamp amplifier at 512 Hz (g.tec Medical Engineering, Austria). In both cases, a [0.1, 100] Hz hardware bandpass filter was enabled.

Each session comprised 1–4 runs (target: 3), with each run containing 15 motor and 15 rest trials presented in random order. Trials lasted 4 seconds and were initiated by an auditory cue in French. During motor trials, patients were instructed to open the affected hand. Rest trials prompted patients to remain still. Trials were separated by a randomised inter-trial interval of 4–6 seconds.

B. Data Analysis and Evaluation

Data were processed on a per-session basis to account for variability in awareness levels across recordings of the same participant. Recorded data were cleaned using the Fully Online and Automated Artifact Removal for BCI (FORCe) method [18]. Each run was treated with DC removal for all channels and a cross-neighbour Laplacian spatial filter was applied to enhance signal localization. When neighboring channels were unavailable, the remaining available neighbors were used for the Laplacian derivation. To evaluate Sensorimotor Rhythms (SMR), the Power Spectral Density (PSD) of the EEG signal was computed for each channel individually in 1 s overlapping, sliding windows with shift of 62.5 ms or 100 ms (for 512 Hz and 500 Hz sampling frequency, respectively), using the Welch periodogram method [19]. The internal Welch PSD parameterization involved 0.5 s windows with 50% over-

lap for a sampling frequency of 512 Hz (g.USBamp) and 40% overlap for 500 Hz (g.Nautilus). This resulted in all cases in a 2 Hz frequency band resolution of the PSD spectrum which was confined to the [4,48] Hz range of interest, covering the μ (8-12 Hz), and β (16-30 Hz) bands, which are neurophysiologically relevant with respect to motor tasks. Towards this end, PSD features undergo a feature selection step before being fed to the training of a machine learning (ML) decoder, replicating the open-loop processing of a Motor Imagery (MI) BCI. Specifically, for each session, the 6 best features, defined as those exceeding the 5th percentile after Fisher Score (FS) ranking were selected for classification. We employed the conventional definition of FS as $FS = \frac{|\mu_1 - \mu_2|}{\sqrt{s_1^2 + s_2^2}}$, where μ_1, μ_2 are the means and s_1, s_2 are the standard deviations of the two mental classes contrasted (MI vs Rest). Also, features corresponding to particularly vulnerable to artifacts channel Fz, and frequencies beyond the physiologically significant for motor tasks μ (8-14 Hz) and low β (18-24 Hz) bands were excluded.

Unlike shallow models, which rely on handcrafted features such as PSD extracted across frequency bands, deep learning models operate directly on raw EEG signals. Consequently, continuous EEG recordings were segmented into MI and rest epochs and provided directly as input to the classifiers. Prior to classification, each epoch was normalized on a per-channel basis across trials using Z-score standardization, which is consistent with standard deep learning preprocessing practices. For each input configuration, data splits were generated using precomputed leave-one-trial-out (LOTO) cross-validation schemes. A trial here refers to a 4-second segment corresponding to a specific task condition (move, rest). By contrast, a sample denotes a 1-second segment extracted from a trial. This distinction was required because both analysis were performed.

In the trial-level approach, the complete 4-second epoch was used directly as a single input instance for classification. In the sample-level approach, each trial was subdivided into overlapping 1-second sliding windows with shifts of 62.5 ms and 100 ms for sampling rates of 512 Hz and 500 Hz, respectively. This procedure generated 32 samples per trial at 512 Hz and 50 samples per trial at 500 Hz. Under this formulation, a trial was defined as the aggregate of all samples originating from the same parent epoch. For each cross-validation fold, a single trial, comprising all 32 or 50 samples, depending on the sampling rate, was reserved as the test set, while the remaining trials were used for training. Predictions were generated for all held-out samples and subsequently aggregated using a softmax operation to obtain a trial-level prediction. Session-level accuracy was then computed by averaging classification accuracies across all folds.

The EEG data were arranged in an array of dimensions (N, C, T) , where N denotes the number of trials, C the number of channels, and T the number of temporal samples. To ensure compatibility with two-dimensional convolutional neural networks, the data were reshaped to $(N, 1, C, T)$.

This corresponds to the standard Conv2D input structure (batch, in-channels, height, width), analogous to image-based inputs. In this representation, the batch dimension corresponds to the number of trials (N), the height to the number of channels (C), and the width to the temporal dimension (T). The additional singleton dimension (1) denotes the input channel, reflecting the fact that EEG constitutes a single-modality signal, analogous to a grayscale image. Each input sample was paired with its corresponding class label, and the sampling frequency was additionally provided as an input to the model.

Training procedures and optimization settings were standardized across all three architectures. Model optimization was performed using the Adam optimizer with a learning rate of 0.0002 and exponential decay coefficients $\beta_1 = 0.05$ and $\beta_2 = 0.999$. Training was conducted with a batch size of 64, whereas inference was performed using a batch size of 1. To reduce overfitting, a dropout rate of 0.5 was applied throughout training, and early stopping with checkpointing was employed to retain the best-performing model weights. To ensure reproducibility, random seeds were fixed across Python, NumPy, and PyTorch, thereby controlling sources of stochasticity such as weight initialisation, data shuffling, and dropout generation. In addition, CUDA Deep Neural Network (CuDNN) was operated in deterministic mode, while Compute Unified Device Architecture (CUDA) acceleration was enabled to support reproducible Graphics Processing Unit (GPU)-based computation. The architectures differed substantially in model complexity. EEGNetv4 was the most lightweight architecture, comprising 4,090 trainable parameters at 500 Hz and 4,202 parameters at 512 Hz. In comparison, EEGConformer contained 225,642 trainable parameters, whereas DeepConvNet was the largest model with 562,927 parameters. All experiments were executed on a single NVIDIA GeForce GTX 1050 GPU (driver version 577.57) using the CUDA backend for accelerated computation. Final model performance is reported as the mean classification accuracy across all cross-validation folds.

III. RESULTS

Figure 1A-C presents the results obtained using windowed EEG segments (samples) as input. This approach yielded markedly overoptimistic performance estimates compared with the shallow-model benchmark, illustrated in Figure 2A. Using EEGNet, 55 sessions (89%) exceeded the binomial significance threshold, compared with 51 sessions (82%) for DeepConvNet and all 62 sessions (100%) for EEGConformer. By contrast, only 17 sessions (27%) exceeded the corresponding permutation-based threshold when analysed using the shallow classifier. Nevertheless, these differences cannot be attributed to improved classification ability of DNNs with respect to LDA, as DNN performance tends to detect latent awareness in virtually all cases, including a coma control patient and most other cases where the standard clinical assessment confidently predicts that patients cannot respond to commands (see figure legend).

A similar pattern is evident for the within-session LOTO cross-validation results shown in Figure 1D-F. All three deep learning architectures produced classification accuracies substantially above the 50% chance level across most sessions, including many sessions that would not be expected to demonstrate reliable command-following based on their clinical diagnosis. When evaluated against the more stringent binomial significance threshold, 29 sessions (47%) exceeded threshold with EEGNet, whereas all 62 sessions (100%) exceeded threshold with both DeepConvNet and EEGConformer. In comparison, the shallow classifier exceeded its permutation-based significance threshold in only 5 sessions (8.1%).

More elaborately, several observations suggest that within-session results in both cases (1 s window samples, 4 s trials) are overly optimistic. For example, the coma control session ($S25, 1$) for trial-wise input exceeded both chance and binomial thresholds when classified using DeepConvNet and EEGConformer, despite there is no doubt that this patient (being in coma for several years at the time of the recording) has no command following ability. Similarly, several DOC patients achieved higher classification accuracy than the healthy control sessions ($S9, 1$ and $S9, 2$), an effect that is not reasonable as MI of responsive patients is known to be inconsistent and marginally separable, thus yielding in general low accuracy even if the neural correlates of different commands can be distinguished from EEG; this effect has not been observed with the shallow classifier. EEGNet, the shallower of the DNN architectures compared, exhibited somewhat more conservative behaviour, failing to classify the healthy control above the binomial threshold in some cases while also producing lower accuracies for several patient sessions. Nevertheless, the overall pattern remained inconsistent with prior clinical expectations. Similar deviations from anticipated, sound outcomes were observed under the within-session and sample-wise scheme, where several DOC sessions achieved higher accuracies than the healthy control and the coma control failed to remain consistently below significance thresholds across architectures.

Figure 1G-I shows the results obtained using the LOSO cross-validation scheme. This strategy was motivated by the hypothesis that training on all remaining sessions may provide sufficient data for the models to generalise to a previously unseen session. However, none of the three deep learning architectures achieved performance exceeding the corresponding chance-level significance thresholds, falsifying this assumption.

IV. DISCUSSION

The aim of this paper has been to compare state-of-the-art deep learning pipelines with a classical shallow model, using the sizeable novel dataset employed in this study and derived by a standard command-following MI paradigm [17], [20]. In other words, we have attempted to study whether the classification accuracy metric extracted with DNN models yields an added value when it comes to inferring CMD in DOC. The attempt to classify a session's data in isolation

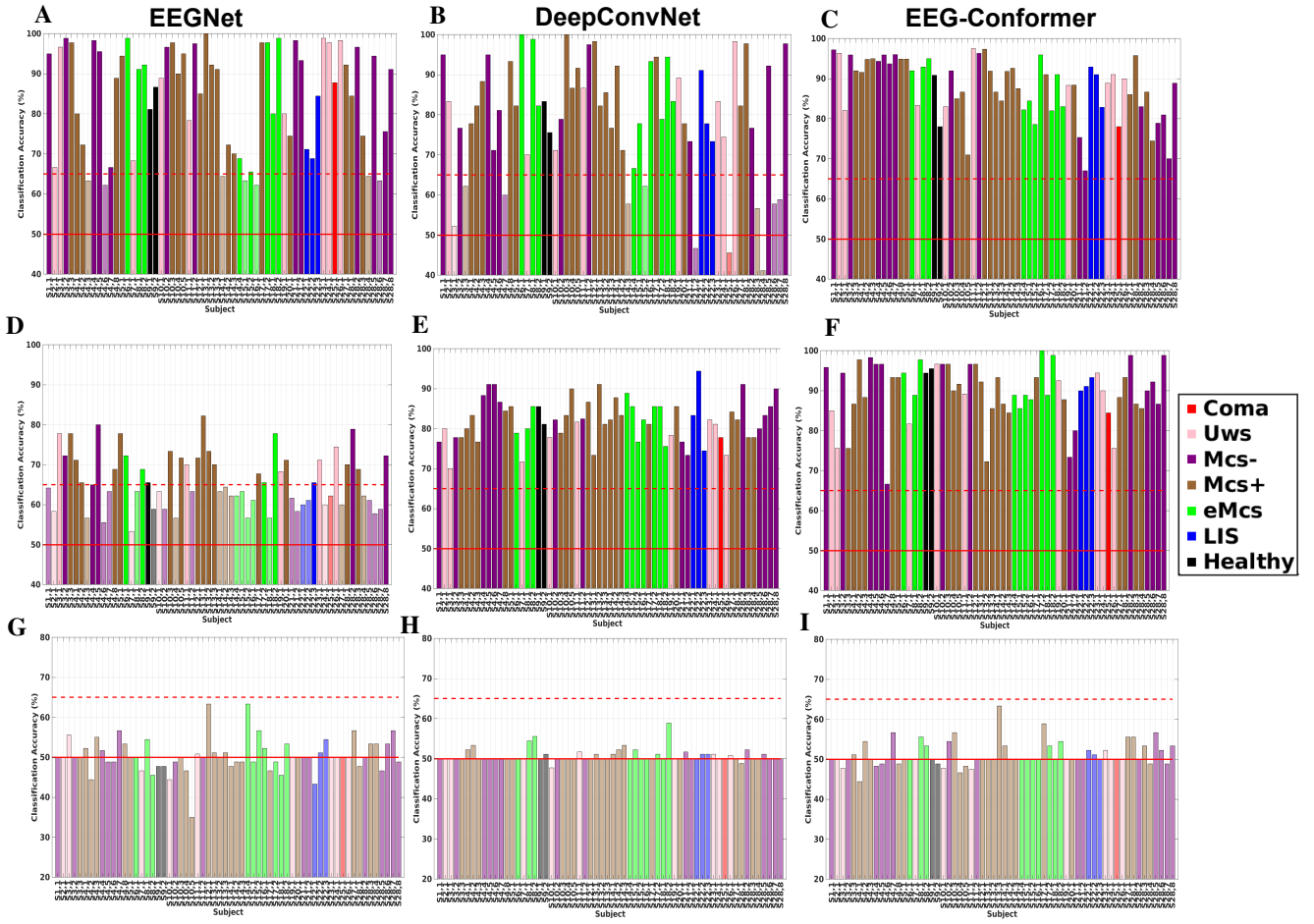


Fig. 1: MI versus Rest classification accuracy across all recording sessions for the three deep learning architectures. Top row: within-session leave-one-trial-out (LOTO) results with sample input for (A) EEGNet, (B) DeepConvNet, and (C) EEGConformer. Middle row: within-session LOTO results with trial input for (D) EEGNet, (E) DeepConvNet, and (F) EEGConformer. Bottom row: leave-one-session-out (LOSO) results with trial input for (G) EEGNet, (H) DeepConvNet, and (I) EEGConformer. The solid red line indicates the theoretical chance level of 50% for binary classification. Where present, error bars represent variability across cross-validation folds.

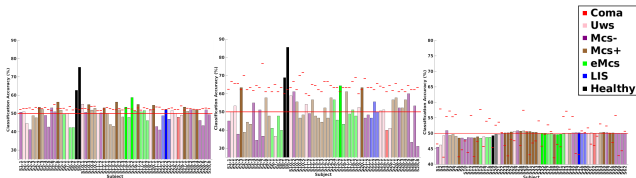


Fig. 2: Trial-wise MI versus Rest classification accuracy across all recording sessions obtained using the LDA classifier. Panels (A)–(C) present results for the session-wise LOTO with sample input (A), trial input (B) and with LOSO (C) validation, respectively. The solid red line indicates the theoretical chance level (50%) for binary classification. The horizontal dashed line denotes the 99% binomial significance threshold for a two-class BCI task with 45 trials, whereas the session-specific lines indicate permutation-derived significance thresholds at the 99% confidence level.

(session-wise), either with full-trial or with sample/window-based input (which multiplies the cardinality by a factor of approximately 50) leads to classification accuracy results that are consistent and exhibit the same characteristics. First, all deep learning models resulted in much higher average accuracy compared to LDA; this is true also for the vast majority of the individual session results. Second, this effect is clearly correlated with the size and expressiveness of the model, with the EEGConformer (which is equipped with both transformer-like, far-sighted, attention-based features and the ability to extract spatial/spectral/temporal dependencies through convolution) outperforming the DeepConvNet (which is only capable of the latter; note that although DeepConvNet has double the weights of EEGConformer, the depth of both these models is at the same order of magnitude), while they both significantly outperform the considerably shallower EEGNet architecture. As a result, the accuracy of most of the individual sessions exceeds the statistical significance threshold for most

of the sessions and, again, increasingly so for the deeper and more expressive models. It must be noted that chance levels were extracted here with a binomial distribution assumption for classification outcomes despite this assumption's known limitations, due to the great computational burden of using random shuffling of labels with deep models.

While at a first, superficial glance the above observations may lead to the conclusion that deep learning can indeed have great impact in diagnosing command-following behaviour in DOC, the other emerging commonalities of these results tell an entirely different story. Specifically, the accuracy extracted with the deep models with respect to the control subject/sessions seems to be erratic and completely unreasonable. On the one hand, in all cases but two (EEGNet trial-wise and DeepConvNet sample-wise) the deep models found the coma control participant, *S25*, 1, to show signs of awareness, as the corresponding accuracy is well above chance. Even more strangely, the coma patient accuracy results are in fact shown to exceed the performance of the healthy control and of patients clinically diagnosed to be somewhere in the MCS spectrum and, thus, aware to some extent. On the other hand, the results become even more counterintuitive when it comes to the healthy control sessions, *S9*, 1 and *S9*, 2. Not only does the EEGNet miss the (beyond any doubt) full-awareness state of the healthy control in the trial-wise classification scenario, but, overall, the sensible outcomes of the shallow model, where the healthy control sessions vastly exceed the accuracy of all other subjects, is not replicated at all by the deep architectures: in all these cases, the performance of the healthy subject lies somewhere close to the average accuracy of the studied population.

First, it is worth highlighting that the results of this paper underline the need for large control groups on both sides of the awareness spectrum (coma, healthy), since, absent these controls, the findings here could have been easily misinterpreted as being favourable to the use of deep learning in DOC. Importantly, these paradoxical control-session findings cannot be blamed on the limitations induced by the binomial assumption to extract chance levels, as the erratic behaviour regards the absolute accuracy values and not only their relation to the chance level definition used.

Clearly, on the grounds of a complete failure of DNN models to correctly classify the control patients, and of their overall overoptimistic tendency to diagnose almost all sessions as CMD these results cannot be taken at face value. It is thus crucial to interpret these outcomes and identify potential causes explaining the unanticipated behaviour. Towards this goal, we propose two hypotheses, which may have acted individually or in tandem to produce these findings.

First, it is observed that the aforementioned correlation between the depth and expressiveness of the models and the average accuracy obtained indicates overfitting caused by data leakage. However, this cannot be a typical case of data leakage, as all results were acquired with LOTO cross-validation (and the corresponding code has been subjected to meticulous testing for bugs), keeping dependent samples in

the same fold and thus preventing data leakage overestimating accuracy, exactly as has been done for the LDA model. Nevertheless, it is possible that data leakage could still have occurred because adjacent and, in general, nearby trials of EEG are only a few seconds apart and thus still somewhat auto-correlated; hence, not entirely independent. While LOTO cross-validation is shown to be sufficient for eliminating any effect of such small dependency for shallow models with low expressiveness (and, thus, no strong tendency to overfit), the profound learning capacity and expressiveness of deep models probably renders them able to discover very subtle similarities of even slightly correlated samples that appear in both the training and testing folds, so that data leakage is not avoided. This hypothesis is consistent with the observed effect where larger models showcased higher accuracy.

Second, we conjecture that the difference of feeding PSD samples to the shallow model and time-domain signals to the deep models may also, partly or fully, explain the ostensible overfitting effect. Furthermore, all 16 recorded channels were fed to the deep models, while the artifact-prone frontal channel *Fz* was excluded from the shallow model analysis, which could have also played a role. The PSD input to the shallow model effectively guarantees that LDA only attempts to classify SMR rhythms between Rest and MI, in accordance with the relevant neurophysiological background in the context of a command-following protocol involving MI tasks. On the contrary, the deep models, being able to process the raw data and implementing their own embedded feature extraction, are not constrained to exploiting any particular feature domain or brain phenomenon. This opens the door to the exploitation of features that may be differentially modulated within periods labelled as either Rest or motor attempt, but in fact have no relation to command following and, consequently, to awareness. For example, remnants of the auditory cue in the trial data — which differ between the two tasks ("Bougez", "Ne bougez pas") — despite the first second of each epoch being discarded for this very reason, may have been classified with high accuracy, giving the false impression that MI is successfully distinguished from rest, even for the coma patient. Additionally, artifacts not fully removed by the FORCe method that are correlated with class labels may generate spurious separability unrelated to the command-following task, particularly given that the artifact-prone *Fz* channel was retained in the deep model pipeline. Class-correlated artifacts are a known issue in SMR BCI: able-bodied participants often adopt specific body postures during MI tasks, inducing muscular and movement artifacts. Similar effects may have occurred in this protocol, particularly as many participants reacted to Functional Electrical Stimulation (FES) stimulation with movements lasting several seconds. While state-of-the-art artifact removal was performed, residual artifacts may have been exploited by the deep models to inflate accuracy estimates in ways that did not affect the LDA classifier.

Another interesting question arises from the fact that LOSO evaluation failed to recognize awareness in all cases, including

the healthy individual that is clearly detected by LDA with PSD features. A possible explanation is that this amount of data, despite being much greater than in session-wise classification, still does not suffice to power DNN models, even at the scale of EEGNet. This issue can only be elucidated when larger DOC datasets become available. However, it is noted that LOSO validation reflects learning attempts across different subjects (it must be highlighted that, in order to avoid leakage, for subjects with multiple sessions, when a session appears as testing/target session, the remaining same-subject sessions are not included in the training set). Thus, a key open question for such a transfer learning approach, is the extent to which SMR patterns are transferable across individuals and, particularly, across DOC patients—an issue that remains unresolved, as current transfer learning and domain adaptation approaches in MI BCI yield only limited improvements [21], [22]. In other words, bigger datasets may not be guaranteed to resolve this issue.

V. CONCLUSION

Deep learning models have demonstrated remarkable success in domains characterised by large, well-annotated datasets. However, in BCI and EEG research, data availability is often severely limited and marked by substantial inter and intra subject variability. The results presented here confirm that application of DL in BCI may not be as seamless as in other fields. While model accuracy scaled with architectural complexity, examination of control sessions revealed a failure to generalise reliably, which may be indicative of implicit data leakage. Furthermore, increasing training data through LOSO cross-validation did not improve robustness; instead, performance became random, reflecting either the need for even more data, or, potentially, the absence of transferable discriminative features within the highly variable DOC EEG data. In this setting, shallow models still demonstrate superior reliability and interpretability than DL architectures.

REFERENCES

- [1] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces,” *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
- [2] R. T. Schirrmester, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggersperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, “Deep learning with convolutional neural networks for eeg decoding and visualization,” *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [3] Y. Song, Q. Zheng, B. Liu, and X. Gao, “Eeg conformer: Convolutional transformer for eeg decoding and visualization,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.
- [4] A. O’Shea, R. Ahmed, G. Lightbody, E. Pavlidis, R. Lloyd, F. Pisani, W. Marnane, S. Mathieson, G. Boylan, and A. Temko, “Deep learning for eeg seizure detection in preterm infants,” *International journal of neural systems*, vol. 31, no. 08, p. 2150008, 2021.
- [5] W. Chen, Y. Luo, and J. Wang, “Three-branch temporal-spatial convolutional transformer for motor imagery eeg classification,” *IEEE Access*, vol. 12, pp. 79 754–79 764, 2024.
- [6] S. De, A. Singh, and A. K. Bhandari, “A novel vision transformer based multimodal fusion approach for clinical mdd diagnosis using eeg and audio signals,” *IEEE Transactions on Computational Biology and Bioinformatics*, 2025.
- [7] M. Khayretdinova, A. Shovkun, V. Degtyarev, A. Kiryasov, P. Pshonkovskaya, and I. Zakharov, “Predicting age from resting-state scalp eeg signals with deep convolutional neural networks on td-brain dataset,” *Frontiers in aging neuroscience*, vol. 14, p. 1019869, 2022.
- [8] F. Del Pup, A. Zanolà, L. F. Tshimanga, A. Bertoldo, L. Finos, and M. Atzori, “The role of data partitioning on the performance of eeg-based deep learning models in supervised cross-subject analysis: a preliminary study,” *Computers in Biology and Medicine*, vol. 196, p. 110608, 2025.
- [9] Y. He, X. Wang, Z. Yang, L. Xue, Y. Chen, J. Ji, F. Wan, S. C. Mukhopadhyay, L. Men, M. C. F. Tong *et al.*, “Classification of attention deficit/hyperactivity disorder based on eeg signals using a eeg-transformer model,” *Journal of Neural Engineering*, vol. 20, no. 5, p. 056013, 2023.
- [10] B. Yang, F. Rong, Y. Xie, D. Li, J. Zhang, F. Li, G. Shi, and X. Gao, “A multi-day and high-quality eeg dataset for motor imagery brain-computer interface,” *Scientific Data*, vol. 12, no. 1, p. 488, 2025.
- [11] H. Nam, J.-M. Kim, W. Choi, S. Bak, and T.-E. Kam, “The effects of layer-wise relevance propagation-based feature selection for eeg classification: a comparative study on multiple datasets,” *Frontiers in Human Neuroscience*, vol. 17, p. 1205881, 2023.
- [12] M.-j. Kim, Y. C. Youn, and J. Paik, “Deep learning-based eeg analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset,” *NeuroImage*, vol. 272, p. 120054, 2023.
- [13] M. C. Tjepkema-Cloostermans, C. da Silva Lourenço, B. J. Ruijter, S. C. Tromp, G. Drost, F. H. Kornips, A. Beishuizen, F. H. Bosch, J. Hofmeijer, and M. J. van Putten, “Outcome prediction in postanoxic coma with deep learning,” *Critical care medicine*, vol. 47, no. 10, pp. 1424–1432, 2019.
- [14] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, M. Liu, F. Chen, and W. Che, “A deep transfer convolutional neural network framework for eeg signal classification,” *IEEE Access*, vol. 7, pp. 112 767–112 776, 2019.
- [15] A. Al-Saegh, S. A. Dawood, and J. M. Abdul-Jabbar, “Deep learning for motor imagery eeg-based classification: A review,” *Biomedical Signal Processing and Control*, vol. 63, p. 102172, 2021.
- [16] J. T. Giacino, K. Kalmar, and J. Whyte, “The JFK coma recovery scale-revised: measurement characteristics and diagnostic utility,” *Archives of physical medicine and rehabilitation*, vol. 85, no. 12, pp. 2020–2029, 2004.
- [17] C. Schneider, S. Perdakis, M. Silva, J. Jöhr, A. Pincherle, J. d. R. Millán, and K. Diserens, “Motor attempt eeg paradigm as a diagnostic tool for disorders of consciousness,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4681–4684.
- [18] I. Daly, R. Scherer, M. Billinger, and G. Müller-Putz, “Force: Fully online and automated artifact removal for brain-computer interfacing,” *IEEE transactions on neural systems and rehabilitation engineering*, vol. 23, no. 5, pp. 725–736, 2014.
- [19] P. Welch, “The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, p. 70–73, Jun. 1967. [Online]. Available: <http://dx.doi.org/10.1109/TAU.1967.1161901>
- [20] I. Amaunam, C. Schneider, M. L. da Silva, J. Jöhr, K. Diserens, and S. Perdakis, “A discussion of statistical criteria for assessing awareness with smr bci after brain injury,” in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 3709–3714.
- [21] C. Sartzetaki, P. Antoniadis, N. Antonopoulos, I. Gkinis, A. Krasoulis, S. Perdakis, and V. Pitsikalis, “Beyond within-subject performance: A multi-dataset study of fine-tuning in the eeg domain,” in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 4429–4435.
- [22] G. Sarafraz, A. Behnamnia, M. Hosseinzadeh, A. Balapour, A. Meghraz, and H. R. Rabiee, “Domain adaptation and generalization of functional medical data: A systematic survey of brain data,” *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–39, 2024.