10.3217/978-3-99161-014-4-081

USING A CNN-LSTM ARCHITECTURE WITH DATA AUGMENTATION TO IMPROVE HD-ECOG SPOKEN SYLLABLE CLASSIFICATION

Mehdi Javani Mirehkoohi[1], Zachary Freudenburg [1], Amira Neumann [1], Nick F Ramsey [1]

[1]Brain Center, Department of Neurology and Neurosurgery, University Medical Center Utrecht, Utrecht, 3584 CX, the Netherlands

E-mail: m.javanimirehkoohi@umcutrecht.nl

ABSTRACT: Brain-Computer Interfaces (BCIs) have emerged as vital tools in understanding and assisting individuals with LIS due to neurological diseases such as ALS. This study focuses on the and feasibility of recognizing spoken syllables from implanted HD-ECoG signals as a platform for Speech BCIs. We propose a hybrid deep learning model, which uses a modified EEGNet as a feature extractor coupled with an LSTM. A primary challenge in this domain is the limited quantity of ECoG data. To address this challenge, we employ window clipping as a data augmentation technique, effectively increasing the amount of training data available for the model. Using a dataset comprising recordings from six subjects implanted with HD-ECoG, we evaluate our proposed method. Results indicate a notable improvement in classification accuracy achieved through the designed hybrid DL model. Furthermore, our findings elucidate the distinctive impact of data augmentation methods in further enhancing the performance of our designed model.

Keywords: HD-ECoG, ECoGNet, CNN, LSTM, Data Augmentation

INTRODUCTION

The realm of Brain-Computer Interface (BCI) systems has revolutionized human-computer interaction, enabling direct communication pathways between the human brain and external devices. Language BCI represent a frontier in assistive technology, designed to empower individuals with communication disabilities by translating recorded brain activity into language. Electrocorticography (ECoG) due to its capabilities in recording a wide range of frequency and also high density recording of a specific areas of the brain which are responsible for specific cognitive task has been widely used in this regard. In recent years, researchers have made many efforts to leverage various methods to decode language, particularly deep learning methods as the most promising method to this aim ([**?**], [**?**], [**?**]). While various deep learning architectures have demonstrated success in decoding of spoken phonemes, words, and sentences with acceptable performances, the task of syllable decoding poses greater challenges. Unlike words, which vary in length and possess distinct sounds, syllables typically exhibit uniform length and share acoustic features. Consequently, decoding syllables presents a formidable hurdle, as neural networks cannot rely solely on length or distinctiveness for classification. Also, syllables can involve overlapping combinations of phonemes making them useful building blocks for language but also less distinct. Despite the complexity, decoding syllables is pivotal, serving as a foundational step towards deciphering spoken words. Addressing this challenge necessitates the development of robust and adaptable neural networks capable of enhancing decoding performance, particularly for individuals with limited data. By exploring the potential of such networks, we aim to push the boundaries of language decoding in BCIs, fostering greater inclusively and effectiveness in communication assistance technologies.

EEGNet efficiently extracts temporal features reflecting short and long-term changes in brain activity. While EEG and ECoG measure brain electrical potential differences using electrodes, they differ in invasiveness and spatial coverage [**?**]. EEG, non-invasive, captures broader spatial coverage with lower density, while ECoG, invasive, offers higher density with narrower spatial coverage. Despite these differences, both methods share preprocessing and feature extraction techniques, often utilizing frequency analysis. Thus, deep neural networks proficient in extracting frequency information from EEG data could enhance ECoG analysis.

Peterson et al. [**?**] introduced a modified version of EEGNet tailored for ECoG, incorporating a mapping layer from individual ECoG electrode positions to a 1D input space. While their approach yielded improved results over traditional EEGNet in binary classification tasks, we sought to explore an alternative mapping paradigm. Thus, we directly adapted EEGNet to investigate this alternative approach.

Our mapping concept involves translating the inherent 2D structure of ECoG data into a standardized grid space, aligning native electrode coordinates. Leveraging this spatial input in 2D, we have tailored a modified variant of EEGNet specifically optimized for the unique characteristics of ECoG data. This adapted network, denoted as ECoGNet, maintains a parallel block structure to EEGNet while accommodating the intricacies of ECoG signal processing.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-081

An intriguing progression entails merging CNN and LSTM networks to build hybrid architectures. This synthesis facilitates the concurrent extraction of spatial and temporal features, correspondingly. This novel approach bears substantial potential for augmenting the capabilities of Deep Learning [?, ?, ?].

Utilizing Deep Learning for ECoG signal classification faces challenges due to limited dataset sizes, particularly in Motor Imagery analysis. K-Fold Cross Validation (K-Fold CV) addresses this issue by partitioning data into 'K' subsets, enabling robust model training and evaluation. Hewaidi et al. [?] leveraged K-Fold CV to enhance their methodology, integrating variational autoencoders, deep autoencoders (DAE), and CNNs for EEG motor imagery classification. Recent literature [?] introduces two key K-Fold CV methods: inter-subject and intra-subject, providing insights into model performance across subjects and within individual subjects, respectively.

Data augmentation methods offer a potent solution to the challenge of limited dataset sizes in Deep Learning. By expanding the training data, these techniques bolster classification stability and accuracy, enabling models to generalize better to new datasets [?]. Moreover, data augmentation addresses class imbalance issues, crucial for classification tasks. Techniques such as geometric transformations and noise introduction effectively diversify datasets, enhancing model robustness. The utilization of sliding windows is a prevalent data augmentation technique across various domains. In neonatal seizure detection, O'Shea et al. [?] employed overlapping windows, with 8-second trials and 50% overlap, to augment seizure instances within EEG signals. Kwak et al. [?] explored different shift lengths, ranging from 10 ms to 60 ms within 2-second windows, revealing superior performance with shorter shifts.

In this paper, we present a comprehensive approach to EEG signal classification, leveraging deep learning models and innovative data augmentation techniques. We begin by introducing the analyzed data in the Data and Materials section, followed by an explanation of the used deep learning architectures. Subsequently, we describe our designed model and detail the methods employed to address the inherent challenges posed by limited dataset sizes. Moving forward, the Results section showcases the outcomes of implementing our model, with particular emphasis on the impact of utilizing data augmentation methods. Finally, we conclude by summarizing the project's findings and highlighting avenues for future research and development in EEG signal classification.

DATA AND MATERIALS

Data and Preprocessing: The dataset was collected at UMC Utrecht and comprises recordings from six subjects. Each subject underwent different trials, and the electrode configurations varied among subjects. Some subjects contributed 180 trials, while others had 90, and due to data collection errors, certain trials were eliminated from the valid dataset. Moreover, the number of electrodes differed among subjects, with some recorded using 128 electrodes and others with 64 electrodes. It's worth noting that not all electrodes provided valid signals, as some were too noisy to convey useful information. The properties of the dataset are summarized in table 1.

The locations of the electrode grids for all participants are illustrated in Figure 1. Due to various restrictions and limitations, such as individual anatomical variations and positioning constraints during data collection, the electrode placements vary in their standard Montreal Neurological Institute (MNI) coordinate system locations across subjects. Here the electrode locations are determined based on spherical components (Phi and Theta), with the center of the component aligning with the center of the brain.

Table 1: Summary of the data of all participants

Participant	n. actual/valid trials	Sampling Freq	n. actual/valid Electrodes
S01	180/177	2000	128/128
S02	180/173	512	64/52
S03	90/85	2000	128/125
S04	90/89	2000	128/120
S05	90/86	2000	128/121
S06	90/87	2000	128/109

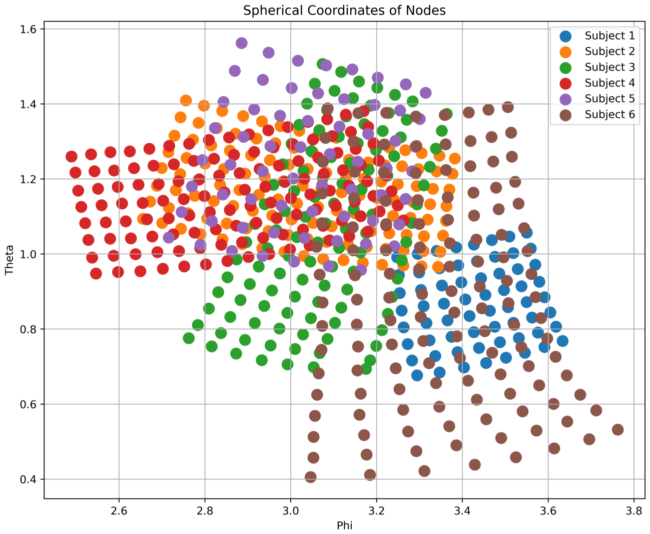

Figure 1: Electrode Grid Placements for Participants

The task entails conducting trials where participants utter one of nine distinct syllables: "mi", "mu", "ma", "ki", "ku", "ka", "zi", "zu", and "za". These syllables exhibit similarities in their articulation and serve as the nine classes we seek to classify. The trial protocol includes randomization of these syllables interspersed with occasional rest trials. Participants are prompted on the screen to perform either 10 or 20 repetitions of each syllable.

10.3217/978-3-99161-014-4-081

The subjects performed the task multiple times leading to a range in trails from 180 trials, to 90 over subjects, and due to data collection errors, certain trials were eliminated from the valid dataset.

Two crucial time points are defined: Cue time marks the initiation of monitoring for the intended syllable, and Voice Onset Time (VoT) indicates when it becomes discernible that the participant starts articulating the syllable. Due to individual differences, participants initiate pronunciation after different durations following the Cue time. Additionally, the duration from VoT varies depending on the syllable and the participant's capabilities.

To standardize trial durations across participants and syllables, a fixed duration of 1 second is set starting from VoT, recognized as the most informative segment of the trial.

CNN: Convolutional Neural Networks (CNNs) have garnered considerable acclaim for their adeptness in extracting robust spatial features from images through deep learning. The architectural underpinnings of CNNs ensure spatial robustness, which revolves around three pivotal elements: local receptive fields, convolutional layers, and pooling layers. By employing small receptive fields, convolutional filters adeptly capture fundamental visual features from distinct regions of the input image. These extracted features undergo progressive amalgamation and enhancement across subsequent layers to discern higher-level features. However, the insertion of pooling layers following convolutional layers, while essential for preventing overfitting and reducing spatial dimensions, can potentially entail a loss of precise spatial information—a concern warranting attention.

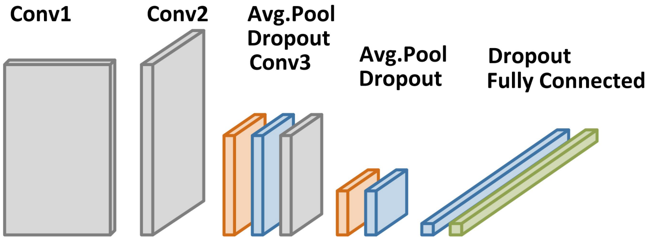

Figure 2: EEGNET Architecture

LSTM:

The LSTM architecture revolves around three primary states: the cell state ($\mathtt{C}_{t-1}$,$\mathtt{C}_{t}$), the input state ($\mathtt{X}_{t}$ and $\mathtt{h}_{t-1}$), and the output state ($\mathtt{h}_{t}$). Additionally, LSTM incorporates four crucial gates: the forget gate (f_t), the input gate (i_t), the new memory gate ($\mathtt{C}'_{t}$), and the output gate (O_t). These gates play pivotal roles in regulating internal operations within the LSTM.

The cell state serves as a memory reservoir that facilitates information flow across LSTM units. Each LSTM unit features skip connections in the form of gates, which intricately control the inflow and outflow of information to and from the cell state. Specifically, the forget gate discerns which information to retain or discard from the prior cell state, while the input gate governs the integration of new information.

The new cell state is crafted by merging the previous cell state with inputs from the input gate and the new memory gate. Finally, the output gate oversees the information contributing to the LSTM unit's output. Leveraging these architectural components, including gates and memory units, empowers the network to capture and retain pertinent information essential for effective learning.

The LSTM's prowess in managing long-term dependencies underscores its versatility and efficacy across a spectrum of deep learning applications.

MODEL ARCHITECTURES (HYBRID CNN/LSTM APPROACH)

To effectively capture the intricate spatial and temporal characteristics inherent in ECoG signals, we propose a sophisticated hybrid neural network architecture that seamlessly integrates Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks. This hybrid approach combines the robustness of CNNs in spatial feature extraction with the proficiency of LSTMs in modeling sequential data, thereby augmenting the analysis and classification of ECoG signals.

In this work we have used EEGNET model as the CNN component of the designed hybrid model. we adopt the EEGNET model as the foundation for our CNN component, tailored with necessary modifications to suit the dimensions of our ECoG data (16/8,8,2000/512). This adaptation ensures optimal utilization of the CNN's capabilities in discerning spatial intricacies within the ECoG signals. We call this network ECoGNet.

EEGNet is a Deep Learning model structured with multiple convolutional blocks, outlined in Figure 2 . The initial block consists of a standard convolutional layer followed by a batch normalization (BN) layer. Subsequently, a depth-wise convolutional layer is utilized in the following block, succeeded by a BN layer, an Exponential Linear Unit (ELU) activation function, and an average pooling layer. Additionally, a dropout layer is introduced at the end of this block. The third block incorporates a separable convolution, a BN layer, an ELU activation, and another average pooling layer. Notably, dropout layers are applied both before and after flattening the data. For the classification stage, a fully connected layer is employed, followed by a softmax function to classify the data into nine distinct classes.

In tandem with the CNN component, we incorporate an LSTM network to capture the nuanced temporal dependencies inherent in ECoG signals. LSTMs, renowned for their prowess in modeling sequential data, prove instrumental in unraveling the temporal dynamics and long-term dependencies embedded within the ECoG signals. By seamlessly integrating an LSTM network, our model gains the ability to discern intricate temporal patterns, thereby enriching the analysis of ECoG signals.

Deep Learning Obstacle:

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-081

Enhancing the proficiency of designed models often relies on providing them with enough data to be learned. However, in certain applications such as the analysis of brain signals, acquiring a sufficient amount of data can be challenging due to constraints imposed by the nature of the data collection process. This scarcity of data poses a significant obstacle for biomedical specialists seeking to train effective models. In this project we have used a couple of methods to deal with this problem, Cross Validation, and Data Augmentation.

To mitigate the data constraint, K-Fold Cross Validation (K-Fold CV) is commonly employed. K-Fold CV divides the data into 'K' subsets, allowing the model to train on different combinations and reducing overfitting while providing robust evaluation of generalization ability. For instance, Hwaidi et al. [?] utilized K-Fold CV to enhance the performance of their approach, integrating variational autoencoders, deep autoencoders, and CNNs for EEG signal classification. In this work, we have also used this method to not only try to mitigate the data restrictions, but also prevent overfitting. Due to the amount of the data we have, the 5-fold CV is chosen.

The second commonly used method to deal with the obstacle, is the Data Augmentation method. Data augmentation is a highly effective method for addressing the challenge of limited dataset size in deep learning. By increasing the quantity and variety of training data, it enhances classification stability and accuracy, enabling models to be more robust and less biased when handling new datasets [?]. Additionally, data augmentation helps mitigate class imbalance in classification tasks. It employs geometric transformations such as translations, rotations, cropping, flipping, and scaling, along with noise introduction, to expand the dataset and generate new instances. Depending on the type of data involved, various augmentation techniques can be applied. In the realm of biomedical signal analysis, window clipping stands out as a widely utilized method. In our project, we have employed window clipping to augment the available data. Nevertheless, this approach encounters challenges, particularly in determining the optimal quantity of clipped windows and their overlapping ranges. During our experimentation, we conducted tests using different numbers of windows ranging from 1 to 4, with each window having a fixed duration of 1 second. Additionally, we explored various overlapping ranges, spanning from 5% to 50%. These parameters were inherently constrained by the duration of the useful signal.

In our proposed hybrid architecture, the LSTM component follows the CNN component. This architectural arrangement facilitates the seamless flow of information from spatial to temporal domains, as the output of the CNN is meticulously fed into the LSTM network. This cohesive integration empowers the model to discern sequential patterns and dependencies within the ECoG signals, thereby enabling a holistic understanding of both spatial and temporal aspects of the data.

By harnessing the collective strengths of CNNs and LSTMs, our hybrid architecture endeavors to exploit spatial and temporal information in tandem, thereby enhancing the discriminative power and interpretability of our proposed model. This comprehensive approach facilitates a nuanced analysis and classification of ECoG signals, paving the way for advancements in neuroscientific research and clinical applications.

In this work, we will analyze the results of our designed model from 2 aspect. First, we want to find out how adding LSTM as a classifier to the ECoGNet model may enhance the accuracy percentage, and then we will test the effect of using data augmentation method to the performance of the designed model, and comparing the designed model's performance when we use different numbers of windows in data augmentation.

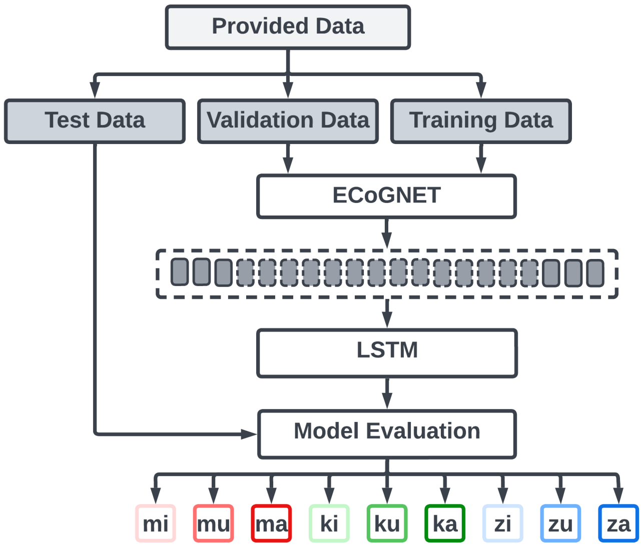

Figure 3: Hybrid Architecture

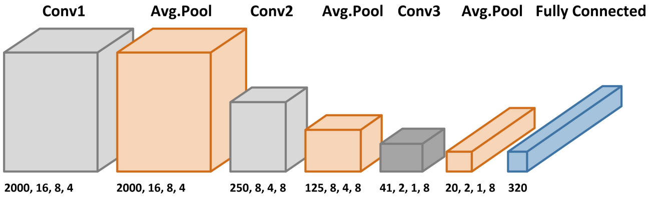

Figure 4: Modified EEGNET Architecture (Imported data shape: 2000 x 16 x 8)

To find out the performance of the designed model, we should provide a baseline model which is strong enough to classify the ECoG data. As a baseline model, we utilized Spatial Match Filters (SMFs), a non-deep learning method commonly employed in BCI research, particularly with ECoG data. SMFs have demonstrated promising results, achieving a classification accuracy of 76% in four phoneme classification tasks [?]. This technique involves a trial-by-trial comparison of activity patterns against mean activity patterns of different conditions. Initially, the signal undergoes wavelet transformation into the time-frequency domain, followed by computation of

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-081

mean values across defined frequency bands and time-points for each electrode, focusing on the high frequency band between 65 and 95 Hz. The resulting mean values represent activity patterns for each electrode. Subsequently, correlation analysis is performed between each trial's activity pattern and the mean activity patterns for nine classes. This correlation computation, conducted in a leave-one-out fashion to ensure unbiased estimates, assigns each trial to the class with the highest correlation pattern. Notably, only electrodes with significant functional responses are included in the correlation computation, further enhancing classification accuracy.

The second baseline model which is used to comparison aims is EEGNET that has been introduced.

RESULTS

The table 2 shows the results of implementing the designed model to classify the 9 syllabus of each participant. the results shows a significant improvement in accuracy. when a LSTM layer is added. This is highlighted by the fact that for all subjects the accuracy is at least 9 percentage points above the theoretical chance level of 11% for the Hybrid DL Model while both ECoGNet and SMF both show 3/6 subjects below or around chance level.

Table 2: Accuracy (%)

Participant	SMF	EEGNet	Hybrid DL Model
S01	39	60	67
S02	24.9	20	20
S03	13.3	12	29.41
S04	0	14.3	33.33
S05	12.5	14.4	21.1
S06	20.2	27.8	27.8
Mean	18.32	24.75	33.11

As it is discussed in previous sections, we have used 5-fold CV and also window clipping method to overcome the limitation of data quantity. During our experimentation, we conducted tests using different numbers of windows and explored various overlapping ranges. These parameters were inherently constrained by the duration of the useful signal.

The results, presented in Table 3, clearly demonstrate the efficacy of data augmentation in improving the model's proficiency in classification tasks.

By leveraging data augmentation techniques, we have successfully enhanced the model's ability to classify biomedical signals. This augmentation strategy not only mitigates the limitations imposed by data scarcity but also contributes to the overall robustness and generalization capability of the model.

Table 3: Accuracy with Different Numbers of Windows

	Windows			
Participant	1w	2w	3w	4w
S01	67	81	62	67.8
S02	20	20	20	20
S03	29.41	32.35	33.3	20.59
S04	33.33	40	43	36.11
S05	21.1	30.6	44.44	47.2
S06	27.8	36.1	27.78	30.56
Mean	33.107	40.01	38.42	37.04

Analyzing the table reveals an intriguing trend: an increase in the number of windows from 1 to 2 correlates with higher accuracy. However, this pattern falters as additional windows are added, resulting in a decline in overall performance by mean. This observation underscores a crucial point: not all segments of each trial contribute equally to classification accuracy. Indeed, the informativeness of added windows varies, with non-informative windows potentially detracting from overall results. Notably, this effect can differ across subjects. For instance, Subject S05 demonstrates an increase in accuracy but experiences a slowdown in processing speed with the introduction of 3 and 4 windows. Conversely, Subject S01 witnesses a decline in accuracy after the incorporation of the third and fourth windows.

DISCUSSION

Such disparities highlight the nuanced interplay between participant concentration levels, physical capabilities, and data quality. Indeed, individual differences among participants can significantly influence the duration of informative data within each trial. Moreover, the expansion of the number of windows necessitates more extensive data processing, demanding higher computational resources and potentially leading to longer computation times. Consequently, a delicate balance must be struck between model accuracy and computational efficiency. In certain scenarios, such as those where resource constraints are paramount, opting for two windows may represent the more optimal choice.

Thus, a thorough consideration of the trade-offs between performance and computational resources is imperative in maximizing the effectiveness of the classification process while ensuring optimal resource allocation.

CONCLUSION

During this work, we proposed a novel hybrid deep learning model that combines a modified EEGNet for feature extraction with a LSTM network for temporal analysis. Our approach addresses the challenge of limited ECoG data through the innovative use of window clipping as a data augmentation technique.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-081

Our experiments, conducted on a dataset comprising recordings from six subjects, demonstrate promising results. We observed a significant enhancement in classification accuracy compared to previous models, affirming the effectiveness of our hybrid model in recognizing the syllabless.

Furthermore, our analysis of data augmentation techniques highlights the importance of optimizing the number of clipped windows to balance classification accuracy and computational efficiency. While increasing the number of windows initially improves accuracy, there is a diminishing return beyond a certain point, emphasizing the need for careful consideration of resource constraints and performance trade-offs.

REFERENCES

[1] Goli P, Mazrooei Rad E. Advantages of deep learning for ecog-based speech recognition. The Hearing Journal. 2019;72(8):10.

[2] Luo S, Rabbani Q, Crone NE. Brain-computer interface: Applications to speech decoding and synthesis to augment communication. Neurotherapeutics. 2022;19(1):263–273.

[3] Pandarinath C *et al.* High performance communication by people with paralysis using an intracortical brain-computer interface. eLife. 2017;6(February):e18554.

[4] Hashiguchi K *et al.* Correlation between scalp-recorded electroencephalographic and electrocorticographic activities during ictal period. Seizure. 2007;16(3):238–247.

[5] Peterson SM, Steine-Hanson Z, Davis N, Rao RPN, Brunton BW. Generalized neural decoders for transfer learning across participants and recording modalities. Journal of Neural Engineering. 2021;18(2):026014.

[6] Zhang R, Zong W, Dou L, Zhao X, Tang Y, Li Z. Hybrid deep neural network using transfer learning for eeg motor imagery decoding. Biomedical Signal Processing and Control. 2021;63:102144.

[7] Khademi Z, Ebrahimi F, Montazery Kordy H. A transfer learning-based cnn and lstm hybrid deep learning model to classify motor imagery eeg signals. Journal of Neural Engineering. 2023;20(2):025004.

[8] Li H, Ding M, Zhang R, Xiu C. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. Biomedical signal processing and control. 2022;72:103342.

[9] Hwaidi JF, Chen TM. Classification of motor imagery eeg signals based on deep autoencoder and convolutional neural network approach. IEEE access. 2022;10:48071–48081.

[10] Wang X, Hersche M, Tömekce B, Kaya B, Magno M, Benini L. 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, 1–6.

[11] Wang F, Zhong Sh, Peng J, Jiang J, Liu Y. Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part II 24. 2018, 82–93.

[12] O'Shea A, Lightbody G, Boylan G, Temko A. Neonatal seizure detection using convolutional neural networks. In: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). 2017, 1–6.

[13] Kwak NS, Müller KR, Lee SW. A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. PloS one. 2017;12(2):e0172578.

[14] Branco MP *et al.* Alice: A tool for automatic localization of intra-cranial electrodes for clinical and high-density grids. Journal of Neuroscience Methods. 2018;301:43–51.

Published by
Verlag der Technischen Universität Graz