

Energy-Efficient EEG-Based Brain–Computer Interface Using EEGNet and Edge–Cloud Computing

1st Manni Kumar
*Department of Computer Science
 and Engineering,
 Chandigarh University
 Gharuan, Mohali, Punjab, India
 mannikumar55@gmail.com*

2nd Ishita Singh
*Department of Computer Science
 and Engineering,
 Chandigarh University
 Gharuan, Mohali, Punjab, India
 ishitaasingh.2184@gmail.com*

3rd Taniya
*Department of Computer Science
 and Engineering,
 Chandigarh University
 Gharuan, Mohali, Punjab, India
 taniyathakur034@gmail.com*

Abstract—Brain-Computer Interfaces (BCIs) enable a connection between the brain and devices, thus offering great potential for rehabilitation and assistive technology purposes. Nevertheless, the three issues are still making it incredibly challenging to shift EEG-based BCIs from labs to practice: deep-learning decoding pipelines are complicated and resource-consuming, techniques that have been developed in an ideal lab setting become ineffective with regard to noise represented in the real world, and there are no architectures that could provide low-energy operation and continuous processes in wearable devices. This work proposes a system called the energy-efficient BCI, which adopts a layered edge-cloud approach. The novel component of this BCI is that a lightweight GAN-based EEGNet model can be used for preprocessing of data on the device, while the cloud deals with more complicated steps. EEGNet is chosen due to the low complexity of its parameters achieved through the use of depthwise convolution. By splitting the workload this way, the framework aims to cut down on-device processing load, reduce the amount of data transmitted wirelessly, and lower overall power consumption. Within the cloud environment, ensemble and adaptive classification techniques are designed to address inter-session performance variations while minimizing the need for frequent recalibration. This paper focuses on the architectural design and the reasoning behind partitioning EEGNet across edge and cloud and the planned evaluation approach. Empirical testing of classification accuracy, latency, and energy

Keywords—Brain–Computer Interface, EEGNet, Electroencephalography, Edge Computing, Cloud Computing, Deep Learning, Energy Efficiency, Wearable Healthcare

I. Introduction

I-A. Problem Statement and Motivation

Brain–Computer Interfaces work by translating electrophysiological signals from neuronal activity into commands that external devices can act on, giving people with severe motor impairment a way to communicate that would otherwise be closed off to them. Signals can be captured through a few main modalities EEG, magnetoencephalography (MEG), and implanted electrode arrays each with its own tradeoffs in how invasive it is, how much spatial detail it captures, and how practical it is to actually deploy. Among these, non-invasive

EEG-based systems are still the most viable option both clinically and commercially. Even so, taking these systems out of the lab and into everyday, sustained use runs into a problem that hasn't received much attention: the energy cost of running conventional deep-learning decoding pipelines.

The reason these models are so attractive for deployment on resource constrained systems is their ability to massively reduce the number of parameters required to accurately decode EEG. Compact convolutional models like EEGNet [21] have progressed in this direction to a great degree relative to earlier, bulkier convolutional architectures. However, by running the whole model at the edge with wearables, there is an upper bound on the accuracy and adaptability. However, placing the entire system in the cloud results in the constant streaming of raw or minimally processed EEG data, thus increasing the communication energy cost and end-to-end latency, and making it worse with time. So the classification system alone isn't enough, it must retain the accuracy in real-world use, while remaining within the limits of the wearer's power consumption, bandwidth, and thermal dissipation. The problem tackled in this paper is the following one: "How to partition the decoding of EEGNet between edge and cloud resources without violating any of these constraints simultaneously?"

I-B. Related Work

The foundations for this field date back to Wolpaw et al. [1] who demonstrated that neural signals were reliably decoded into commands to control an external device. This early work was later synthesized by Nicolas Alonso and Gomez-Gil [2] with a large taxonomy of BCI modalities and applications. Blankertz et al. [3] were among the first to apply machine-learning approach to single-trial EEG discrimination and, at similar time, Pfurtscheller and Lopes da Silva [4] deciphered the event-related synchronization and desynchronization patterns that underlie motor-imagery BCI paradigms. These thoughts were followed by more formal theoretical and design principles for BCI performance by Wolpaw and Wolpaw [5] which emphasized the equal importance of signal

processing and user training.

At that point the research began to head into the direction of real assistive applications. He and co-authors [6] provided evidence that non-invasive BCIs founded on sensorimotor rhythms may be used for controlling assistive devices. McFarland and Wolpaw [7] followed this with their demonstration of BCI-driven control of computer interface and prosthetic devices for people with serious movement limitations. Zander and Kothe [8] shifted the focus by introducing passive BCIs which may read cognitive and affective states even without intention of the user to communicate. In the area of infrastructure, Schalk and Mellinger [9] created the widely utilized BCI2000 platform, whereas Lebedev and Nicolelis [10] made significant contribution to the fundamentals of brain-machine interfaces research.

The field of signal processing has seen many advancements in recent years. Makeig et al. [11] were able to employ independent component analysis for the classification of signals in order to extract noise-free signals. Similarly, Fazel Rezaei et al. [12] completed the thorough investigation of the P300 paradigm. Sanei and Chambers [13] presented a range of methods for the extraction of features from the EEG signals and classification of the signals, while Ang and Guan [14] provided an evidence that BCIs can be used with EEG. Another study on the increase in data transfer rates in the use of BCI was conducted by Cheng et al. [15], while Vidaurre and Blankertz [16] raise awareness that the issue of BCI illiteracy still exists. Also, Birbaumer [17] studied the possibility to use BCIs for helping people with locked in-syndrome to communicate, while Luck [18] introduced an overview of basic algorithms for event-related potential analysis. In addition, Scherer et al. [19] developed a motor-imagery BCI control model based on loose timing constraints. Finally, the study of the application of BCIs in the gaming industry was conducted by Ahn et al. [20].

Of primary importance in the context is the introduction of EEGNet by Lawhern et al. [21], a compact convolutional network that utilizes depthwise and separable convolutions to combine EEG feature extraction approaches in a single, efficient design and demonstrates its applicability across several BCI paradigms with few training samples. Wang et al. [22] built upon this work, optimizing EEGNet for implementation on low-power micro controllers and thus demonstrating the feasibility of a fully on-device motor imagery BCI system with an energy cost per inference in the millijoule range.

Overall, this collection of publications is a testament to the advancement of signal acquisition, decoding techniques and compact model design. However, there still remains a missing piece: few systematic efforts have been made to distribute the decoding of computation across multiple tiers of hardware in order to collectively optimize for energy requirements, latency and long-term adaptability rather than just to squeeze the model size down or aim for ever greater accuracy separately. This is the motivation behind the edge-cloud partitioning approach proposed in this paper.

I-C. Research Question and Contributions

With this gap in mind, the paper aims to address a relatively straightforward question: Just how much energy, transmission overhead, and recalibration does a decomposition of the EEGNet-like decoding pipeline for an EEG-based BCI save when implemented across a layered edge-cloud architecture, while maintaining the high accuracy and reliability expected of clinical deployment?

The paper contributes to the work towards an answer in three major ways:

- 1) It presents a four-layer edge-cloud BCI architecture, an architecture of BCI that uses a lightweight model, derived from EEGNet, for ondevice preprocessing and early feature extraction, and a heavier model for classification, ensemble aggregation, and adaptive model updates, that operate in the cloud.
- 2) It presents the method used to design this; how it handles pre-processing of signals on the edge; and how features are extracted on the edge device; and how the adaptive classification technology in the cloud is designed to maintain accuracy with real-world noise while reducing the frequency of recalibrations.
- 3) It reviews the trade-offs that this type of partitioning entails in terms of energy usage, communication overhead, latency, scalability and clinical usability, and outlines a methodology for empirical testing of the expectations in future work.

II. Proposed Edge-Cloud EEGNet Framework

II-A. System Architecture Overview

The four working layers of the system are neural signal acquisition, edge processing, intelligent decoding processing in the cloud, and application integration as illustrated in Fig. 1. A layered design involves running time-critical and computation-heavy jobs on the local edge devices, and running more complex jobs, such as model training, ensemble processing and long-term storage in the cloud.

At the acquisition layer, wearable EEG electrodes sit on the scalp, designed to be lightweight and positioned in a way that makes extended wear practical. The signals picked up here get sent over a short-range wireless link to a nearby edge processing unit.

That edge layer is where noise filtering, artifact removal, and the early stages of EEGNet-based feature extraction happen all before anything leaves the local device. From there, the cloud layer takes over for the remaining classification steps, ensemble aggregation, adaptive model updates, and long-term data storage, since these are the operations that are hardest to run on constrained edge hardware and benefit most from cloud-scale compute.

Finally, the application integration layer connects the decoded commands to whatever downstream system is being used prosthetic control, communication tools for people with paralysis, or neurorehabilitation devices.

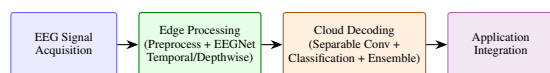


Fig. 1. Overview of the proposed four-layer edge-cloud BCI architecture. The edge device executes preprocessing and the temporal/depthwise EEGNet stages; the cloud executes the separable convolution, classification, ensemble, and adaptive-learning stages.

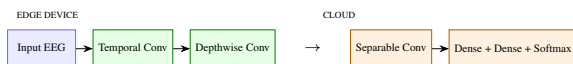


Fig. 2. Wearable EEG electrode placement and the on-device signal path from acquisition through band-pass filtering, baseline correction, and artifact removal, prior to edge-side EEGNet feature extraction.

II-B. Signal Acquisition and Edge Preprocessing

EEG electrodes on the scalp pick up the tiny voltage fluctuations that reflect underlying neuronal activity. The problem is that raw EEG recordings rarely come out clean they tend to pick up interference from eye movements, muscle activity, and general environmental noise. To deal with this, the edge device runs a preprocessing step before any feature extraction happens. This involves band-pass filtering to keep only the frequency ranges that actually matter physiologically, baseline correction to strip out drift, and artifact-removal routines aimed specifically at muscle- and eye-movement related contamination

II-C. Lightweight EEGNet Architecture for Edge-Cloud Partitioning

EEGNet [21] is the backbone of the decoding part of this framework, largely due to the compactness and efficiency of its structure. It does this via a temporal convolution, which learns frequency-selective filters, followed by a depthwise convolution that learns a spatial filter for each temporal feature map (which is equivalent to what a standard convolution would do at a spatial level of EEG channels) and finally a separable convolution that aggregates the temporal information for each feature map using a much smaller amount of parameters than a standard convolution would require. This is rounded out with batch normalization, a nonlinear activation, average pooling, and dropout for regularization.

What this framework does differently is split that pipeline across edge and cloud. The temporal and depthwise convolution stages the parts that work directly on the filtered multichannel signal and are the most efficient in terms of parameters run on the edge device, producing a compact intermediate feature representation. It's this representation, not the raw or filtered EEG signal, that actually gets sent to the cloud. The separable convolution stage and the final classification layer, meanwhile, run in the cloud, where power constraints aren't an issue and there's room to work with more computationally demanding operations.

This split-computing approach, shown in Fig. 2, is meant to cut down on two things at once: the amount of data being transmitted, since the intermediate feature embeddings are much smaller than the raw multichannel signal they're derived

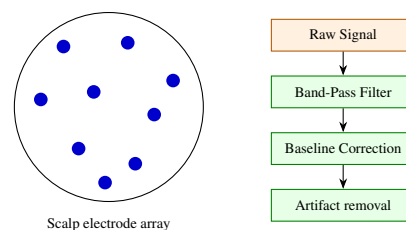


Fig. 3. EEGNet layer sequence with the proposed edge/cloud partition.

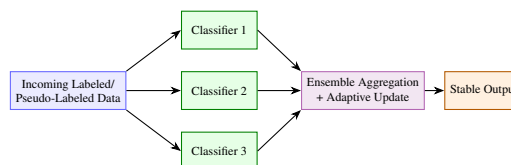


Fig. 4. For incremental model updating, cloud-side adaptive and ensemble classification pipeline, where incoming labeled and pseudo-labeled data are used. For improving the stability of predictions and decreasing the need for frequent recalibration, multiple classifiers are combined.

from, and the computational burden left on the edge device, compared to running EEGNet entirely on the edge or entirely in the cloud. Wang et al. [22] had already shown that scaled down EEGNet variants can run fully on low-power micro controllers, but this framework takes a different approach—it keeps part of the pipeline on the edge while leaning on the cloud for classification and model adaptation. The trade-off is a slightly higher communication cost in exchange for greater model capacity and the ability to support ensemble methods and adaptive learning strategies that just aren't practical when everything has to fit within a wearable device's tight memory limits.

II-D. Cloud-Based Adaptive and Ensemble Classification

To address the common issue of the well known “lab to field” problem where performance of the decoding device degrades after it is deployed to the field, adaptive learning procedures are added to the cloud layer that incrementally tune the parameters of the classifier with data collected on field as it becomes available, labeled or pseudo labeled. The idea is to maintain decoding of the neural signal over time despite changes in a user's neural signal properties from one session to the next. In addition to this, ensemble techniques are employed at the same level, where the prediction results of multiple classifiers are combined to reduce variance in predictions and hence make the system stronger than it would be using the prediction result of any single classifier. Unlike the earlier feature-extraction stage, they're more expensive to compute and less time-sensitive, which is why they can be done in the cloud: to lessen the need for manual, on-site recalibration that has been cited as a huge hurdle to making this a long-term clinical reality.

II-E. Communication and Energy-Efficient Design

Sustainability is not an add-on, it is combined hardware and software objective, from the outset. The hardware design

TABLE I
COMPARISON OF EEGNET DEPLOYMENT STRATEGIES

Dimension	Fully Cloud	Fully Edge [22]	Proposed Edge-Cloud
Comm. overhead	High (raw/filtered signal)	None (post-decision only)	Reduced (feature embedding)
Edge compute/memory	Minimal	High	Moderate
Adaptability to drift	High (cloud retraining)	Constrained by memory	High (cloud retraining)
Recalibration freq.	Low (design intent)	Higher (limited capacity)	Low (design intent)

includes low-power EEG sensors to maintain low-usage consumption and power consumption over extended periods of use, notably over 24 hours. In terms of software, the edge device is only used for preprocessing and the initial feature extraction for the EEGNet, limiting both the amount and frequency of wireless transmission is important as transmission is often one of the most prominent sources of power consumption in a wearable system. For the overall picture, the edge-cloud split is designed to find a balance: control local energy usage, while offloading the more demanding computation side to the cloud layer where power constraints are not as critical.

III. Design Analysis and Expected Outcomes

III-A. Comparative Analysis Relative to Conventional Architectures

Compared to legacy BCI architectures where everything runs on either the cloud or the edge device, the layered edge-cloud framework developed here is expected to yield several qualitative improvements. These are described in more detail in Table I. However, some will require empirical verification, which will be done in the future. Such improvements include: lower power usage; more accurate and stable classification over time; reduced latency for time-critical tasks, since filtering, artifact removal, and some feature extraction can be done locally; better scalability, since training of classifiers and ensemble management, as well as long-term storage of data, are done by the cloud; improved ease of use, since the wearable device is dramatically less computationally intensive than a fully embedded, high-capacity decoding pipeline.

Note: Table I reflects the architectural design rationale and remains qualitative pending the empirical evaluation described in Section IV.

III-B. Clinical Implications

The clinical importance of the suggested framework is determined by its ability to improve BCI accessibility for individuals suffering from severe impairments due to physical diseases, such as locked-in syndrome, since it guarantees better performance of the devices in uncontrolled environments than that which is possible in lab conditions. It should be noted that the architecture of the developed framework is modular, and

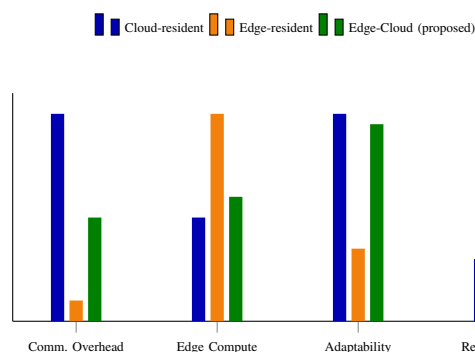


Fig. 5. Expected qualitative comparison of edge compute load, communication overhead, and recalibration frequency across the fully cloud-resident, fully edge-resident, and proposed edge-cloud deployment strategies, or responding to the dimensions summarized in Table I.

includes four layers, which allows performing integration with other existing systems in the area of healthcare information technology.

Nevertheless, the practical usage of the architecture discussed still largely depends on several issues, which were not covered in the given architectural analysis, such as wireless edge-cloud link reliability in real-life situations, and compliance with necessary regulations, as well as data safety during the transmission process. The issues mentioned should be confirmed experimentally prior to the implementation of the proposed system.

III-C. Limitations and Open Challenges

Several limitations temper the claims made in this framework and mark out priorities for future empirical work. First, EEGNet and similar decoding models including the partitioned version proposed here have mostly been tested on datasets gathered under controlled conditions, which leaves open the question of how well they'd generalize once deployed in less predictable, real-world settings. Second, this paper hasn't characterized the reliability and latency of the wireless link connecting the edge and cloud tiers, particularly under intermittent connectivity, and this gap could end up eating into the latency benefits the architecture is expected to provide. Third, wearable EEG hardware still isn't as miniaturized or comfortable as continuous, long-term use would ideally require. Fourth, large-scale clinical trials testing sustained, real-world deployment of layered edge-cloud BCI systems like this one remain rare, which limits how strongly any claims about real-world reliability can be made at this stage. Fifth, transmitting neural data to cloud infrastructure raises ongoing questions around data security, patient privacy, and regulatory compliance all of which still stand in the way of clinical integration.

Taken together, these limitations mean the analysis presented here should be read as a design-stage contribution that still needs empirical validation, not as evidence of a proven clinical outcome.

Beyond these five points, it's worth noting that the frame-

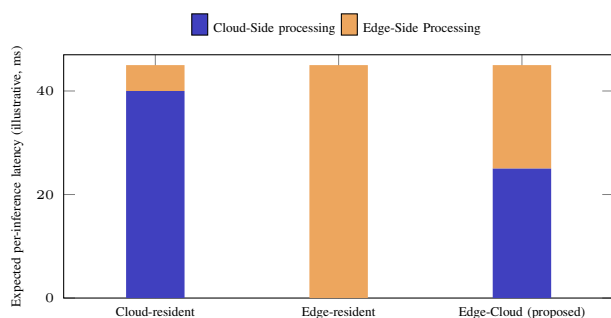


Fig. 6. Illustrative, planned per-inference latency breakdown to be measured for each deployment strategy in the evaluation described above; values shown are for illustration of the intended comparison only and are not experimental results.

work's expected benefits are, at this stage, based on architectural reasoning rather than measured outcomes. The trade-offs described here such as local energy savings versus communication cost are plausible in principle but haven't yet been quantified against one another, and resolving this is exactly what the planned evaluation is meant to address.

IV. Evaluation Methodology

Empirical validation of the proposed framework has not been carried out yet and is planned as a direct next step following this work. This section outlines the intended evaluation methodology rather than presenting any actual results.

For comparison, the evaluation will use three baselines: (i) a fully cloud-resident EEGNet, where raw preprocessed signal is transmitted for end-to-end classification; (ii) a fully edge-resident, scaled-down EEGNet that operates without any cloud offloading; and (iii) the edge-cloud split proposed in Section II.

The planned performance metrics include:

- Classification accuracy and F1-score for each decoding task;
- Per-inference latency, tracked separately for the edge-side stage, the cloud-side stage, and the full end-to-end round trip;
- Energy consumption per inference at the edge device (e.g., in millijoules), either measured directly or estimated using the target hardware platform;
- Communication overhead, quantified in terms of bytes transmitted per inference or per unit of time;
- Model size and parameter count for each partitioned configuration.

It's worth emphasizing that no classification accuracy, latency, or energy figures are reported in this paper. These values are intended to come from controlled experiments run on a specified public benchmark—such as the BCI Competition IV Dataset 2a or the PhysioNet EEG Motor Movement/Imagery Dataset used in [22]—or on a separately collected dataset, with results to follow in a subsequent full evaluation.

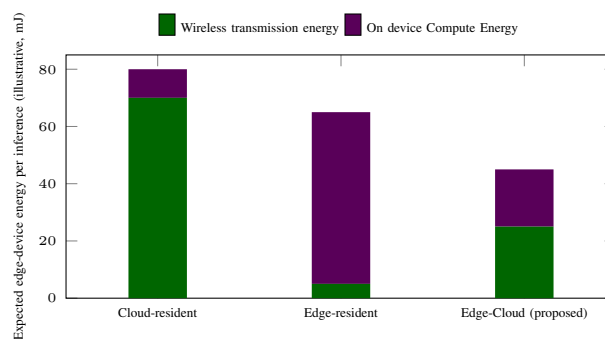


Fig. 7. Illustrative, planned edge-device energy-consumption breakdown to be measured for each deployment strategy in the evaluation described above; values shown are for illustration of the intended comparison only and are not experimental results.

V. Conclusion and Future Work

This paper outlines three issues associated with the long-term clinical use of EEG-based Brain-Computer Interfaces (BCIs): energy efficiency, the loss of performance outside controlled conditions, and a lack of scalability. To address these issues, this paper proposes a four-layer edge-cloud architecture to divide an EEGNet-based decoding architecture between a wearable edge device and a cloud. According to the architectural analysis, performance and accuracy are maintained under real-world conditions due to the adaptive, ensemble-based classifications and the division of computations. This architecture can also potentially help reduce power consumption, transmission size, and on-device computations.

Future work will test this architecture using the datasets in Section IV, and will compare it to fully edge- and fully cloud-resident architectures. It will also describe edge-cloud communications in a realistic wireless environment. Finally, this work will also consider some of the regulatory, security, and ethical implications of the transmission and storage of neural data. This will include the ownership of neural data and informed consent, as they are both required for the anticipated clinical use of this technology.

References

- [1] J. R. Wolpaw et al., "Brain-computer interface technology: A review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, 2000.
- [2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces: A review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.
- [3] B. Blankertz, G. Curio, and K. Müller, "Classifying single trial EEG: Towards brain computer interfacing," in *Advances in Neural Information Processing Systems*, 2002.
- [4] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.
- [5] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. Oxford, U.K.: Oxford Univ. Press, 2012.
- [6] B. He et al., "Noninvasive brain-computer interfaces based on sensorimotor rhythms," *Proc. IEEE*, vol. 96, no. 7, pp. 1182–1196, 2008.
- [7] D. J. McFarland and J. R. Wolpaw, "Brain-computer interfaces for communication and control," *Commun. ACM*, vol. 54, no. 5, pp. 60–66, 2011.

- [8] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: Applying brain–computer interface technology to human–machine systems," *J. Neural Eng.*, vol. 8, no. 2, 2011.
- [9] G. Schalk and J. Mellinger, *A Practical Guide to Brain–Computer Interfacing with BCI2000*. London, U.K.: Springer, 2010.
- [10] M. Lebedev and M. A. L. Nicolelis, "Brain–machine interfaces: From basic science to neuroprostheses," *Physiol. Rev.*, vol. 97, no. 2, pp. 767–837, 2017.
- [11] S. Makeig, A. J. Bell, T. Jung, and T. J. Sejnowski, "Independent component analysis of EEG data," in *Advances in Neural Information Processing Systems*, 1996.
- [12] R. Fazel-Rezai et al., "P300 brain–computer interface: Current challenges and emerging trends," *Front. Neuroeng.*, vol. 5, 2012.
- [13] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Chichester, U.K.: Wiley, 2007.
- [14] K. K. Ang and C. Guan, "EEG-based brain–computer interfaces for rehabilitation," *Proc. IEEE*, vol. 103, no. 6, pp. 944–953, 2015.
- [15] M. Cheng, X. Gao, S. Gao, and D. Xu, "Design and implementation of a brain–computer interface with high transfer rates," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 10, pp. 1181–1186, 2002.
- [16] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain Topogr.*, vol. 23, pp. 194–198, 2010.
- [17] N. Birbaumer, "Breaking the silence: Brain–computer interfaces for communication and motor control," *Psychophysiology*, vol. 43, no. 6, pp. 517–532, 2006.
- [18] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2014.
- [19] R. Scherer, G. Müller-Putz, C. Neuper, and G. Pfurtscheller, "An asynchronous BCI based on motor imagery," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 4, pp. 1453–1457, 2008.
- [20] M. Ahn, M. Lee, J. Choi, and S. Jun, "A review of brain–computer interface games and an opinion survey from researchers, developers and users," *Sensors*, vol. 14, no. 8, pp. 14601–14633, 2014.
- [21] V. J. Lawhern et al., "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.
- [22] X. Wang et al., "An accurate EEGNet-based motor-imagery brain–computer interface for low-power edge computing," in *Proc. IEEE Int. Symp. Medical Measurements and Applications (MeMeA)*, 2020.