Edge AI-Brain–Computer Interfaces System: A Survey

Manh-Dat Nguyen, *Student Member, IEEE*, Thomas Do, Xuan-The Tran, Quoc-Toan Nguyen, and Chin-Teng Lin, *Fellow, IEEE*

***Abstract*—Edge artificial intelligence (Edge AI) has emerged as a transformative paradigm for enhancing the performance, portability, and autonomy of brain–computer interface (BCI) systems. By integrating advanced AI capabilities directly into electroencephalography (EEG)-based devices, Edge AI enables real-time signal processing, reduces dependence on external computational resources, and improves data privacy. However, deploying AI on resource-constrained hardware introduces challenges related to computational capacity, power consumption, and system latency. This survey provides a comprehensive examination of Edge AI–enabled BCI systems, covering the full pipeline from EEG hardware specifications and on-device data acquisition to signal preprocessing techniques and lightweight deep learning models optimized for embedded platforms. We review existing frameworks, specialized hardware accelerators, and energy-efficient AI approaches that facilitate real-time BCI processing at the edge. Furthermore, the paper reviews state-of-the-art solutions, examines key technical challenges, and outlines future research directions in hardware–software co-design and application development. This work aims to serve as a reference for researchers and practitioners seeking to design efficient, portable, and practical Edge AI–powered BCI systems.**

***Index Terms*—Brain–computer interfaces (BCIs), electroencephalography (EEG), edge artificial intelligence (Edge AI), embedded systems, deep learning, real-time processing, on-device AI, neural signal processing.**

I. Introduction

BRAIN Computer Interface (BCI) systems have seen remarkable advancements in recent years, expanding their role beyond traditional applications in assistive technology [1], [2], [3] to emerging fields such as image and video reconstruction [4], [5], [6], text and speech synthesis [7], [8], [9], and action [10] decoding from brain activity. Additionally, BCIs have demonstrated significant potential in neurorehabilitation [11], [12], [13], cognitive driving [14], [15], physical navigation [16], [17], and human-computer interaction [18]. These systems enable direct communication between the human brain and external devices, facilitating control without needing conventional motor functions [19]. The integration of artificial intelligence (AI) and embedded computing has further enhanced BCI performance [20], [21], improving response times for communication with external devices and increasing the overall portability of the system. However, the deployment of AI-powered BCI systems presents several challenges, particularly concerning hardware constraints, real-time processing requirements, and the balance between computational efficiency and power consumption. To address these limitations, AI-integrated BCIs must be optimized for real-time processing and resource efficiency. This paper reviews the current advancements in edge AI-powered BCIs, highlighting recent progress, existing challenges, and future research directions. Integration of AI models directly within embedded BCI devices enhances real-time signal processing, improves classification accuracy, and enables more efficient brain-controlled interfaces for medical and consumer applications.

A typical BCI system consists of several essential elements, including signal acquisition, preprocessing, feature extraction, classification, and output generation. Among non-invasive methods, electroencephalography (EEG) has emerged as the predominant technique for recording neural activity due to its high temporal resolution, ease of implementation, and ability to provide a safe and effective alternative to invasive approaches [22]. The quality of EEG data acquisition is critically dependent on hardware specifications, necessitating an analog-to-digital conversion (ADC) resolution of at least 16 bits and a minimum sampling rate of 250 Hz for research applications. Preprocessing is a fundamental step in BCI pipelines, as EEG signals are inherently susceptible to noise from muscle artifacts, power-line interference, and environmental factors. Filtering techniques are employed to isolate specific frequency bands of interest, including theta (4–8 Hz), alpha (8–12 Hz) [23], and beta (13–30 Hz) [24],

Received 29 April 2025; revised 29 August 2025; accepted 1 October 2025. Date of publication 8 October 2025; date of current version 16 October 2025. This work was supported in part by Australian Research Council (ARC)discovery under Grant DP220100803 and Grant DP250103612, in part by the ARC Research Hub for Human-Robot Teaming for Sustainable and Resilient Construction (ITRH) under Grant IH240100016, in part by the Australian National Health and Medical Research Council (NHMRC) Ideas under Grant APP2021183, in part by the Australia Advanced Strategic Capabilities Accelerator (ASCA) under Contract P18-650825 and Contract ASCA EDT DA ID12994, and in part by the Australian Defence Science Technology Group (DSTG) under Agreement 12549. *(Corresponding author: Thomas Do.)*

The authors are with the Computational Intelligence and Brain Computer Interfaces Laboratory, Australian AI Institute, School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: thomas.do@uts.edu.au).

Digital Object Identifier 10.1109/TNSRE.2025.3618688

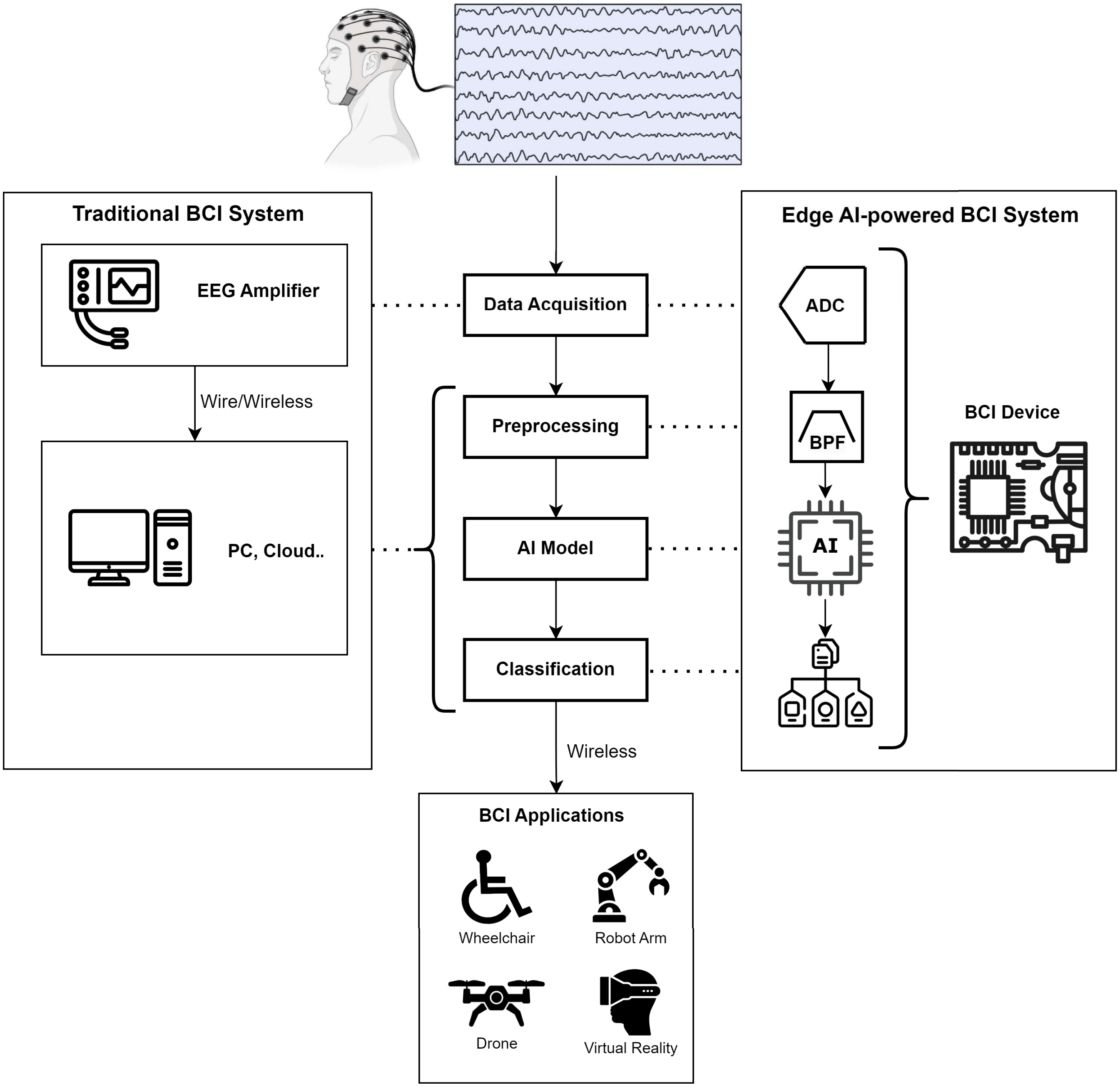

Fig. 1. Edge AI – Brain-Computer Interfaces: A Conceptual Overview. The figure illustrates key components of an Edge AI-powered BCI system. Abbreviations: BCI - Brain-Computer Interfaces, EEG - Electroencephalogram, AI - Artificial Intelligence, ADC – Analog-to-Digital Converter, BPF – Band-Pass Filter.

which are commonly associated with cognitive and motor functions. Recent advancements in AI-based feature extraction and classification have demonstrated significant improvements in BCI performance. However, deploying these models on resource-constrained platforms, such as microcontroller units (MCUs), introduces additional considerations regarding computational complexity and energy efficiency. A comparative assessment of MCUs and single-board computers (SBCs) highlights the trade-offs between low-power operation and enhanced processing capabilities, influencing the selection of appropriate hardware for real-time BCI applications. As illustrated in Figure 1, Edge AI enables a shift from traditional BCI architectures, where signal processing and classification are performed on external PCs, to a more compact and efficient system. By integrating these tasks into an Edge AI module, the overall system size is reduced, enhancing portability and usability, while also preserving data privacy by avoiding the transmission of sensitive neural data to external servers. However, running AI workloads on resource-constrained MCUs must be carefully optimized to operate within strict computational and power budgets.

This paper presents a thorough examination of AI-powered BCI systems, encompassing hardware design, data acquisition, preprocessing methodologies, and deep learning

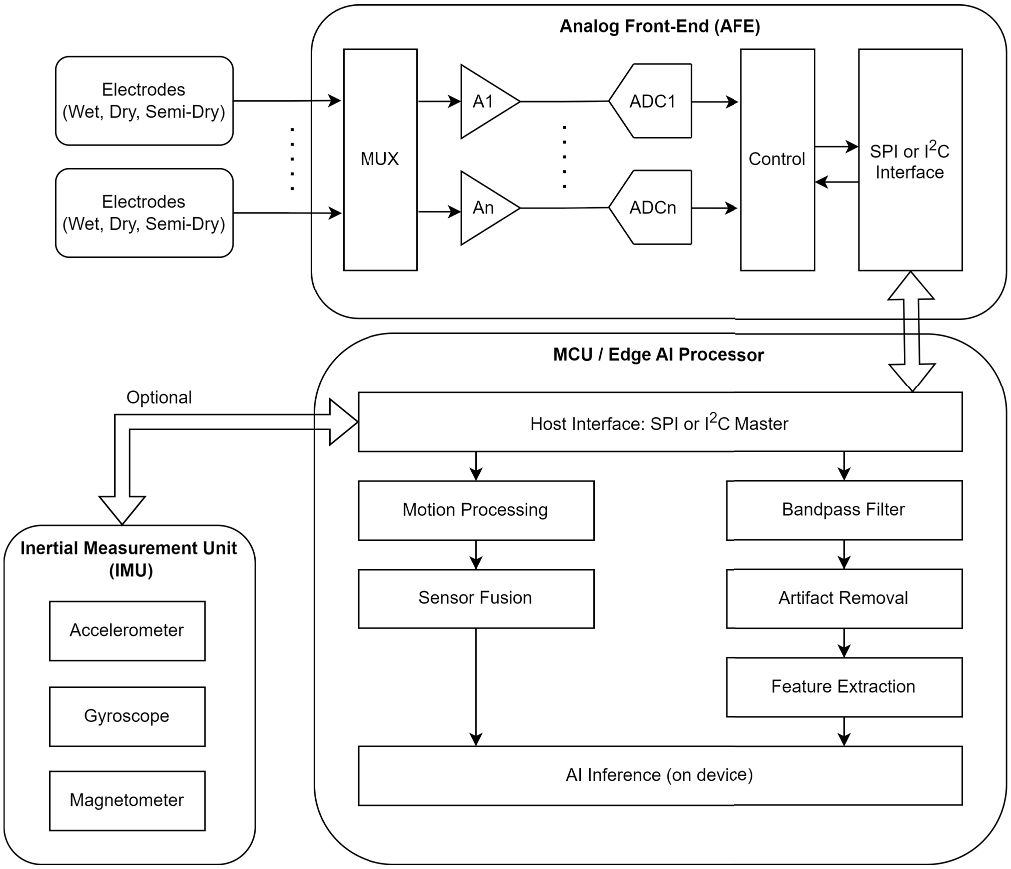

Fig. 2. System-level overview of EEG acquisition and on-device inference in Edge AI-powered BCIs.

models tailored for real-time applications. section II outlines system-level considerations for EEG acquisition, emphasizing how specifications such as channel count, sampling rate, and ADC resolution influence the quality of neural data and the feasibility of on-device AI inference. section III discusses data acquisition and preprocessing techniques, focusing on noise reduction and feature extraction.section IV highlights advancements in edge AI for efficient BCI processing, focusing on emerging frameworks and on-device computation techniques that enhance system performance and portability. section V provides a comprehensive review of recent advancements in on-device AI for BCI, summarizing key research efforts, novel hardware-software co-design strategies, and real-world demonstrations of TinyML-enabled BCI applications. section VI presents real-world applications of edge AI-powered BCI systems, showcasing their potential to enable real-time processing, reduce reliance on external infrastructure, and enhance privacy. Finally, section VII concludes with key findings and potential directions for further investigation.

II. System-Level Considerations for EEG Acquisition in Edge AI-Powered BCIs

A. System-Level Overview

Figure 2 illustrates the end-to-end acquisition pipeline in a typical Edge AI-powered BCI system. Each stage in the signal chain, including electrodes, analog front-end (AFE), digitization, and embedded processing, impacts the fidelity, latency, and interpretability of neural signals used for AI-based inference.

Edge AI-powered BCI systems introduce additional design constraints compared to traditional laboratory setups. These systems must operate under limited power budgets, maintain compact hardware footprints, and preserve robust signal fidelity in mobile or wearable contexts. To meet these requirements, designers must carefully consider both the AFE characteristics and the configuration of the ADC.

Three analog performance metrics such as input impedance, input-referred noise, and amplifier input range significantly affect signal fidelity prior to digitization. These metrics are influenced not only by the design of the AFE but also by how programmable gain and ADC resolution are configured in practice. High input impedance (typically greater than 100 MΩ, and often in the GΩ range for modern dry or semi-dry electrodes [25]) helps prevent signal attenuation caused by variable skin–electrode contact. Input-referred noise determines the minimum resolvable signal amplitude and should remain below 1 μV_{rms} across the EEG bandwidth to preserve microvolt-level features [26], [27], [28]. The amplifier input range, determined by gain and reference voltage settings, must be sufficiently wide to accommodate both the EEG signals and larger electrode-induced offsets or artifacts. Practical guidelines recommend a minimum of ±50 mV in DC-coupled mode [29], while classical designs specify tolerance of up to

±200 mV without saturation [30]. Ensuring such headroom prevents clipping from large offsets and allows optimal utilization of the ADC's dynamic range [31], [32].

Aside from analog fidelity, three critical ADC-related factors, including number of channels, sampling rate, and resolution, directly influence the quality and quantity of neural data available for on-device inference. Systems with higher channel counts and sampling rates offer richer spatial and temporal resolution, but they also increase data volume, energy consumption, and processing complexity. Conversely, low-density systems with modest sampling rates can better support real-time inference on constrained hardware by reducing latency and computational load. These digital acquisition parameters are tightly coupled with analog fidelity metrics and collectively determine the overall quality of data available for Edge AI processing.

A comprehensive survey by Niso et al. [33] systematically compares wireless EEG acquisition systems across consumer, research, and clinical domains. The review details key hardware specifications, including electrode type, number of channels, sampling frequency, ADC resolution, and connectivity protocol, making it a valuable reference for selecting acquisition platforms compatible with Edge AI-powered BCI applications.

B. Number of Channels

The number of EEG channels is a critical factor in determining how many scalp locations can be simultaneously monitored. A higher channel count improves the spatial resolution of brain activity data, enabling more detailed insights into neural dynamics. However, it also increases system complexity and resource requirements. Channel selection is often applied to reduce dimensionality and improve computational efficiency by excluding redundant or noisy signals, thereby minimizing overfitting in machine learning pipelines [34].

Low-density EEG (LD-EEG), typically using 8 to 32 channels [35], [36], is sufficient for many research and clinical applications, including sleep studies and general neurological assessments. In contrast, high-density EEG (HD-EEG) systems may employ 64, 128, or even 256 electrodes [37], [38], and are ideal for applications requiring fine-grained spatial resolution, such as localizing epileptogenic zones [39] or mapping motor areas in cognitive tasks involving physical movement like walking [40] and driving [14], [41].

The placement of the reference electrode also plays a crucial role, as all recorded potentials are measured relative to it. Improper placement can introduce artifacts or obscure signal clarity, significantly affecting signal interpretability [42].

Increased channel count enhances decoding potential but also amplifies the volume of data to be processed and transmitted. This places greater demands on memory, processing throughput, and power consumption, which are factors that must be carefully balanced in edge AI systems with limited hardware resources.

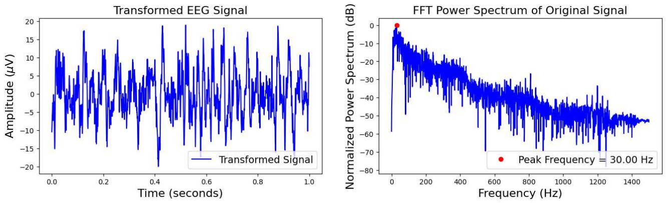

Fig. 3. The transformed EEG signal in both the time and its corresponding Fast Fourier Transform (FFT) power spectrum.

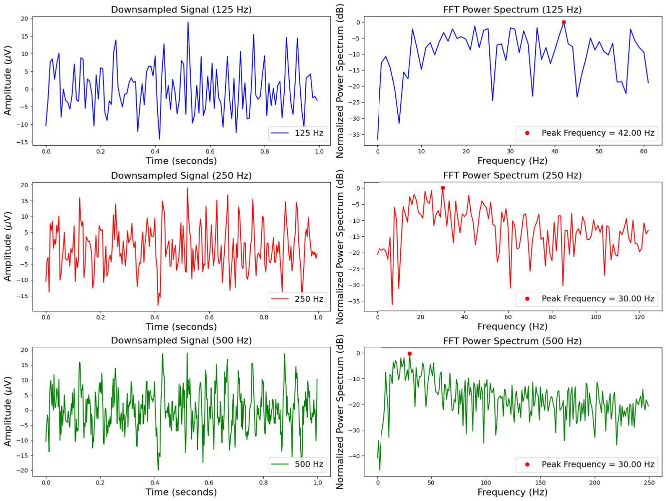

Fig. 4. Comparison of EEG signals sampled at 125Hz, 250Hz, and 500 Hz. **Left:** Time-domain EEG signals at different sampling rates. **Right:** Corresponding power spectra computed using FFT.

C. Sampling Rate

The sampling rate of an EEG system defines how frequently the signal is digitized, directly influencing both temporal resolution and frequency resolution. A higher sampling rate allows for more precise capture of rapid time-domain fluctuations and more accurate representation of high-frequency components.

To illustrate the effects of resampling, Figure 3 presents a simulated EEG signal derived from a 30-second segment of slow-wave sleep data originally sampled at 100 Hz [43]. This signal was resampled to 3000 Hz and compressed to 1 second, which proportionally shifted the dominant frequency from 1 Hz to 30 Hz. The amplitude was also adjusted to 20μ V to simulate beta-band activity. This transformation highlights how sampling rate affects both temporal compression and spectral scaling, and serves as a basis for evaluating signal processing parameters such as sampling rate and ADC resolution.

Figure 4 compares EEG recordings sampled at 125 Hz, 250 Hz, and 500 Hz. The signal acquired at 125 Hz exhibits broadened and distorted spectral peaks, whereas those sampled at 250 Hz and 500 Hz more accurately preserve the waveform morphology and spectral characteristics.

Although higher sampling rates improve frequency resolution and temporal detail, they also increase data volume and computational load during signal acquisition and preprocessing. In edge AI applications, where latency and energy consumption are critical design factors, it is important to

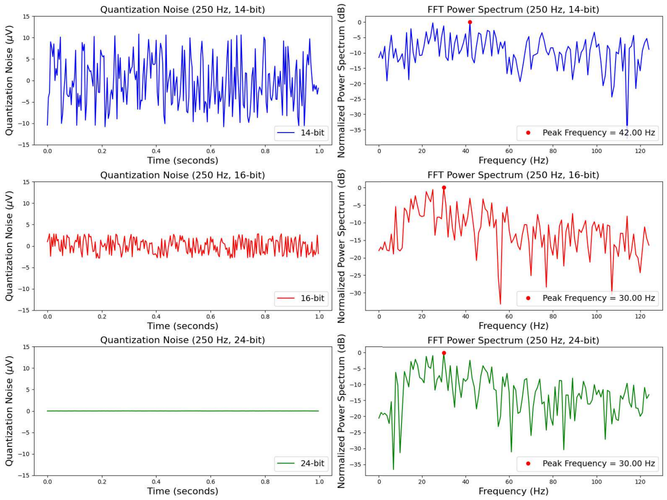

Fig. 5. Quantization effects at different ADC resolutions (14-bit, 16-bit, and 24-bit). **Left:** Time-domain representation of the quantization noise, computed as the difference between the original signal and its quantized version. **Right:** Frequency-domain spectrum (FFT) of the quantized signal, illustrating distortion introduced by quantization noise.

select the lowest sampling rate that still preserves task-relevant spectral information. This ensures a balance between signal fidelity and system efficiency.

D. ADC Resolution

Analog-to-digital converter (ADC) resolution defines the number of discrete levels used to digitize continuous EEG voltages. Higher resolutions reduce quantization noise and preserve subtle signal variations, which is particularly important when working with low-amplitude EEG features. Lower resolution not only increases quantization noise but also directly reduces the effective signal-to-noise ratio (SNR), potentially masking meaningful neural activity in the digitized signal.

Figure 5 shows the impact of quantization across 14-bit, 16-bit, and 24-bit ADCs. The 14-bit resolution introduces noticeable time-domain distortion and frequency-domain artifacts, including spurious components near 42 Hz that obscure the original spectral peak at 30 Hz. The mathematical derivation of quantization error and its relationship to ADC bit depth is provided in Appendix. Lower resolution not only increases quantization noise but also directly reduces the effective signal-to-noise ratio (SNR), potentially masking meaningful neural activity in the digitized signal.

While 16-bit or 24-bit resolution is recommended for high-fidelity EEG acquisition, higher resolution also results in more data per sample. This increases memory usage, transfer bandwidth, and computational overhead, which can stress the limited resources of edge AI platforms unless efficient data handling strategies are implemented

E. System-Level Trade-Offs in Edge AI-Powered BCI Systems

The acquisition parameters discussed in the previous subsections, namely the number of EEG channels, sampling rate, ADC resolution, and trial length, influence not only signal quality and decoding performance but also the overall feasibility of deploying BCI models on edge AI hardware. To meet real-time requirements within limited computational resources, these parameters must be co-optimized with model complexity and processing latency.

In embedded BCI systems, the total processing time per trial includes the time required for signal acquisition $T_{\text{acquisition}}$, preprocessing T_{preproc}, and model inference $T_{\text{inference}}$. This total latency T_{total} must not exceed the trial duration T_{trial} in order to maintain continuous real-time operation. This constraint can be expressed as follows:

$$T_{\text{total}} = T_{\text{acquisition}} + T_{\text{preproc}} + T_{\text{inference}} \leq T_{\text{trial}} \tag{1}$$

Reducing the trial length is a common strategy to improve information transfer rate and responsiveness. However, shorter trials also reduce the time available for data handling and inference. This places a practical lower bound on trial duration, particularly in systems with limited compute or memory resources.

The complexity of the AI model further influences feasibility. Models with a large number of parameters and a high count of multiply-accumulate operations require more processing cycles, which increases inference latency. In such cases, signal acquisition parameters may need to be adjusted. For example, reducing the number of EEG channels or lowering the sampling rate can help keep data volume within the processing capacity of the hardware.

The total volume of raw data acquired during each trial can be estimated as follows:

$$\begin{aligned} \text{Data Volume(bits)} = T_{\text{trial}} \times \Big(& N_{\text{EEG}} \cdot f_s \cdot R_{\text{ADC}} \\ & + \underbrace{N_{\text{IMU}} \cdot f_{\text{IMU}} \cdot R_{\text{IMU}}}_{\text{optional if IMU is used}} \Big) \end{aligned} \tag{2}$$

where:

- T_{trial}: trial duration in seconds
- N_{EEG}: number of EEG channels
- f_s: EEG sampling rate in Hz
- R_{ADC}: EEG ADC resolution in bits
- N_{IMU}: total number of IMU data channels (e.g., 3 per accelerometer, 3 per gyroscope, 3 per magnetometer)
- f_{IMU}: IMU sampling rate in Hz
- R_{IMU}: IMU resolution in bits

These relationships highlight that trial length, signal acquisition settings, and model architecture are tightly interdependent. Reducing trial duration, increasing channel count, or raising the sampling rate all increase the volume of data that must be handled within a fixed latency budget. Meanwhile, selecting a more complex model reduces the available margin for acquisition and preprocessing. Achieving real-time performance on embedded hardware requires a co-design approach that jointly optimizes signal fidelity, computational efficiency, and responsiveness across the entire data pipeline.

III. EEG Data Acquisition and Preprocessing

Accurate EEG data acquisition and effective preprocessing are critical steps in ensuring the reliability and responsiveness

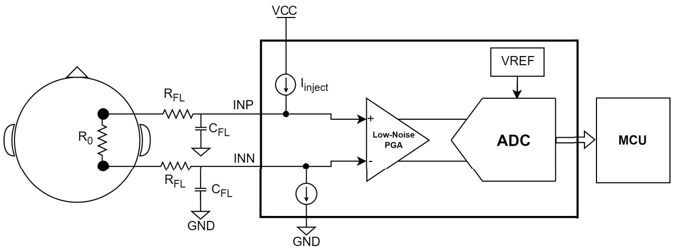

Fig. 6. Lead-off Detection diagram.

TABLE I
IMPEDANCE THRESHOLDS FOR WET AND DRY ELECTRODES

Sensor Type	Acceptable Impedance	Faulty Threshold
Wet	Below 5 kΩ [48] to 10 kΩ [44], [45]	Above 10 kΩ
Dry	Below 150 kΩ [46], [47]	Above 150 kΩ

of BCI systems. While hardware characteristics govern signal fidelity (as outlined in Section II), this section focuses on maintaining data quality during signal acquisition and transforming raw EEG data into clean, structured formats suitable for downstream AI processing. Raw EEG signals are frequently contaminated by muscle activity (EMG), eye movements (EOG), or electrical interference, which poses challenges for real-time or edge deployment. To address these issues, preprocessing methods such as filtering, artifact rejection, and normalization are employed to enhance signal quality and improve model compatibility. This section first discusses signal integrity at the electrode interface, then examines AI learning strategies for EEG decoding, followed by an overview of widely used preprocessing pipelines with an emphasis on their feasibility for Edge AI applications.

A. Electrode Interface and Impedance Monitoring

While Section II described the complete signal acquisition chain, this section focuses specifically on the interface between electrodes and the scalp, and how impedance affects signal integrity. The quality of the electrode–skin contact can degrade over time due to factors such as movement, sweat, or drying conductive gel, which may lead to increased impedance and adversely affect the quality of recorded signals.

To ensure the reliability of data acquisition, EEG systems must incorporate a lead-off detection feature, also referred to as electrode-off detection, which continuously monitors the electrode connections. This feature provides users with real-time feedback on the validity of the recorded data by identifying whether an electrode is functioning properly. The lead-off detection works by injecting a small excitation current into each electrode and measuring the resulting voltage. The impedance is then calculated to assess the electrode's status.

If the impedance exceeds a threshold value, typically 10 kΩ [44], [45] for wet electrodes and 150 kΩ [46], [47] for dry electrodes, the electrode connection is identified as faulty. These threshold values are based on established practices and supported by prior studies, as summarized along with their corresponding References in Table I. This ensures transparency regarding the source of these thresholds and reinforces their use in evaluating electrode performance. The table also helps determine when an electrode is functioning properly or needs attention, with thresholds varying depending on the sensor type and the equipment used. This mechanism ensures that users can detect and address broken leads, thereby maintaining the quality of EEG recordings. The principal diagram of this detection process is depicted in Figure 6.

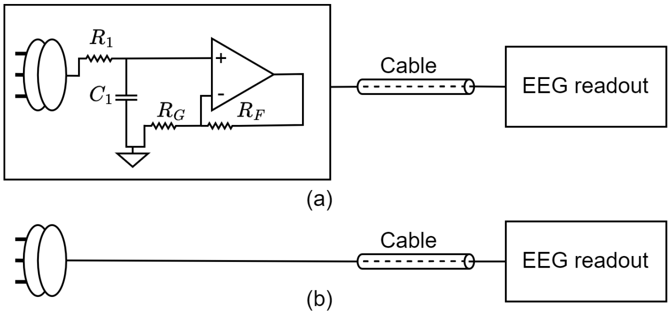

Fig. 7. The typical block diagram of (a) active electrode and (b) passive electrode.

When the lead-off detection feature is enabled, a small excitation current is injected into the circuit. This current flows through a filter resistor at the positive input pin (INP), passes through the impedance between two electrodes (representing the patient-electrode connection), and then returns via a filter resistor at the negative input pin (INN). If the connection between the electrodes is intact, the impedance has a finite value, allowing the current to flow. The resulting voltage difference between the INP and INN pins reflects the impedance of the electrode connection and is converted into a digital value by the ADC. This process can be enhanced by injecting a small excitation current with a fixed frequency, enabling additional information to be extracted from the EEG data. By applying a band-pass filter tuned to the frequency of the excitation signal, the root mean square (RMS) voltage of the response signal caused by the injected current can be calculated. Consequently, the impedance can be continuously measured in real-time over a specific window of EEG samples. This impedance is then compared against a predefined threshold to determine if the connection is valid. However, if the connection is broken (e.g., due to poor electrode contact), the impedance between the two electrodes approaches infinity. In this case, no current flows, and the voltage difference between INP and INN is dictated primarily by the circuit's open-loop behavior. The ADC output will indicate an abnormally high impedance, signaling that either the INP or the INN is disconnected. To minimize interference with EEG signals, an excitation current amplitude of 6 nA is typically used for lead-off detection [49]. This careful management of excitation currents not only preserves signal integrity but also influences the choice of electrode type for EEG acquisition [50]. Depending on the design and application, different electrode configurations, such as active or passive, can further enhance signal quality and stability.

Active electrodes have a pre-amplifier circuit mounted on the electrode that amplifies signals before they are passed on to an amplifier. On the other hand, passive electrodes collect signals at the scalp and then transmit them to an amplifier integrated into the EEG readout device. The typical design of designing an active and passive electrode has been shown in Figure 7. In both diagrams, a cable connects the electrode to the EEG readout device. A long cable can act as an antenna, picking up electromagnetic interference as noise, which can significantly affect the transmitted signals. For passive electrodes, both the EEG signal and the noise introduced by the cable are amplified at the EEG device. This makes the signal more susceptible to noise and degradation, especially over long cable lengths. In contrast, active electrodes address this limitation by buffering the signal at the source using a unity-gain amplifier, thereby minimizing the transmission of the non-buffered, high-impedance signal between the electrode and the processing circuitry. This impedance buffering reduces susceptibility to noise and motion artifacts introduced by long cables. Additionally, the amplifier's low output impedance reduces the impact of cable motion artifacts, eliminating the need for shielded cables and lowering overall system cost [51]. Even while active electrodes' signal quality has greatly improved, both active and passive electrodes operate similarly in optimal laboratory environments when conductive gel is used to create low electrode impedance [52]. However, adding the preamplifier circuit to the electrode makes it heavier, which may limit participant and device mobility. Active electrodes also require additional energy to power the preamplifier circuits, leading to higher power consumption. Furthermore, the inclusion of these circuits increases the cost of the electrodes. Therefore, the choice between active and passive electrodes depends on the specific EEG application, the available budget, and the mobility requirements of the system.

B. Machine Learning Approaches in AI-Powered BCI Systems

Within the domain of AI-driven BCI systems, this section elucidates the pivotal contributions of machine learning (ML) and deep learning (DL) algorithms to the decoding of EEG data. Two predominant methodologies govern the application of ML and DL to EEG signal processing: feature-based techniques and end-to-end deep learning frameworks [53]. Feature-based approaches involve the extraction of multifaceted features from EEG signals spanning temporal, spectral, and spatial domains, which are subsequently input into conventional classifiers or regressors optimized for specific decoding objectives. This methodology necessitates rigorous preprocessing of EEG data to optimize feature integrity [54], [55] and suppress artifactual interference [56], rendering it particularly advantageous in contexts characterized by constrained EEG datasets, such as those with limited subject cohorts or trial numbers, and restricted computational resources.

This advantage of feature-based methods becomes more apparent when considering that, unlike vision or speech domains, EEG datasets are often small and heterogeneous due to inter-subject variability, session-to-session differences, and the cost of extensive data collection [57]. This scarcity poses a significant limitation for deploying high-capacity deep learning models in BCI applications, as large labeled datasets are typically required to achieve robust performance. The lack of sufficient data can limit model generalization and robustness, particularly for end-to-end architectures directly adapted from other fields. To mitigate this challenge, strategies such as transfer learning, data augmentation, and domain adaptation have been explored to enable effective model training and deployment, including in resource-constrained Edge AI scenarios.

Conversely, end-to-end deep learning paradigms minimize preprocessing demands, frequently leveraging raw EEG data [58] to achieve robust decoding performance. The streamlined data processing architecture inherent to DL-based systems enhances their applicability to real-time BCI implementations [59]. Nevertheless, these approaches are contingent upon extensive EEG datasets and substantial computational infrastructure for both training and inference phases, posing significant challenges for deployment in resource-limited settings. The ensuing section will present a comprehensive review, concentrating predominantly on deep learning methodologies within BCI systems, and delineating their recent advancements and prospective capabilities.

IV. Edge AI for Efficient BCI Processing

The incorporation of AI capabilities into BCI systems has significantly improved both the response time of BCIs when interacting with external devices and the overall portability of the system. Traditional AI models are often trained and executed on high-performance computing platforms. However, with the rise of edge computing, deploying AI models directly on embedded devices has become increasingly feasible. To achieve efficient on-device inference, several lightweight AI frameworks have been developed, optimizing deep learning models for resource-constrained hardware.

A. Edge AI Framework

ESP-DL (launched on April 2, 2021) is a lightweight and efficient neural network inference framework tailored for Espressif's ESP series System on Chips (SoCs), making AI application development simple and fast [60]. It provides intuitive APIs for loading, debugging, and running AI models, while seamlessly integrating with other Espressif SDKs. With ESP-PPQ, ESP-DL supports quantizing models from ONNX, PyTorch, and TensorFlow into its efficient ESP-DL Standard Model Format, which uses FlatBuffers for lightweight, zero-copy deserialization. The framework is optimized for performance with features like efficient operator implementation, a static memory planner that allocates resources based on internal RAM size, and dual-core scheduling for computationally intensive tasks such as Conv2D, Gemm. Additionally, ESP-DL accelerates inference by implementing activation functions (except ReLU and PReLU) using an 8-bit Look-Up Table (LUT), ensuring high efficiency without compromising flexibility. Designed specifically for resource-constrained environments, ESP-DL is the ideal choice for deploying AI models on ESP series chips.

TABLE II
Comparison of Edge AI Frameworks

Key parameters	**ESP-DL**	**STM32.AI**	**TFLite Micro**	**Edge Impulse**	**ARM CMSIS-NN**
Provider	Espressif Systems	STMicroelectronics	Google	Edge Impulse	ARM
Target Hardware	ESP32-S3, ESP32-P4, ESP32, ESP32-C3	STM32 MCUs and MPUs	Cross-platform edge devices, including STM32, ESP, Nordic, Texas Instruments, and NXP platforms	Cross-platform edge devices, including STM32, ESP, Nordic, Infineon, Raspberry Pi, Texas Instruments	ARM Cortex-M microcontrollers
Model Format	Pytorch, TensorFlow Lite and ONXX	Pytorch, TensorFlow Lite, Keras, ONNX, Scikit learn	TensorFlow Lite	TensorFlow, TensorFlow Lite, ONNX	TensorFlow, TensorFlow Lite
Quantization Support	INT8, INT16 [68]	INT8 [69]	INT8, FP16, Mixed (INT8: weights, INT16: activations) [70]	INT8 [71]	INT8, INT16 [67]
Ease of Use	Moderate	Moderate	Easy	Easy	Diffcult
Applications	ESP32-specific applications, such as Image and object recognition	Applications on STM32 platform such as Object detection, Audio event detection, Image classification	General-purpose ML applications on resource-constrained devices; suitable for a wide array of IoT applications	Rapid prototyping and deployment of ML models in IoT and edge applications	Performance-critical applications where maximizing the efficiency of neural network inference on ARM Cortex-M devices is essential
Licensing	Free and Open-source	Free but Proprietary	Free and Open-Source	Free for Individuals, Paid for Enterprise	Free and Open-Source
Suitability for BCI	Suitable for embedded BCI systems with strict power constraints on ESP32 platforms	Suitable for lightweight BCI tasks (e.g., motor imagery decoding) on STM32 with good power efficiency	Suitable for lightweight BCI tasks across platforms with moderate optimization and high flexibility	Highly suitable for prototyping lightweight BCI tasks (e.g., motor imagery) across platforms	Highly suitable for performance-critical BCI on Cortex-M with high efficiency

STM32.AI (introduced in 2019) is a framework developed by STMicroelectronics to facilitate the deployment of machine learning models on STM32 microcontrollers [61]. It supports model conversion from popular frameworks like TensorFlow, PyTorch, and ONNX, optimizing them for STM32 hardware accelerators such as Chrom-ART and DSP units. STM32.AI integrates seamlessly with STM32CubeMX and STM32Cube.AI, providing developers with tools for model conversion, validation, and performance benchmarking. This framework is ideal for applications requiring efficient AI deployment on STM32 devices.

TensorFlow Lite (first committed on April 9, 2021) for Microcontrollers (TFLM) is a lightweight version of TensorFlow Lite designed to run machine learning models on microcontrollers and other devices with limited resources [62]. It supports post-training quantization (INT8) and pruning, which help reduce model size and improve inference speed, making it suitable for low-power, resource-constrained devices. TFLite Micro is hardware-agnostic and portable, supporting a wide range of microcontrollers, including those from NXP [63], Texas Instruments [64], and Espressif Chipsets [65]. The framework provides an interpreter to efficiently run models on-device and tools for model conversion and optimization.

Edge Impulse (founded in 2019) is an end-to-end platform for developing and deploying machine learning models on edge devices [66]. It provides cloud-based tools for data collection, labeling, model training, and deployment, making it accessible even to those without extensive machine learning expertise. Edge Impulse supports a wide range of hardware platforms and is particularly suited for rapid prototyping and development of edge AI applications. The platform also provides support for various deployment targets, including microcontrollers and other resource-constrained devices.

ARM Common Microcontroller Software Interface Standard neural network (CMSIS-NN, introduced in 2017) is a collection of highly optimized neural network kernels designed for ARM Cortex-M microcontrollers [67]. As part of the CMSIS, it supports INT8 and INT16 quantization and focuses on low-latency, low-memory inference. CMSIS-NN provides efficient implementations of common neural network operations, enabling developers to achieve significant performance improvements in AI applications on ARM Cortex-M devices. It is particularly beneficial for applications requiring real-time processing with minimal resource consumption.

As Edge AI frameworks advance to allow AI models to operate effectively on embedded devices, the next stage of innovation prioritizes creating specialized hardware and optimized architectures to boost BCI applications. Current trends highlight the importance of real-time AI processing on devices, bespoke AI accelerators, and energy-efficient AI models tailored for low-power BCI devices. To assess the adequacy of current frameworks in meeting these requirements, Table II provides a comparison of major Edge AI tools, namely ESP-DL, STM32.AI, TFLite Micro, Edge Impulse and ARM CMSIS-NN, focusing on their compatibility with various hardware, their optimization methods, and their suitability for BCI applications. These frameworks guarantee low-latency processing, enhanced privacy, and real-time adaptability by harnessing their distinct advantages, unlocking novel pathways for on-device AI applications in BCIs.

TABLE III
RECENT ADVANCES IN ON-DEVICE AI FOR BRAIN-COMPUTER INTERFACES

Study	System Components	Framework	Model	Accuracy	Latency	Data Volume & Task
[87]	Custom Device: • SoC: nRF52840 (MCU @64MHz & BLE 5.3) • AFE: ADS1299 (24-bit), 250 SPS • Dry sensors, 4 channels	TFLM	Quantized CNN	96.11%	25.44 ms	Data Volume: 24 (kB) Task: Driver Drowsiness Detection (DDD)
[88]	Custom Device: • SoC: nRF52832 (MCU @64MHz & BLE 5.2) • AFE: ADS1298 (24-bit), 1 kSPS • IMU: TDK InvenSense MPU-9150 (16-bit) • Dry sensors, 3 channels	-	Nearest Centroid Classifier (NCC)	83%	-	Data Volume: - Task: Drowsiness detection using EEG and IMU signals
[89]	Custom Device: • MCU: GAP9 (10 cores @240 MHz) • BLE: nRF52811 • AFE: ADS1298 (24-bit), 500 SPS • Active sensors, 8 channels	DORY	MI-BMInet (lightweight CNN)	90.88%	21.5 ms	Data Volume: 46.875 (kB) Task: Motor Movement / Motor Imagery Classification
[90]	Custom Device: • MCU: GAP9 (10 cores @240 MHz) • BLE: nRF52811 • AFE: ADS1298 (24-bit), 500 SPS • Semi-dry sensors, 8 channels	DORY	EPIDENET	99.74%	4.58 ms	Data Volume: 46.875 (kB) Task: EEG-based biometric subject recognition (alpha waves, SSVEP, and motor movement classification)
[91]	Custom Device: • MCU: GAP9 (10 cores @240 MHz) • BLE: nRF52811 • AFE: ADS1298 (24-bit), 500 SPS • Dry sensors, 8 channels	DORY	VOWELNET	42.8%	40.9 ms	Data Volume: 58.594 (kB) Task: 13-class imagined speech classification (vowels, commands, rest)
[92]	Custom Device: • MCU: STM32L476RG (ARM Cortex M4 @80 MHz) • AFE: ADS1299 (24-bit), 250 SPS • Dry sensors, 2 channels	STM32.AI	1D-CNN	99.3%	200 ms	Data Volume: 1.172 (kB) Task: Mobile Robot Control via Eyeblinks and Winks

Note: Data volume values are expressed in kilobytes (kB), converted from bits as $\mathrm{kB} = \dfrac{\mathrm{bits}}{8 \times 1024}$.

While Table II outlines the capabilities of popular edge AI frameworks and their compatibility with various embedded platforms, it is important to assess whether these frameworks are practically sufficient for real-time BCI applications. To address this, we surveyed recent studies where AI models were deployed on embedded platforms for specific BCI tasks. These studies, summarized in Table III, provide empirical evidence of the feasibility of on-device AI in real-world BCI scenarios. They illustrate how frameworks such as TFLite Micro, STM32.AI, and GAP SDKs (e.g., DORY) have been used to achieve low-latency and energy-efficient inference across various use cases, including motor imagery classification, driver drowsiness detection, and biometric identification. This connection between general-purpose frameworks and real-world BCI deployment helps demonstrate the evolving readiness of Edge AI for BCI tasks.

B. On-Device AI for Real-Time BCI Processing

Traditional BCI systems often rely on local computers or cloud-based computing to process neural signals and generate predictions. However, the shift towards on-device AI is becoming increasingly prominent, allowing real-time inference with minimal latency [72]. By processing neural signals directly on the device, on-device AI eliminates the need to transmit data to remote servers, enabling real-time inference. This is particularly critical for applications such as prosthetic control or neurofeedback, where even minor delays can disrupt user experience or compromise functionality. Additionally, privacy and security are greatly enhanced, as sensitive neural data remains on the device and is not exposed to potential breaches during transmission or cloud storage [73]. This is especially important for medical and personal BCI applications, where data confidentiality is paramount. Furthermore, modern microcontrollers and AI-enabled chips are designed to be energy-efficient [74], [75], making them ideal for wearable BCI devices that require long battery life. Finally, on-device AI enables offline functionality, allowing BCI systems to operate in environments with limited or no internet connectivity, such as remote areas or industrial settings.

Implementing on-device AI for BCI systems faces various challenges. Typically, microcontrollers and edge devices are constrained by limited memory, processing capabilities, and energy availability. As a result, hardware restrictions are a primary limitation, making it more challenging to deploy sophisticated AI models due to the need for extensive processing with high-dimensional neural data. To address this issue, approaches such as model optimization through quantization [76], pruning [77], and distillation [78] are used to diminish the size and computational demands of AI models without sacrificing their accuracy. Another challenge is ensuring real-time processing, as BCI applications require consistent and rapid inference to provide meaningful feedback to users. Additionally, adaptability is a concern, as neural signals can vary significantly between individuals [79]. Developing AI models that can generalize across users while also personalizing to individual neural patterns remains an ongoing area of research.

Recent advancements in hardware and software have paved the way for on-device AI in BCI systems. One key innovation is the development of AI-enabled chips, such as neuromorphic processors [80], [81] (e.g., Intel's Loihi), Tensor Processing Units (TPUs) by Google [82], and Neural Processing Units (NPUs) [83] found in Qualcomm's Snapdragon platforms. These specialized hardware platforms are designed to perform AI inference efficiently, even on resource-constrained devices, enabling real-time processing of neural signals. Another breakthrough is the emergence of Tiny Machine Learning (TinyML), a field focused on deploying machine learning models on ultra-low-power microcontrollers like the Arm Cortex-M series or Espressif's ESP32. Frameworks such as TensorFlow Lite for Microcontrollers and Edge Impulse have become popular for developing TinyML applications, allowing BCI systems to leverage AI capabilities without compromising battery life or device portability. Additionally, federated learning offers a promising approach to training AI models across multiple devices without sharing raw data [84]. This decentralized method, supported by frameworks like PySyft [85] and TensorFlow Federated [86], enhances privacy while improving model performance, making it particularly suitable for BCI applications where data sensitivity is a concern. These technologies collectively represent significant achievements in enabling efficient, secure, and scalable on-device AI for real-time BCI processing.

V. Recent Works On-Device AI for BCI

This section highlights recent advancements in on-device AI for BCI, focusing on its transformative applications and benefits across various domains, such as driver drowsiness detection (DDD), motor movement classification, and robotic control.

One of the most promising applications of on-device AI in BCI is driver drowsiness detection. For instance, Nguyen et al. [87] proposed a novel behind-the-ear (BTE) EEG-based DDD system that leverages TinyML for on-device processing. The system utilizes a wearable headband device equipped with dry sensors to collect EEG signals from four BTE locations. These signals are preprocessed on-device, and Welch's method is applied to extract the relative power spectral density ratio of theta, alpha, and beta EEG bands. Two neural network approaches, namely a multilayer perceptron (MLP) and a CNN, were developed and evaluated by the authors. These were compared with a support vector machine (SVM) for detecting drowsiness. The CNN model, which was quantized and implemented on an nRF52840 SoC (64 MHz MCU) utilizing TensorFlow Lite for Microcontrollers (TFLM), showed a remarkable accuracy of 96.11% with a latency of merely 25.44 ms. This system exemplifies the potential of on-device AI for low-power, real-time, and privacy-responsible drowsiness detection, providing a viable approach for enhancing road safety in everyday scenarios. In addition, Kartsch et al. [88] introduced an energy-efficient wearable platform for drowsiness detection that prioritizes low power consumption and minimal latency. The system integrates dry EEG sensors alongside IMU sensors to capture both neural and behavioral signals. The processing is performed using Mr. Wolf, an 8-core ultra-low-power digital platform, running a Nearest Centroid Classifier (NCC) trained with a semi-supervised algorithm. While the system achieved a slightly lower accuracy of 83%, it offers a key advantage in real-time detection with minimal latency, making it highly suitable for real-world deployment where immediate responsiveness is critical. Furthermore, its design emphasizes wearability and extended battery life, addressing common limitations of EEG-based drowsiness detection systems.

Several studies have leveraged the GAP9 microcontroller and the DORY deployment framework to enable real-time, energy-efficient on-device AI for diverse BCI tasks. Mei et al. [89] demonstrated high-accuracy motor imagery classification (90.88%) using a lightweight CNN (MI-BMInet), achieving only 21.5 ms latency and 0.45 mJ per inference. Frey et al. [90] introduced GAPSes, a wearable smart glasses platform that achieved 99.74% accuracy in EEG-based biometric recognition using just 8 channels, with an ultra-low energy footprint of 121 μJ per inference. Extending the capabilities of this ecosystem, Ingolfsson et al. [91] tackled the complex task of imagined speech decoding. Their system used a low-channel dry EEG headset and a compact neural model (VOWELNET) to classify 13 speech imagery classes (vowels, commands, and rest), achieving 42.8% average accuracy with 40.9 ms latency while consuming only 25.93 mW. These studies showcase the flexibility and practicality of GAP9-based platforms for supporting a wide spectrum of cognitive and motor BCI applications in real-world settings.

Outside the GAP9 ecosystem, other platforms have also demonstrated compelling applications of on-device AI for BCI. For instance, Chepyk et al. [92] proposed a novel system that uses four types of electrooculography (EOG) signals—left and right winks, voluntary blinks, and involuntary blinks—to control a remote robotic platform. The system differentiates between voluntary and involuntary blinks to avoid unintended commands, ensuring robust and accurate control. A tinyML algorithm is employed to analyze and interpret the EOG signals in real-time, making the system suitable for resource-constrained environments. The proposed solution includes an event detection algorithm to select signal segments and a 1D CNN for classification. The entire system is embedded on a custom-made board featuring an STM32L476RG MCU, which handles all processing without the need for external devices. This setup achieves an impressive 99.3% average classification accuracy for the four classes of EOG signals, enabling precise three-degrees-of-freedom control of a robotic platform. The system has been tested by multiple users, who reported high accuracy and ease of use when controlling a three-wheeled robot.

These advancements underscore the transformative potential of on-device AI in BCI systems, paving the way for innovative applications that improve safety, accessibility, and human-machine interaction. Table III summarizes the key contributions from recent studies, including system components, frameworks, models, accuracy, latency, and BCI tasks.

VI. Real-World Applications of Edge AI-Powered BCIs

The deployment of edge AI-powered BCI systems opens new possibilities across various fields by enabling real-time processing, reducing reliance on external infrastructure, and enhancing privacy. These advantages are particularly critical in applications where latency, data security, and operational independence are paramount. Below, we explore how these benefits manifest in human-robot, medical, consumer, and industrial applications.

A. Human-Robot Teaming

Edge AI-powered BCI systems revolutionize human-robot teaming by enabling seamless collaboration between humans and robots through direct neural control. This eliminates the need for traditional interfaces like joysticks or keyboards. This collaborative effort merges the intuitive sense, adaptability, and decision-making capabilities of humans with the precision, strength, and stamina of robots. The principal benefits of edge AI, namely real-time processing, diminished dependency on external infrastructure, and enhanced privacy, are crucial in scenarios requiring swift responses, operational independence, and data protection, rendering these systems revolutionary across various sectors. In the manufacturing sector, employees using BCIs can manage robotic arms for activities like precise assembly and material handling. Edge AI facilitates rapid interpretation of neural signals, enabling swift and precise robotic actions that enhance efficiency and reduce mistakes [93]. The independence from external networks enables uninterrupted operation in large factories or areas with limited connectivity, enhancing scalability and reliability. Additionally, on-device processing protects proprietary data, such as worker performance metrics, from external breaches. In space exploration, BCIs enable astronauts to control robots for tasks such as sample collection or maintenance, particularly in long-duration missions where communication delays with Earth can be significant. For example, research emphasizes the potential of BCIs to reduce astronauts' mental burden, thereby aiding space exploration missions and improving overall health [94]. In search and rescue, responders deploy robots into hazardous environments, such as collapsed buildings or disaster zones, to locate survivors or assess damage. Edge AI-powered BCIs enable real-time navigation and decision-making, crucial for time-sensitive missions [95]. These applications underscore how edge BCI devices optimize human-robot teaming by providing swift, protected, and self-sufficient functionalities, significantly boosting performance and coordination across space exploration, manufacturing, and search and rescue.

B. Emotion-Aware BCI Applications

Affective computing focuses on the development and interaction with systems that can recognize, interpret, respond to, and even influence human emotions [96]. As a core component of affective computing, emotion recognition plays a vital role in enabling machines to understand emotional states, which, despite its complexity, provides significant value across various real-world domains, including healthcare, education, and security [96], [97], [98], [99], [100].

Emotion recognition has advanced considerably with the emergence of wearable BCI technologies, which provide a practical and non-invasive means of capturing neural responses in everyday settings [100], [101]. Recent research increasingly emphasizes integrating these systems into real-time, portable platforms to improve usability and responsiveness.

Several studies have proposed algorithmic innovations to enhance EEG-based emotion recognition. For instance, Li et al. proposed a model that combines attention mechanisms with Bidirectional Long Short-Term Memory (BiLSTM) networks to extract temporal dynamics from EEG signals, aligned with Russell's circumplex model of affect [102]. Their system achieved promising binary classification results, with accuracies of 0.833 for arousal and 0.794 for valence. Haipeng et al. later introduced a hybrid 1DCNN-BiLSTM model that improved performance further, achieving 0.916 for arousal and 0.915 for valence [103]. Similarly, Shadi et al. developed the SS4-STANN architecture, reporting accuracies of 0.830 for arousal and 0.827 for valence [104].

On-device and system-level implementations are gaining increasing attention due to their relevance to real-world deployment. For example, Mai et al. developed a real-time, on-chip machine learning-based wearable EEG system positioned behind the ear for continuous emotion monitoring, enabling high performance within a compact, low-power embedded platform [105]. Luo et al. presented a portable and affordable four-channel EEG system that leverages self-supervised learning for efficient emotion recognition with minimal labeled data, demonstrating feasibility for edge computing environments [106]. Similarly, Li et al. proposed a real-time wireless emotion-aware system based on a Body Area Network (BAN), optimized for low-latency processing in IoMT applications [107]. In another hardware-oriented approach, Ezilarasan and Leung implemented emotion classification on an FPGA, highlighting the potential of reconfigurable platforms for power-efficient EEG signal analysis in embedded systems [108].

Other advanced methods have also pushed algorithmic boundaries. Garcia-Moreno et al. applied Gradient Boosting (GB) and showed that wearable BCIs can achieve performance comparable to traditional wet-electrode systems [98]. Wang et al. employed an Extra Trees Classifier for emotion polarity classification with an accuracy of 0.883 [109].

Mai et al. also introduced the EEER framework using Vision Transformers (ViT) with Spatial-Temporal Processing (STP) and Locality Self-Attention (LSA), achieving an accuracy of 0.9239 [110].

Crucially, recent studies emphasize that generic deep learning models from domains such as computer vision or speech may not capture the neurophysiological dynamics inherent in EEG signals. To address this limitation, brain-inspired model architectures, such as graph neural networks (GNNs), are being explored. For example, Li et al. introduced BF-GCN, a cognition-inspired graph learning framework that integrates both data-driven features and functionally informed brain network priors. This hybrid model models the brain's cognitive pathways and spatial connectivity, achieving robust and interpretable EEG-based emotion recognition [111]. The incorporation of such neurophysiological priors marks a significant step toward models that are not only accurate but also aligned with the underlying structure of emotional processing in the brain.

Together, these advances reflect a broader shift in Emotion-Aware BCI applications, moving from algorithm-centric studies toward real-world, on-device, and neurophysiologically informed systems. This progression supports the vision of practical, user-centric, and context-aware affective computing platforms.

C. Medical Application

Edge BCI devices facilitate real-time neurorehabilitation and assistive communication without depending on external systems. This independence is critical for patients with motor impairments or neurological conditions, especially in environments with unstable internet connectivity. By leveraging on-device AI processing, these systems ensure low-latency responses, data security, and continuous functionality, making them indispensable in modern healthcare. In assistive technologies for disabilities, BCIs allow individuals with paralysis, amyotrophic lateral sclerosis, or spinal cord injuries to control prosthetic limbs, wheelchairs, or communication devices using brain signals [112], [113], [114]. Edge AI ensures low-latency interpretation of neural data, enabling seamless interaction, while on-device processing guarantees uninterrupted functionality in areas with poor connectivity and safeguards sensitive neural data from breaches [115]. For neurorehabilitation, BCIs monitor brain activity during stroke recovery, providing real-time feedback to therapists and patients [116]. Edge AI eliminates the need for cloud-based processing, ensuring continuous operation and compliance with healthcare privacy regulations [73]. In epilepsy and seizure detection, wearable BCIs with edge AI detect and predict seizures in real-time, offering immediate alerts even in remote settings [117]. By processing data locally, these systems reduce reliance on external infrastructure and protect sensitive health information [118]. Together, these applications highlight how edge AI-powered BCIs are transforming healthcare by delivering real-time, secure, and independent solutions for patients and practitioners alike.

D. Consumer Electronics

Wearable edge BCI systems enable seamless interaction with augmented/virtual environments, allowing users to control applications through brain signals without requiring continuous internet access or external computing power. This on-device processing capability improves user mobility and responsiveness, particularly in gaming and augmented reality (AR) contexts. In gaming and entertainment, edge AI ensures low-latency interpretation of brain signals, enabling instantaneous control and immersive experiences, while eliminating the need for cloud-based processing [119]. For wearable devices, such as smart headbands and EEG-based wearables, edge AI monitors focus, stress, and sleep patterns in real-time, providing users with actionable insights without relying on external servers [120]. This offline functionality ensures accessibility in areas with limited connectivity, while localized processing protects sensitive user data. In personalized learning and training, BCIs adapt educational content based on the user's cognitive state, with edge AI delivering real-time feedback and maintaining data security through on-device computation [121]. Additionally, in brain-controlled smart home devices, edge AI enables users to control appliances like lights and thermostats through brain signals, reducing reliance on external infrastructure and ensuring privacy [122]. These applications demonstrate how edge AI-powered BCIs are enhancing user experiences by delivering real-time, secure, and independent solutions in consumer electronics.

E. Industrial Use

Edge BCIs improve worker monitoring in high-risk environments, enhancing focus assessment and reducing accidents. The elimination of external processing requirements minimizes latency, enabling more accurate real-time interventions in industrial applications. In worker safety and fatigue monitoring, edge AI analyzes cognitive states to detect fatigue or stress, providing instant alerts to prevent accidents [123]. By processing data on-device, these systems ensure continuous operation in remote or hazardous environments with unstable connectivity while safeguarding sensitive biometric data. For human-machine collaboration, BCIs allow workers to control machinery or robots using brain signals, with edge AI enabling low-latency responses for precise operations. The elimination of external processing requirements ensures reliability and reduces dependency on cloud infrastructure [124]. In training and skill development, edge AI-powered BCIs provide real-time feedback during employee training by analyzing cognitive engagement, and accelerating skill acquisition while maintaining data confidentiality [125]. Finally, in remote operation of equipment, BCIs enable operators to control machinery in hazardous environments (e.g., mining, oil rigs) using brain signals. Edge AI ensures real-time control and minimizes latency, enhancing operational safety and efficiency without relying on external networks [126]. These applications highlight how edge AI-powered BCIs are improving productivity, safety, and efficiency in industrial settings through real-time, secure, and independent solutions.

VII. Conclusion and Future Outlook

This paper has provided a comprehensive review of AI-powered BCI systems, covering key components such as hardware design, data acquisition, preprocessing techniques, and deep learning models optimized for real-time, on-device applications. A critical factor in BCI performance is the acquisition of high-fidelity EEG signals, which typically requires an ADC resolution of at least 16 bits and a minimum sampling rate of 250 Hz. Preprocessing techniques, including noise reduction and feature extraction, remain essential for improving signal quality and ensuring accurate classification. Recent advancements in AI, edge computing, and specialized hardware are enabling next-generation BCI systems that are more portable, energy-efficient, and responsive.

From an embedded systems perspective, the challenge is not simply to "fit" a model on a device, but to design a closed-loop pipeline where every stage, from electrode–skin contact and analog front-end configuration to feature extraction and inference, is co-optimized for latency, accuracy, and power. This calls for a shift from maximizing accuracy in unconstrained environments to optimizing accuracy under the realities of mobile, resource-limited hardware.

Future research should prioritize several directions. First, optimizing AI models for embedded systems through quantization, pruning, architectural simplification, and EEG-specific accelerators will enable efficient deployment in resource-constrained environments. Second, improving real-time signal processing and data streaming pipelines will enhance responsiveness and reliability, particularly in wearable or mobile use cases. Third, personalization through continual on-device learning, transfer learning, and adaptive interfaces can address inter-subject and session-to-session variability while preserving user privacy by avoiding cloud-based retraining.

Looking ahead, the next breakthroughs in AI-on-device BCIs will likely be driven by three converging trends:

1) EEG-aware hardware accelerators: Custom low-power NPUs and neuromorphic processors tuned for the temporal–spectral characteristics of EEG could run compact convolutional or transformer-based models at milliwatt-level power. Research prototypes already demonstrate FPGA-based CNN inference accelerators for real-time EEG classification [127], hybrid CNN–LSTM EEG decoders for drowsiness detection [128], and ASIC-like EEGNet processors optimized for low-power deployment [129]. These platforms integrate both signal conditioning and inference on a single chip, offering notable gains in latency and energy efficiency.
2) Continual, privacy-preserving learning: Lightweight adaptation methods such as federated learning, few-shot transfer, or session-specific calibration layers can keep BCIs accurate over time without transmitting sensitive neural data to external servers. These approaches could transform BCIs from static classifiers into adaptive partners that grow with the user.
3) Sensor fusion at the edge: Combining EEG with IMU, EMG, or eye-tracking sensors directly on-device can significantly improve robustness in uncontrolled environments. Embedded-friendly fusion techniques, from early-layer multimodal networks to low-latency filtering, are emerging as practical solutions for real-time, context-aware BCIs.

In the longer term, Edge AI-powered BCIs may evolve from research prototypes into ubiquitous, wearable companions that adapt in real time, operate seamlessly in the background, and provide intuitive control over digital systems without frequent recalibration. Achieving this vision will require sustained collaboration between neuroscience, embedded systems engineering, AI research, and human–computer interaction. By embracing hardware–software co-design, efficient modeling, and user-centered adaptability, the field can move toward BCIs that are smaller, smarter, always available, and deeply integrated into everyday life.

Appendix
ADC Quantization Derivation

To understand why higher resolution leads to better accuracy, the size of one code (LSB) in ADC can be calculated as shown in Eq. 3.

$$\mathrm{LSB} = \frac{V_{\mathrm{ref}}}{2^N} \tag{3}$$

where V_{ref} is the reference voltage of the ADC, and N is the ADC resolution. The quantized output voltage ($V_{\mathrm{quantized}}$) captured by ADC can be expressed as Eq. 4.

$$V_{\mathrm{quantized}} = \mathrm{round}\left(\frac{V_0}{\mathrm{LSB}}\right) \times \mathrm{LSB} \tag{4}$$

Here, V_0 represents the real input voltage. Hence, the quantization error E_q can be defined in Eq. 5:

$$E_q = V_0 - V_{\mathrm{quantized}} \tag{5}$$

Since the quantization error is limited to a range of $\pm\frac{\mathrm{LSB}}{2}$, therefore the boundary of E_q can be defined as Eq. 6:

$$0 \leq |E_q| \leq \frac{\mathrm{LSB}}{2} \tag{6}$$

Substituting the LSB from Eq. 3 to Eq. 6, the relationship between quantization error bound and ADC resolution is obtained in Eq. 7.

$$0 \leq |E_q| \leq \frac{V_{\mathrm{ref}}}{2^{N+1}} \tag{7}$$

Taking the limit as $N \to \infty$:

$$0 \leq \lim_{N\to\infty} |E_q| \leq \lim_{N\to\infty} \frac{V_{\mathrm{ref}}}{2^{N+1}} = 0 \tag{8}$$

As the ADC resolution (N) increases, the quantization error decreases, improving the accuracy of EEG signal recording (Eq. 8). In the theoretical limit where $N \to \infty$, the quantization error approaches zero, allowing for near-perfect signal reconstruction.

References

[1] M. S. Willsey et al., "A high-performance brain–computer interface for finger decoding and quadcopter game control in an individual with paralysis," *Nature Med.*, vol. 31, no. 1, pp. 1–9, 2025.

[2] J. D. R. Millán, "Combining brain–computer interfaces and assistive technologies: State-of-the-art and challenges," *Frontiers Neurosci.*, vol. 1, Sep. 2010, Art. no. 161.

[3] F. Cincotti et al., "Non-invasive brain–computer interface system: Towards its application as assistive technology," *Brain Res. Bull.*, vol. 75, no. 6, pp. 796–803, Apr. 2008.

[4] B. Zeng et al., "Controllable mind visual diffusion model," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 6935–6943.

[5] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14453–14463.

[6] Y. Song, B. Liu, X. Li, N. Shi, Y. Wang, and X. Gao, "Decoding natural images from EEG for object recognition," in *Proc. 12th Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2023, pp. 15475–15492.

[7] J. Lévy et al., "Brain-to-text decoding: A non-invasive approach via typing," 2025, *arXiv:2502.17480*.

[8] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[9] J. Zhou et al., "Pretraining large brain language model for active BCI: Silent speech," 2025, *arXiv:2504.21214*.

[10] M. Zhang et al., "From thought to action: How a hierarchy of neural dynamics supports language production," 2025, *arXiv:2502.07429*.

[11] L. L. Oganesian and M. M. Shanechi, "Brain–computer interfaces for neuropsychiatric disorders," *Nature Rev. Bioeng.*, vol. 2, no. 8, pp. 653–670, Jun. 2024.

[12] S. N. Flesher et al., "A brain–computer interface that evokes tactile sensations improves robotic arm control," *Science*, vol. 372, no. 6544, pp. 831–836, May 2021.

[13] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain–computer interfaces for communication and rehabilitation," *Nature Rev. Neurol.*, vol. 12, no. 9, pp. 513–525, 2016.

[14] T. N. Do, C.-H. Chuang, S.-J. Hsiao, C.-T. Lin, and Y.-K. Wang, "Neural comodulation of independent brain processes related to multitasking," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1160–1169, Jun. 2019.

[15] T.-T. N. Do, Y.-K. Wang, and C.-T. Lin, "Increase in brain effective connectivity in multitasking but not in a high-fatigue state," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 3, pp. 566–574, Sep. 2021.

[16] T.-T. N. Do, C.-T. Lin, and K. Gramann, "Human brain dynamics in active spatial navigation," *Sci. Rep.*, vol. 11, no. 1, p. 13036, Jun. 2021.

[17] T.-T. N. Do, T.-P. Jung, and C.-T. Lin, "Retrosplenial segregation reflects the navigation load during ambulatory movement," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 488–496, 2021.

[18] K. M. Patrick-Krueger, I. Burkhart, and J. L. Contreras-Vidal, "The state of clinical trials of implantable brain–computer interfaces," *Nature Rev. Bioengineering*, pp. 50–67, 2024.

[19] C.-T. Lin and T.-T.-N. Do, "Direct-sense brain–computer interfaces and wearable computers," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 298–312, Jan. 2021.

[20] X. Zhang et al., "The combination of brain–computer interfaces and artificial intelligence: Applications and challenges," *Ann. Transl. Med.*, vol. 8, no. 11, p. 712, Jun. 2020.

[21] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain–computer interfaces," *Trends Cognit. Sci.*, vol. 25, no. 8, pp. 671–684, Aug. 2021.

[22] B. J. Edelman et al., "Non-invasive brain–computer interfaces: State of the art and trends," *IEEE Rev. Biomed. Eng.*, vol. 18, pp. 26–49, 2025.

[23] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Res. Rev.*, vol. 29, nos. 2–3, pp. 169–195, Apr. 1999.

[24] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw, "Mu and beta rhythm topographies during motor imagery and actual movements," *Brain Topography*, vol. 12, no. 3, pp. 177–186, Mar. 2000.

[25] L.-D. Liao, I.-J. Wang, S.-F. Chen, J.-Y. Chang, and C.-T. Lin, "Design, fabrication and experimental validation of a novel dry-contact sensor for measuring electroencephalography signals without skin preparation," *Sensors*, vol. 11, no. 6, pp. 5819–5834, May 2011.

[26] T. J. Sullivan, S. Deiss, G. Cauwenberghs, and T. Jung, "A low-noise low-power EEG acquisition node for scalable brain-machine interfaces," *Proc. SPIE*, vol. 6592, pp. 42–49, May 2007.

[27] T. Moy et al., "An EEG acquisition and biomarker-extraction system using low-noise-amplifier and compressive-sensing circuits based on flexible, thin-film electronics," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 309–321, Jan. 2017.

[28] G. Huang, T. Yin, H. Yang, and X. Cai, "A 0.8 VRMS 8-channel front-end for EEG recording," *Analog Integr. Circuits Signal Process.*, vol. 99, no. 2, pp. 427–436, May 2019.

[29] A. Ortiz and J. Mínguez, "Main features of EEG amplifiers," Bitbrain, Zaragoza, Spain, Tech. Rep., 2024.

[30] A. C. MettingVanRijn, A. Peper, and C. A. Grimbergen, "Amplifiers for bioelectric events: A design with a minimal number of parts," *Med. Biol. Eng. Comput.*, vol. 32, no. 3, pp. 305–310, May 1994.

[31] R. R. Harrison and C. Charles, "A low-power low-noise CMOS for amplifier neural recording applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

[32] Y. M. Chi, T.-P. Jung, and G. Cauwenberghs, "Dry-contact and non-contact biopotential electrodes: Methodological review," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 106–119, 2010.

[33] G. Niso, E. Romero, J. T. Moreau, A. Araujo, and L. R. Krol, "Wireless EEG: A survey of systems and studies," *NeuroImage*, vol. 269, Apr. 2023, Art. no. 119774.

[34] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP J. Adv. Signal Process.*, vol. 2015, Aug. 2015, Art. no. 66.

[35] M. Seeck et al., "The standardized EEG electrode array of the IFCN," *Clin. Neurophysiol.*, vol. 128, no. 10, pp. 2070–2077, Oct. 2017.

[36] A. Soler, P. A. Muñoz-Gutiérrez, M. Bueno-López, E. Giraldo, and M. Molinas, "Low-density EEG for neural activity reconstruction using multivariate empirical mode decomposition," *Frontiers Neurosci.*, vol. 14, Feb. 2020, Art. no. 175.

[37] G. Lantz, R. Grave de Peralta, L. Spinelli, M. Seeck, and C. M. Michel, "Epileptic source localization with high density EEG: How many electrodes are needed?," *Clin. Neurophysiol.*, vol. 114, no. 1, pp. 63–69, Jan. 2003.

[38] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. G. De Peralta, "EEG source imaging," *Clin. Neurophysiol.*, vol. 115, no. 10, pp. 2195–2222.

[39] C. M. Michel, G. Lantz, L. Spinelli, R. G. de Peralta, T. Landis, and M. Seeck, "128-channel EEG source imaging in epilepsy: Clinical yield and localization precision," *J. Clin. Neurophysiology*, vol. 21, no. 2, pp. 71–83, Mar. 2004.

[40] C. A. T. Cortes, C.-T. Lin, T. N. Do, and H.-T. Chen, "An EEG-based experiment on VR sickness and postural instability while walking in virtual environments," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2023, pp. 94–104.

[41] C.-T. Lin et al., "Effects of multisensory distractor interference on attentional driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10395–10403, Aug. 2022.

[42] S. Hu, Y. Lai, P. A. Valdes-Sosa, M. L. Bringas-Vega, and D. Yao, "How do reference montage and electrodes setup affect the measured scalp EEG potentials?," *J. Neural Eng.*, vol. 15, no. 2, Apr. 2018, Art. no. 026013.

[43] R. Vallat, "Bandpower of an EEG signal," Ph.D. dissertation, Personal Academic Website, Univ. California, Berkeley, Berkeley, CA, USA,May 2018. Accessed: Oct. 9, 2025. [Online]. Available: https://raphaelvallat.com/bandpower.html

[44] S. R. Sinha et al., "American clinical neurophysiology society guideline 1: Minimum technical requirements for performing clinical electroencephalography," *J. Clin. Neurophysiology*, vol. 33, no. 4, pp. 303–307, 2016.

[45] T. C. Ferree, P. Luu, G. S. Russell, and D. M. Tucker, "Scalp electrode impedance, infection risk, and EEG data quality," *Clin. Neurophysiol.*, vol. 112, no. 3, pp. 536–544, Mar. 2001.

[46] R. B. Damalerio, R. Lim, Y. Gao, T.-T. Zhang, and M.-Y. Cheng, "Development of low-contact-impedance dry electrodes for electroencephalogram signal acquisition," *Sensors*, vol. 23, no. 9, p. 4453, May 2023.

[47] Q. Liu, L. Yang, Z. Zhang, H. Yang, Y. Zhang, and J. Wu, "The feature, performance, and prospect of advanced electrodes for electroencephalogram," *Biosensors*, vol. 13, no. 1, p. 101, Jan. 2023.

[48] J. Górecka and P. Makiewicz, "The dependence of electrode impedance on the number of performed EEG examinations," *Sensors*, vol. 19, no. 11, p. 2608, Jun. 2019.

[49] *ADS1299-X Low-Noise, 4-, 6-, 8-Channel, 24-Bit, Analog-to-Digital Converter for EEG and Biopotential Measurements*, Texas Instruments, Dallas, TX, USA, 2017.

[50] E. Habibzadeh Tonekabony Shad, M. Molinas, and T. Ytterdal, "Impedance and noise of passive and active dry EEG electrodes: A review," *IEEE Sensors J.*, vol. 20, no. 24, pp. 14565–14577, Dec. 2020.

[51] A. C. MettingVanRijn, A. P. Kuiper, T. E. Dankers, and C. A. Grimbergen, "Low-cost active electrode improves the resolution in biopotential recordings," in *Proc. 18th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Oct./Nov. 1996, pp. 101–102.

[52] S. Laszlo, M. Ruiz-Blondet, N. Khalifian, F. Chu, and Z. Jin, "A direct comparison of active and passive amplification electrodes in the same amplifier system," *J. Neurosci. Methods*, vol. 235, pp. 298–307, Sep. 2014.

[53] L. A. W. Gemein et al., "Machine-learning-based diagnostics of EEG pathology," *NeuroImage*, vol. 220, Oct. 2020, Art. no. 117021.

[54] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.

[55] Y. Zhang et al., "Improving EEG decoding via clustering-based multitask feature learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3587–3597, Aug. 2022.

[56] J. A. Urigüen and B. García-Zapirain, "EEG artifact removal—State-of-the-art and guidelines," *J. Neural Eng.*, vol. 12, no. 3, 2015, Art. no. 031001.

[57] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 4–19, Mar. 2022.

[58] M. Xia, Y. Zhang, Y. Wu, and X. Wang, "An end-to-end deep learning model for EEG-based major depressive disorder classification," *IEEE Access*, vol. 11, pp. 41337–41347, 2023.

[59] E. Lashgari, J. Ott, A. Connelly, P. Baldi, and U. Maoz, "An end-to-end CNN with attentional mechanism applied to raw EEG in a BCI classification task," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 0460e3.

[60] *GitHub-Espressif/ESP-DL: Espressif Deep-Learning Library for AIoT Applications*, Espressif Systems, Shanghai, China, 2021.

[61] *STM32Cube.AI-STMicroElectronics-STM32 AI*, STMicroelectronics, Geneva, Switzerland, 2019.

[62] *GitHub-Tensorflow/Tflite-Micro: Infrastructure to Enable Deployment of ML Models to Low-Power Resource-Constrained Embedded Targets*, Google LLC, Tensorflow, Mountain View, CA, USA, 2021.

[63] *EIQT Inference With TF Lite Micro*, NXP Semiconductors, Eindhoven, The Netherlands, 2021.

[64] *TI Deep Learning Library User Guide: TIDL-TI Deep Learning Library*, Texas Instruments Incorporated, Dallas, TX, USA, 2021.

[65] *GitHub-Espressif/ESP-Tflite-Micro: TensorFlow Lite Micro for Espressif Chipsets*, Espressif Systems (Shanghai) Co., Ltd., Shanghai, China, 2022.

[66] *Edge Impulse—The Leading Edge AI Platform*, Edge Impulse, Inc., San Jose, CA, USA, 2019.

[67] *GitHub-ARM-Software/CMSIS-NN: CMSIS-NN Library*, Arm Limited, Cambridge, U.K., 2017.

[68] *Check Out ESP-DL!*, Espressif Systems (Shanghai) Co., Ltd., Shanghai, China, 2021.

[69] *Quantized Model Support—ST Edge AI Core Technology*, STMicroelectronics, Geneva, Switzerland, 2025.

[70] *Model testing—Edge Impulse Documentation*, Edge Impulse, Inc., San Jose, CA, USA, 2025.

[71] *Model Testing — Edge Impulse Documentation*, Edge Impulse.

[72] S. S. Saha, S. S. Sandha, and M. Srivastava, "Machine learning for microcontroller-class hardware: A review," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21362–21390, Nov. 2022.

[73] K. Xia et al., "Privacy-preserving brain–computer interfaces: A systematic review," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 5, pp. 2312–2324, May 2022.

[74] S. Bian et al., "On-device learning of EEGNet-based network for wearable motor imagery brain–computer interface," in *Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, Oct. 2024, pp. 9–16.

[75] T. Schneider, X. Wang, M. Hersche, L. Cavigelli, and L. Benini, "Q-EEGNet: An energy-efficient 8-bit quantized parallel EEGNet implementation for edge motor-imagery brain-machine interfaces," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Sep. 2020, pp. 284–289.

[76] P.-E. Novac, G. Boukli Hacene, A. Pegatoquet, B. Miramond, and V. Gripon, "Quantization and deployment of deep neural networks on microcontrollers," *Sensors*, vol. 21, no. 9, p. 2984, Apr. 2021.

[77] Y. Jiang et al., "Model pruning enables efficient federated learning on edge devices," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10374–10386, Dec. 2023.

[78] C. Sun, Q. Tong, W. Yang, and W. Zhang, "DiReDi: Distillation and reverse distillation for AIoT applications," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 748–760, 2024.

[79] D. P. Subha, P. K. Joseph, U. R. Acharya, and C. M. Lim, "EEG signal analysis: A survey," *J. Med. Syst.*, vol. 34, no. 2, pp. 195–212, 2008.

[80] Y. Sandamirskaya, M. Kaboli, J. Conradt, and T. Celikel, "Neuromorphic computing hardware and neural architectures for robotics," *Sci. Robot.*, vol. 7, no. 67, Jun. 2022, Art. no. eabl8419.

[81] M. Davies et al., "Advancing neuromorphic computing with loihi: A survey of results and outlook," *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.

[82] *Introduction to Cloud TPU*, Google LLC, Mountain View, CA, USA, 2025.

[83] *What is an NPU? And Why is It Key to Unlocking On-Device Generative AI?*, Qualcomm Technologies, Inc., San Diego, CA, USA, 2024.

[84] C. Herath, X. Liu, S. Lambotharan, and Y. Rahulamathavan, "Enhancing federated learning convergence with dynamic data queue and data-entropy-driven participant selection," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 6646–6658, Mar. 2025.

[85] *GitHub-OpenMined/PySyft: Perform Data Science on Data That Remains in Someone Else's Server*, OpenMined, San Francisco, CA, USA, 2025.

[86] *TensorFlow Federated*, Google LLC, Mountain View, CA, USA, 2025.

[87] H.-T. Nguyen, N.-D. Mai, B. G. Lee, and W.-Y. Chung, "Behind-the-ear EEG-based wearable driver drowsiness detection system using embedded tiny neural networks," *IEEE Sensors J.*, vol. 23, no. 19, pp. 23875–23892, Oct. 2023.

[88] V. Kartsch, S. Benatti, M. Guermandi, F. Montagna, and L. Benini, "Ultra low-power drowsiness detection system with BioWolf," in *Proc. 9th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Mar. 2019, pp. 1187–1190.

[89] L. Mei et al., "An ultra-low power wearable BMI system with continual learning capabilities," *IEEE Trans. Biomed. Circuits Syst.*, vol. 19, no. 3, pp. 1–12, Jun. 2025.

[90] S. Frey et al., "GAPses: Versatile smart glasses for comfortable and fully-dry acquisition and parallel ultra-low-power processing of EEG and EOG," *IEEE Trans. Biomed. Circuits Syst.*, vol. 19, no. 3, pp. 1–11, Jun. 2025.

[91] T. M. Ingolfsson, V. Kartsch, L. Benini, and A. Cossettini, "A wearable ultra-low-power system for EEG-based speech-imagery interfaces," *IEEE Trans. Biomed. Circuits Syst.*, vol. 19, no. 4, pp. 743–755, Aug. 2025.

[92] O. Chepyk et al., "An embedded EOG-based brain computer interface system for robotic control," in *Proc. 8th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Jun. 2023, pp. 1–6.

[93] Z. Bi, A. Mikkola, W. H. Ip, K. L. Yung, and C. Luo, "Brain computer interface (BCI) for shared controls of unmanned aerial vehicles (UAVs)," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 4, pp. 3860–3871, Apr. 2024.

[94] C. De Negueruela, M. Broschart, C. Menon, and J. D. R. Millán, "Brain–computer interfaces for space applications," *Pers. Ubiquitous Comput.*, vol. 15, no. 5, pp. 527–537.

[95] S. Liu et al., "Remote-oriented brain-controlled unmanned aerial vehicle for IoT," *IEEE Internet Things J.*, vol. 11, no. 17, pp. 29202–29214, Sep. 2024.

[96] A. Bayro and H. Jeong, "A systematic review of experimental protocols: Towards a uniform framework in virtual reality affective research," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 1–16, Jul. 2025.

[97] J. Shen et al., "UA-DAAN: An uncertainty-aware dynamic adversarial adaptation network for EEG-based depression recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 3, pp. 2130–2141, Jul. 2025.

[98] F. M. Garcia-Moreno, M. Badenes-Sastre, F. Expósito, M. J. Rodriguez-Fortiz, and M. Bermudez-Edo, "EEG headbands vs caps: How many electrodes do i need to detect emotions? The case of the MUSE headband," *Comput. Biol. Med.*, vol. 184, Jan. 2025, Art. no. 109463.

[99] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102019.

[100] Y. S. Can and C. Ersoy, "Smart affect monitoring with wearables in the wild: An unobtrusive mood-aware emotion recognition system," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2851–2863, Oct. 2023.

[101] X. Li, H. Deng, J. Ouyang, H. Wan, W. Yu, and D. Wu, "Act as what you think: Towards personalized EEG interaction through attentional and embedded LSTM learning," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 3741–3753, May 2024.

[102] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102185.

[103] H. Liu et al., "EEG emotion recognition via a lightweight 1DCNN-BiLSTM model in resource-limited environments," *IEEE Sensors J.*, vol. 25, no. 3, pp. 5723–5730, Feb. 2025.

[104] S. Sartipi, M. Torkamani-Azar, and M. Cetin, "A hybrid end-to-end spatiotemporal attention neural network with graph-smooth signals for EEG emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 16, no. 2, pp. 732–743, Apr. 2024.

[105] N.-D. Mai, H.-T. Nguyen, and W.-Y. Chung, "Real-time on-chip machine-learning-based wearable behind-the-ear electroencephalogram device for emotion recognition," *IEEE Access*, vol. 11, pp. 47258–47271, 2023.

[106] H. Luo, H. Li, W. Tao, Y. Yang, C.-I. Ieong, and F. Wan, "A portable and affordable four-channel EEG system for emotion recognition with self-supervised feature learning," *Mathematics*, vol. 13, no. 10, p. 1608, May 2025.

[107] C. Li, Y. Mao, Q. Huang, W. Xie, X. He, and J. Wu, "A real-time emotion-aware system based on wireless body area network for IoMT applications," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 41182–41193, Dec. 2024.

[108] M. R. Ezilarasan and M.-F. Leung, "An efficient EEG signal analysis for emotion recognition using FPGA," *Information*, vol. 15, no. 6, p. 301, May 2024.

[109] Y. Wang et al., "Wearable wireless dual-channel EEG system for emotion recognition based on machine learning," *IEEE Sensors J.*, vol. 23, no. 18, pp. 21767–21775, Sep. 2023.

[110] N.-D. Mai, H.-T. Nguyen, and W.-Y. Chung, "Deep learning-based wearable ear-EEG emotion recognition system with superlets-based signal-to-image conversion framework," *IEEE Sensors J.*, vol. 24, no. 7, pp. 11946–11958, Apr. 2024.

[111] C. Li et al., "An efficient graph learning system for emotion recognition inspired by the cognitive prior graph of EEG brain network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 1–15, Apr. 2025.

[112] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: Communication and restoration of movement in paralysis," *J. Physiol.*, vol. 579, no. 3, pp. 621–636, 2007.

[113] Y. Jiang, K. Li, Y. Liang, D. Chen, M. Tan, and Y. Li, "Daily assistance for amyotrophic lateral sclerosis patients based on a wearable multimodal brain–computer interface mouse," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 33, pp. 150–161, 2025.

[114] R. Rupp, "Challenges in clinical applications of brain computer interfaces in individuals with spinal cord injury," *Frontiers Neuroengineering*, vol. 7, Sep. 2014, Art. no. 38.

[115] R. Rupp, S. C. Kleih, R. Leeb, J. del R. Millan, A. Kübler, and G. R. Müller-Putz, *Brain–Computer Interfaces and Assistive Technology*. Dordrecht, The Netherlands: Springer, 2014, pp. 7–38.

[116] R. Mane, T. Chouhan, and C. Guan, "BCI for stroke rehabilitation: Motor and beyond," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 041001.

[117] T. M. Ingolfsson et al., "BrainFuseNet: Enhancing wearable seizure detection through EEG-PPG-accelerometer sensor fusion and efficient edge deployment," *IEEE Trans. Biomed. Circuits Syst.*, vol. 18, no. 4, pp. 720–733, Aug. 2024.

[118] R. Zanetti, A. Arza, A. Aminifar, and D. Atienza, "Real-time EEG-based cognitive workload monitoring on wearable devices," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 1, pp. 265–277, Jan. 2022.

[119] H. Y. Zhu, N. Q. Hieu, D. T. Hoang, D. N. Nguyen, and C.-T. Lin, "A human-centric metaverse enabled by brain–computer interface: A survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 2120–2145, 3rd Quart., 2024.

[120] C. V. F. Pereira, E. M. de Oliveira, and A. D. de Souza, "Machine learning applied to edge computing and wearable devices for healthcare: Systematic mapping of the literature," *Sensors*, vol. 24, no. 19, p. 6322, Sep. 2024.

[121] N. Beauchemin et al., "Enhancing learning experiences: EEG-based passive BCI system adapts learning speed to cognitive load in real-time, with motivation as catalyst," *Frontiers Human Neurosci.*, vol. 18, Oct. 2024, Art. no. 1416683.

[122] N. Kosmyna, F. Tarpin-Bernard, N. Bonnefond, and B. Rivet, "Feasibility of BCI control in a realistic smart home environment," *Frontiers Human Neurosci.*, vol. 10, p. 416, Aug. 2016.

[123] P. Li, R. Meziane, M. J.-D. Otis, H. Ezzaidi, and P. Cardou, "A smart safety helmet using IMU and EEG sensors for worker fatigue detection," in *IEEE Int. Symp. Robotic Sensors Environments (ROSE) Proc.*, Oct. 2014, pp. 55–60.

[124] Y. An, J. Wong, and S. H. Ling, "Development of real-time brain–computer interface control system for robot," *Appl. Soft Comput.*, vol. 159, Jul. 2024, Art. no. 111648.

[125] L. Carelli et al., "Brain–computer interface for clinical purposes: Cognitive assessment and rehabilitation," *BioMed Res. Int.*, vol. 2017, pp. 1–11, 2017.

[126] J. Meng, S. Zhang, A. Bekyo, J. Olsoe, B. Baxter, and B. He, "Noninvasive electroencephalogram based control of a robotic arm for reach and grasp tasks," *Sci. Rep.*, vol. 6, no. 1, p. 38565, Dec. 2016.

[127] X. Yu, T. Majoros, and S. Oniga, "Hardware implementation of CNN based on FPGA for EEG signal patterns recognition," in *Proc. Int. Conf. E-Health Bioeng. (EHB)*, Nov. 2021, pp. 1–4.

[128] R. M. R. Yanamala and M. Pullakandam, "FPGA-accelerated hybrid CNN-LSTM system for efficient EEG-based drowsiness recognition," *J. Supercomput.*, vol. 81, Jan. 2025, Art. no. 453.

[129] J. Cao et al., "An optimized EEGNet processor for low-power and real-time EEG classification in wearable brain–computer interfaces," *Microelectron. J.*, vol. 145, Mar. 2024, Art. no. 106134.