electronics

MDPI

Article

Ultra-Low-Power Signal Processing and Edge-AI Hardware for Brain–Computer Interfaces: A Survey from Analog Front-Ends to On-Chip Inference

Tuan-Anh Tran [1,*], Thithu-Huyen Tran [2] and Cong-Kha Pham [1,*]

[1] Department of Computer and Network Engineering, The University of Electro-Communications (UEC), 1-5-1 Chofugaoka, Tokyo 182-8585, Japan
[2] The Institute of Biomedicine and Pharmacy (IBP), 222 Phung Hung Street, Ha Dong Ward, Hanoi 12123, Vietnam; tranthuhuyen.cnsh@gmail.com
* Correspondence: tuananh@vlsilab.ee.uec.ac.jp (T.-A.T.); phamck@uec.ac.jp (C.-K.P.)

Abstract

The evolution of Brain–Computer Interfaces (BCIs) toward unconstrained, daily life deployment requires a fundamental shift from cloud-tethered processing to low-power, wearable System-on-Chip (SoC) architectures. This survey presents a narrative, full-stack taxonomy of the acquisition-to-inference pipeline, examining the co-design space spanning neural signal acquisition, hardware-friendly feature extraction, and on-chip AI acceleration for edge BCIs. We first evaluate high-fidelity acquisition technologies, detailing low-noise Analog Front-Ends (AFEs) and non-invasive dry-electrode arrays optimized for signal integrity under motion artifacts. Next, we categorize lightweight neural network topologies—including compact Convolutional Neural Networks (CNNs), Spiking Neural Networks (SNNs), and Hyperdimensional Computing (HDC)—tailored for real-time decoding under limited memory constraints. To bridge algorithmic complexity with battery-constrained edge execution, we analyze state-of-the-art model compression methodologies and dedicated hardware acceleration paradigms across FPGA and ASIC platforms. Finally, we highlight the monolithic integration of mixed-signal analog conditioning and high-efficiency inference engines into specialized "EEG-aware" SoCs, offering a quantitative comparison of power efficiency (power per channel), decoding latency, and silicon area. We conclude by identifying hardware–algorithm co-design as the critical path toward autonomous, ultra-low-power neural prosthetics and closed-loop bio-interactive systems.

Keywords: brain–computer interface; analog front-end BCI; CNN; edge AI BCI; SoC; ASIC

Check for updates

Academic Editors: Marina Prvan, Josip Musić and Duje Čoko

Received: 2 September 2026
Revised: 20 October 2026
Accepted: 21 October 2026
Published: 23 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. BCIs in Daily Life and the Shift Toward Edge-AI Solutions

Brain–computer interfaces (BCIs) establish a communication and control pathway that bypasses the peripheral nervous system, translating electrophysiological activity directly into commands for external devices [1]. For most of their history, they have been laboratory instruments, developed primarily to restore communication and movement for people with severe neuromuscular impairment: P300 spellers for augmentative communication [2] and motor-imagery control of a hand orthosis for tetraplegic users [3] are representative of that era. Those systems paired multi-channel wet-electrode caps with desktop-class computing, typically under a general-purpose software platform such as BCI2000 [4]: an arrangement that yields a favorable signal-to-noise ratio but requires skin preparation, degrades as the gel dries, and is unsuited to unconstrained daily use [5].

Two developments are moving BCIs out of that setting. First, acquisition hardware has become wearable: miniaturized wireless multimodal front-ends [6], ear-worn decoding ICs [7], dry-electrode headsets [8] and complete sub-10-mW wearable platforms [9] now record outside the laboratory. The price is a harder signal-conditioning problem, since dry contacts raise the electrode impedance and lose statistical power relative to wet electrodes [10], and motion renders recordings unreliable in precisely the ambulatory conditions that a daily life BCI must survive [11]. Second, decoding itself has moved onto the device. Streaming raw biopotentials to a remote server imposes a telemetry and latency cost that closed-loop interaction cannot absorb, and it exposes neural data in transit; decoding is therefore increasingly performed next to the sensor, on dedicated accelerators [12] and reconfigurable biomedical processors [13].

These two trends meet at a single engineering problem. The decoder, the mixed-signal acquisition chain and the silicon that executes them can no longer be designed independently because the information rate a wearable BCI delivers is ultimately bounded by the power it is allowed to dissipate [14]. Figure 1 contrasts the server-dependent laboratory arrangement with the integrated wearable alternative that this constraint has produced. This survey follows that pipeline end to end—from the electrode and the analog front-end, through hardware-friendly feature extraction, to the on-chip engine that performs inference—and examines the design space at each stage.

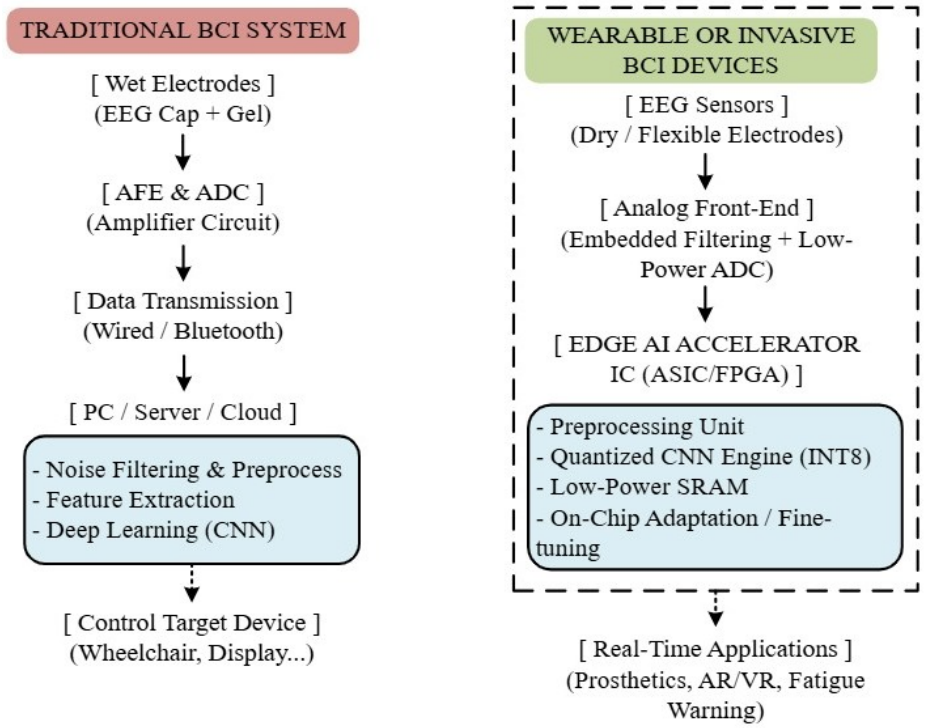


Figure 1. Architectural comparison: traditional lab-based BCI vs. wearable edge AI BCI.

1.2. Scope and Literature Selection

This article is a narrative review. Its scope is the hardware path that a neural signal travels in a wearable or implantable brain–computer interface: the electrode interface and the mixed-signal acquisition chain, the feature extraction and decoding algorithms that are realisable under an embedded power budget, and the FPGA and ASIC substrates on which those decoders are executed. Within this span, we give priority throughout to works reporting results measured on physical hardware, and we report such results as categorically distinct from efficiency claims resting only on parameter counts, floating-point operation counts or server-GPU timing. Outside that span lie three bodies of work that this survey

does not attempt to cover: clinical efficacy and patient-outcome studies, decoding algorithms with no demonstrated or prospective embedded implementation, and acquisition modalities other than EEG, ECoG and intracortical recording, which appear only where a comparison requires them.

The corpus was assembled through purposive, expertise-guided selection rather than by an exhaustive protocol. Candidate works were discovered through queries issued through Google Scholar combining an application axis (BCI, EEG, iEEG, ECoG, neural recording), a technology axis (analog front-end, low-noise amplifier, SAR and $\Delta\Sigma$ conversion, feature extraction, CNN, SNN, hyperdimensional computing, Riemannian decoding, foundation model) and an implementation axis (ASIC, FPGA, SoC, accelerator, wearable, implantable, ultra-low power, quantization, pruning, knowledge distillation). Records were then retrieved and verified on the publishers' own platforms, predominantly IEEE Xplore, which accounts for the majority of the corpus, together with IOPscience, Wiley Online Library, MDPI, the ACM Digital Library, SpringerLink, ScienceDirect and Nature Portfolio; a smaller number of sources lie outside these platforms, namely open-access journals, the proceedings of machine learning conferences that do not publish through a commercial platform, and two book chapters. Three criteria governed inclusion. First, the work is peer-reviewed and written in English, and where a preprint and an archival version both exist, the archival version is cited. Second, the work either reports measured or post-layout hardware results, or it defines an algorithm, dataset or benchmark that has subsequently been implemented in hardware. Third, among the candidates satisfying the first two criteria, established works were selected on the basis of their standing in the field, indicated by citation counts and by recurrent appearance in the reference lists of the works already selected; recent publications, which cannot yet have accumulated citations, were instead admitted on the strength of the venue and on whether they report results measured on physical hardware. The reference lists of the selected works were then followed backward, both to recover the foundational studies they identify as such and to reach implementations that the keyword queries had not surfaced.

We state explicitly that this procedure is neither exhaustive nor reproducible in the sense required of a systematic review. Selection on citation standing favors older work and under-represents very recent contributions; the concentration of the corpus on a small number of publisher platforms, and on IEEE Xplore in particular, reflects where the hardware literature of this field is published but will have left relevant work elsewhere unexamined, and the final corpus reflects the authors' judgement of which works carry hardware evidence. The corpus should therefore be read as a representative rather than a complete map of the field, and the absence of a work from it should not be taken as an assessment of that work.

Because the designs surveyed were validated on different datasets, paradigms and window lengths and fabricated in nodes spanning 0.5 µm to 28 nm, their published figures are not directly commensurable, and three conventions govern every comparison that follows. Metrics are reported exactly as published, with the dataset, task and channel count stated alongside them, and cells the original authors did not report are marked "n.r." rather than estimated; values we derive from other reported quantities are marked "†", and the derivation is given in the caption. Where a common basis is needed, we prefer intrinsically normalized figures of merit rather than raw numbers. Published values are never rescaled to a common technology node because such extrapolation is unreliable for mixed-signal front-ends dominated by noise and matching rather than by digital scaling; the comparative tables are consequently an order-of-magnitude map of the design space, and the trends drawn from them rest on differences of at least one order of magnitude rather than on rankings between adjacent entries.

1.3. Key Contributions

The principal contributions of this survey are as follows:

- A unified front-end-to-inference roadmap. We formulate an end-to-end architectural framework covering the complete BCI pipeline, from physical biopotential sensing and mixed-signal conditioning to on-chip decoding, and we make the couplings between these layers explicit rather than treating them as independent design problems.
- An explicit statement of scope and selection criteria. We identify explicitly the limitation that follows from purposive selection so that readers can judge what the corpus does and does not cover (Section 1.2).
- Quantitative comparison of the sensing layer. We compare published CMOS acquisition ICs on dedicated numerical axes, namely supply voltage, power per channel, CMRR, input-referred noise with the associated NEF/PEF, bandwidth, on-chip ADC resolution, and silicon area per channel, and we complement this with a quantitative comparison of wet, dry, and semi-dry electrodes in terms of contact impedance, agreement with simultaneous wet recordings, and behavior under motion.
- Taxonomy of decoding paradigms, from handcrafted features to foundation models. We contrast spectral, temporal, and spatial feature extraction with the three lightweight topologies that dominate edge implementations, namely compact CNNs, spiking neural networks, and hyperdimensional computing, and we extend the taxonomy to the paradigms that have emerged most recently: Riemannian-geometry decoders on the manifold of covariance matrices, EEG foundation models, and the open benchmarking infrastructure used to compare them (Section 3.5).
- Hardware-aware compression and acceleration analysis. We analyze quantization-aware training, structured pruning, channel selection, and knowledge distillation; we quantify the accuracy cost that the reviewed works report for quantization; and we map these techniques onto their realizations on FPGA and custom ASIC substrates. Throughout, results measured on physical hardware are reported as categorically distinct from efficiency claims based only on parameter counts, floating-point operations, or server-GPU timing.
- Cross-cutting trade-off analysis. We place the reviewed implementations on common axes of energy per inference, latency, silicon area, model complexity, and accuracy; propose composite figures of merit that couple energy to information throughput and to decoding quality; and translate the resulting power figures into the constraints that actually bound deployment, namely tissue-heating limits, battery autonomy, and device yield (Section 3.6).
- Categorization of real-world and clinical applications. We analyze closed-loop deployments in neuroprosthetic control, rehabilitation robotics, epileptic seizure detection, and cognitive-state monitoring, and we identify on-chip adaptation as the open problem that separates laboratory demonstrations from chronic daily life use.

How to read this survey. Figure 2 maps the structure of this article, which is organized to be consulted selectively rather than read continuously. Readers concerned with the acquisition layer will find the analysis of CMOS analog front-ends in Sections 2.4 and 2.5, with the quantitative comparison of published acquisition ICs and of electrode technologies in there. Readers concerned with decoding algorithms will find handcrafted feature extraction in Section 3, the three lightweight topologies that dominate edge implementations in Section 3.3, and the more recently emerged paradigms in Section 3.5. Readers concerned with system-level implementation will find model compression and hardware acceleration in Section 3.4, the standardized comparison of reported accelerators given there, and the cross-cutting analysis of energy, latency, area and accuracy, together with the

thermal, battery and yield constraints that bound deployment, in Section 3.6. Each table is self-contained and followed by a discussion of what its standardized metrics imply.

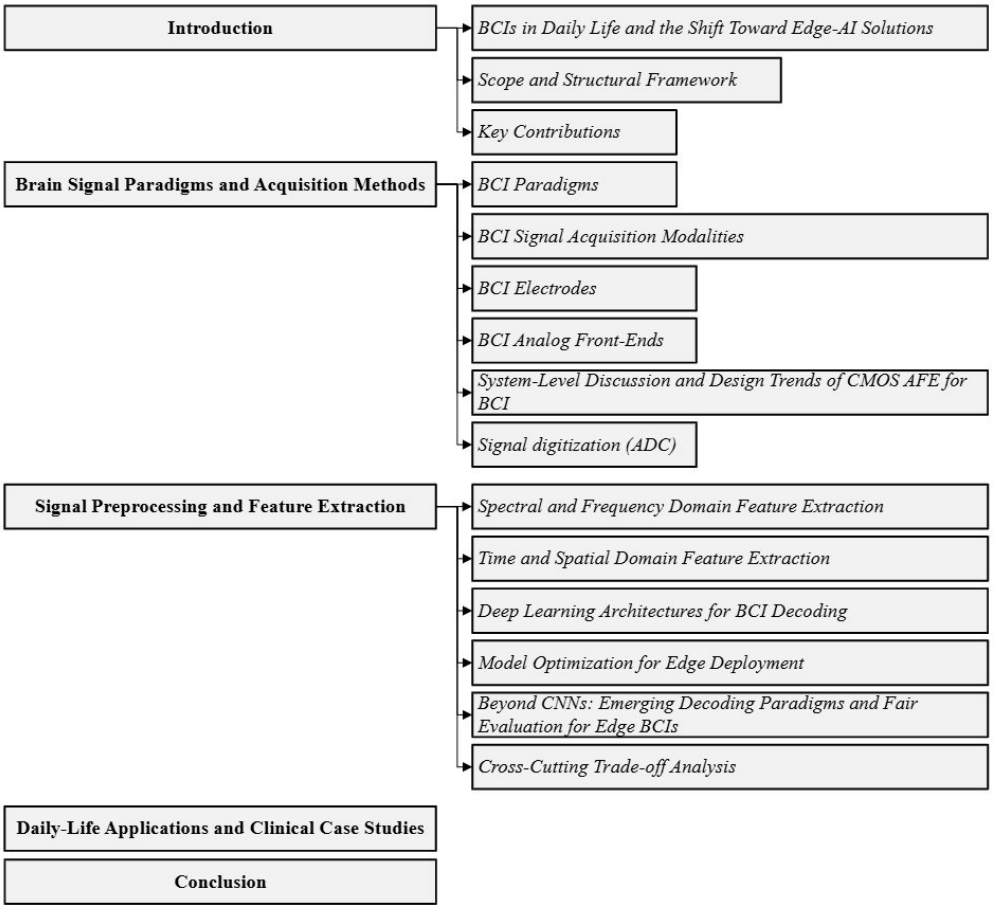


Figure 2. Structure of paper.

2. Brain Signal Paradigms and Acquisition Methods

2.1. BCI Paradigms

EEG-based BCIs establish a non-muscular communication pathway by interpreting neural oscillations and event-related potentials [2], which are conventionally organized by frequency band and by the paradigm used to elicit them [1,15]. The five canonical rhythms are conventionally delimited as follows [3,15]: delta (<4 Hz), tied to deep sleep and easily confounded with slow motion artifacts; theta (4–8 Hz), associated with drowsiness and relaxed states; alpha (8–13 Hz), prominent during relaxed wakefulness, whose sensorimotor variant, the mu rhythm, is the primary carrier of motor intent; beta (13–30 Hz), reflecting alertness and actual or imagined movement; and gamma (>30 Hz), linked to higher cognitive functions and, in its high-gamma range (80–160 Hz), a rich source of movement information that is accessible mainly through invasive recordings. From a hardware perspective, the relevant consequence is that the informative content of scalp EEG lies almost entirely below 100 Hz at amplitudes of a few tens of microvolts, which sets the bandwidth and noise budget of the acquisition chain analyzed in Section 2.3.

The paradigm determines whether neural patterns are produced voluntarily (endogenous) or in response to external stimulation (exogenous) [2], and each choice propagates directly into the requirements placed on the decoder. Motor imagery is endogenous and intuitive and requires no stimulator, but it relies on event-related desynchronization of the alpha and beta bands, is strongly user-dependent, non-stationary, and ineffective for a substantial minority of users, which is why it demands the most elaborate decoding

pipelines [16–19]. Steady-state visual evoked potentials are exogenous and offer the highest information transfer rate with almost no user training, since the visual cortex reproduces the flicker frequency and its harmonics, at the cost of eye strain and reliable gaze control [20,21]. The P300 event-related potential, elicited roughly 300 ms after an attended oddball stimulus, generalizes well across users with minimal training [2] but is fatiguing during prolonged use. Passive BCIs instead monitor spontaneous cognitive states such as workload, fatigue, or stress without interrupting the user, but they operate at a very low signal-to-noise ratio and with slow estimation times. Motor imagery and the two evoked paradigms therefore dominate the hardware literature surveyed here, the former because of its decoding complexity and the latter because of their throughput.

2.2. BCI Signal Acquisition Modalities

Recordable cortical fields originate from the synchronized postsynaptic currents of axially oriented pyramidal-cell dipoles, and the depth at which these fields are sampled defines the three acquisition tiers illustrated in Figure 3: non-invasive, semi-invasive, and invasive [22–24]. Scalp EEG senses 5–300 μV potentials below roughly 100 Hz after volume conduction through brain tissue, cerebrospinal fluid, meninges, skull, and scalp; it is safe, inexpensive, and ambulatory, but the skull attenuates and spatially low-pass filters the signal, and the recording is vulnerable to ocular and muscular artifacts and to variable electrode–skin impedance. Electrocorticography places subdural or epidural grids on the cortical surface without penetrating tissue, thereby bypassing the skull barrier and delivering a markedly better signal-to-noise ratio, wider bandwidth, and greater chronic stability than penetrating arrays, at the price of a craniotomy. Penetrating microelectrode arrays reach single units, resolving action potentials from 100 Hz to 7 kHz with amplitudes up to 500 μV alongside local field potentials up to 200 Hz and 1 mV, which provides the highest information density but entails surgical risk and progressive signal loss from the chronic tissue response. Each tier therefore trades clinical safety against spatial–temporal bandwidth, and the resulting characteristics are synthesized in Table 1.

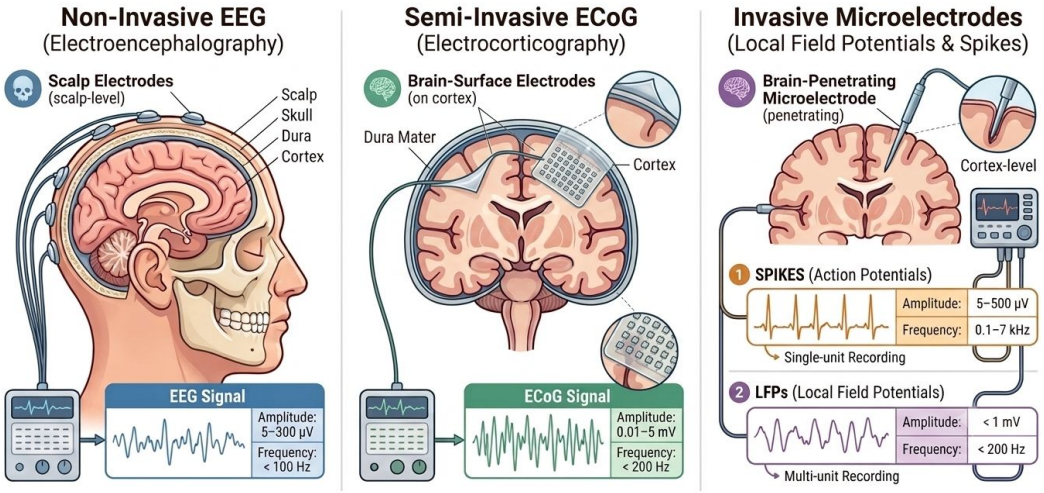


Figure 3. BCI main signal acquisition modalities.

The individual modalities compared in Table 1 differ mainly in the physical quantity they transduce. Scalp EEG records the electrical fields of synchronized neuronal populations with sub-millisecond temporal resolution at low cost and high portability, but with centimeter-scale spatial resolution and strong sensitivity to ocular and muscular artifacts [1,25]. Functional near-infrared spectroscopy instead measures hemoglobin oxygenation optically, which makes it robust to motion but introduces a hemodynamic delay of several seconds that precludes deterministic real-time control [26]. Functional magnetic

resonance imaging offers millimeter-scale whole-brain coverage, including deep structures, yet it shares the same hemodynamic latency and requires stationary infrastructure, while magnetoencephalography combines EEG-like temporal resolution with better source localization but depends on shielded rooms and cryogenic sensors; neither is compatible with wearable operation [27]. Electrocorticography bypasses the skull and yields a high signal-to-noise ratio over a broad bandwidth [15], at the cost of a craniotomy and limited chronic stability, whereas intracortical microelectrode arrays resolve local field potentials and single-unit spikes with unmatched precision for demanding decoding tasks such as speech or dexterous motor reconstruction, but face tissue trauma, infection risk, and signal degradation from glial scarring [1]. For the remainder of this survey, we focus on electrical modalities, and primarily on EEG, since they alone satisfy the latency, power, and form-factor constraints of edge BCI hardware.

Table 1. Comparison of brain signal acquisition methods.

Method	Invasiveness	Spatial Resolution	Temporal Resolution	Mobility	Primary Application
EEG	Non-invasive	Low (1–2 cm)	Very high (<1 ms)	Very High (Wearable)	Consumer BCI, Gaming, Fatigue Monitoring
fNIRS	Non-invasive	Moderate (0.5–1 cm)	Low (1–5 s)	High	Cognitive Workload, Hybrid BCI Systems
fMRI/MEG	Non-invasive	Very high (1 mm)	Low to Moderate	Infeasible (Stationary)	Neuroscience, High-precision Diagnosis
ECoG	Semi-invasive	High (1–2 mm)	Very High (<1 ms)	Low (Clinical)	Rehabilitation, Epilepsy Mapping
Invasive	Fully invasive	Very high (Single-neuron)	Very high (<0.1 ms)	Moderate	Clinical Probing, Neuroprosthetics

2.3. BCI Electrodes

Electrodes serve as the essential hardware interface and the key precondition for detecting and delivering electrical brain activity within brain–computer interface (BCI) systems [28–30]. At the biological level, recordable neural signals primarily originate from the collective postsynaptic potentials and the flow of ions (such as Na^+, K^+, Ca^{2+}, and Cl^-) across neuronal membranes, which generate miniature electric dipoles. Scalp EEG monitors these electrical potentials as voltage differences recorded between active and reference electrodes placed on the head [31]. Beyond passive signal acquisition, high-performance electrode technologies facilitate bidirectional communication between the brain and external devices, enabling both neural signal capture and functional electrical stimulation (FES) to modulate neural pathways.

However, conventional neural interfaces suffer from a severe mechanical mismatch. Brain tissues are extremely soft, exhibiting an elastic (Young's) modulus in the kilopascal (kPa) range, whereas traditional rigid metal and silicon electrodes are several orders of magnitude stiffer, possessing moduli in the range of tens of gigapascals (GPa). This mechanical mismatch induces chronic shear stress and micromotion at the tissue–electrode interface, frequently triggering neuroinflammatory responses and glial encapsulation by microglia within months. This tissue encapsulation progressively degrades the signal-to-noise ratio (SNR) and limits long-term recording stability. Therefore, transitioning to flexible electrodes fabricated from soft polymers and hydrogels with a low elastic modulus is vital to achieve mechanical compliance, conformal tissue adaptation, and stable long-term biocompatibility.

Extracranial electrodes are categorized as wet, dry, or semi-dry according to how they establish the ionic–electronic path across the stratum corneum. Wet electrodes remain the gold standard, since an electrolytic gel or saline buffer hydrates the skin and yields the lowest contact impedance and the best motion robustness [11], but gel application is tedious, may irritate the skin, and the contact degrades as the gel dehydrates [5]; flexible conductive films, such as aerogels based on multi-walled carbon nanotubes and cellulose nanocrystal fibres, are being explored to improve water retention and biocompatibility. Dry electrodes remove the conductive medium altogether and instead rely on geometry to reach the scalp through hair, using flexible micro-needle arrays that pierce only the non-vascular stratum corneum, silver-coated bristles that distribute contact pressure, or claw-like multipin structures of thermoplastic polyurethane coated with Ag/AgCl [32,33]; the price is a contact impedance often in the megaohm range and a pronounced sensitivity to motion artifacts. Semi-dry electrodes are the compromise, releasing on the order of 10–20 µL/h of electrolyte from a reservoir through porous polymer, sponge, or ceramic wicks [34], while pre-gelled and ear-worn variants with replaceable conical hydrogel probes offer rapid setup, contact impedance below 30 kΩ, and no post-recording residue [30,35]. For the analog front-end designed in Section 2.4, the decisive consequence of this choice is the electrode impedance and the electrode DC offset that the amplifier must tolerate: the megaohm-scale, drift-prone impedance of dry contacts is what motivates the high input impedance, chopper stabilization, and DC-servo loops described there.

Because the choice of electrode propagates directly into the specification of the analog front-end, Table 2 summarizes the quantities that the published comparisons actually measure. Three findings recur. First, contact impedance differs by two to three orders of magnitude between technologies: in a controlled ear-EEG study, the same electrode sites gave 5.1 kΩ at high frequency and 34 kΩ at low frequency when wet, against 230 kΩ and 1.2 MΩ when dry, with a standard deviation of the same magnitude as the mean [36]; semi-dry electrodes that release a controlled amount of electrolyte recover most of this gap, reaching 17.8 ± 3.7 kΩ at 10 Hz [34,37]. Second, the resulting signal degradation is measurable but modest at rest: semi-dry recordings correlate with simultaneous wet recordings at 0.95 ± 0.07 in time and above 0.90 in coherence beyond 10 Hz, with BCI classification accuracies that are statistically indistinguishable [34], whereas active dry electrodes still resolve the expected event-related potentials but with higher root-mean-square noise and lower statistical power than wet electrodes [10]. Third, motion is where the technologies separate. Dry and wet systems show comparable signal-to-noise ratios while the subject is seated, but the dry system yields more rejected channels, higher inter-trial variance of the P300 component, and fails to record reliably during walking [11], which is precisely the condition that a daily life BCI must survive.

These numbers set the specification of the front-end analyzed in Section 2.4 rather than merely motivating it. A source impedance that may reach 1 MΩ and varies by a factor of two between electrodes and over time requires an input impedance of the order of 1 GΩ if attenuation and the impedance-imbalance-induced degradation of the common-mode rejection ratio are to remain negligible, which is why the acquisition ICs in Table 3 reports CMRR values above 105–120 dB together with boosted input impedance, and why several of them integrate electrode-impedance monitoring. Because motion artifacts appear as low-frequency modulations of that impedance rather than as additive noise, they overlap in the delta and theta bands and cannot be filtered out without removing signal; mitigating them is therefore a system-level task, combining mechanically compliant electrodes, high and stable input impedance, and artifact-aware decoding, rather than a task for the amplifier alone.

For intracranial applications, flexible electrode development requires meticulous co-design of substrates, conductive traces, and bio-interfacial properties:

Table 2. Quantitative comparison of extracranial electrode technologies, as reported by controlled comparison studies. (Impedances are given at the frequency stated by the original work; agreement with the wet baseline is reported in the terms used by each original study. The Wet column serves as the baseline against which the agreement and the BCI accuracy of the other two technologies are expressed; "— (baseline)" therefore denotes that no relative value applies rather than a missing value.)

Property	Wet (gel, Ag/AgCl)	Dry (Metal, Polymer, Bristle)	Semi-Dry (Controlled Electrolyte Release)
Contact impedance	5.1 kΩ at high frequency; 34 kΩ at low frequency (ear sites) [36]	230 kΩ at high frequency; 1.2 MΩ at low frequency, with a standard deviation comparable to the mean [36]	17.8 ± 3.7 kΩ at 10 Hz [34]; typically tens of kΩ [37]
Agreement with wet	— (baseline)	Event-related potentials preserved, but higher RMS noise and lower statistical power [10]	Temporal correlation 0.95 ± 0.07; coherence > 0.90 above 10 Hz [34]
BCI accuracy	— (baseline)	Usable; high-speed SSVEP demonstrated with dry arrays [33]	Statistically indistinguishable from wet across five paradigms [34]
Behaviour under motion	Robust; gel maintains the ionic path	Comparable SNR when seated, but more rejected channels, higher inter-trial variance, and unreliable recording while walking [11]	Intermediate; electrolyte release stabilizes the interface, 10–20 µL/h [37]
Practical cost	Gel application and skin preparation; contact degrades as the gel dries [5,38]	No consumable; contact quality depends on applied force and hair [32,33]	Small reservoir to refill; no post-recording residue [30,35]
Front-end implication	Standard AC-coupled input	Input impedance of the order of 1 GΩ, chopper stabilization, DC-servo loop, impedance monitoring	Relaxed input-impedance requirement; offset drift remains the limiting factor

Flexible Substrates and Conducting Materials: Biocompatible polymers such as polyimide (PI, Young's modulus ≈ 3 GPa), polydimethylsiloxane (PDMS, Young's modulus ≈ 10^{-3} GPa), and parylene C are utilized as insulating substrates to minimize mechanical mismatches with brain tissues. PDMS acts as an excellent soft elastomer but is prone to swelling in physiological fluids, whereas PI offers high-temperature tolerance (over 400 °C) and microfabrication stability but lacks stretchability. Common conducting elements are deposited as thin films and include inert metals (Pt, Au, TiN, IrO_x) or nanomaterials like Ti_3C_2 MXene and graphene. Conductive polymers such as PEDOT:PSS are highly favored as electrode coatings due to their outstanding electrochemical stability and low impedance.

Interface Adhesion Engineering: To resolve the severe risk of delamination between noble metals and conductive polymer coatings under chronic hydration, intermediate chemical adhesion layers are engineered. For instance, electropolymerized poly(5-nitroindole) (PIN-5NO_2) establishes robust amino-group covalent bonds at the interface, improving PEDOT adhesion to gold substrates by nearly seven-fold. On-skin and on-brain interfaces also utilize catechol-functionalized hydrogels or dopamine-hybridized dPEDOT nanoparticles to enable robust, biomimetic tissue adhesion [38].

Biocompatibility and Smart Bio-materials: Utilizing natural polymers like silk fibroin, cellulose, alginate, and chitosan mimics the extracellular matrix, facilitating synaptogenesis and reducing chronic foreign body responses [38]. Hot-pressed nanocellulose paper substrates exhibit high wettability, recyclability, and superior biocompatibility compared to synthetic PET films. To assist surgical insertion of highly compliant, ultra-thin meshes, transient bio-resorbable stiffeners are used as temporary scaffolds. Once mounted on the cortex, the silk fibroin dissolves and is absorbed by the cerebrospinal fluid, allowing the ultra-thin electrode mesh to wrap spontaneously around brain folds driven strictly by capillary forces [28].

Transparent and Multimodal Nodes: For advanced optogenetic and neuro-imaging applications, transparent conducting films (such as AgNWs/IZO or PEDOT:PSS-EG) on

flexible PET substrates allow simultaneous electrical recording and optical interrogation or fluorescent imaging with minimum light-induced artifacts.

2.4. BCI Analog Front-Ends

Traditional neural acquisition systems based on bulky, discrete components suffer from excessive power dissipation, large physical form factors, and high susceptibility to environmental noise [39]. The physical realization of BCI on custom ASICs or highly integrated SoCs represents a fundamental technological cornerstone for moving neural interfaces from laboratory environments to ambulatory, daily life applications [40,41]. By migrating the entire analog processing chain onto a single silicon CMOS die, designers can achieve massive reductions in current consumption and active silicon area. This extreme power-and-area optimization is the primary enabling factor for integrating high-density neural interfaces into unobtrusive, wearable form factors or surgically housing them within an enclosure, called the "skull unit", to be safely implanted directly into the skull [42,43]. To prevent localized tissue damage or health hazards to the patient, all implantable circuits must be designed to operate under strict power budgets—frequently biased in the MOSFET weak inversion (WI) region to maximize power efficiency and minimize heat dissipation, while maintaining high gain and low noise operation [44,45]. Figure 4 depicts the detailed block diagram of an exemplary multi-channel brain signal acquisition front-end, demonstrating the complete pipeline from biopotential capture to low-power digital serialization.

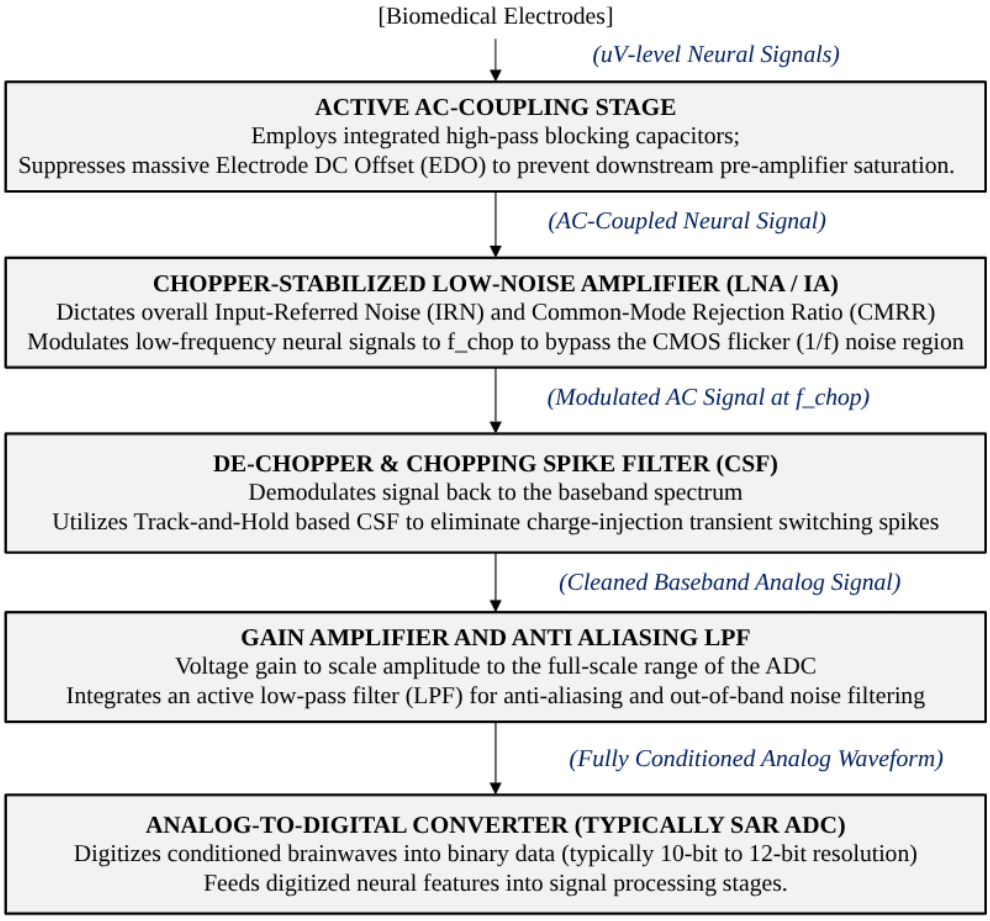


Figure 4. A general block diagram of a BCI acquisition CMOS AFE channel.

At the tissue–silicon boundary, the electrochemical reaction between the recording electrodes and the skin or cortical tissue inevitably generates a massive, unpredictable Electrode DC Offset (EDO) or Differential Electrode Offset (DEO) [46]. Because these

offsets are several orders of magnitude larger than the micro-volt biopotentials of interest, input AC-coupling capacitors are conventionally deployed to block the DC baseline. The AC-coupled input interface not only rejects the low-frequency drift but also provides a critical layer of electrical safety and isolation between the patient's brain and the active amplifier circuitry. To prevent these large offsets from saturating the subsequent high-gain amplification stages, state-of-the-art architectures integrate an active feedback servo loop. This servoloop dynamically senses and suppresses the EDO directly at the pre-amplifier's virtual ground, ensuring the system maintains high input impedance and superior linearity without experiencing analog saturation.

The Low-Noise Instrumentation Amplifier (LNA) or Instrumentation Amplifier (IA) is the most critical and power-consuming building block of the front-end, as it directly defines the overall common-mode rejection ratio (CMRR) and input-referred noise (IRN) floor of the system. Sub-micron CMOS transistors exhibit dominant flicker ($1/f$) noise at the low frequencies characteristic of neural signals (typically below 100 Hz). To bypass this noisy region, an Input Chopper modulates the low-frequency neural signal to a higher chopping carrier frequency where thermal noise dominates. Pre-amplification is subsequently carried out by a highly power-efficient amplifier utilizing cross-coupled active loads or current-feedback topologies biased in the MOSFET weak inversion region to maximize transconductance gain (g_m/I_D) under a sub-micro-ampere current budget. This chopper-stabilized current-feedback topology eliminates the need for power-hungry active operational amplifiers, achieving exceptional noise efficiency and an outstanding common-mode rejection ratio (CMRR) to reject severe common-mode interferences.

Following pre-amplification at the modulated carrier frequency, an Output De-Chopper demodulates the amplified neural signal back to its original baseband spectrum while simultaneously modulating the amplifier's internal $1/f$ noise and DC offset to the chopping frequency. Because modern high-density BCI SoCs record from dozens or hundreds of channels, transferring raw analog waveforms out of the chip is extremely constrained by physical wiring. To facilitate better input–output cable management, the outputs of the amplifier array are multiplexed in time using a high-speed serializer. This stage utilizes non-overlapping clock signals to ensure that only one amplifier is connected to the output buffer at a time during channel switchover, thereby eliminating switching transient crosstalk and dramatically reducing the required number of output telemetry wires.

Because neural signals exhibit significant dynamic range fluctuations across different physiological states—varying from sub-100 µV scalp EEG to millivolt-level ECoG potentials—dynamic signal conditioning is mandatory. The Programmable Gain Stage (PGS) or Programmable Gain Amplifier (PGA) adjusts the overall channel gain via a digital control interface to scale the signal amplitude to the full-scale input of the analog-to-digital converter. Concurrently, the PGA integrates an active Low-Pass Filter (LPF). This stage functions as an anti-aliasing filter to suppress out-of-band noise, high-frequency chopping harmonics, and high-frequency folding during the subsequent digitization and multiplexing processes, maintaining a clean analog output.

The finalized, fully conditioned analog brainwaves are converted into high-fidelity digital streams via an integrated, multi-channel Analog-to-Digital Converter (ADC). To bridge the sensitive analog front-end with noisy digital processing units, a partition design or isolated substrate layout is critical to avoid substrate noise coupling from disturbing the micro-volt biopotentials. The Successive Approximation Register (SAR) ADC is widely established as the gold standard for wearable and implantable BCIs due to its charge-sharing capacitive DAC architecture. By executing a highly efficient binary search algorithm, the on-chip SAR ADC achieves low-energy conversions without consuming static bias currents, providing low-latency, digitized neural features to the edge AI cores.

Table 3. Quantitative comparison of state-of-the-art CMOS analog front-ends (AFEs) and acquisition ICs for BCI. (Values are as reported by the original works; "n.r." denotes not reported. IRN: input-referred noise; NEF/PEF: noise/power efficiency factor.)

Study & Year	Tech.	Ch.	Supply (V)	Power/ch	CMRR (dB)	IRN (NEF/PEF)	Bandwidth	On-Chip ADC	Techniques, Area & Design Trade-Offs
Yazicioglu et al. (2007) [46]	0.5 µm CMOS	1	3.0	60 µW (IA: 11.1 µA)	>120	NEF 9.2	Configurable (EEG/ECG/EMG)	None	AC-coupled chopped IA (ACCIA), current-feedback IA, chopping spike filter; rejects ±50 mV DEO. Single channel and no digitizer limit channel density; area n.r.
Yazicioglu et al. (2008) [47]	0.5 µm CMOS	8	3.0	25 µW (200 µW total; IA: 2.3 µA)	>120	0.57 μV_{rms} (NEF 4.3)	0.5–100 Hz	11-bit SAR	Coarse–fine servoloop, 1 GΩ input impedance, electrode-impedance measurement. High 3 V supply and large input capacitors limit area per channel.
Ng & Chan (2005) [41]	0.5 µm CMOS	8 (mux)	1.5	485 µA total at 1.5 V	>115 (0–1 kHz)	0.86 μV_{rms} (0.3–150 Hz; NEF n.r.)	Externally set	None	Digitally programmable rail-to-rail IA. Very high current consumption; requires external passives to define band corners.
Mollazadeh et al. (2009) [48]	0.5 µm CMOS	16	3.3	AFE: 12.2 µA; ADC: 76 µW at 3.3 V	>76 (1–10 kHz)	1.65 μV_{rms} (NEF 3.2)	Tunable bandpass (LFP to spikes)	10-bit incremental (continuous-time)	Two-stage capacitor-feedback amplifier with tunable corners for multi-scale recording. High ADC power and large area in 0.5 µm technology.
Chen et al. (2014) [49]	0.18 µm CMOS	8	1.8–10 (Adjustable)	2.8 mW	n.r.	NEF 1.77	HPF 0.1–10 Hz; LPF 0.8–7 kHz	10-bit DMSAR (ENOB 9.57)	Closed-loop stimulation (30 µA), on-chip 8051 MCU/SRAM, wireless power and data. Stimulator integration raises power; low channel count.

Table 3. *Cont.*

Study & Year	Tech.	Ch.	Supply (V)	Power/ch	CMRR (dB)	IRN (NEF/PEF)	Bandwidth	On-Chip ADC	Techniques, Area & Design Trade-Offs
Muller et al. (2015) [23]	65 nm CMOS	64	0.5/1.0	2.3 µW (IC total 225 µW)	88	1.2 μV_{rms} (NEF 4.76/PEF 11.3)	~1–500 Hz	VCO-based, 13-bit range at 1 kS/s	DC-coupled chopper front-end with mixed-signal servo DAC (±50 mV offset cancellation); 0.025 mm^2/ch. Restricted dynamic range and 88 dB CMRR versus chopper-stabilized AC-coupled designs.
Chandrakumar & Marković (2018) [50]	40 nm CMOS	1	1.2	7.3 µW	n.r.	15.2 ENOB over 200 mV_{pp} input range	LFP/spike band	3rd-order CT-ΔΣM with 6-bit SAR quantizer	8×-gain CCIA, time-varying degeneration, data-weighted averaging. Single-channel prototype; high complexity and area per channel prevent dense scaling.
Pu et al. (2022) [51]	180 nm HV CMOS	32 rec./ 4 stim.	HV (±20 V stim. compliance)	n.r.	88	1.03 μV_{rms} (2.37/5.62)	50–160 Hz (recording)	Dual-mode DAQ (full-band + baseband)	2 mV time-based charge balancing, stimulation-artifact contour shaping, 340 mV_{pp} linear input swing, 14 mA/8-bit stimulator. Large HV switch fabric and complex feedback loops dominate area.
Jang et al. (2024) [24]	28 nm CMOS	1024	1.0	268 nW	n.r.	7.4 μV_{rms} (2.84/8.07)	300 Hz–5 kHz	Global single-slope ADC with PPM digital pixels	Wired-OR data-compressive readout (146× output-rate reduction); 36 × 36 μm^2 (0.0013 mm^2) per channel. Narrow band excludes LFP/EEG; wired-OR collisions degrade resolution.
Karimi-Bidhendi et al. (2017) [40]	130 nm/ 180 nm	64/4	0.4	0.216 µW	greater than 60/74	2.19 μV_{rms} (4.65/11.7)	ECoG/EEG band	None (analog output)	Weak-inversion amplifiers with cross-coupled active loads and channel serializers; 0.044 mm^2/ch. No on-chip digitizer; susceptible to power-line interference.

2.5. System-Level Discussion and Design Trends of CMOS AFE for BCI

Table 3 makes the quantitative design trends explicit. First, supply voltage has scaled from 3 V in 0.5-µm designs [46,47] to 1.2 V, 1.0 V, and even 0.4–0.5 V in deep-submicron implementations [23,24,40,50], which is the single largest contributor to the reduction in power per channel, from tens of microwatts down to 268 nW/channel [24]. Second, this reduction is not obtained for free: because the input-referred noise of a weak-inversion amplifier scales as $v_{n,\text{in}}^2 \propto 4kT\gamma/g_m$ and $g_m \propto I_D$, lowering the bias current degrades noise unless chopping or current reuse is employed. The reported figures illustrate the trade-off directly, with 0.57 μV_{rms} at 25 µW/channel in an ambulatory EEG ASIC [47] against 7.4 μV_{rms} at 268 nW/channel in a 1024-channel spike-recording array [24]; the latter is acceptable only because action potentials are two orders of magnitude larger than scalp EEG. The NEF and PEF columns normalize this trade-off and show that the efficiency of the analog core has improved far more slowly (NEF of roughly 3.7–4.8 across two decades) than the raw power figures suggest. Third, CMRR remains the discriminator between non-invasive and implantable designs: chopper-stabilized AC-coupled front-ends for scalp EEG maintain more than 120 dB [46,47], and bi-directional cortical interfaces exceeding 105 dB [51], whereas the DC-coupled, area-optimized implant front-end achieves 88 dB [23]. Finally, the area per channel decreases by more than an order of magnitude with technology scaling, from 0.044 mm^2 in 130 nm [40] to 0.025 mm^2 in 65 nm [23] and 0.0013 mm^2 in 28 nm [24], which is what ultimately enables 1024-channel arrays and explains the migration toward compressive, digital-pixel readout architectures. Bandwidth and ADC resolution follow the target modality rather than the technology node: 0.5–100 Hz with 11-bit conversion is sufficient for EEG [47], whereas spike recording requires 300 Hz–5 kHz [24], and closed-loop iEEG systems use tunable corners between 0.1 Hz and 7 kHz [49].

A chronological and architectural evaluation of the state-of-the-art neural readout ASICs reveals a distinct correlation between semiconductor technology scaling and the achievable channel density. Early pioneering designs, such as the works by [41,46,47] were typically fabricated in mature 0.5 µm CMOS processes. While these architectures successfully achieved micro-watt power consumption per channel, their channel scaling was strictly constrained by the physical area occupied by passive components, particularly the large integrated capacitors required for AC-coupling and low-frequency filtering. As neural interfaces scaled toward high-density spatial mapping, researchers migrated to sub-micron and deep-submicron CMOS nodes. This scaling trajectory is highlighted by the transition to 180 nm nodes for bi-directional closed-loop systems (e.g., [49,51] to 65 nm nodes for high-density ECoG matrices (e.g., [23]) to 40 nm nodes for single-channel direct-digitization front-ends (e.g., [50]), and ultimately to the state-of-the-art 28 nm CMOS process featured in the work of [24]. This relentless technological scaling has enabled an unprecedented reduction in the physical footprint per channel, shrinking the active pixel area down to a microscopic scale of 36 × 36 μm^2, which in turn allows for the parallel integration of up to 1024 recording channels on a single chip.

Beneath the physical scaling of these ASICs lies a fundamental, classic mixed-signal design trade-off between AC-coupling and DC-coupling input topologies. At the electrode–tissue interface, electrochemical half-cell potentials routinely generate a massive Electrode DC Offset (EDO) or Differential Electrode Offset (DEO) of up to ±300 mV. Traditional architectures (e.g., [40,46,47]) employ AC-coupling input capacitors to physically block this DC baseline, thereby preventing the high-gain pre-amplifier from saturating. However, these input capacitors must be sufficiently large to minimize input signal attenuation and maintain a low high-pass corner frequency, which consumes a prohibitive amount of silicon real estate. To bypass this area bottleneck, refs. [23,52] pioneered the DC-coupled neural front-end architecture, completely eliminating the bulky AC-coupling capacitors to achieve

an active pixel area of just 0.013 mm^2 [23,52]. Nevertheless, this DC-coupled paradigm introduces severe design challenges, as the front-end becomes highly vulnerable to baseline drift and electrode DC offsets. To maintain a linear dynamic range and prevent saturation under a low 0.5-V supply, DC-coupled systems must integrate complex active DC servo loops (DSLs) or background offset compensation networks, shifting the design burden from passive components to active feedback circuitry.

Finally, the paradigm of neural signal conditioning is undergoing a major transition from traditional analog-dominant chains toward direct digitization and in situ data compression architectures. Historically, the standard readout chain consisted of a low-noise pre-amplifier, a programmable gain stage (PGA) to maximize the signal's dynamic range, and an anti-aliasing filter before feeding into a centralized ADC. However, as the channel count expands into the thousands, transmitting high-bandwidth raw digitized waveforms off-chip creates an unsustainable data-bandwidth and power-dissipation bottleneck, risking local tissue heating beyond the safe 1 °C limit. To resolve this tension, contemporary designs have explored direct digitization at the front-end, as demonstrated by [50] who utilize a continuous-time ΔΣ-ADC with a large 200 mVpp linear input range to directly digitize the electrode signal without a dedicated pre-amplifier stage. Furthermore, to combat the data deluge in massively parallel arrays, ref. [24] introduced an event-driven, compressive wired-OR analog-to-digital readout array. By executing spatial data compression directly at the pixel level, this compressive sensing approach discards redundant baseline data and only transmits sparse neural action potentials (spikes). Although this compression scheme introduces a trade-off by slightly reducing the absolute amplitude resolution and introducing spike collision risks, it dramatically lowers the power-per-channel down to an extraordinary 268 nW, representing the vanguard of energy-efficient, high-bandwidth neural interfaces.

The study by [46] focuses on the implementation of an on-chip ASIC analog readout front-end capable of flexible configuration for various biosignals, such as EEG (brain), ECG (heart), and EMG (muscle). Traditional instrumentation amplifiers (IAs) often struggle to balance low power consumption with low noise, while also being susceptible to 1/f noise (flicker noise) and differential electrode offset (DEO) from the electrodes. The authors propose an AC-Coupled Chopped Instrumentation Amplifier (ACCIA) architecture incorporating two key techniques: a chopping technique that shifts the signal to a higher frequency range to eliminate CMOS transistor 1/f noise and enhance the Common-Mode Rejection Ratio (CMRR); and an AC-coupling technique that allows for the rejection of DEO (up to 50 mV) from standard Ag/AgCl electrodes without saturating the amplifier. The resulting design, fabricated using a 0.5 µm CMOS process, achieves an ultra-low total power consumption of 60 µW (with an ACCIA current draw of only 11.1 µA from a 3 V supply), an input-referred noise of 57 nV/$\sqrt{\text{Hz}}$, and an impressive CMRR exceeding 120 dB. A highlight of the study by Yazicioglu et al. is the optimization of the trade-off between noise and power consumption—quantified by the Noise Efficiency Factor (NEF)—as expressed in Equation (1), which was introduced by Steyaert and Sansen [53] for micropower instrumentation amplifiers and was subsequently established as the standard figure of merit for neural recording amplifiers by Harrison and Charles [54].

$$\text{NEF} = V_{\text{in,rms}} \cdot \sqrt{\frac{2 \cdot I_{\text{tot}}}{\pi \cdot U_T \cdot 4kT \cdot \text{BW}}} \quad (1)$$

where $V_{in,rms}$ is total input-referred RMS noise voltage integrated over the operating bandwidth [V_{rms}], I_{tot} is the total current consumption of the entire amplifier circuit, [A], U_T is the thermal voltage ($U_T \approx 25.9$ mV at a room temperature of 300 K), k is the Boltzmann constant, T is the absolute temperature in Kelvin, $4kT$ is the thermal noise spectral density, and BW is the operating bandwidth of the amplifier circuit in Hertz [Hz]. A lower NEF

value indicates a design that is more optimized in terms of energy efficiency for a given noise level. This study achieved an actual NEF value of 9.2 (and subsequently improved to 4.3 in the 2008 version [47]).

The study by [41] addresses the challenge of operating a system at low supply voltages without compromising signal quality or dynamic range. The SoC architecture integrates an 8-to-1 analog multiplexer, a novel rail-to-rail instrumentation amplifier (IA), a programmable gain amplifier (PGA), a low-pass filter (LPF), and an output scaling amplifier. To handle 1/f noise and the substantial DC offset arising from the electrode-skin interface (which can reach up to 300 mV), the authors employ a Differential Difference Amplifier (DDA) structure combined with a chopper stabilization technique. The system achieves a stable maximum AC gain of 10,000 V/V, a Common-Mode Rejection Ratio (CMRR) exceeding 115 dB (from 0 to 1 kHz), and an input-referred noise of 0.86 μV rms (within the 0.3 Hz–150 Hz bandwidth) using a 0.5 μm CMOS fabrication process. The study's most significant contribution is the design of an amplifier capable of DC-offset suppression, thereby preventing output saturation during high-gain amplification. This characteristic is defined by the IA's transfer function, as shown in Equation (2).

$$A_V(s) = \frac{V_{\text{out}}(s)}{V_{\text{in}}(s)} \approx \begin{cases} 1 & \text{when } f \to 0 \text{ (DC)} \\ 40 \text{ V/V} & \text{when } f > 0.1 \text{ Hz} \end{cases} \quad (2)$$

Typically, with electrode DC offsets reaching hundreds of millivolts, a high-gain amplifier would immediately saturate (exceeding the 1.5 V supply voltage). The ingenious mechanism here is that the chip treats the DC offset as a signal with a gain of unity (avoiding saturation) while simultaneously amplifying weak EEG signals above 0.1 Hz by a factor of 40 at the initial IA stage. The combination of a rail-to-rail input voltage range and this offset cancellation capability is the key factor enabling Ng and Chan's system to operate stably at ultra-low power levels, paving the way for modern wearable BCI devices. The study by [7] presents a BCI control system fully integrated into an earplug, utilizing a dedicated SoC chip to decode in-ear EEG (ear-EEG) signals. Key challenges addressed in this study include the fact that ear-EEG signals have amplitudes 10–15 dB lower than scalp EEG signals due to the thickness of the ear bone structure and the use of dry electrodes generates a significant DC offset (EDO)—up to 350 mV—which can easily saturate the amplifier. They proposed a Current-Reusing Low-Noise Amplifier (CRLNA), which employs a current-reuse technique to achieve low noise levels while minimizing power consumption. A Bootstrapping DC Servo Loop (BDSL) serves as an intelligent feedback loop, enabling the system to measure micro-signals even in the presence of EDOs up to 350 mV without increasing the noise floor. A Dual-Mode Programmable Gain Amplifier (DMPGA) is also introduced, featuring two operating modes: a "Sleep mode" consuming only 5.7 μW (monitoring only for blink signals to trigger wake-up) and an "Active mode" consuming 82.9 μW. The study's most significant contribution is the simultaneous optimization of noise, current, and supply voltage. The Power Efficiency Factor (PEF), calculated using Equation (3), is used to compare the energy efficiency of various front-end chip designs.

$$\text{PEF} = \text{NEF}^2 \cdot V_{\text{DD}} \quad (3)$$

In this study, the NEF achieved is 3.32, and the supply voltage (VDD) is very low, at just 0.8 V. The use of PEF instead of NEF reflects the authors' effort to reduce the supply voltage (VDD) below 1 V while maintaining very low noise (0.38 μV_{rms}). The PEF value of 8.8 achieved by [7] is significantly lower than that of traditional IA designs, confirming the hardware optimization advantages for wearable BCI devices.

The realization of these diverse BCI visions necessitates the adoption of specialized Integrated Circuits (ICs) and System-on-Chip architectures to replace bulky, laboratory-

bound instrumentation. By integrating neural acquisition front-ends, signal processing units, and machine learning classifiers onto a single silicon die, these SoCs achieve the radical miniaturization and ultra-low power consumption required for long-term autonomy. Furthermore, this "Silicon" approach enables high-efficiency on-chip computation, which is essential for reducing communication bandwidth, ensuring neural data privacy, and providing the millisecond-level latency required for real-time interaction in daily life.

2.6. Signal Digitization (ADC)

In modern BCI systems, the signal conversion stage is no longer merely a standalone ADC block; it has evolved into a mixed-signal processing system capable of dynamically adjusting resolution and sampling rates based on the requirements of the downstream digital processing algorithms, with the ultimate goal of extending the device's battery life.

The SAR ADC is the dominant ADC architecture in BCI systems, primarily due to its exceptional energy efficiency within the low-frequency range (from a few hundred Hz to a few kHz) characteristic of EEG and ECoG signals [47,49]. This architecture employs a binary search algorithm to determine the signal's digital value, enabling ultra-low power consumption (in the range of a few μW) and a compact chip footprint by eliminating the need for bulky filters. The SAR ADC operation begins by sampling the signal and performing a sequential bit-cycling process, starting from the Most Significant Bit (MSB). At each comparison step, the system compares the input voltage against a reference voltage generated by a capacitive DAC to determine the search direction and set the current bit to '1' or '0' via a charge addition or subtraction mechanism. This process repeats until all N bits have been resolved, yielding an accurate digital representation of the original analog signal with superior energy efficiency [47]. The relationship between Signal-to-Noise and Distortion Ratio (SNDR) and Effective Number of Bits (ENOB) serves as the most practical metric for ADC accuracy, as illustrated in Equation (4).

$$\mathrm{ENOB} = \frac{\mathrm{SNDR} - 1.76}{6.02} \tag{4}$$

where 1.76 is the ideal quantization noise floor constant derived for a theoretical sinusoidal input over a full-scale range, and 6.02 is the theoretical scaling factor converting power ratios in decibels (dB) to resolution bits ($20 \cdot \log_{10}(2) \approx 6.02$). In BCI systems, the actual ENOB value typically ranges from 9.5 to 10.5 bits to ensure that subtle brainwave features are not lost amidst background noise.

When the application requires high flexibility across the diverse spectrum of neural signals—from single-unit spikes to low-frequency EEG—the Incremental Delta-Sigma (ΔΣ) ADC architecture emerges as a superior, multi-modal alternative. This architecture is distinguished by its continuous-time integration, typically realized through a Gm-C input stage that converts differential voltage signals into currents. A primary advantage of this continuous-time integration is that it naturally provides anti-aliasing filtering, which effectively obliterates the need for complex, area-consuming external sampling circuitry. The "incremental" nature of this converter allows it to reset at the beginning of each conversion cycle, making it ideal for dedicated per-channel digitization without the crosstalk issues often found in high-speed multiplexed systems. Furthermore, this architecture offers unparalleled reconfigurability through software; by adjusting the quantization interval, the system can dynamically trade bandwidth for resolution. For example, it can be configured to provide a 7-bit resolution at 16,000 samples/s for capturing rapid neural spikes, or a 12-bit resolution at 500 samples/s to accommodate the fine amplitude requirements of EEG signals [48]. To further enhance the dynamic range, these ADCs often incorporate digital gain modulation through duty cycle modulation in the current feedback loop, providing precise gain ranging from 1 to 4096. Additionally, to combat the low-frequency noise inherent in

CMOS processes, the internal comparator typically utilizes a Correlated Double Sampling (CDS) scheme, which significantly reduces the dominant 1/f noise floor. This combination of software-defined resolution and robust noise suppression makes the incremental ΔΣ ADC a powerful "Silicon" tool for modern, adaptable BCI interfaces.

3. Signal Preprocessing and Feature Extraction

Biopotential signals captured from the cortical surface or the scalp (such as EEG, ECoG, or iEEG) are characterized by extremely low amplitudes (typically in the microvolt range), low-frequency profiles, and high susceptibility to both physiological artifacts (such as EMG and EOG) and CMOS-inherent 1/f noise [23]. Since these signals must propagate through multiple highly dispersive biological layers—including cerebrospinal fluid, skull, and scalp—they suffer from heavily degraded signal-to-noise ratios (SNRs) and high spatial correlation. Consequently, streaming continuous raw multi-channel neural data across a wireless link introduces severe power and bandwidth bottlenecks, which limits the operational longevity of wearable or implantable BCI platforms [49]. To address this challenge, the feature extraction stage acts as a crucial mixed-signal or digital gateway, aiming to denoise, compress, and translate high-dimensional raw data into dense, physiologically meaningful biomarkers [24].

To maximize BCI decoding accuracy—such as in movement intent classification or closed-loop seizure detection—the extracted features must capture neural variations across the spectral (frequency), spatial (channel), and temporal (time) domains. In dedicated ASICs and SoCs platforms, traditional handcrafted feature extractors are hardwired or coarsely accelerated on silicon to achieve low-power operation:

- Spectral/Band Energy Extraction: Realized through a bank of bandpass filters (BPFs) to isolate physiological bands, followed by squaring multipliers and integrators to calculate signal power over a specific window. Alternatively, highly optimized FFT cores are used to compute power spectral density (PSD) [55].
- Approximate Entropy (ApEn): Quantifies the randomness and complexity of iEEG signals, serving as an effective biomarker for on-chip closed-loop seizure detection [49].
- Adaptive Wavelet Transform: Offers simultaneous time-frequency localization, using customized mother wavelets to detect the onset of motor tasks with one second latency [56].
- Autoregressive (AR) Modeling: Models temporal characteristics parametrically, either by solving the Yule–Walker equations through the hardware-efficient Levinson–Durbin recursion or by estimating the coefficients directly with the Burg algorithm [55].
- Canonical Correlation Analysis (CCA): The gold standard for steady-state visual evoked potential (SSVEP) target identification, which finds the maximum correlation between multi-channel EEG and sine-cosine reference templates, often requiring dedicated hardware accelerators for QR decomposition and singular value decomposition (SVD) [9,14].

Recently, the development of domain-specific accelerators has enabled the integration of light deep learning models, such as EEGNet, directly at the edge [57]. These networks utilize separable convolutions, where depthwise convolutions (dCONVs) act as adaptive spatial filters to combine electrode channels, and pointwise convolutions (pCONV) mix temporal-spectral features. However, deploying these deep learning models within the stringent power budgets of wearable body area networks (BAN)—typically restricted to less than 200 mW—demands a rigorous algorithm–hardware co-design approach. Crucial hardware-friendly optimization strategies, including embedded channel selection to reduce input dimensionality, weight/activation quantization (such as binarized or product quantization) to minimize SRAM footprint, and the adaptive reuse of processing elements (PEs), are

essential to eliminate computational redundancy while preserving decoding fidelity [58,59]. While traditional methods like BPF and AR modeling offer hardware-friendly simplicity, modern paradigms like CCA and deep learning achieve superior classification accuracy at the expense of computational density, thereby necessitating advanced hardware–algorithm co-design. A summary of dominant feature extraction techniques tailored for BCI SoC architectures is presented in Table 4.

Table 4. Comparison of feature extraction methods in BCI SoCs.

Comparison Criteria	Spectral Analysis/ Band Energy (FFT/BPF)	Wavelet Transform (DWT/CWT)	Autoregressive (AR) Modeling	Canonical Correlation Analysis (CCA)	Deep Learning Approaches (CNN)
Representation Domain	Frequency (Power Spectrum)	Joint Time-Frequency	Parametric Time-Domain	Spatial-Frequency	Automated Multi-Domain (End-to-End)
Decoding Accuracy	Moderate (highly sensitive to noise, low SNR)	Good (efficiently captures rapid transient signals)	Good (well-suited for rhythmic and periodic signals)	Excellent (specifically for SSVEP targets with low inter-subject variability)	Very High (given a sufficiently large and diverse training dataset)
Hardware Complexity (Area & Power)	Lowest; Easily implemented using static hardware filter banks or dedicated FFT accelerators.	Moderate; Highly optimizable in logic footprint if the lifting scheme is utilized instead of convolution.	Moderate to High; Resource-intensive due to recursive calculations and hardware-expensive division operations.	High; Requires specialized real-time linear algebra engines for QR decomposition and SVD.	Very High; Involves massive floating-point or fixed-point MAC operations and severe memory bandwidth constraints.
Buffer Memory Requirements	Low; Requires only small registers for static coefficients and accumulation variables.	Moderate; Requires temporal sliding windows to store samples for multi-level decomposition.	Low; Only stores a compact vector of model coefficients (model order m).	High; Requires substantial storage for both the multi-channel EEG sample matrix and reference templates.	Very High; Demands advanced quantization and network pruning techniques (e.g., Brevitas, FINN) to minimize SRAM overhead at the edge.

3.1. Spectral and Frequency Domain Feature Extraction

Spectral and frequency domain feature extraction is the most heavily utilized method for decoding brain states, given that neural activities are intrinsically characterized by rhythmic oscillations. Neural signals are typically classified into distinct physiological frequency bands—such as Delta, Theta, Alpha, Beta, and Gamma—each of which correlates with specific cognitive or pathological events, including motor execution, eye movement, or epileptic seizures. To extract these spectral characteristics on-chip under stringent power and area constraints, hardware-friendly mathematical approximations are used in place of computationally expensive continuous Fourier transforms.

In energy-constrained SoCs for BCI, spectral-domain feature extraction using a parallel bandpass filter (BPF) bank combined with energy integration is the premier paradigm for isolating physiological rhythms. By localizing and accumulating signal power within target sub-bands on-chip, this methodology achieves massive data compression, significantly reducing both wireless transmission bandwidth and overall system power [15].

The SoC proposed by Verma et al. (2010) provides an exemplary digital pipeline for on-chip spectral feature extraction [60]. Digitized intracranial EEG (iEEG) samples (12-bit, sampled at 600 Hz) are first down-sampled using a decimation filter (factor of 8) to reduce computational overhead. The decimated stream is then distributed to a parallel bank of seven modulated bandpass filters (BPF1–7) linearly spaced from 0 to 18 Hz with a center frequency spacing (Δf_c) of 3 Hz. Each filter possesses a narrow −3 dB bandwidth of 3 Hz and a stop-band attenuation of −25 dB. To eliminate the need for power-hungry hardware multipliers, each BPF is followed by a magnitude summation block ($\sum |x|$). The accumulated energy of each bin is represented as a 16-bit number computed over a sliding

two-second window (T_{int} = 2s). The seven 16-bit sub-band energies are multiplexed to generate a highly compressed 112-bit feature vector. By executing this feature extraction locally and transmitting only these feature vectors to a remote machine learning classifier (e.g., Linear SVM), the digital processor reduces wireless data transmission overhead, consuming merely 2.1 µW at 600 Hz.

In contrast to digital filtering, the EPIC (ECoG/EEG Processing Integrated Circuit) chip by Zhang et al. (2011) performs spectral decomposition entirely in the analog domain before digitization [15]. The band energy extractor for each channel consists of a fourth-order programmable switched-capacitor (SC) BPF, an analog multiplier that squares the signal, and a lossless SC integrator that sums the power over a 25ms window (T_{int} = 25 ms). The fourth-order BPF response is achieved by cascading two second-order SC biquad sections that are stagger-tuned to obtain a flat Butterworth frequency response across four independently selectable physiological bands: Alpha (8–12 Hz), Beta (18–26 Hz), low-Gamma (30–50 Hz), and Gamma (70–100 Hz). A major architectural advantage of EPIC's analog sub-banding approach is that it compresses the dynamic range of the neural signal prior to the ADC. By extracting the band power before digitization, the resolution requirement of the downstream ADC is relaxed by approximately 3 bits, significantly lowering the power consumption of the ADC drivers. However, this analog implementation demands a larger silicon footprint (2.8 mm^2 active area) due to the extensive on-chip capacitor banks required for SC tuning.

In the design of energy-efficient and highly reliable BCIs, Approximate Entropy (ApEn) serves as a vital non-linear time-domain biomarker to evaluate the randomness, complexity, and regularity of non-stationary neural time series. First formulated mathematically by S. M. Pincus, ApEn measures the logarithmic likelihood that runs of patterns that are close for a designated window length remain close on subsequent incremental observations [61].

The practical utility of this non-linear complexity extraction is demonstrated in the fully integrated closed-loop neural-prosthetic CMOS SoC developed by [49] for real-time epileptic seizure control. Digitized intracranial EEG (iEEG) signals from an on-chip 10-bit delta-modulated SAR (DMSAR) ADC are routed to a Biomedical Signal Processor (BSP). The BSP houses a hardware-efficient ApEn encoder alongside a power-efficient FFT core and a Linear Least Square (LLS) classifier [49].

3.2. Time and Spatial Domain Feature Extraction

In modern BCIs, raw neural recordings acquired from non-invasive or invasive electrodes suffer from low spatial resolution, poor signal-to-noise ratios (SNRs), and severe volume conduction artifacts. While traditional spectral analysis isolates specific neurological rhythms, it fails to capture the transient temporal dynamics and localized spatial variations crucial for decoders. Consequently, spatiotemporal and joint time-frequency representation (TFR) techniques are utilized to map multi-channel, non-stationary neural time series into compact, highly discriminative feature vectors. To achieve this under strict power and area constraints, BCI hardware architectures primarily leverage three dominant mathematical paradigms: Discrete Wavelet Transform (DWT) for multi-resolution time-frequency localization, Autoregressive (AR) Modeling for parametric temporal characterization, and Canonical Correlation Analysis (CCA) for spatial-frequency target identification.

The DWT is a powerful multi-resolution analysis tool that addresses the limitations of the classic Fourier transform by providing simultaneous localization in both the time and frequency domains. This temporal–spectral resolution makes DWT exceptionally well-suited for capturing transient, non-stationary neural anomalies such as epileptic spikes or motor imagery sensorimotor rhythms. The DWT decomposes a discrete neural signal $x[n]$ by passing it through a pair of complementary quadrature mirror filters—a low-pass filter

$g[n]$ and a high-pass filter $h[n]$—followed by dyadic down sampling. This decomposition yields approximate (coarse) coefficients $y_g[n]$ and detail coefficients $y_h[n]$, respectively [62]:

$$y_g[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[2n-k]; \quad y_h[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[2n-k] \tag{5}$$

In BCI processing (especially for motor imagery and epilepsy), the low-frequency approximate coefficients are preferred because they preserve the significant physiological energy bands while filtering out high-frequency somatic and electrode noise. DWT architectures can be realized using either convolution-based or lifting-based schemes. For wearable biomedical processors, the lifting-based scheme (using Haar or Daubechies D6 wavelets) is highly favored over convolution because it reduces arithmetic complexity, eliminates redundant multiplications, and cuts down register requirements, allowing efficient sharing of a central arithmetic core.

Autoregressive (AR) modeling is a parametric time-domain technique that approximates the dynamic characteristics of an EEG signal by treating it as the output of an all-pole linear time-invariant system. Instead of storing large segments of raw time series, the system represents the signal's temporal evolution using a compact vector of model coefficients, providing excellent data compression. An AR process of order m expresses a discrete-time neural sample $y(n)$ as a linear combination of its m immediate preceding samples plus a prediction error $e(n)$:

$$y(n) = -\sum_{i=1}^{m} a_i y(n-i) + e(n) \tag{6}$$

where a_i represent the AR coefficients (or Linear Prediction Coefficients-LPC), which serve as the feature vector. To estimate these coefficients on-chip, designers employ two paradigms: Non-Adaptive AR: The signal is divided into short stationary windows (typically a 1 s epoch), over which the coefficients are obtained either by solving the Yule–Walker equations with the Levinson–Durbin recursion or, more robustly for short epochs, with the Burg algorithm, which instead minimizes the sum of the forward and backward prediction-error powers in a lattice structure and guarantees a stable, minimum-phase model; Adaptive AR (AAR): Coefficients are updated recursively for every new sample point using Least Mean Squares (LMS) or Recursive Least Squares (RLS) algorithms, which bypasses the need for sample buffering and is ideal for real-time, low-latency applications. Although non-adaptive AR estimation is highly accurate, the Burg algorithm requires recursive multiplication, addition, and division blocks, placing a heavier burden on the central processing core. Conversely, adaptive AR (AAR) requires no buffer memory, making it highly compatible with ultra-low-power, streaming digital architectures.

Canonical Correlation Analysis (CCA) is a powerful multi-dimensional, spatial-frequency filtering method that serves as the gold standard for Steady-State Visual Evoked Potential (SSVEP) target identification. Rather than analyzing single-channel waveforms, CCA calculates the maximal correlation between a multi-channel EEG signal and predefined multi-harmonic reference templates, rendering it robust against background EEG activity and individual subject variability [9].

Given a multi-channel EEG data matrix $X \in \mathbb{R}^{N_p \times N_c}$ (where N_p is the number of samples and N_c is the number of channels) and a multi-harmonic reference matrix $Y \in \mathbb{R}^{N_p \times 2N_h}$ containing the sine–cosine templates of a stimulus frequency f, CCA seeks to find the spatial and frequency weight vectors W_X and W_Y that maximize the canonical correlation $\rho(x, y)$ between their linear combination [8]:

$$\max_{W_X, W_Y} \rho(x, y) = \frac{E\left[W_X^T X Y^T W_Y\right]}{\sqrt{E\left[W_X^T X X^T W_X\right] E\left[W_Y^T Y Y^T W_Y\right]}} \tag{7}$$

The resulting canonical correlation coefficients ρ are then combined (typically using an Euclidean norm) to classify the target stimulus. Integrating CCA onto edge-computing MCUs or accelerators (such as the parallel ultra-low-power BioWolf platform) presents a significant linear algebra bottleneck. The algorithm must be decomposed into QR factorizations (using Householder reduction to extract orthogonal matrices Q_x and Q_y) and Singular Value Decomposition (SVD) via the Golub–Reinsch algorithm or Givens rotations. Because SVD has tight sequential data dependencies, parallel architectures must utilize task-level parallelism (allocating individual target frequencies to separate processor cores) to maintain millisecond-level real-time latency within a micro-watt power envelope.

3.3. Deep Learning Architectures for BCI Decoding

3.3.1. Compact CNNs for Edge Decoding

Traditionally, BCI pipelines relied heavily on handcrafted features (such as CSP or wavelet coefficients) mapped to shallow classifiers. However, the advent of Deep Learning (DL) has revolutionized this field by enabling end-to-end learning directly from raw or minimally preprocessed EEG signals, thereby automating the extraction of hierarchical, highly discriminative spatiotemporal representations. To process multi-channel EEG time-series, which resemble 2D arrays of (channels × time), CNNs serve as the foundation of modern deep BCI modeling due to their weight-sharing and localized receptive fields. Three architectures are widely recognized as the standard baselines in the literature:

- ShallowConvNet: Developed by Schirrmeister et al., this architecture is designed specifically for classifying oscillatory brain signals (such as motor imagery sensorimotor rhythms). It mimics traditional Filter Bank Common Spatial Pattern (FBCSP) pipelines by employing consecutive temporal and spatial convolutions, followed by a squaring non-linearity, average pooling, and a logarithmic activation to directly extract log band-power features.
- DeepConvNet: Also proposed by Schirrmeister et al., DeepConvNet represents a general-purpose, deep-layered CNN. It uses multiple sequential convolutional and pooling layers to hierarchically construct abstract and complex spatial–temporal feature representations, making it highly effective at detecting subtle, multi-layered neural dynamics.
- EEGNet: Formulated by Lawhern et al., EEGNet is the premier compact and specialized CNN for multi-paradigm BCI classification. Its elegance lies in the strategic use of depthwise and separable convolutions. Specifically, Block 1 utilizes a 2D temporal convolution to act as a digital bandpass filter bank. Next, a depthwise convolution is applied along the channel axis (with filter size $C \times 1$) to learn frequency-specific spatial filters. Finally, separable convolutions are deployed to summarize temporal patterns within each feature map and optimally mix them together. This compact design drastically reduces trainable parameters, making it highly robust against overfitting on small BCI datasets.

Rather than operating as isolated, independent designs, contemporary state-of-the-art convolutional neural networks (CNNs) developed for brain–computer interfaces (BCIs) represent structural descendants, regularization variants, or hybrid extensions of these three foundational backbones. These ancestral architectures established the core neuroscientific and mathematical principles of decoupling temporal frequency filtering from spatial electrode mapping. Consequently, subsequent variations adapt these decoders by either introducing sparse group regularization to enhance clinical interpretability, deploying multi-branch pointwise convolutions to combat cross-subject variability, or fusing localized convolutional feature extractors with self-attention mechanisms to model global temporal dependencies. To provide a systematic taxonomic overview, Table 5 classifies and

synthesizes these landmark architectures, tracing their engineering lineages back to the three foundational archetypes.

Table 5. Comparison of pioneering CNN models in EEG decoding (compute figures quoted as published, under the convention of the sources).

CNN Model	Architectural Features	Strengths	Limitations & Design Trade-Offs	Target EEG/BCI Applications
EEGNet [63]	2D temporal convolutions; Depthwise spatial convolutions; Separable convolutions.	Compact parameter footprint (0.8 k–2.3 k params); Low risk of over-fitting; High generalization capacity.	Limited temporal dependency learning; Performance degrades significantly with excessive channel pruning.	Generalized BCI, P300, ERN, MRCP, SMR, ear-EEG SSVEP, and Motor Imagery.
TSGL-EEGNet [18]	EEGNet backbone integrated with a Temporally Constrained Sparse Group Lasso (TCSGL) regularization term.	Enhances feature sparsity; Mathematically equivalent to spatial filters like CSP; High interpretability.	High computational and search complexity during regularization parameter optimization.	Motor Imagery EEG classification.
ShallowConvNet [64]	Tailored for band-power features; Consecutive temporal/spatial CONV; Squaring and log activations; Input adapted to 6 s (600 samples).	Achieved 84.5% accuracy (90.5% specificity, 77.3% sensitivity) on TUH Abnormal Corpus; Directly extracts log band-power.	High parameter count compared to EEGNet; Poor generalization on non-oscillatory event-related potentials (ERPs).	Oscillatory signal classification, Motor Imagery (SMR), and clinical EEG pathology detection.
DeepConvNet [64]	5 temporal-spatial CONV blocks; ELU, BatchNorm, and Max Pooling; Pooling strides moved to CONV layer for 6 s input acceleration.	Achieved 85.4% accuracy (94.1% specificity, 75.1% sensitivity) on TUH Abnormal Corpus; Learns highly abstract hierarchical features.	Large parameter footprint (>150 k params); High overfitting risk; requires cropped training or massive datasets.	Multi-class neural state decoding, driver drowsiness, motor imagery, and clinical EEG pathology detection.
CS-CNN [65]	Maps raw scalp-space potentials to brain source-space via EEG Source Imaging (ESI) prior to convolutional layers.	Mitigates inter-subject and inter-session variability by projecting signals back to anatomical regions of interest (ROIs).	Prohibitive computational overhead due to the real-time inverse mathematical solver of ESI.	Motor Imagery EEG classification.
PAMCNet [66]	Multi-domain branches: PreTrain, Pre-active (spatial, frequency, joint time-frequency), and Multi-conv blocks.	Outstanding spatiotemporal-frequency mapping without raw feature engineering.	Complex multi-branch structure with parallel pathways, presenting on-device edge processing bottlenecks.	Motor Imagery decoding.
CNNViT-MILF-a [67]	Late-fusion (LF) strategy joining localized CNN feature maps (receptive fields of 3×3) with global ViT self-attention tokens.	Captures both local spatiotemporal correlations and long-range global dependencies; Outperforms advanced baseline models by +0.74% to +2.50%.	Prohibitive model complexity (13.2 M parameters, 7.55 GFLOPs, 6.85 G MACs) making edge deployment highly power-demanding.	High-fidelity Motor Imagery decoding, sleep stage monitoring, emotion recognition, and clinical rehabilitation.
SWCNet [68]	Points temporal convolutions integrated with a sliding-window data augmentation scheme.	Outstanding cross-subject generalization with extremely low inter-subject variability under subject-independent setups.	Focuses heavily on temporal pointwise convolutions, potentially ignoring complex non-linear spatial correlations.	Subject-independent and subject-dependent Motor Imagery (MI-EEG) decoding.

The dedicated processing engines exploit structural pruning, fixed-point quantization, and dynamic data-flow optimization to span a wide latency–power design space, from sub-millisecond inference on watt-class FPGA platforms to inference times of a few to a few tens of milliseconds on milliwatt-class ASICs, as quantified in Table 6. Two observations follow from the standardized metrics in this table. First, the reporting practice of the field is itself a limitation: only two of the six designs state an energy per inference directly, three report power without a supply voltage, and one reports power only as a relative improvement

over a desktop GPU, which is why several entries above had to be derived or marked as not reported. Second, once the metrics are placed on a common basis, the gap between platform classes becomes explicit: the 65 nm EEGNet ASIC needs 0.267 mJ per inference [12] and the Spartan-6 engine under 0.15 mJ [69], whereas the watt-class FPGA and ARM + FPGA deployments need 0.38 mJ [70] and 19.1 mJ [71] for the same class of network. Latency alone is therefore a misleading figure of merit, since the fastest design in the table is also one of the least energy-efficient; the energy per inference, together with the channel count that sets the input dimensionality, is the metric that determines whether a decoder can run continuously from a wearable battery.

While CNNs excel at capturing local spatial and short-term temporal features, they are inherently limited in their ability to model long-range sequential dynamics [72]. To overcome this constraint, researchers introduced hybrid networks that merge CNNs with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) blocks. Yi-Kai Chen, Jia-Yu Li, and Wai-Chi Fang (2023) developed a dedicated Edge AI hardware accelerator based on a hybrid Long-Term Recurrent Convolutional Network (LRCN) model. The implemented LRCN classifier features a multi-layered hybrid topology, comprising three temporal convolutional layers, three average pooling layers, a Long Short-Term Memory (LSTM) recurrent layer to model long-term temporal dependencies, a fully connected layer, and a softmax layer. The physical system was implemented on a Xilinx Kintex-7 (xc7k160t) FPGA platform, where a software–hardware co-design strategy was deployed. A RISC-V CPU core is responsible for executing real-time EEG signal preprocessing and feature extraction (using an 8–45 Hz bandpass filter and Short-time Fourier Transform), while the computationally intensive convolutions and LSTM matrix-vector multiplications are offloaded via a dedicated protocol to the custom AI hardware accelerator. To minimize silicon area and maximize hardware reusability for different input dimensions, the authors optimized the accelerator by integrating hardware sharing and advanced data-reuse datapaths. The custom accelerator block alone utilizes only 12,364 LUTs, 7544 FFs, and 20 BRAMs. When the entire system (including the RISC-V CPU and interconnects) is mapped, it consumes 29,438 LUTs, 16,313 FFs, and 122 BRAMs. Validated on the benchmark SEED dataset using only 8 EEG channels, the system achieved a high classification accuracy of 88.6% for two-class classification and up to 69.31% for three-class classification. Most notably, due to channel reduction and optimized parallel execution pathways, the hybrid processor achieves an exceptionally low latency of only 0.014 ms per input frame while drawing a modest 0.18 W of power, yielding an outstanding energy efficiency of 9.26 GOPS/W. This seminal study demonstrates that through meticulous algorithm–hardware co-design, highly complex hybrid sequential deep learning models like CNN–LSTMs can be tightly embedded on edge silicon for low-power, real-time BCI wearables.

Table 6. Standardized comparison of hardware-accelerated CNN implementations for EEG decoding. Metrics are reported as given by the original works; values marked "†" are derived as (power × latency) or from the reported throughput and energy efficiency, and "n.r." denotes not reported. Latency is normalized to milliseconds per inference.

Study & Year	Platform/ Technology	EEG ch.	Supply/ Power	Latency (ms/inf.)	Energy per Inference	Accuracy	Model, Key Strengths and Design Trade-Offs
Feng et al. (2022) [12]	ASIC, 65 nm CMOS (0.23 M gates)	n.r. (embedded channel selection)	n.r.; 10.97 mW	24.4	0.267 mJ	79.79% (ERN)	Compressed EEGNet (embedded channel selection, normalization merging, product quantization); PE/MAC reuse and clock gating give 87.22% smaller area and 20.77% better energy efficiency than Q-EEGNet on the Mr. Wolf RISC-V SoC; the hardwired datapath is EEGNet-specific and thus loses general programmability, and product quantization introduces a small precision loss.

Table 6. *Cont.*

Study & Year	Platform/ Technology	EEG ch.	Supply/ Power	Latency (ms/inf.)	Energy per Inference	Accuracy	Model, Key Strengths and Design Trade-Offs
Flood et al. (2022) [59]	FPGA, Xilinx Zynq UltraScale+ XCZU7EV	62	n.r.; 16.9× lower active power than an Nvidia GTX 1080	<2	n.r.	61.8% (54 subjects)	DeepConvNet quantized to a uniform 2-bit precision with Brevitas and compiled with the Xilinx FINN dataflow compiler; sub-2 ms latency is compatible with closed-loop control, but the absolute power is not reported and the design flow is tied to a specific vendor toolchain.
Cai et al. (2024) [70]	ARM + FPGA heterogeneous SoC (Zynq)	64 (SSVEP)	1.91 W	0.2	≈0.38 mJ †	93.3%/94.4% (SSVEP)	Quantized EEGNet with the ARM core handling dataflow and the programmable logic accelerating convolutions; 31.5× faster than a conventional CPU for only 0.7% accuracy loss, but the watt-level power is two orders of magnitude above custom ASICs and the achievable ITR is limited by the number of stimuli.
Pacini et al. (2024) [71]	FPGA, Xilinx UltraScale + ZU7EV (FPG-AI flow)	64	1.47 W	12.97	≈19.1 mJ †	77%	EEGNet deployed automatically with the technology-independent FPG-AI tool flow using 4-bit weights and 8-bit activations; 98% smaller memory footprint (476 Mb to 6.37 Mb) and 89%/71% lower power than the CPU/embedded-GPU baselines, but 96% of the clock cycles are spent in the first convolutional layer, so latency is higher than on the embedded GPU.
Gonzalez et al. (2020) [69]	FPGA, Xilinx Spartan-6 (Digilent Atlys)	14	≈0.15 W † (1.65 GOps at 11 GOps/W)	<1	<0.15 mJ †	77.57% (DEAP)	BioCNN, a handcrafted pipelined CNN engine with aggressive hardware parallelism and resource reuse that minimizes the memory footprint; latency is far below the 150 ms human-interaction bound, but accuracy remains modest relative to uncompressed offline models and the montage is limited to 14 channels.
Shi et al. (2025) [73]	ASIC, 40 nm CMOS, RISC-V core	4 (emotion)	0.7–1.0 V (logic), 0.9–1.1 V (SRAM)/ 180.5 mW;	3.23 (emotion); 1.82 (seizure)	n.r.	94.98% (4-channel emotion)	EEGNet accelerated through eight extended macro instructions (Custom-3 opcodes) that fuse dozens of operations and bypass memory bottlenecks; an approximate multiplier (HB = 8) reduces area by 28.80% and power by 28.34%, at the cost of truncation error, a 10 MHz clock ceiling and added complexity in cross-domain level shifting.

3.3.2. Spiking Neural Networks (SNNs) for Event-Driven Ultra-Low-Power Inference

To address the stringent power and area constraints of wearable and implantable brain–computer interfaces, SNNs have emerged as a premier lightweight neural network topology. Unlike conventional artificial neural networks that rely on continuous-valued activations and dense multi-bit MAC operations, SNN topologies exploit spatio-temporal signal sparsity to execute asynchronous, event-driven processing using hardware-efficient, single-bit accumulations (ACCs). Architecturally, edge-AI SNN topologies can be categorized into two structural paradigms based on their input interface and spiking depth:

(1) Partially Spiking Neural Networks (PSNNs): Conventional SNN conversions often construct PSNNs, where only hidden and output layers operate on discrete spikes, while the input layer still processes continuous analog or high-resolution multi-bit values via standard MAC units. To quantify the severe computational bottleneck of the first layer, the First Layer Quota (FLQ) assumes an exponential spike decay factor r ($0 < r < 1$) across an n-layer network as:

$$FLQ = \frac{1 - r}{1 - r^n} \tag{8}$$

Equation (8) demonstrates that the first layer handles the vast majority of total network computations, rendering PSNNs computationally inefficient for microwatt edge implants.

(2) Fully Spiking Neural Networks (FSNNs): To eliminate the first-layer MAC bottleneck, FSNN topologies integrate a temporal spike-encoding scheme prior to the first layer, allowing the entire network architecture—including input feature extractors—to operate strictly on binary spikes. Exemplifying this FSNN paradigm, Muneeb and Kassiri developed a customized FSNN that directly interfaces with asynchronous Level-Crossing ADCs (LC-ADCs) [74]. Raw EEG signals are digitized into sparse temporal spikes at the sensory boundary, bypassing continuous analog processing entirely. At the neuron level, the hardware implements an Integrate-and-Fire (IF) dynamic model, updating the membrane potential $V(t)$ as:

$$V(t+1) = V(t) + I(t) \tag{9}$$

where $I(t)$ denotes the incoming synaptic current. When $V(t+1)$ reaches the firing threshold θ, a spike is emitted, and the potential is reset via threshold subtraction ($V(t+1) - \theta$) to preserve inter-layer temporal synchronization. By parameterizing weight and bias quantization to 10-bit fixed-point precision, this lightweight FSNN topology compresses the model storage footprint down to a mere 9.75 KB (a 4-fold memory reduction) [74]. Synthesized in a standard 130 nm CMOS process at 1.8 V, the custom spiking CNN occupies an active area of 3.3 mm^2 and dissipates only 1.37 mW of dynamic power while maintaining an average sensitivity of 92.2% and specificity of 97.3% on the CHB-MIT EEG dataset with only a 1% sensitivity drop compared to its 32-bit floating-point baseline [74].

3.3.3. Hyperdimensional Computing (HDC) as a Non-Neural Lightweight Alternative

Alongside compact CNNs and SNNs, HDC has emerged as a third lightweight decoding paradigm that is particularly well matched to ultra-low-power silicon. Instead of learning real-valued weights by gradient descent, HDC maps each input feature to a pseudo-random, quasi-orthogonal binary vector of very high dimension D (typically $D = 1000$–$10{,}000$) drawn from an item memory, and it combines these hypervectors with three cheap, massively parallel operations: binding (element-wise XOR), which associates a channel identifier with its quantized feature value; bundling (element-wise majority), which superposes the bound vectors of all channels and time steps into a single vector; and permutation, which encodes temporal order [75,76]. Formally, an epoch with N_c channels is encoded as

$$\mathbf{H} = \left[\bigoplus_{t=1}^{T} \rho^t \left(\left[\bigoplus_{c=1}^{N_c} (\mathbf{ID}_c \otimes \mathbf{V}(x_{c,t})) \right] \right) \right], \tag{10}$$

where $\otimes$ denotes binding, $\bigoplus$ bundling with majority thresholding $[\cdot]$, ρ the permutation operator, $\mathbf{ID}_c$ the channel hypervector, and $\mathbf{V}(\cdot)$ the level hypervector of the quantized sample. Training amounts to bundling the encoded epochs of each class into a single prototype hypervector, and inference to selecting the prototype with the smallest Hamming distance. The hardware consequences are substantial: the entire datapath consists of XOR gates, popcount trees, and shift registers; there are no multipliers, no floating-point units, and no iterative gradient computation; the model memory is $K \times D$ bits for K classes; and single-pass, one-shot training can be executed on-chip, which directly addresses the non-stationarity problem discussed in Section 1.2.

These properties have been demonstrated on biosignals. HDC classifiers achieve short-latency, false-alarm-free seizure detection from long-term human iEEG recordings while using only binary operations, and an ensemble formulation additionally performs automatic electrode selection, which reduces both the acquisition and the computation load [77,78]; a tutorial treatment of the iEEG use case is given in [79]. At the circuit level, a standard-cell-memory-based configurable HDC accelerator operates at 5 µW for always-on smart sensing [80], and replacing the costly item memory with an on-the-fly cellular automaton yields 39.1 nJ per prediction on a >200-channel emotion-recognition workload,

a 4.9× energy-efficiency improvement (9.5× per channel) over the previous state-of-the-art HDC processor [81]. Because the encoder is dominated by memory access rather than arithmetic, HDC also maps naturally onto in-memory computing substrates [82], and its cheap incremental update rule enables in-sensor adaptive learning, as demonstrated in a wearable biosensing system that retrains itself on the device [83]. The trade-off is accuracy: for complex, low-signal-to-noise-ratio tasks such as multiclass motor imagery, HDC typically remains below well-tuned CNNs unless it is combined with a stronger feature extractor, so it is best viewed as the preferred choice for always-on detection and wake-up stages operating under a microwatt budget, with a CNN or Riemannian decoder invoked only when the detector fires.

3.4. Model Optimization for Edge Deployment

To enable the physical deployment of computationally intensive deep learning models—such as EEGNet or DeepConvNet—on highly resource-constrained wearable and implantable BCI edge nodes, model quantization has emerged as an indispensable optimization paradigm [12]. While neural networks are conventionally trained using high-precision 32-bit floating-point (FP32) representation to ensure gradient convergence, running FP32 inference on lightweight embedded platforms (such as MCUs or low-power FPGAs) introduces severe latency, memory, and power bottlenecks due to the limited availability or complete absence of dedicated floating-point units (FPUs) [84]. To bypass these hardware constraints, quantization compresses the model's parameters (weights and activations) by mapping continuous real-valued tensors into discrete, lower-precision fixed-point representations (typically 8-bit or 16-bit integers) [85]. The mathematically rigorous formulation of this transformation is established through an affine mapping of real-valued mathematical numbers r to quantized integer representations q:

$$r = S(q - Z) \tag{11}$$

where the positive real number S (the scale factor) represents the step size of the quantizer, and the integer Z (the zero-point) of the same data type as q corresponds exactly to the real value 0.0, thereby ensuring that essential operations like zero-padding do not introduce quantization noise. When executing a core matrix multiplication $r_3 = r_1 r_2$ in the quantized domain, this mapping allows the output quantized values q_3 to be calculated in an integer-arithmetic-only fashion. The mathematical formulation scales the discrete sum-of-products through a constant multiplier M, which is defined solely by the offline-precomputed quantization scales:

$$M = \frac{S_1 S_2}{S_3} \tag{12}$$

Because M empirically falls within the interval $(0, 1)$, it can be efficiently represented in a hardware-friendly normalized form $M = 2^{-n} M_0$ (where $M_0 \in [0.5, 1)$ is a fixed-point integer and n is a non-negative integer), which effectively replaces expensive floating-point arithmetic with highly efficient bit-shifts and integer multiplications on silicon. By shifting from FP32 to 8-bit integer (INT8) arithmetic, this scheme reduces the active model footprint and memory bandwidth by up to 4 times while accelerating inference by 2 to 4 times on standard hardware with negligible degradation in neural decoding accuracy. Furthermore, to bridge the performance gap caused by finite-precision arithmetic, developers utilize quantization-aware training (QAT) to simulate these rounding effects in the forward pass of training, allowing the model to adapt early and preserve near-identical classification performance upon physical edge deployment.

On resource-constrained wearable and implantable BCI edge platforms, to deploy multi-layered deep learning models, model optimization techniques are indispensable to overcome severe on-chip memory (BRAM) and processing power constraints. Network pruning

reduces hardware overhead by systematically cutting redundant synaptic connections or entire channels, thereby decreasing both the memory footprint required to store weights and the total number of arithmetic operations. While unstructured pruning introduces random, irregular sparsity that can challenge the execution efficiency of regular hardware datapaths, structured pruning—such as embedded channel selection—is highly favored in BCI SoCs to directly reduce input dimensionality and hardware complexity. Mathematically, the importance of each electrode channel ic is quantified using the l_1-norm of its corresponding spatial weights, sorting them to select the top C_s channels for subsequent retraining:

$$l_{1,ic} = \sum_{j} \|w_{ic}^{j}\| \tag{13}$$

where $w_{j,ic}$ represents the weight matrix associated with channel ic.

A third family of compression techniques, complementary to quantization and pruning, is knowledge distillation (KD). Rather than shrinking a trained network directly, KD transfers the behavior of a large, accurate teacher to a compact student that is small enough for the edge node by training the student on the softened class posteriors of the teacher [86]. With logits z_i and temperature τ, the teacher distribution $p_i^{\mathcal{T}} = \mathrm{softmax}(z_i^{\mathcal{T}}/\tau)$ carries the inter-class similarity structure that a one-hot label discards, and the student is optimized with:

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\,\mathcal{H}\big(y, p^{\mathcal{S}}\big) \;+\; \alpha\,\tau^2\,\mathcal{D}_{\mathrm{KL}}\big(p^{\mathcal{T}} \,\|\, p^{\mathcal{S}}\big), \tag{14}$$

where $\mathcal{H}$ is the cross-entropy with the ground-truth label y, $\mathcal{D}_{\mathrm{KL}}$ is the Kullback–Leibler divergence between teacher and student distributions, and α balances the two terms. From a hardware standpoint, KD has a property that quantization and pruning do not: it is entirely an offline training-time cost. The teacher is discarded after training, so the silicon area, memory footprint, and energy per inference are those of the student alone, and the student can be an architecture chosen for the target datapath rather than a pruned residue of the teacher. KD is therefore especially attractive for BCI, where the teacher may be a deep network, an ensemble, or a large pre-trained model that could never be deployed on a wearable node: teacher–student frameworks have been reported for epilepsy prediction on edge devices [87], and distilling large pre-trained EEG models into compact students is one of the routes discussed in Section 3.5 for bringing such models within a milliwatt budget. In practice, the three techniques are combined: distillation fixes the architecture and the accuracy target, structured pruning removes residual redundancy, and quantization-aware training maps the result onto the fixed-point datapath of the accelerator.

3.5. Beyond CNNs: Emerging Decoding Paradigms and Fair Evaluation for Edge BCIs

The preceding subsections reviewed decoders whose physical realizations are already mature, namely handcrafted spectral, temporal, and spatial feature extractors and compact deep networks mapped onto FPGAs and ASICs. In recent years, three developments have reshaped how EEG decoders are designed and compared, and they increasingly determine which algorithm is worth committing to silicon: (i) Riemannian-geometry decoders, which classify spatial covariance matrices directly on the manifold of symmetric positive-definite (SPD) matrices [88,89]; (ii) EEG foundation models (EEG-FMs), which are pre-trained in a self-supervised manner on large unlabeled corpora and then adapted to downstream tasks [90]; and (iii) open, large-scale benchmarking infrastructures, most notably the Mother of All BCI Benchmarks (MOABB) [91]. These three topics differ in nature—a decoder family, a modeling paradigm, and an evaluation methodology—but from the front-end-to-inference perspective of this survey, they raise a common question: what does each approach cost in memory, latency, and energy on a wearable or implantable platform, and how can this cost be weighed against decoding performance under fair and reproducible

conditions? This subsection addresses that question. Throughout, we explicitly distinguish results measured on physical hardware from efficiency claims based only on parameter counts, floating-point operations (FLOPs), or server-GPU timing, because the latter do not reliably predict the energy consumption of memory-constrained microcontrollers (MCUs) or custom accelerators.

3.5.1. Riemannian-Geometry Decoders: Formulation and Hardware Cost

Let $\mathbf{X}_i \in \mathbb{R}^{N_c \times T}$ denote a band-pass-filtered, zero-mean EEG epoch with N_c channels and T samples. Its spatial covariance matrix

$$\mathbf{P}_i = \frac{1}{T-1}\mathbf{X}_i\mathbf{X}_i^\top \in \mathcal{S}_{N_c}^{++} \tag{15}$$

lies on the SPD manifold $\mathcal{S}_{N_c}^{++}$, on which Euclidean averaging and distances are geometrically inappropriate [88,89]. The affine-invariant Riemannian distance between two SPD matrices is

$$\delta_R(\mathbf{P}_1, \mathbf{P}_2) = \left\|\log\left(\mathbf{P}_1^{-1/2}\mathbf{P}_2\mathbf{P}_1^{-1/2}\right)\right\|_F = \left[\sum_{n=1}^{N_c} \log^2 \lambda_n\right]^{1/2}, \tag{16}$$

where λ_n are the eigenvalues of $\mathbf{P}_1^{-1}\mathbf{P}_2$, and the Riemannian (Karcher) mean of a set of matrices is $\bar{\mathbf{P}} = \arg\min_{\mathbf{P}} \sum_i \delta_R^2(\mathbf{P}, \mathbf{P}_i)$. Two classifier families follow directly. The minimum distance to Riemannian mean (MDRM) classifier assigns an epoch to the class whose mean is closest,

$$\hat{y} = \arg\min_{k \in \{1,\ldots,K\}} \delta_R(\mathbf{P}, \bar{\mathbf{P}}_k), \tag{17}$$

whereas tangent-space (TS) methods project each matrix onto the tangent plane at a reference point $\mathbf{P}_{\text{ref}}$ (usually the global Riemannian mean) and vectorize it,

$$\mathbf{s}_i = \text{upper}\left(\log\left(\mathbf{P}_{\text{ref}}^{-1/2}\,\mathbf{P}_i\,\mathbf{P}_{\text{ref}}^{-1/2}\right)\right) \in \mathbb{R}^{N_c(N_c+1)/2}, \tag{18}$$

where upper$(\cdot)$ retains the upper-triangular entries with off-diagonal terms weighted by $\sqrt{2}$, so that any Euclidean linear classifier (LDA, SVM, or logistic regression) can subsequently be applied. On the four-class BCI Competition IV-2a dataset, Barachant et al. showed that tangent-space mapping combined with LDA raised the mean accuracy from 65.1% (multiclass CSP + LDA) to 70.2% without any explicit spatial filtering [88]. A large-scale study on 109 subjects subsequently confirmed the robustness of the simple MDRM baseline, which reached up to 81.5% for motor execution and 76.4% for motor imagery [92].

After training, all model parameters reduce to the offline-computed whitening matrix $\mathbf{P}_{\text{ref}}^{-1/2}$ and the weights of a linear classifier, so inference consists of band-pass filtering, covariance estimation, one symmetric eigenvalue decomposition (EVD) per frequency band for the matrix logarithm, and a dot product. The EVD dominates and scales as $O(N_c^3)$, which makes channel reduction far more rewarding than for CNNs: sensor-subset selection on the SPD manifold decreased classification time by 82% on average while improving accuracy [93], and LGL-BCI combines SPD channel selection with a geometric network of only 64.9 k parameters [94]. Because this kernel is shared with the QR/SVD engines already required by CCA-based SSVEP detection (Section 3.2), a single linear-algebra coprocessor, for instance, a parallel Jacobi/CORDIC eigenvalue architecture [95], could serve both decoder families.

Hardware implementations. The most complete embedded realization reported to date originates from the parallel ultra-low-power (PULP) ecosystem that also produced the BioWolf platform [9]. Starting from a multiscale temporal and spectral Riemannian classifier [96], Wang et al. first quantized a single-window multispectral Riemannian classifier (MRC) to mixed precision and parallelized it on the Mr. Wolf SoC, requiring 1.304 mJ and

33.39 ms per classification [97]. The refined implementation on the multi-core Vega SoC [98] employs 18 fourth-order Butterworth IIR bands realized as quantized second-order sections, Householder tridiagonalization for the EVD, and a combination of 8-, 16-, and 32-bit fixed- and floating-point representations. It attains 75.1% accuracy on the four-class BCI IV-2a task (76.4% with subject-specific classifiers), loses only 1–1.4% after quantization, and remains up to 4.1% more accurate than the state-of-the-art embedded CNN. Each classification requires 16.9 ms and 198 µJ, and continuous operation with 50% overlapping windows consumes only 85 µW [98]. These results demonstrate that a Riemannian decoder can match or exceed compact CNNs on the accuracy–energy trade-off at the MCU level. At the embedded-Linux level, a CSP plus tangent-space pipeline has been timed on Raspberry Pi and Jetson Nano boards at 0.08–0.3 s per decision with approximately 15 MB of memory [99]. Dedicated hardware for the closely related CSP pipeline, which shares the covariance and eigen-decomposition kernels, provides useful architectural precedents: an early Stratix-IV FPGA implementation consumed roughly 301 mJ per decision [98,100], whereas a recent register-transfer-level design of an IIR + CSP + SVM chain targeting a 45 nm ASIC flow reports 77% accuracy with a 0.63 ms response time and approximately 2.5 µJ per trial [101]. To the best of our knowledge, however, no ASIC dedicated to Riemannian tangent-space or MDRM decoding has yet been reported.

Geometric deep learning and on-device adaptation. Deep Riemannian networks stack SPD-preserving layers, such as the bilinear mapping $\mathbf{P} \mapsto \mathbf{W}^{\top}\mathbf{P}\mathbf{W}$, eigenvalue rectification, and a final matrix-logarithm layer [102]. Recent EEG-specific variants add manifold-aware channel selection [94] or self-attention built on the Bures–Wasserstein metric, which is better conditioned than the affine-invariant metric for ill-conditioned covariance matrices [103]. Because every layer requires an eigen-decomposition in both the forward and backward passes, the hardware cost of these networks remains unexplored, and none of the cited works reports an implementation beyond CPU or GPU platforms. In contrast, Riemannian geometry offers an exceptionally inexpensive mechanism for on-device adaptation. Re-centering each session or subject such that its reference becomes the identity, $\tilde{\mathbf{P}} = \mathbf{P}_{\text{ref}}^{-1/2}\mathbf{P}\,\mathbf{P}_{\text{ref}}^{-1/2}$, compensates for much of the inter-session shift [104], and the reference can be tracked online along the geodesic

$$\mathbf{P}_{\text{ref}} \leftarrow \mathbf{P}_{\text{ref}}^{1/2}\left(\mathbf{P}_{\text{ref}}^{-1/2}\,\mathbf{P}_{t}\,\mathbf{P}_{\text{ref}}^{-1/2}\right)^{\alpha}\mathbf{P}_{\text{ref}}^{1/2}, \qquad 0 < \alpha \leq 1, \tag{19}$$

which reuses the same EVD kernel as inference and requires neither gradient computation nor activation storage. Domain-specific SPD batch normalization extends this idea to deep Riemannian networks [105]. Adaptation is nevertheless not universally beneficial: on a 109-subject dataset, supervised, unsupervised, and re-biasing strategies did not outperform the baseline MDRM classifier [92].

3.5.2. EEG Foundation Models Toward the Edge

EEG foundation models learn general-purpose representations by pre-training an encoder f_θ on large unlabeled corpora, typically by reconstructing masked signal patches,

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}}\left\|\mathbf{x}_j - g_\phi\left(f_\theta(\mathbf{x}_{\setminus\mathcal{M}})\right)_j\right\|_2^2, \tag{20}$$

where $\mathcal{M}$ is the set of masked patch indices and g_ϕ is a lightweight decoder discarded after pre-training; the encoder is subsequently fine-tuned, linearly probed, or adapted with parameter-efficient methods. Representative models include Neuro-GPT [106], Neuro-BERT [107], the graph-enhanced GEFM [108], BIOT [109], LaBraM [110], EEGPT [111], and CBraMod [112]. A critical review of early EEG foundation models found that most adopt Transformer backbones trained with masked temporal reconstruction, that their

evaluations remain heterogeneous and largely limited, and that substantial scaling effects have yet to be demonstrated [90].

Why foundation models are difficult to place on silicon. Model size spans several orders of magnitude, from the order of 10^3 parameters for EEGNet (Table 5) to LaBraM variants comprising from 5.8 M to 369 M parameters [110]. Even more critical for edge hardware is the scaling with input length. When an epoch is tokenized into $L = N_c N_p$ spatio-temporal patches of embedding dimension d, self-attention requires $O(L^2 d)$ operations and $O(L^2)$ activation memory per layer, which rapidly exceeds the on-chip SRAM of wearable MCUs. State-space models such as Mamba replace attention with the linear recurrence

$$\mathbf{h}_t = \bar{\mathbf{A}}\,\mathbf{h}_{t-1} + \bar{\mathbf{B}}\,\mathbf{x}_t, \qquad \mathbf{y}_t = \mathbf{C}\,\mathbf{h}_t, \qquad \bar{\mathbf{A}} = \exp(\Delta\mathbf{A}), \;\; \bar{\mathbf{B}} \approx \Delta\mathbf{B}, \tag{21}$$

whose cost is $O(L\,d\,N)$ for a state size N and whose streaming memory footprint is constant in L, a property that matches continuous EEG processing on MCUs.

First measured deployment on an ultra-low-power MCU. Hardware evidence for EEG foundation models has emerged only very recently. FEMBA, a 7.8 M-parameter bidirectional-Mamba model pre-trained on more than 21,000 h of EEG, is, to our knowledge, the first state-space EEG foundation model deployed on a parallel RISC-V MCU, namely the nine-core GAP9 SoC operating at 370 MHz with 1.5 MB of L2 SRAM [113]. Quantization-aware training with 2-bit weights and 8-bit activations compresses the model from 7.8 MB in INT8 format to approximately 2 MB, whereas conventional post-training quantization degrades accuracy by roughly 30%. On GAP9, inference over a 5 s window requires 1.70 s and 75 mJ at an average power of 44.1 mW while achieving 81.95% balanced accuracy and an AUROC of 0.893 for abnormal-EEG detection on the TUAB corpus [113]. Table 7 places this result alongside the Riemannian implementations discussed above. The comparison highlights the current gap: whereas a Riemannian decoder classifies a motor-imagery trial in 16.9 ms at 198 µJ [98], a compressed foundation model requires seconds and tens of millijoules per window, that is, two to three orders of magnitude more energy, albeit for a different and considerably broader clinical task.

Table 7. Measured hardware implementations of emerging decoding paradigms. Tasks, datasets, and window lengths differ; the table is intended for an order-of-magnitude comparison of cost rather than as a ranking. "n.r.": not reported; "†" denotes a value derived as (power × latency).

Work	Decoder	Platform	Latency	Energy/ Power	Task (Performance)
Wang et al. (2021) [97]	Multispectral Riemannian (TS + SVM)	Mr. Wolf PULP SoC	33.39 ms	1.304 mJ	Four-class MI
Wang et al. (2021) [98]	Multispectral Riemannian (TS + SVM)	Vega PULP SoC	16.9 ms	198 µJ; 85 µW continuous	Four-class MI, BCI IV-2a (75.1%)
Shyu et al. (2023) [99]	CSP + tangent space + SVM	Raspberry Pi/ Jetson Nano	0.08–0.3 s	n.r. (15 MB memory)	MI, BCI IV-2a (78.55%)
Vasei et al. (2024) [101]	IIR + CSP + SVM (covariance/EVD)	RTL, 45 nm ASIC flow	0.63 ms	4 mW; ≈2.5 µJ †	MI (77%)
FEMBA (2026) [113]	Mamba EEG foundation model, 2-bit QAT	GAP9 RISC-V MCU	1.70 s per 5 s window	75 mJ; 44.1 mW	Abnormal EEG, TUAB (81.95% balanced accuracy)

Adaptation and compression paths. Because foundation models are pre-trained on clinical or laboratory corpora, adaptation to a wearer's electrode montage and physiology is the step most likely to be performed on the device. Parameter-efficient fine-tuning of a pre-trained EEGPT backbone has been proposed for lightweight sleep-stage classification, updating only a small fraction of the model parameters [114], and hardware-aware neural architecture search with explicit model-size constraints offers a complementary route toward compact, deployable decoders [115]. These mechanisms directly complement the compression techniques analyzed in Section 3.4. Finally, the evidence currently available does not support replacing compact decoders with foundation models in subject-specific wearable BCIs: reported evaluations of EEG foundation models remain heterogeneous and rarely include on-target latency or energy measurements [90], so their additional silicon cost is justified mainly for multi-task, cross-subject, or continuous clinical-monitoring scenarios in which their generality can be amortized.

3.5.3. Fair and Hardware-Relevant Evaluation: MOABB

Committing an algorithm to an ASIC is expensive and largely irreversible, so the decision should rest on evidence that generalizes beyond the one or two datasets on which a method was originally proposed. MOABB addresses this requirement by decomposing any analysis into a dataset, a paradigm (event selection, filtering, and epoching), an evaluation (within-session, cross-session, or cross-subject splitting together with the metric), and a scikit-learn-compatible pipeline [91]. Because datasets differ in channel count, trial number, and recording hardware, statistical tests are conducted within each dataset—a permutation-based paired t-test for fewer than 20 subjects and a Wilcoxon signed-rank test otherwise—and the resulting one-sided p-values are combined using Stouffer's method weighted by the square root of the number of subjects n_j:

$$Z = \frac{\sum_j w_j Z_j}{\sqrt{\sum_j w_j^2}}, \qquad w_j = \sqrt{n_j}, \qquad Z_j = \Phi^{-1}(1 - p_j). \tag{22}$$

The original analysis of 12 motor-imagery datasets comprising more than 250 subjects showed that many previously reported improvements did not generalize beyond the datasets on which they had been validated, and that tangent-space classification was the best of the tested pipelines for single-session decoding [91].

Riemannian versus deep decoders under fair conditions. A systematic MOABB comparison across 13 datasets and 80 subjects, covering motor imagery, P300, and SSVEP paradigms, found no statistically significant difference between EEGNet and tangent-space pipelines in within-session, cross-session, or cross-subject evaluation while observing that training EEGNet required substantially higher computational resources than the Riemannian pipelines [116]. For hardware designers, this result is significant: if decoding performance is comparable, the decoder with the smaller memory footprint, the simpler datapath, and the lower calibration cost should be preferred, and the resources thereby saved can be reinvested elsewhere in the signal chain. This conclusion is reinforced under the pseudo-online protocol of MOABB, which slides overlapping windows over the continuous recording and adds an idle class: there, classical Riemannian pipelines outperformed deep networks, and several offline rankings were reversed [117].

Toward hardware-aware benchmarking. Existing benchmarks report decoding scores obtained on workstation or GPU hardware and, at most, host-side timing. We therefore recommend that future edge-BCI studies report, alongside decoding scores obtained under a MOABB-compatible protocol, the window and hop lengths, the parameter count, the peak

memory, and the measured latency and energy per decision E_{dec} on the target platform. A useful combined figure of merit is the energy per transferred bit,

$$E_{\text{bit}} = \frac{E_{\text{dec}}}{B}, \tag{23}$$

where B is the number of bits transferred per decision, as given by the information transfer rate. This metric penalizes decoders that save energy at the expense of information throughput and allows Riemannian, CNN, and foundation-model decoders to be compared on a common physical scale.

3.5.4. Open Challenges and Hardware Roadmap

The preceding analysis suggests the following research directions.

- Dedicated linear-algebra accelerators for Riemannian decoding. No ASIC for tangent-space or MDRM decoding has been reported to date. Fixed-point symmetric EVD and matrix-function accelerators (logarithm, square root, and power), for instance based on parallel Jacobi/CORDIC schemes [95], could serve Riemannian decoders, CCA-based SSVEP detection, and online re-centering (Equation (19)) with a single co-processor. Their numerical robustness for the ill-conditioned covariance matrices produced by dry or in-ear electrodes remains an open issue [103].
- Hardware support for deep Riemannian networks. SPD networks [94,102,103] currently lack any hardware implementation, because each layer requires eigen-decompositions in both the forward and backward passes. Eigen-free approximations, such as iterative matrix square roots, are promising but have not yet been evaluated on silicon.
- Closing the gap between foundation models and sub-milliwatt budgets. The first measured foundation-model deployment operates at tens of milliwatts with a latency of 1.70 s per window [113], whereas dedicated CNN accelerators reach sub-millijoule energy per inference [12]. Further progress will likely require accelerators tailored to state-space recurrences, extreme low-bit quantization-aware training, distillation of foundation models into compact students, and hardware-aware architecture search [115].
- Low-cost on-device adaptation. Riemannian re-centering [104] and parameter-efficient fine-tuning [114] define a spectrum of adaptation mechanisms with markedly different memory and energy costs. Characterizing this trade-off on identical hardware and data is an essential step toward the adaptive "EEG-aware" SoCs discussed in Section 5.
- Hardware-aware benchmarks. Extending MOABB-style protocols [91,117] with on-target latency, memory, and energy measurements, and with metrics such as Equation (23), would allow algorithm and circuit designers to select decoders on the basis of evidence that jointly reflects decoding quality and silicon cost.

3.6. Cross-Cutting Trade-Off Analysis

Sections 3.1–3.5 compared designs within each layer of the pipeline. Placing them on common axes exposes the trade-off that ultimately governs the choice of an edge decoder, and it separates the two frontiers that bound it: the hardware frontier, defined by energy, latency, and silicon area, and the algorithmic frontier, defined by model complexity and achievable accuracy.

3.6.1. Energy, Latency, and Area Across Platform Classes

Figure 5 plots energy per inference against latency for the motor-imagery decoders in Tables 6 and 7, this being the only paradigm for which several implementations report both

quantities measured on real hardware; works that report power only in relative terms are necessarily excluded.

Three observations follow. First, the spread in energy exceeds three orders of magnitude, from 2.5 µJ for a synthesized CSP datapath in a 45 nm flow and 267 µJ for a 65 nm EEGNet ASIC to 19.1 mJ for the same class of network deployed on a watt-class FPGA, whereas the spread in accuracy across the same designs is only a few percentage points. Energy, not accuracy, is therefore the discriminating variable at the edge. Second, latency and energy are not monotonically related: the designs that occupy the lower-left corner of the figure are custom datapaths that are simultaneously fast and efficient, whereas the platforms that shorten latency through brute-force parallelism pay for it in watts, so reporting latency alone is misleading. Third, silicon area governs the same trade-off one layer below: the per-channel area of the acquisition front-ends in Table 3 falls from 0.044 mm^2 in 130 nm to 0.025 mm^2 in 65 nm and 0.0013 mm^2 in 28 nm, which is what permits the channel counts that in turn set the input dimensionality, and hence the energy, of the decoder that follows.

Because the four axes cannot be optimized independently, we suggest reporting the energy per correct decision normalized by the input dimensionality,

$$\mathrm{FoM} = \frac{E_{\mathrm{inf}}}{N_c\,(P - P_{\mathrm{chance}})}, \tag{24}$$

where E_{inf} is the measured energy per inference, N_c the number of input channels, P the decoding accuracy, and $P_{\mathrm{chance}} = 1/K$ for K classes. This figure of merit penalizes both an energy-hungry datapath and a decoder operating close to chance level, and it makes designs with different montages comparable. Applied to the entries for which all three quantities are published, it separates custom silicon from general-purpose programmable platforms by roughly two orders of magnitude.

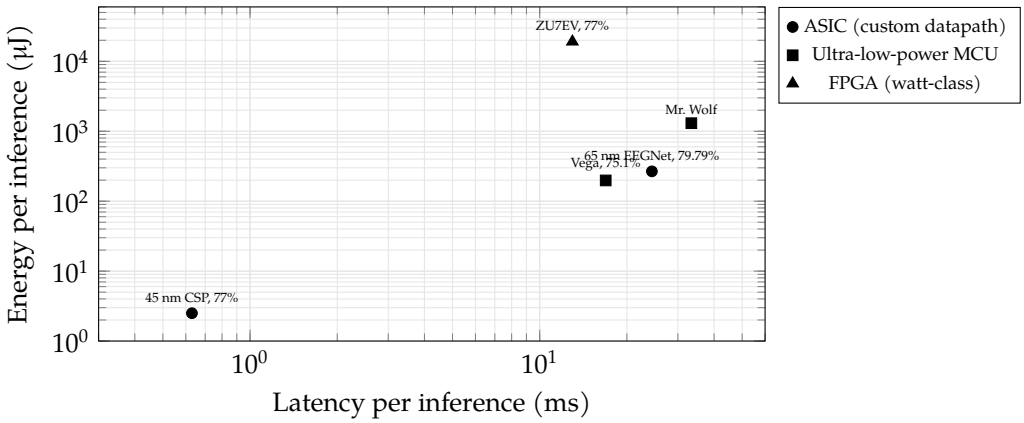


Figure 5. Energy per inference versus latency for motor-imagery decoders with metrics measured on the target hardware, from Tables 6 and 7. Markers are annotated with the platform and, where reported, the decoding accuracy. Entries whose energy was derived as (power × latency) are flagged in Table 6; designs reporting only a relative power improvement are omitted.

3.6.2. Accuracy–Complexity Frontier of Lightweight Models and Compression

The hardware frontier above is bounded from above by an algorithmic one. Figure 6 places the decoders surveyed in Sections 3.3.1–3.5 on axes of model complexity, measured as the number of parameters, and reported decoding accuracy.

The frontier exhibits three regimes. Below 10^4 parameters, compact convolutional decoders such as EEGNet, with 1 k to 2.3 k parameters (Table 5), reach the accuracy typical of motor-imagery decoding with a memory footprint that fits entirely in on-chip SRAM; this is the regime in which every ASIC of Table 6 operates. Between 10^4 and 10^5 parameters, geometric models trade a modest memory increase for a substantial accuracy gain: LGL-BCI

reaches 82.54% with 64.9 k parameters, against 62.22% for a baseline SPD network of 183.7 k parameters [94]. Above 10^6 parameters, foundation models occupy a regime in which the accuracy advantage does not scale with size for the data-scarce, subject-specific tasks that characterize wearable BCI, while memory and energy grow by three orders of magnitude, from the kilobytes of a compact CNN to the 2 MB of a 2-bit-quantized 7.8 M-parameter state-space model [113].

Compression moves a design horizontally along this frontier, and the reviewed works quantify its price in accuracy. Uniform 2-bit quantization of a deep convolutional decoder costs 2.5 percentage points with respect to the 32-bit baseline [59]; product quantization combined with channel selection and normalization merging costs 0.80 points [12]; 4-bit weights with 8-bit activations cost about 4 points relative to the floating-point CPU baseline while reducing the memory footprint by 98%, from 476 Mb to 6.37 Mb [71]; and a quantized heterogeneous ARM-plus-FPGA deployment costs 0.7 points [70]. For state-space foundation models, the frontier is discontinuous rather than smooth: post-training quantization degrades performance by roughly 30%, whereas quantization-aware training at 2-bit weights preserves it while compressing the model from 7.8 MB to about 2 MB [113]. Two design rules follow. Aggressive low-bit compression is viable only when it is folded into training rather than applied afterwards, and once a decoder is already below 10^4 parameters, structured reduction of the input dimensionality through channel selection is more effective than further weight compression because the cost of the acquisition chain, of the covariance estimation, and of the first convolutional layer all scale with the channel count [71,93,94].

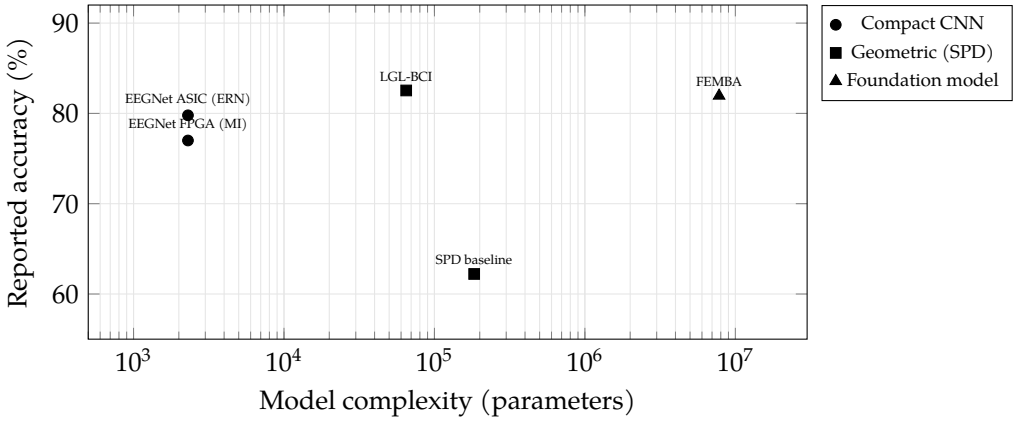


Figure 6. Accuracy–complexity frontier for edge BCI decoders with measured hardware deployments. Tasks and datasets differ between markers, so the figure indicates the regime occupied by each model class rather than a direct ranking; the values are those of Tables 5–7.

3.6.3. Deployment Constraints: Thermal Safety, Energy Autonomy, and Yield

Thermal safety. For implanted and body-worn electronics, power is bounded by tissue heating before it is bounded by the battery. Chronic heating studies place the safe steady-state heat flux at approximately 40 mW/cm^2, above which measurable tissue remodeling occurs, and report temperature rises below 1 °C at 15.5 mW/cm^2 against as much as 3 °C at 62 mW/cm^2 [118]. For a cortical implant, the finite-element model of an active Utah array validated in vivo predicts a temperature rise that grows linearly with dissipated power at 0.029 °C/mW [119], so the widely used 1–2 °C safety margin corresponds to a total budget of roughly 35–70 mW for the entire implanted system, acquisition, decoding and telemetry included.

Placed against the measured figures of Tables 6 and 7, this budget is decisive. The Riemannian classifier, operating at 85 µW [98], and the 65 nm EEGNet ASIC at about 11 mW [12] sit two to three orders of magnitude and a factor of three below it, respectively;

a quantized state-space foundation model at 44.1 mW [113] consumes the whole implant budget by itself and is therefore realistic only for a body-worn, not an implanted, node; and the watt-class FPGA deployments exceed it by more than an order of magnitude, which excludes them from any implanted or skin-contact scenario irrespective of their latency. For wearable devices, the corresponding constraint is the limit that IEC 60601-1 places on the temperature of parts in prolonged skin contact, which for a sealed plastic enclosure of a few square centimeters translates into a comparable, though less stringent, ceiling of a few hundred milliwatts.

Energy autonomy. Where thermal safety bounds instantaneous power, battery capacity bounds its integral. For a node that acquires N_c channels and issues f decisions per second, the autonomy is

$$T_{\mathrm{op}} = \frac{E_{\mathrm{batt}}}{N_c P_{\mathrm{ch}} + f\, E_{\mathrm{inf}} + P_{\mathrm{static}}}, \tag{25}$$

where E_{batt} is the usable battery energy, P_{ch} the acquisition power per channel from Table 3, E_{inf} the energy per inference from Tables 6 and 7, and P_{static} the power drawn by the platform when it is not computing. Table 8 evaluates Equation (25) for a 100 mAh lithium-polymer cell at 3.7 V, that is 370 mWh, of the size used in earbuds and headbands, at one decision per second.

The result spans more than three orders of magnitude and reorders the designs relative to a ranking by latency or accuracy. A continuously operating Riemannian classifier on an ultra-low-power MCU runs for roughly 180 days on this cell, whereas the same class of network on a watt-class FPGA exhausts it in about a quarter of an hour. The 65 nm EEGNet ASIC and the synthesized 45 nm CSP datapath consume so little per decision, 0.267 mJ and 2.5 µJ, respectively, that their duty-cycled decoding power falls below that of the acquisition chain feeding them: sixty-four channels at the 2.3 µW per channel of a state-of-the-art front-end already amount to 147 µW. Beyond this point, further decoder optimization is futile, and the system-level bottleneck migrates back to the analog front-end, the radio, and ultimately battery self-discharge. Two consequences follow. First, on programmable platforms, it is P_{static} rather than E_{inf} that dominates, since an FPGA or an application processor cannot be power-gated between decisions as a custom accelerator or a duty-cycled MCU can; the watt-level entries of Table 8 should therefore be read as a property of the platform, not of the network it executes. Second, the cost of a foundation model lies in its duty cycle rather than in its energy per window: at 75 mJ per 5 s window, the quantized state-space model averages about 15 mW and yields roughly one day of continuous monitoring, which suits clinical use but not an always-on consumer device unless inference is gated by a low-power detector such as the spiking or hyperdimensional front-ends of Sections 3.3.2 and 3.3.3.

Yield and device variability. A third constraint is rarely quantified in the literature surveyed here. Published front-ends and accelerators almost always report a single measured die, or a single FPGA build, without the distribution across wafers, so the yield implications of the architectures compared in Table 3 cannot be assessed from the published record. This omission matters because the trends identified above act directly against yield: reducing the supply voltage towards 0.4 V pushes the input pair into weak inversion, where the mismatch-induced offset and noise spread widen rather than narrow; and increasing the channel count to 1024 multiplies by the same factor the number of analog cells that must all meet specification on a single die [24]. Existing designs already absorb part of this variability at the circuit level, through per-channel DC-servo loops and offset-canceling digital-to-analog converters [23] and through digitally assisted calibration, but these mechanisms consume area and power that are seldom attributed to yield in the reported budgets. We therefore consider the reporting of per-channel yield, of calibration overhead, and of

the distribution of key metrics across dies to be a necessary complement to the single-die figures that currently populate the comparison tables of this field.

Table 8. Battery autonomy predicted by Equation (25) for a 100 mAh, 3.7 V cell (370 mWh) at one decision per second, using the measured figures of Tables 6 and 7. "Duty-cycled" assumes the decoder is power-gated between decisions; "always-on" assumes the platform remains powered. Acquisition, radio, and self-discharge are excluded.

Decoder/ Platform	Energy per Inference	Average Power at 1 s^{-1}	Autonomy (Duty-Cycled)	Autonomy (Always-On)
CSP datapath, 45 nm ASIC flow [101]	2.5 µJ	2.5 µW	>1 year (acquisition-limited)	4 mW: ≈4 days
Compressed EEGNet, 65 nm ASIC [12]	0.267 mJ	0.267 mW	≈58 days	—
Riemannian MRC, Vega MCU [98]	198 µJ	85 µW (continuous, 50% overlap)	≈180 days	—
Riemannian MRC, Mr. Wolf MCU [97]	1.304 mJ	1.30 mW	≈12 days	—
Mamba foundation model, GAP9 MCU [113]	75 mJ per 5 s window	≈15 mW (34% duty cycle)	≈1 day	—
EEGNet, ARM + FPGA SoC [70]	0.38 mJ	0.38 mW	≈40 days	1.91 W: ≈12 min
EEGNet, ZU7EV FPGA [71]	19.1 mJ	19.1 mW	≈19 h	1.47 W: ≈15 min

3.6.4. On-Chip Adaptation: Quantifying Drift, Gain, and Update Cost

Non-stationarity is the motivation invoked for on-chip adaptive learning throughout this survey, yet the argument is usually made qualitatively. Three quantities decide whether adaptation is worth its silicon: the magnitude of the drift it must compensate, the accuracy it recovers, and the cost of one update.

How large is the drift? Reproducible benchmarking makes this measurable. Under the MOABB protocol, moving from within-session to cross-session evaluation on the same four-class motor-imagery dataset lowers the normalized Matthews correlation coefficient of a minimum-distance-to-mean classifier from 0.81 to 0.59 and that of an augmented tangent-space SVM from 0.86 to 0.73 [117]. A loss of 0.13–0.22 between a calibrated session and the next one is far larger than the differences between the decoder families compared in Section 3.6.2, and larger still than the 0.8–4 percentage points that quantization costs (Section 3.6.2). For a deployed device, adaptation is therefore not a refinement but the dominant accuracy lever.

How much is recovered, and at what cost? The adaptation mechanisms reviewed here span four orders of magnitude in update cost. At the cheapest end, Riemannian re-centering replaces the reference matrix by the geodesic update of Equation (19), which reuses the eigenvalue decomposition already performed for inference, stores one $N_c \times N_c$ matrix, and requires neither labels nor gradients [104]. Hyperdimensional classifiers occupy

the same regime: a class prototype is updated by bundling a single encoded example, which is why in-sensor systems can retrain themselves on the device [83]. In the middle, parameter-efficient fine-tuning of a pre-trained backbone updates only a small fraction of the weights and therefore avoids storing optimizer state for the whole model [114]. At the expensive end, full back-propagation requires storing activations for every layer, which is what dedicated adaptive-learning silicon addresses: the reconfigurable biomedical processor of [13] integrates an on-chip adaptive-learning architecture specifically to absorb patient-to-patient variation, in 65 nm and within a wearable power budget, and reports the lowest energy per classification among the comparable single-task designs.

What the literature does not yet report. Adaptation is not universally beneficial, and the evidence is mixed rather than absent: on 109 subjects, supervised, unsupervised, and re-biasing strategies failed to outperform the baseline minimum-distance classifier [92], whereas domain-specific normalization of the covariance statistics does improve unsupervised transfer for deep Riemannian networks [105]. What is systematically missing from the hardware literature is the third quantity, convergence. Almost no implementation reports how many trials or how much time separate the first adapted decision from a stabilized one, nor the energy integrated over that interval, and none of the designs in Tables 6 and 7 report an accuracy-versus-time curve under session-to-session drift. Without those numbers, an adaptive engine cannot be compared with the alternative strategy that part of the field has adopted instead, namely eliminating retraining altogether through subject-independent decoders. We therefore recommend that studies claiming on-chip adaptation report, at a minimum, the number of labeled or unlabeled trials required to reach steady-state accuracy, the memory and energy of a single update relative to one inference, and the residual accuracy gap with respect to a fully recalibrated model.

4. Daily Life Applications and Clinical Case Studies

To transition brain–computer interfaces from constrained laboratory environments into unconstrained daily life applications, hardware development has prioritized unobtrusive, user-friendly form factors. A prime exemplar is the BioWolf platform, which integrates multi-channel acquisition, on-board decoding and wireless connectivity into a single miniaturized, battery-powered device so that a complete BCI can be worn during normal activity rather than kept at a desk [9]. Similarly, the in-ear BCI platform proposed by Lee et al. is housed within a personalized earplug, utilizing ear-canal anatomy to acquire robust auditory-evoked potentials [7]. Operating in a dual-mode configuration, the controller lets the wearer wake it from ultra-low-power sleep with an intentional eye blink and make binary selections while it sits in the ear [7]. Furthermore, the dry-electrode framework developed by Na et al. eliminates conductive gel preparation, with the control of smart-home appliances and Internet-of-Things devices among the daily life scenarios it targets [8].

For individuals recovering from severe stroke, spinal cord injury (SCI), or neurodegenerative disorders, translating voluntary motor intent into mechanical actuation is essential for restoring functional independence. Motor-imagery control of motorized wheelchairs is among the functions that such wearable decoding engines are built to support. In upper-limb neuroprosthetics, the RAMAN tinyML accelerator decodes continuous hand kinematics rather than discrete movement classes, demonstrated on recordings from macaque motor cortex [58]. Similarly, the MasterMind SoC brings onto the device the domain-adaptation and distribution-alignment computation on which a motor-cortex decoder depends, mapping primary motor cortex (M1) activity into control commands [120]. To actively promote functional recovery, the protocol designed by Benzy et al. decodes imagined hand-movement direction in order to drive a motorized arm support for stroke rehabilitation [121]. Pairing

central intent with peripheral mechanical actuation in this way is the rationale of such protocols, whose aim is to reinforce sensorimotor loops across consecutive therapeutic sessions.

In deep-tissue clinical interventions, bi-directional implantable SoCs are transforming the treatment of chronic neurological disorders [122]. The closed-loop neuroprosthetic SoC realized by Chen et al. monitors intracranial EEG and closes the loop on chip, detecting seizure onset and triggering localized electrical stimulation in response; the loop has been validated in animal experiments rather than in patients [49]. Concurrently, bi-directional cortical chipsets such as the architecture presented by Pu et al.—which combines recording with a configurable, charge-balanced stimulator on one die [51]—are intended to shape electric fields across motor and sensory cortices for the restoration of function after spinal cord injury. For cognitive and psychiatric monitoring, wearable processors such as the BioCNN inference engine [69] and the LRCN accelerator [72] decode affective states directly from non-invasive EEG streams, so far on benchmark emotion datasets rather than with patients.

At the bedside, integrating BCIs with physical neurorehabilitation is establishing new standards for active functional recovery. Passive rehabilitation, wherein a physical therapist manually mobilizes a paretic limb without active patient engagement, yields limited neuroplastic reorganization. In contrast, the Neuro-AI Therapy System developed by Amudha et al. detects movement initiation from non-invasive EEG on the device so that stimulation can be delivered in synchrony with the patient's own intent [123]. For motor-imagery decoding developed with rehabilitation in view, the hybrid CNN–ViT framework of Zhao et al., produced in collaboration with a hospital rehabilitation department, improves decoding accuracy on public and self-collected motor-imagery datasets, although it is evaluated offline rather than in a deployed system [67].

5. Conclusions

The migration of brain–computer interfaces from server-tethered laboratory instrumentation to autonomous, daily life devices is, at its core, a hardware problem. This survey has traced the complete path from microvolt-level cortical biopotentials to their decoding in silicon, and the quantitative comparisons assembled along that path support four conclusions.

First, the sensing layer sets the boundary conditions for everything that follows. Wearable form factors such as in-ear and dry-contact arrays are viable only because low-noise front-ends suppress electrode DC offsets of hundreds of millivolts and $1/f$ noise while preserving fidelity, and the measured record shows what this costs: supply voltages have fallen from 3 V in 0.5 µm designs to 0.4–1.2 V in deep-submicron ones, power per channel from tens of microwatts to 268 nW, and area per channel from 0.044 mm^2 to 0.0013 mm^2, while CMRR above 100 dB remains the property that separates non-invasive from area-optimized implanted front-ends. The electrode choice propagates directly into these specifications, since contact impedance differs by two to three orders of magnitude between wet and dry interfaces and degrades primarily under motion, which is the condition a daily life system must survive.

Second, at the computational layer, no single decoder family dominates. Handcrafted spectral and spatial features remain attractive where the paradigm is narrow; compact convolutional networks, spiking networks, and hyperdimensional classifiers occupy the sub-10^4-parameter regime in which every ASIC reviewed here operates; Riemannian decoders match compact CNNs under fair, large-scale benchmarking at a fraction of the calibration cost; and foundation models, which have only very recently been demonstrated on an ultra-low-power microcontroller, buy generality at the price of three orders of magnitude in memory and energy without a commensurate accuracy gain in the data-scarce, subject-specific setting that characterizes wearable BCIs. Compression moves a design along

this frontier at a measurable price, from 0.7 to about 4 percentage points of accuracy for the quantization schemes reviewed, provided that aggressive low-bit compression is folded into training rather than applied afterwards.

Third, once the reviewed implementations are placed on common axes, energy rather than accuracy emerges as the discriminating variable. Energy per inference spans more than three orders of magnitude across platform classes, while accuracy varies by a few percentage points, latency and energy are not monotonically related, and the resulting power figures map directly onto the physical limits of deployment: a 35–70 mW thermal budget for an implanted system, and a battery autonomy that ranges from months for a microwatt-class Riemannian decoder to minutes for a watt-class programmable platform. For the most efficient custom accelerators, the decoding power already falls below that of the acquisition chain feeding them, which means that further decoder optimization is futile and that the system-level bottleneck has migrated back to the analog front-end, the radio, and the battery.

Finally, the frontier lies in adaptation. Reproducible benchmarking shows that moving from a calibrated session to the next one costs far more accuracy than any choice of decoder family or quantization scheme, so static, offline-trained decoders will inevitably degrade in chronic use. The mechanisms available span four orders of magnitude in update cost, from gradient-free Riemannian re-centering and one-shot hyperdimensional updates to parameter-efficient fine-tuning and full on-chip back-propagation, but their convergence behavior is almost never reported. These constraints are not abstract: the deployments surveyed in Section 4—neuroprosthetic control, rehabilitation robotics, closed-loop seizure control and cognitive-state monitoring—already operate under them, and it is in chronic use that the adaptation gap becomes a clinical rather than a benchmarking problem.

Limitations. Three caveats bound the comparisons on which these conclusions rest, and all three are properties of the published record rather than of the designs themselves: energy per inference is stated directly by only a minority of the works compared here, so several entries in our tables had to be derived; the front-ends and accelerators surveyed almost always report a single measured die or a single build, without the distribution across wafers that a yield assessment would require; and no implementation reports an accuracy-versus-time curve under session-to-session drift. The rankings presented here should therefore be read as the best available ordering of a record that the field does not yet report uniformly.

Closing that gap, together with the reporting of on-target latency, energy, memory, and yield alongside decoding scores, is what will turn adaptive, "EEG-aware" systems-on-chip from passive decoders into dependable long-term companions.

Author Contributions: Conceptualization, T.-A.T., T.-H.T. and C.-K.P.; methodology, T.-A.T. and T.-H.T.; software, T.-A.T. and T.-H.T.; validation, T.-A.T., T.-H.T. and C.-K.P.; formal analysis, T.-A.T. and T.-H.T.; investigation, T.-A.T. and T.-H.T.; resources, C.-K.P.; data curation, T.-A.T. and T.-H.T.; writing—original draft preparation, T.-A.T.; writing—review and editing, T.-A.T., T.-H.T. and C.-K.P.; visualization, T.-A.T.; supervision, C.-K.P.; project administration, C.-K.P.; funding acquisition, C.-K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST NEXUS, Japan, Grant Number JPMJNX25D4.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the reported results are available from the corresponding authors on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Wolpaw, J.; Birbaumer, N.; Heetderks, W.; McFarland, D.; Peckham, P.; Schalk, G.; Donchin, E.; Quatrano, L.; Robinson, C.; Vaughan, T. Brain-computer interface technology: A review of the first international meeting. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 164–173. [CrossRef] [PubMed]
2. Donchin, E.; Spencer, K.; Wijesinghe, R. The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 174–179. [CrossRef] [PubMed]
3. Pfurtscheller, G.; Neuper, C. Motor imagery and direct brain-computer communication. *Proc. IEEE* **2001**, *89*, 1123–1134. [CrossRef]
4. Mellinger, J.; Schalk, G. BCI2000: A General-Purpose Software Platform for BCI. In *Toward Brain-Computer Interfacing*; Dornhege, G., Millán, J.d.R., Hinterberger, T., McFarland, D.J., Müller, K.R., Eds.; MIT Press: Cambridge, MA, USA, 2007; pp. 359–368. [CrossRef]
5. Grozea, C.; Voinescu, C.; Fazli, S. Bristle-sensors—Low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications. *J. Neural Eng.* **2011**, *8*, 025008. [CrossRef] [PubMed]
6. von Lühmann, A.; Wabnitz, H.; Sander, T.; Müller, K.R. M3BA: A Mobile, Modular, Multimodal Biosignal Acquisition Architecture for Miniaturized EEG-NIRS-Based Hybrid BCI and Monitoring. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1199–1210. [CrossRef] [PubMed]
7. Lee, J.; Lee, K.R.; Ha, U.; Kim, J.H.; Lee, K.; Gweon, S.; Jang, J.; Yoo, H.J. A 0.8-V 82.9-µW In-Ear BCI Controller IC with 8.8 PEF EEG Instrumentation Amplifier and Wireless BAN Transceiver. *IEEE J. Solid-State Circuits* **2019**, *54*, 1185–1195. [CrossRef]
8. Na, R.; Zheng, D.; Sun, Y.; Han, M.; Wang, S.; Zhang, S.; Hui, Q.; Chen, X.; Zhang, J.; Hu, C. A Wearable Low-Power Collaborative Sensing System for High-Quality SSVEP-BCI Signal Acquisition. *IEEE Internet Things J.* **2022**, *9*, 7273–7285. [CrossRef]
9. Kartsch, V.; Tagliavini, G.; Guermandi, M.; Benatti, S.; Rossi, D.; Benini, L. BioWolf: A Sub-10-mW 8-Channel Advanced Brain–Computer Interface Platform with a Nine-Core Processor and BLE Connectivity. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *13*, 893–906. [CrossRef] [PubMed]
10. Mathewson, K.E.; Harrison, T.J.L.; Kizuk, S.A.D. High and Dry? Comparing Active Dry EEG Electrodes to Active and Passive Wet Electrodes. *Psychophysiology* **2017**, *54*, 74–82. [CrossRef] [PubMed]
11. Oliveira, A.S.; Schlink, B.R.; Hairston, W.D.; König, P.; Ferris, D.P. Proposing Metrics for Benchmarking Novel EEG Technologies Towards Real-World Measurements. *Front. Hum. Neurosci.* **2016**, *10*, 188. [CrossRef] [PubMed]
12. Feng, L.; Shan, H.; Zhang, Y.; Zhu, Z. An Efficient Model-Compressed EEGNet Accelerator for Generalized Brain-Computer Interfaces with Near Sensor Intelligence. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *16*, 1239–1249. [CrossRef] [PubMed]
13. Liu, J.; Fan, J.; Zhong, Z.; Qiu, H.; Xiao, J.; Zhou, Y.; Zhu, Z.; Dai, G.; Wang, N.; Liu, Q.; et al. An Ultra-Low Power Reconfigurable Biomedical AI Processor with Adaptive Learning for Versatile Wearable Intelligent Health Monitoring. *IEEE Trans. Biomed. Circuits Syst.* **2023**, *17*, 952–967. [CrossRef] [PubMed]
14. Byun, W.; Je, M.; Kim, J.H. Advances in Wearable Brain-Computer Interfaces from an Algorithm-Hardware Co-Design Perspective. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *69*, 3071–3077. [CrossRef]
15. Zhang, F.; Mishra, A.; Richardson, A.G.; Otis, B. A Low-Power ECoG/EEG Processing IC with Integrated Multiband Energy Extractor. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2011**, *58*, 2069–2082. [CrossRef]
16. Zhong, Y.; Wang, Y.; Farina, D.; Yao, L. A Closed-Loop Tactile Stimulation Training Protocol for Motor Imagery-Based BCI: Boosting BCI Performance for BCI-Deficiency Users. *IEEE Trans. Biomed. Eng.* **2025**, *72*, 3019–3029. [CrossRef] [PubMed]
17. Milanés Hermosilla, D.; Trujillo Codorniú, R.; López Baracaldo, R.; Sagaró Zamora, R.; Delisle Rodriguez, D.; Llosas Albuerne, Y.; Álvarez, J.R.N. Shallow Convolutional Network Excel for Classifying Motor Imagery EEG in BCI Applications. *IEEE Access* **2021**, *9*, 98275–98286. [CrossRef]
18. Deng, X.; Zhang, B.; Yu, N.; Liu, K.; Sun, K. Advanced TSGL-EEGNet for Motor Imagery EEG-Based Brain-Computer Interfaces. *IEEE Access* **2021**, *9*, 25118–25130. [CrossRef]
19. Maria Muna, U.; Mehedi Hasan Shawon, M.; Jobayer, M.; Akter, S.; Rahman Sabuj, S. SSTAF: Spatial-Spectral-Temporal Attention Fusion Transformer for Motor Imagery Classification. *IEEE Access* **2025**, *13*, 158855–158869. [CrossRef]
20. Lin, B.S.; Lin, B.S.; Yen, T.H.; Hsu, C.C.; Wang, Y.C. Design of Wearable Headset with Steady State Visually Evoked Potential-Based Brain Computer Interface. *Micromachines* **2019**, *10*, 681. [CrossRef] [PubMed]
21. Nakanishi, M.; Wang, Y.; Chen, X.; Wang, Y.T.; Gao, X.; Jung, T.P. Enhancing Detection of SSVEPs for a High-Speed Brain Speller Using Task-Related Component Analysis. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 104–112. [CrossRef] [PubMed]

22. Smith, W.A.; Mogen, B.J.; Fetz, E.E.; Sathe, V.S.; Otis, B.P. Exploiting Electrocorticographic Spectral Characteristics for Optimized Signal Chain Design: A 1.08 W Analog Front End with Reduced ADC Resolution Requirements. *IEEE Trans. Biomed. Circuits Syst.* **2016**, *10*, 1171–1180. [CrossRef] [PubMed]
23. Muller, R.; Le, H.P.; Li, W.; Ledochowitsch, P.; Gambini, S.; Bjorninen, T.; Koralek, A.; Carmena, J.M.; Maharbiz, M.M.; Alon, E.; et al. A Minimally Invasive 64-Channel Wireless µECoG Implant. *IEEE J. Solid-State Circuits* **2015**, *50*, 344–359. [CrossRef]
24. Jang, M.; Hays, M.; Yu, W.H.; Lee, C.; Caragiulo, P.; Ramkaj, A.T.; Wang, P.; Phillips, A.J.; Vitale, N.; Tandon, P.; et al. A 1024-Channel 268-nW/Pixel 36×36 µm^2/Channel Data-Compressive Neural Recording IC for High-Bandwidth Brain–Computer Interfaces. *IEEE J. Solid-State Circuits* **2024**, *59*, 1123–1136. [CrossRef] [PubMed]
25. Xu, J.; Mitra, S.; Van Hoof, C.; Yazicioglu, R.F.; Makinwa, K.A.A. Active Electrodes for Wearable EEG Acquisition: Review and Electronics Design Methodology. *IEEE Rev. Biomed. Eng.* **2017**, *10*, 187–198. [CrossRef] [PubMed]
26. Tanveer, M.A.; Khan, M.J.; Qureshi, M.J.; Naseer, N.; Hong, K.S. Enhanced Drowsiness Detection Using Deep Learning: An fNIRS Study. *IEEE Access* **2019**, *7*, 137920–137929. [CrossRef]
27. Aricò, P.; Borghini, G.; Di Flumeri, G.; Sciaraffa, N.; Colosimo, A.; Babiloni, F. Passive BCI in Operational Environments: Insights, Recent Advances, and Future Trends. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 1431–1436. [CrossRef] [PubMed]
28. Wang, J.; Wang, T.; Liu, H.; Wang, K.; Moses, K.; Feng, Z.; Li, P.; Huang, W. Flexible Electrodes for Brain–Computer Interface System. *Adv. Mater.* **2023**, *35*, 2211012. [CrossRef] [PubMed]
29. Volosyak, I.; Valbuena, D.; Malechka, T.; Peuscher, J.; Gräser, A. Brain–computer interface using water-based electrodes. *J. Neural Eng.* **2010**, *7*, 066007. [CrossRef] [PubMed]
30. Zhao, H.; Zheng, L.; Yuan, M.; Wang, Y.; Gao, X.; Liu, R.; Pei, W. Optimization of ear electrodes for SSVEP-based BCI. *J. Neural Eng.* **2023**, *20*, 046011. [CrossRef] [PubMed]
31. Yuan, H.; Li, Y.; Yang, J.; Li, H.; Yang, Q.; Guo, C.; Zhu, S.; Shu, X. State of the Art of Non-Invasive Electrode Materials for Brain–Computer Interface. *Micromachines* **2021**, *12*, 1521. [CrossRef] [PubMed]
32. Guger, C.; Krausz, G.; Allison, B.Z.; Edlinger, G. Comparison of Dry and Gel Based Electrodes for P300 Brain–Computer Interfaces. *Front. Neurosci.* **2012**, *6*, 60. [CrossRef] [PubMed]
33. Xing, X.; Wang, Y.; Pei, W.; Guo, X.; Liu, Z.; Wang, F.; Ming, G.; Zhao, H.; Gui, Q.; Chen, H. A High-Speed SSVEP-Based BCI Using Dry EEG Electrodes. *Sci. Rep.* **2018**, *8*, 14708. [CrossRef] [PubMed]
34. Wang, F.; Li, G.; Chen, J.; Duan, Y.; Zhang, D. Novel semi-dry electrodes for brain–computer interface applications. *J. Neural Eng.* **2016**, *13*, 046021. [CrossRef] [PubMed]
35. Pei, W.; Wu, X.; Zhang, X.; Zha, A.; Tian, S.; Wang, Y.; Gao, X. A Pre-Gelled EEG Electrode and Its Application in SSVEP-Based BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 843–850. [CrossRef] [PubMed]
36. Kappel, S.L.; Kidmose, P. Study of Impedance Spectra for Dry and Wet EarEEG Electrodes. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 3161–3164. [CrossRef] [PubMed]
37. Li, G.L.; Wu, J.T.; Xia, Y.H.; He, Q.G.; Jin, H.G. Review of Semi-Dry Electrodes for EEG Recording. *J. Neural Eng.* **2020**, *17*, 051004. [CrossRef] [PubMed]
38. He, G.; Dong, X.; Qi, M. From the perspective of material science: A review of flexible electrodes for brain-computer interface. *Mater. Res. Express* **2020**, *7*, 102001. [CrossRef]
39. Memon, S.A.; Waheed, A.; Başaklar, T.; Ider Ider, Y.Z. Low-Cost Portable 4-Channel Wireless EEG Data Acquisition System for BCI Applications. In Proceedings of the 2018 Medical Technologies National Congress (TIPTEKNO), Famagusta, Cyprus, 8–10 November 2018; pp. 1–4. [CrossRef]
40. Karimi-Bidhendi, A.; Malekzadeh-Arasteh, O.; Lee, M.C.; McCrimmon, C.M.; Wang, P.T.; Mahajan, A.; Liu, C.Y.; Nenadic, Z.; Do, A.H.; Heydari, P. CMOS Ultralow Power Brain Signal Acquisition Front-Ends: Design and Human Testing. *IEEE Trans. Biomed. Circuits Syst.* **2017**, *11*, 1111–1122. [CrossRef] [PubMed]
41. Ng, K.; Chan, P. A CMOS analog front-end IC for portable EEG/ECG monitoring applications. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2005**, *52*, 2335–2347. [CrossRef]
42. Huang, X.; Muratore, D.G. Recording Front-End Electronics for Large-Scale Implantable Brain-Computer Interfaces: A Design Perspective. In Proceedings of the 2025 IEEE Custom Integrated Circuits Conference (CICC), Boston, MA, USA, 13–17 April 2025; pp. 1–8. [CrossRef]
43. van Oosterhout, K.; Chilundo, A.; Branco, M.P.; Aarnoutse, E.J.; Timmermans, M.; Fattori, M.; Ramsey, N.F.; Cantatore, E. Brain-Computer Interfaces Using Flexible Electronics: An a-IGZO Front-End for Active ECoG Electrodes. *Adv. Sci.* **2025**, *12*, 2408576. [CrossRef] [PubMed]
44. Liu, X.; Zhang, M.; Xiong, T.; Richardson, A.G.; Lucas, T.H.; Chin, P.S.; Etienne-Cummings, R.; Tran, T.D.; Van der Spiegel, J. A Fully Integrated Wireless Compressed Sensing Neural Signal Acquisition System for Chronic Recording and Brain Machine Interface. *IEEE Trans. Biomed. Circuits Syst.* **2016**, *10*, 874–883. [CrossRef] [PubMed]

45. Ando, H.; Takizawa, K.; Yoshida, T.; Matsushita, K.; Hirata, M.; Suzuki, T. Wireless Multichannel Neural Recording with a 128-Mbps UWB Transmitter for an Implantable Brain-Machine Interfaces. *IEEE Trans. Biomed. Circuits Syst.* **2016**, *10*, 1068–1078. [CrossRef] [PubMed]
46. Yazicioglu, R.F.; Merken, P.; Puers, R.; Van Hoof, C. A 60 μW 60 nV/√Hz Readout Front-End for Portable Biopotential Acquisition Systems. *IEEE J. Solid-State Circuits* **2007**, *42*, 1100–1110. [CrossRef]
47. Yazicioglu, R.F.; Merken, P.; Puers, R.; Van Hoof, C. A 200 μW Eight-Channel EEG Acquisition ASIC for Ambulatory EEG Systems. *IEEE J. Solid-State Circuits* **2008**, *43*, 3025–3038. [CrossRef]
48. Mollazadeh, M.; Murari, K.; Cauwenberghs, G.; Thakor, N. Micropower CMOS Integrated Low-Noise Amplification, Filtering, and Digitization of Multimodal Neuropotentials. *IEEE Trans. Biomed. Circuits Syst.* **2009**, *3*, 1–10. [CrossRef] [PubMed]
49. Chen, W.M.; Chiueh, H.; Chen, T.J.; Ho, C.L.; Jeng, C.; Ker, M.D.; Lin, C.Y.; Huang, Y.C.; Chou, C.W.; Fan, T.Y.; et al. A Fully Integrated 8-Channel Closed-Loop Neural-Prosthetic CMOS SoC for Real-Time Epileptic Seizure Control. *IEEE J. Solid-State Circuits* **2014**, *49*, 232–247. [CrossRef]
50. Chandrakumar, H.; Markovic, D. A 15.2-ENOB continuous-time ΔΣ ADC for a 7.3 μW 200mV$_{pp}$-linear-input-range neural recording front-end. In Proceedings of the 2018 IEEE International Solid-State Circuits Conference—(ISSCC), San Francisco, CA, USA, 11–15 February 2018; pp. 232–234. [CrossRef]
51. Pu, H.; Malekzadeh-Arasteh, O.; Danesh, A.R.; Nenadic, Z.; Do, A.H.; Heydari, P. A CMOS Dual-Mode Brain-Computer Interface Chipset with 2-mV Precision Time-Based Charge Balancing and Stimulation-Side Artifact Suppression. *IEEE J. Solid-State Circuits* **2022**, *57*, 1824–1840. [CrossRef]
52. Muller, R.; Gambini, S.; Rabaey, J.M. A 0.013 mm^2 5 μW DC-coupled neural signal acquisition IC with 0.5 V supply. In Proceedings of the 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, USA, 20–24 February 2011; pp. 302–304. [CrossRef]
53. Steyaert, M.; Sansen, W. A micropower low-noise monolithic instrumentation amplifier for medical purposes. *IEEE J. Solid-State Circuits* **1987**, *22*, 1163–1168. [CrossRef]
54. Harrison, R.; Charles, C. A low-power low-noise CMOS amplifier for neural recording applications. *IEEE J. Solid-State Circuits* **2003**, *38*, 958–965. [CrossRef]
55. Serasinghe, D.; Piyasena, D.; Pasqual, A. A Novel Low-Complexity VLSI Architecture for an EEG Feature Extraction Platform. In Proceedings of the 2018 21st Euromicro Conference on Digital System Design (DSD), Prague, Czech Republic, 29–31 August 2018; pp. 472–478. [CrossRef]
56. Chamanzar, A.; Shabany, M.; Malekmohammadi, A.; Mohammadinejad, S. Efficient Hardware Implementation of Real-Time Low-Power Movement Intention Detector System Using FFT and Adaptive Wavelet Transform. *IEEE Trans. Biomed. Circuits Syst.* **2017**, *11*, 585–596. [CrossRef] [PubMed]
57. Chen, F.; Song, L.; Li, H.H.; Chen, Y. RAISE: A Resistive Accelerator for Subject-Independent EEG Signal Classification. In Proceedings of the 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), Virtual, 1–5 February 2021; pp. 340–343. [CrossRef]
58. Krishna, A.; Ramanathan, V.; Yadav, S.S.; Shah, S.; van Schaik, A.; Mehendale, M.; Thakur, C.S. A Sparsity-driven tinyML Accelerator for Decoding Hand Kinematics in Brain-Computer Interfaces. In Proceedings of the 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), Toronto, ON, Canada, 19–21 October 2023; pp. 1–5. [CrossRef]
59. Flood, D.; Robinson, N.; Shreejith, S. FPGA-based Deep-Learning Accelerators for Energy Efficient Motor Imagery EEG classification. In Proceedings of the 2022 IEEE International Conference on Omni-Layer Intelligent Systems (COINS), Barcelona, Spain, 1–3 August 2022; pp. 1–6. [CrossRef]
60. Verma, N.; Shoeb, A.; Bohorquez, J.; Dawson, J.; Guttag, J.; Chandrakasan, A.P. A Micro-Power EEG Acquisition SoC with Integrated Feature Extraction Processor for a Chronic Seizure Detection System. *IEEE J. Solid-State Circuits* **2010**, *45*, 804–816. [CrossRef]
61. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [CrossRef] [PubMed]
62. Imran, S.M.; Talukdar, M.T.F.; Sakib, S.K.; Pathan, N.S.; Fattah, S.A. Motor imagery EEG signal classification scheme based on wavelet domain statistical features. In Proceedings of the 2014 International Conference on Electrical Engineering and Information & Communication Technology, Dhaka, Bangladesh, 10–12 April 2014; pp. 1–4. [CrossRef]
63. Lawhern, V.; Solon, A.; Waytowich, N.; Gordon, S.; Hung, C.; Lance, B. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef] [PubMed]
64. Schirrmeister, R.; Gemein, L.; Eggensperger, K.; Hutter, F.; Ball, T. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. In Proceedings of the 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 2 December 2017; pp. 1–7. [CrossRef]
65. Hu, Y.; Liu, Y.; Zhang, S.; Zhang, T.; Dai, B.; Peng, B.; Yang, H.; Dai, Y. A Cross-Space CNN with Customized Characteristics for Motor Imagery EEG Classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 1554–1565. [CrossRef] [PubMed]

66. Xiao, M.; Gu, Z.; Zhi, H. Multi-Domain Convolution with Attention: A Deep Learning Framework for BCI Motor Imagery Decoding. In Proceedings of the 2025 International Conference on Embodied Intelligence and Large Models (EILM '25), New York, NY, USA, 19–21 December 2025; pp. 187–190. [CrossRef]
67. Zhao, Z.; Cao, Y.; Yu, H.; Yu, H.; Huang, J. CNNViT-MILF-a: A Novel Architecture Leveraging the Synergy of CNN and ViT for Motor Imagery Classification. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 1153–1165. [CrossRef] [PubMed]
68. Singh, K.; Singha, N.; Jaswal, G.; Bhalaik, S. A Novel CNN With Sliding Window Technique for Enhanced Classification of MI-EEG Sensor Data. *IEEE Sens. J.* **2025**, *25*, 4777–4786. [CrossRef]
69. Gonzalez, H.A.; Muzaffar, S.; Yoo, J.; Elfadel, I.M. BioCNN: A Hardware Inference Engine for EEG-Based Emotion Detection. *IEEE Access* **2020**, *8*, 140896–140914. [CrossRef]
70. Cai, Z.; Li, P.; Cheng, L.; Yuan, D.; Li, M.; Li, H. A high performance heterogeneous hardware architecture for brain computer interface. *Biomed. Eng. Lett.* **2024**, *15*, 217–227. [CrossRef] [PubMed]
71. Pacini, F.; Pacini, T.; Lai, G.; Zocco, A.; Fanucci, L. Design and Evaluation of CPU-, GPU-, and FPGA-Based Deployment of a CNN for Motor Imagery Classification in Brain-Computer Interfaces. *Electronics* **2024**, *13*, 1646. [CrossRef]
72. Chen, Y.K.; Li, J.Y.; Fang, W.C. An Edge AI Accelerator of LRCN Model with RISC-V Platform for EEG-based Emotion Real-time Detection System. In Proceedings of the 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), Toronto, ON, Canada, 19–21 October 2023; pp. 1–5. [CrossRef]
73. Shi, W.; Qin, H.; Wu, J.; Mai, J. Design of a CNN Accelerator for Multitask EEG Signal Classification Based on RISC-V. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2025**, *33*, 2630–2638. [CrossRef]
74. Muneeb, A.; Kassiri, H. Customized Development and Hardware Optimization of a Fully-Spiking SNN for EEG-Based Seizure Detection. In Proceedings of the 2024 IEEE Biomedical Circuits and Systems Conference (BioCAS), Xi'an, China, 24–26 October 2024; pp. 1–5. [CrossRef]
75. Kanerva, P. Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. *Cogn. Comput.* **2009**, *1*, 139–159. [CrossRef]
76. Rahimi, A.; Kanerva, P.; Benini, L.; Rabaey, J.M. Efficient Biosignal Processing Using Hyperdimensional Computing: Network Templates for Combined Learning and Classification of ExG Signals. *Proc. IEEE* **2019**, *107*, 123–143. [CrossRef]
77. Burrello, A.; Cavigelli, L.; Schindler, K.; Benini, L.; Rahimi, A. Laelaps: An Energy-Efficient Seizure Detection Algorithm from Long-Term Human iEEG Recordings without False Alarms. In Proceedings of the Design, Automation and Test in Europe Conference (DATE), Florence, Italy, 25–29 March 2019; pp. 752–757. [CrossRef]
78. Burrello, A.; Benatti, S.; Schindler, K.; Benini, L.; Rahimi, A. An Ensemble of Hyperdimensional Classifiers: Hardware-Friendly Short-Latency Seizure Detection with Automatic iEEG Electrode Selection. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 935–946. [CrossRef] [PubMed]
79. Schindler, K.A.; Rahimi, A. A Primer on Hyperdimensional Computing for iEEG Seizure Detection. *Front. Neurol.* **2021**, *12*, 701791. [CrossRef] [PubMed]
80. Eggimann, M.; Rahimi, A.; Benini, L. A 5 µW Standard Cell Memory-Based Configurable Hyperdimensional Computing Accelerator for Always-on Smart Sensing. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2021**, *68*, 4116–4128. [CrossRef]
81. Menon, A.; Sun, D.; Sabouri, S.; Lee, K.; Aristio, M.; Liew, H.; Rabaey, J.M. A Highly Energy-Efficient Hyperdimensional Computing Processor for Biosignal Classification. *IEEE Trans. Biomed. Circuits Syst.* **2022**, *16*, 524–534. [CrossRef] [PubMed]
82. Karunaratne, G.; Le Gallo, M.; Hersche, M.; Cherubini, G.; Benini, L.; Sebastian, A.; Rahimi, A. Energy Efficient In-Memory Hyperdimensional Encoding for Spatio-Temporal Signal Processing. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**, *68*, 1725–1729. [CrossRef]
83. Moin, A.; Zhou, A.; Rahimi, A.; Menon, A.; Benatti, S.; Alexandrov, G.; Tamakloe, S.; Ting, J.; Yamamoto, N.; Khan, Y.; et al. A Wearable Biosensing System with In-Sensor Adaptive Machine Learning for Hand Gesture Recognition. *Nat. Electron.* **2021**, *4*, 54–63. [CrossRef]
84. Ma, X.; Zheng, W.; Peng, Z.; Yang, J. FPGA-Based Rapid Electroencephalography Signal Classification System. In Proceedings of the 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 18–20 October 2019; pp. 223–227. [CrossRef]
85. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2704–2713. [CrossRef]
86. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
87. Trabelsi, S.; Gontara, S.; Ben Khalifa, K.; Mtibaa, A. Designing a Teacher–Student Knowledge Distillation Framework for Accurate and Efficient Epilepsy Prediction on Edge Devices. In Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Ras Al-Khaimah, United Arab Emirates, 29–31 October 2025; pp. 1–6. [CrossRef]
88. Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 920–928. [CrossRef] [PubMed]

89. Congedo, M.; Barachant, A.; Bhatia, R. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Comput. Interfaces* **2017**, *4*, 155–174. [CrossRef]
90. Kuruppu, G.; Wagh, N.; Kremen, V.; Varatharajah, Y. EEG foundation models: A critical review of current progress and future directions. *J. Neural Eng.* **2026**, *23*, 021001. [CrossRef] [PubMed]
91. Jayaram, V.; Barachant, A. MOABB: Trustworthy algorithm benchmarking for BCIs. *J. Neural Eng.* **2018**, *15*, 066011. [CrossRef] [PubMed]
92. Shuqfa, Z.; Belkacem, A.N.; Lakas, A. Decoding Multi-Class Motor Imagery and Motor Execution Tasks Using Riemannian Geometry Algorithms on Large EEG Datasets. *Sensors* **2023**, *23*, 5051. [CrossRef] [PubMed]
93. Kalaganis, F.P.; Laskaris, N.A.; Chatzilari, E.; Nikolopoulos, S.; Kompatsiaris, I. A Riemannian Geometry Approach to Reduced and Discriminative Covariance Estimation in Brain Computer Interfaces. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 245–255. [CrossRef] [PubMed]
94. Lu, J.; Tian, Y.; Zhang, Y.; Sheng, Q.Z.; Zheng, X. LGL-BCI: A Motor-Imagery-Based Brain-Computer Interface with Geometric Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2025**, *9*, 1–28. [CrossRef]
95. Bravo, I.; Vázquez, C.; Gardel, A.; Lázaro, J.L.; Palomar, E. High Level Synthesis FPGA Implementation of the Jacobi Algorithm to Solve the Eigen Problem. *Math. Probl. Eng.* **2015**, *2015*, 870569. [CrossRef]
96. Hersche, M.; Rellstab, T.; Schiavone, P.D.; Cavigelli, L.; Benini, L.; Rahimi, A. Fast and Accurate Multiclass Inference for MI-BCIs Using Large Multiscale Temporal and Spectral Features. In Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 1690–1694. [CrossRef]
97. Wang, X.; Schneider, T.; Hersche, M.; Cavigelli, L.; Benini, L. Mixed-Precision Quantization and Parallel Implementation of Multispectral Riemannian Classification for Brain–Machine Interfaces. In Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; pp. 1–5. [CrossRef]
98. Wang, X.; Cavigelli, L.; Schneider, T.; Benini, L. Sub-100 µW Multispectral Riemannian Classification for EEG-Based Brain–Machine Interfaces. *IEEE Trans. Biomed. Circuits Syst.* **2021**, *15*, 1149–1160. [CrossRef] [PubMed]
99. Shyu, K.K.; Huang, S.C.; Tung, K.J.; Lee, L.H.; Lee, P.L.; Chen, Y.H. Common Spatial Pattern and Riemannian Manifold-Based Real-Time Multiclass Motor Imagery EEG Classification. *IEEE Access* **2023**, *11*, 139457–139465. [CrossRef]
100. Belwafi, K.; Romain, O.; Gannouni, S.; Ghaffari, F.; Djemal, R.; Ouni, B. An embedded implementation based on adaptive filter bank for brain–computer interface systems. *J. Neurosci. Methods* **2018**, *305*, 1–16. [CrossRef] [PubMed]
101. Vasei, T.; Saber, M.A.; Nahvy, A.; Navabi, Z.; Ko, S.B. An Efficient RTL Design for a Wearable Brain–Computer Interface. *IET Comput. Digit. Tech.* **2024**, *2024*, 5596468. [CrossRef]
102. Huang, Z.; Gool, L.V. A riemannian network for SPD matrix learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017*; AAAI Press: Washington, DC, USA, 2017; pp. 2036–2042. [CrossRef]
103. Jin, S.; Zhou, T.; Wang, R.; Chen, Z.; Luo, X.; Wu, X.J.; Kittler, J. Towards Robust EEG Decoding Based on Riemannian Self-Attention. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), New York, NY, USA, 9–13 August 2026; pp. 11225–11235. [CrossRef]
104. Zanini, P.; Congedo, M.; Jutten, C.; Said, S.; Berthoumieu, Y. Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 1107–1116. [CrossRef] [PubMed]
105. Kobler, R.; Hirayama, J.I.; Zhao, Q.; Kawanabe, M. SPD Domain-Specific Batch Normalization to Crack Interpretable Unsupervised Domain Adaptation in EEG. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6219–6235. Available online: https://dl.acm.org/doi/10.5555/3600270.3600720 (accessed on 20 October 2026). [CrossRef]
106. Cui, W.; Jeong, W.; Thölke, P.; Medani, T.; Jerbi, K.; Joshi, A.A.; Leahy, R.M. Neuro-GPT: Towards A Foundation Model For EEG. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; pp. 1–5. [CrossRef]
107. Wu, D.; Li, S.; Yang, J.; Sawan, M. Neuro-BERT: Rethinking Masked Autoencoding for Self-Supervised Neurological Pretraining. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 970–980. [CrossRef] [PubMed]
108. Wang, L.; Suzumura, T.; Kanezashi, H. GEFM: Graph-Enhanced EEG Foundation Model. In Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 14–17 July 2025; pp. 1–7. [CrossRef] [PubMed]
109. Yang, C.; Westover, M.B.; Sun, J. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 78240–78260. Available online: https://dl.acm.org/doi/10.5555/3666122.3669542 (accessed on 20 October 2026). [CrossRef]
110. Jiang, W.; Zhao, L.; Lu, B.L. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
111. Wang, G.; Liu, W.; He, Y.; Xu, C.; Ma, L.; Li, H. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 39249–39280. Available online: https://neurips.cc/virtual/2024/poster/93793 (accessed on 20 October 2026). [CrossRef]

112. Wang, J.; Zhao, S.; Luo, Z.; Zhou, Y.; Jiang, H.; Li, S.; Li, T.; Pan, G. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025. Available online: https://openreview.net/forum?id=NPNUHgHF2w (accessed on 20 October 2026).
113. Tegon, A.; Lehmann, N.; Li, Y.; Cossettini, A.; Benini, L.; Ingolfsson, T.M. FEMBA on the Edge: Physiologically-Aware Pre-Training, Quantization, and Deployment of a Bidirectional Mamba EEG Foundation Model on an Ultra-Low Power Microcontroller. *IEEE Trans. Biomed. Eng.* **2026**, *73*, 2746–2757. [CrossRef] [PubMed]
114. Cheng, D.; Shen, S.; Zhu, X. Lightweight Sleep EEG Representation Learning and Automatic Sleep Staging with an EEG Foundation Model. In Proceedings of the 2026 7th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), Shenzhen, China, 26–28 June 2026; pp. 901–904. [CrossRef]
115. Zhu, L.; Wang, S.; Jin, X.; Zheng, K.; Zhang, J.; Sun, S.; Tian, F.; Cai, R.; Hu, B. MDH-NAS: Accelerating EEG Signal Classification With Mixed-Level Differentiable and Hardware-Aware Neural Architecture Search. *IEEE Internet Things J.* **2025**, *12*, 23866–23878. [CrossRef]
116. Eder, M.; Xu, J.; Grosse-Wentrup, M. Benchmarking brain–computer interface algorithms: Riemannian approaches vs convolutional neural networks. *J. Neural Eng.* **2024**, *21*, 044002. [CrossRef] [PubMed]
117. Carrara, I.; Papadopoulo, T. Pseudo-online framework for BCI evaluation: A MOABB perspective using various MI and SSVEP datasets. *J. Neural Eng.* **2024**, *21*, 016003. [CrossRef] [PubMed]
118. Wolf, P.D. Thermal Considerations for the Design of an Implanted Cortical Brain–Machine Interface (BMI). In *Indwelling Neural Implants: Strategies for Contending with the In Vivo Environment*; Reichert, W.M., Ed.; CRC Press: Boca Raton, FL, USA, 2008. Available online: https://www.ncbi.nlm.nih.gov/books/NBK3932/ (accessed on 20 October 2026).
119. Kim, S.; Tathireddy, P.; Normann, R.A.; Solzbacher, F. Thermal Impact of an Active 3-D Microelectrode Array Implanted in the Brain. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2007**, *15*, 493–501. [CrossRef] [PubMed]
120. Eichler, G.; Piccolboni, L.; Giri, D.; Carloni, L.P. MasterMind: Many-Accelerator SoC Architecture for Real-Time Brain-Computer Interfaces. In Proceedings of the 2021 IEEE 39th International Conference on Computer Design (ICCD), Virtual, 24–27 October 2021; pp. 101–108. [CrossRef]
121. Benzy, V.K.; Vinod, A.P.; Subasree, R.; Alladi, S.; Raghavendra, K. Motor Imagery Hand Movement Direction Decoding Using Brain Computer Interface to Aid Stroke Recovery and Rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 3051–3062. [CrossRef] [PubMed]
122. Uehlin, J.P.; Smith, W.A.; Pamula, V.R.; Perlmutter, S.I.; Rudell, J.C.; Sathe, V.S. A 0.0023 mm^2/ch. Delta-Encoded, Time-Division Multiplexed Mixed-Signal ECoG Recording Architecture With Stimulus Artifact Suppression. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 319–331. [CrossRef] [PubMed]
123. Amudha, V.; Balachander, B.; Kumar, G.H.; Divya, K.; Ravichandran, G.; Nithish, V. CNN based Movement Intention Detection based on EEG signals for Neuro Rehabilitation. In Proceedings of the 2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET), Chennai, India, 1–2 August 2025; pp. 1–8. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.