Journal of Electronic Research and Application, 2026, Volume 10, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

AI Large Model-Driven Adaptive Evolution of Brain-Computer Interface Chips: Technical Architecture, Challenges, and Future Directions

Borui Cui*

School of Mathematics, Tianjin University, Tianjin 300350, China

**Corresponding author: Borui Cui, infolklore2025@163.com*

Abstract: This paper focuses on how AI large models, such as Transformers and meta-learning can empower brain-computer interface (BCI) chips to achieve dynamic adaptation, thereby overcoming the limitations of traditional fixed decoding models that struggle to adapt to individual neural plasticity and dynamic changes in brain states. It analyzes pathways to enhance chip generalization and real-time performance across three technical dimensions: hardware architecture, algorithm optimization, and multimodal fusion. The paper also explores core challenges like data privacy and energy-efficiency tradeoffs. Building on this foundation, it proposes a neuromorphic computing design framework for next-generation chips to advance the intelligent and personalized development of BCI in medical rehabilitation and human-computer interaction.

Keywords: Brain-computer interface; Artificial intelligence; Chip; Large model-driven; Meta-learning

Online publication: February 27, 2026

1. Introduction

Brain-computer interface (BCI), as a critical interdisciplinary technology bridging the human brain and external devices, has garnered immense attention in fields such as medical rehabilitation and human-computer interaction. It enables direct communication between the central nervous system and external equipment by decoding neural signals, offering revolutionary solutions for patients with motor disabilities to restore motor functions and promoting the evolution of intelligent human-computer interaction modes. However, the performance of current BCI systems is largely constrained by the underlying chip technology, particularly the limitations of traditional fixed decoding models.

Neural plasticity, the inherent ability of the human brain to reorganize neural connections in response to external stimuli and internal states, coupled with the dynamic changes in brain states caused by factors such as fatigue, emotion, and task switching, poses significant challenges to traditional BCI chips. Fixed decoding models,

designed based on pre-set neural signal features, lack the flexibility to adapt to individual differences in neural activity and real-time changes in brain states, resulting in compromised decoding accuracy, poor generalization across individuals, and inadequate real-time performance. These drawbacks severely hinder the practical application and further development of BCI technology, making it urgent to explore innovative technical paths to enhance the adaptive capacity of BCI chips.

In recent years, the rapid advancement of artificial intelligence (AI) large models, represented by Transformers and meta-learning, has brought new opportunities for breaking through the bottlenecks of traditional BCI chips. These AI models possess powerful capabilities in feature extraction, adaptive learning, and pattern recognition, which can be leveraged to empower BCI chips with dynamic adaptation capabilities. By integrating AI large models into BCI chip design, it becomes feasible to realize real-time adjustment of decoding strategies according to individual neural characteristics and dynamic brain state changes, thereby significantly improving the generalization and real-time performance of BCI systems.

Against this backdrop, this paper focuses on the empowerment mechanism of AI large models on BCI chips. It systematically analyzes the technical pathways to enhance BCI chip performance from three core dimensions: hardware architecture optimization tailored for AI model deployment, algorithm improvement to strengthen adaptive decoding capabilities, and multimodal fusion to complement neural signal information. Meanwhile, critical challenges in the integration process, including data privacy protection of neural signals (which involve highly sensitive personal biological information) and the tradeoff between chip energy efficiency and computational performance, are also explored in depth. On this basis, a neuromorphic computing design framework for next-generation BCI chips is proposed, aiming to provide a theoretical and technical foundation for promoting the intelligent, personalized, and practical development of BCI technology in medical rehabilitation, human-computer interaction, and other key fields.

2. Principles of AI large model-driven adaptive technology

2.1. Dynamic decoding model construction

The core of achieving adaptive BCI chips lies in constructing decoding models capable of dynamic evolution. This approach overcomes individual differences among users and the diversity of brain states that change with different contexts. Such models enable the chip to continuously learn and optimize during actual use, thereby maintaining high performance over the long term.

2.1.1. Online learning architecture based on incremental learning

Incremental learning aims to enable intelligent systems to “continuously learn” like humans, absorbing knowledge from new data while retaining prior learning without retraining on the entire historical dataset [1]. This characteristic offers an ideal solution for addressing the dynamic nature of electroencephalograms (EEG) signals. Since EEG signals drift with users’ physiological states (e.g., fatigue, attention), long-term usage habits, and even neuroplasticity, fixed models often face performance degradation. When integrated into chip-based systems, incremental learning algorithms enable continuous monitoring of EEG signals alongside corresponding operational outcomes. Upon detecting a decline in decoding accuracy or shifts in signal patterns, the system applies small, targeted adjustments to model parameters using newly acquired mini-batch data. This allows chip-deployed decoding models to achieve “accompanied growth” maintaining high-precision alignment with the user’s current

brain state and establishing an algorithmic foundation for long-term adaptability.

2.1.2. Rapid adaptation capability based on meta-learning

Meta-learning, or "learning to learn" aims to enhance a model's generalization and rapid adaptation capabilities on new tasks, enabling efficient adjustments with minimal samples [2]. This capability directly addresses a core user experience bottleneck in traditional BCI: model calibration for new users typically requires lengthy, tedious data collection and training processes. Chips based on algorithms like Model-Agnostic Meta-Learning (MAML) leverage pre-learned cross-user prior knowledge to rapidly adapt to new users with minimal calibration data. In BCI applications, meta-learning reduces initial calibration time from hours to minutes, significantly improving device usability and user experience.

2.2. Multimodal signal fusion

Single-modality EEG signals are susceptible to interference and have limited informational dimensions, making it challenging to ensure the decoding accuracy of adaptive BCI systems. By integrating complementary physiological signals from diverse sources, multimodal fusion significantly enhances the accuracy and environmental adaptability of decoding models while meeting system requirements for low power consumption and high efficiency.

2.2.1. Chip-level fusion architecture design

To achieve low-latency parallel processing of multimodal data, dedicated multimodal fusion processing units must be designed at the chip level. This unit integrates hardware accelerators to perform real-time parallel processing of multimodal data such as EEG, fNIRS (functional near-infrared spectroscopy), and eye tracking. It incorporates specialized preprocessing circuits for different signals, enabling preliminary signal cleaning and feature extraction at the chip front end. Fusion occurs directly within the unit, significantly reducing data transfer overhead between storage and computation units. This design meets the stringent power and speed requirements for real-time interaction. Research by Shi *et al.* indicates that feature-level fusion can integrate complementary characteristics from peripheral physiological signals like EEG and ECG at the stage "after feature extraction and before classifier input," forming a unified joint feature vector to preserve deep intermodal correlations [3]. This conclusion provides critical justification for designing an "EEG-ECG feature concatenation unit" in chip front-end architecture, helping avoid data upload delays and better adapt to real-time brain-computer interaction scenarios.

2.2.2. Cross-modal attention mechanism

At the algorithmic level, a Transformer-based cross-modal attention mechanism can be introduced to dynamically assign weights to information from different modalities. Addressing the issues of "disconnected modal interactions and redundant model parameters" in current multimodal BCI algorithms, Guo *et al.* demonstrated that cross-modal attention mechanisms can integrate information from all input modalities simultaneously (rather than sequentially pairing them) and optimize training stability through techniques like layer normalization and residual connections [4]. This approach can guide the design of efficient multimodal fusion modules, such as simultaneously integrating EEG's temporal features, ECG's heart rate variability features, and eye-tracking's spatial features. This enables deeper exploration of complementarity between modalities, significantly enhancing robustness in emotion or intent recognition.

2.3. Few-shot learning and model compression

Deploying complex AI models onto resource-constrained BCI chips presents a fundamental challenge: the high complexity of models versus the strict limitations of chip computing power, memory, and power consumption. To overcome this bottleneck, few-shot learning and model compression techniques must be introduced to enable efficient and accurate model operation in low-resource environments.

2.3.1. Few-shot learning mechanism

Traditional deep learning heavily relies on massive labeled datasets, whereas practical BCI applications (such as new user calibration) often provide only a minimal number of samples. Few-shot learning aims to enable models to achieve efficient learning and accurate predictions even under conditions of scarce training data. Its core lies in utilizing algorithms like meta-learning to perform pre-training on rich multi-user data before chip production. This endows the model with the intrinsic ability to capture individual signal variations and rapidly adapt to new users. For instance, the meta-learning framework proposed by Li *et al.* provides a key approach, where during the meta-training phase, the model learns across numerous simulated small-sample tasks (composed of a "support set" and a "query set") to obtain a set of highly adaptable universal initialization parameters. During the meta-testing phase (when encountering new users), the model converges rapidly using only a small number of calibration samples (e.g., 5–10 sets of EEG data) to construct an effective personalized classifier [5]. This "learning to learn" capability significantly reduces the time and data costs associated with user calibration.

2.3.2. Knowledge distillation and model compression

Model compression is a core technology for addressing the challenges of deploying large models. Its goal is to significantly reduce model size and computational requirements while maximizing the retention of the original model's performance. As a representative technique, knowledge distillation employs a "teacher-student framework" to achieve model lightweighting: a large, accurate "teacher model" guides the training of a lightweight "student model." The "student model" not only learns actual task labels but also inherits the teacher's strong generalization capabilities by mimicking its softened output probability distribution (using techniques like temperature scaling and KL divergence) and aligning feature representations across intermediate layers [6].

2.3.3. Applications of neural architecture search technology

Traditional neural network design relies on expert experience, making it difficult to automatically achieve optimal matching with target hardware. Neural architecture search (NAS) technology automates the exploration of design spaces, enabling the discovery of optimal neural network architectures directly tailored to specific hardware constraints (e.g., computational power, power consumption, latency) and task objectives (e.g., EEG decoding accuracy). This achieves a paradigm shift from "manually designing models to fit hardware" to "automatically co-searching for optimal model-hardware pairings." Notably, recent advancements have further extended NAS frameworks toward trustworthy and secure deployment, critical aspects in handling sensitive data such as in BCI. For instance, a blockchain-enhanced trustworthy NAS method has been proposed for medical image classification, which integrates secure data provenance and reliable model prediction transmission while maintaining high search performance [7]. This approach underscores how NAS can be synergized with security mechanisms to protect data confidentiality and ensure interpretability in automated model development. Although applied in pneumonia image diagnosis, such a trustworthy NAS framework offers a promising paradigm for BCI systems where neural data

security and reliable decoding are equally vital.

3. Adaptive hardware implementation of brain-computer interface chips

3.1. Reconfigurable computing architecture

To support the dynamic adaptive capabilities of large AI models while meeting the stringent demands of BCI for high energy efficiency and computational real-time performance, reconfigurable computing architectures offer a critical pathway to overcome traditional chip energy efficiency bottlenecks.

3.1.1. FPGA dynamic reconfiguration

As a chip capable of defining its hardware functionality through configuration software after deployment, field-programmable gate arrays (FPGAs) combine the high speed of hardware parallel processing with the flexibility of software programmability. FPGA dynamic reconfiguration technology enables real-time modification, updating, or restructuring of the functionality of partial or complete logic resources within the chip without interrupting overall system operation. This capability provides core support for enabling flexible switching between different processing tasks and efficient reuse of hardware resources in BCI chips. Research indicates that this technology "enables dynamic reprogramming of FPGAs during runtime to alter their functionality online" and "significantly conserves on-chip resources, reserving more space for other operations" [8]. This means a single BCI chip can rapidly switch between multiple hardware acceleration cores, such as those for signal preprocessing, feature extraction, or specific decoding algorithms, by loading different partial configuration files. This "time-division multiplexing" strategy significantly enhances the chip's functional diversity and scenario adaptability under resource-constrained conditions.

3.1.2. Compute-in-memory architecture

By embedding computational units within memory, the compute-in-memory architecture enables operations to be performed directly at the data storage location. This fundamentally reduces latency and power consumption associated with data movement between storage and computational units. To achieve the integration of high energy efficiency and high-precision computing, advanced architectures such as high-precision hybrid floating-point in-memory computation (Hy-FPCIM) for complex models (e.g., Vision Transformers) have been proposed. By efficiently executing decomposed exponent and mantissa operations in-memory, these approaches significantly enhance computational energy efficiency and area efficiency with near-lossless precision [9]. Research indicates that the CIM architecture has demonstrated broad application prospects in fields such as AI and the IoT. Its high energy efficiency and low latency characteristics offer novel technical solutions for tackling challenges in BCI systems, including real-time signal processing, lightweight model deployment, and stringent power constraints. Incorporating the CIM architecture into BCI chip design is expected to drive system development toward higher real-time performance, enhanced energy efficiency, and improved integration.

3.2. Neuromorphic computing integration

To fundamentally approximate the bio-brain's high energy efficiency and adaptive properties, neuromorphic computing simulates the brain's spiking communication and synaptic plasticity mechanisms, endowing next-generation BCI chips with low power consumption and intrinsic learning capabilities.

3.2.1. Spiking neural networks

Spiking neural networks (SNNs) utilize discrete spike sequences as information carriers and perform event-driven computations, exhibiting a mechanism highly analogous to the human brain. When deployed on chips, SNNs offer dual advantages of event-driven efficiency and ultra-low power consumption. However, due to the susceptibility of EEG signals to noise and individual variations, enhancing SNN robustness and their ability to model underlying signal dynamics is a critical prerequisite for practical application. Recent advances, such as the finite-difference physics-informed spiking neural network (FPSNN), address core challenges by successfully integrating physics-informed learning paradigms with SNN architectures, thereby enhancing the network's capacity to capture and generalize the dynamical characteristics of time-series signals like EEG [10]. This approach demonstrates that incorporating prior physical or dynamic constraints can significantly improve SNN performance over purely data-driven methods. Leveraging the efficient data access characteristics of the CIM architecture, SNNs empowered by such physics-aware learning frameworks can serve as the core for low-power, high-robustness computation in BCI chips, further enhancing system practicality and reliability in real-world complex environments.

3.2.2. Brain-inspired chip design

The biomimetic design of brain-inspired chips has advanced to the cellular and molecular levels. By integrating dendritic computing units, advanced brain-inspired chips can perform complex spatiotemporal signal integration, thereby enhancing the computational complexity of individual neuron nodes. This enables smaller-scale networks to achieve powerful pattern recognition capabilities. This approach not only helps reduce chip area and power consumption but also improves robustness when processing the dynamic characteristics of neural signals. A hierarchical multi-core architecture combined with a quasi-event-driven mechanism effectively supports the spatiotemporal signal processing of SNNs, improving the chip's real-time performance and energy efficiency during dynamic tasks. By implementing synaptic plasticity through devices like memristors, BCI chips can simulate lifelong learning at the physical hardware level. This allows the chip to autonomously optimize internal connection weights based on the user's neural activity patterns without requiring high-level software intervention, laying the hardware foundation for achieving "adaptive evolution." Neuromorphic chips supporting on-chip learning can achieve efficient synaptic weight updates through mechanisms like direct feedback alignment. These chips exhibit strong spatio-temporal locality and hardware-friendliness, making them suitable for real-time learning scenarios at the edge.

3.3. Balancing low power consumption and real-time performance

BCI chips designed for implantable or wearable applications must achieve millisecond-level real-time responsiveness under stringent power constraints, yet high-performance computing often conflicts with low power consumption. This section explores how to achieve an optimal balance of system energy efficiency while ensuring real-time performance.

3.3.1. Dynamic voltage and frequency scaling technology

Dynamic voltage and frequency scaling (DVFS) technology achieves intelligent energy efficiency management by continuously monitoring the computational load of chips and dynamically adjusting their operating voltage and clock frequency. The core of this technology lies in precisely matching power supply to computational power based on real-time performance demands, thereby minimizing power consumption while meeting task

requirements. To achieve this goal, research efforts have shifted toward developing advanced scheduling algorithms that optimize energy efficiency while ensuring real-time system performance, thereby reducing the risks of delays and instability caused by frequent frequency and voltage adjustments [11]. The operational intent of BCI users exhibits intermittent and time-varying characteristics, leading to significant fluctuations in chip workload. To address this, an intelligent frequency adjustment mechanism can be developed by incorporating load prediction models based on the AVG algorithm and Hotplug multi-core management strategies. When the system is idle or under low load (e.g., during user rest periods), the chip automatically switches to low-power mode while immediately scaling up to peak performance upon detecting user intent for complex operations. This "power-on-demand" strategy effectively extends battery life for implantable devices, highlighting the practical value of DVFS technology in low-power embedded systems.

3.3.2. Application of approximate computation in EEG signal processing

BCI tasks typically exhibit a degree of fault tolerance, enabling the introduction of approximate computation to trade off for improved energy efficiency. Approximate computation significantly reduces power consumption and hardware resource usage by simplifying computational processes or circuit designs within acceptable accuracy loss margins. As a representative model of brain-inspired computing, spiking neural networks inherently possess a distributed architecture and training methodology that endows them with tolerance for partial computational errors, thereby establishing a theoretical foundation for applying approximate computation in EEG signal processing. In practical implementation, strategies such as applying sensitive average relative error metrics and approximate adder selection can be adopted. By leveraging the input distribution characteristics of addition operations in EEG signal processing, approximate computing units can be selected to achieve an optimal balance between maintaining system recognition accuracy and realizing significant energy efficiency gains. Experiments demonstrate that this approach achieves 37.32% power savings and 31.26% area reduction while only sacrificing approximately 3.47 percentage points in classification accuracy. This points to a highly promising design direction for future implantable BCI chips to maintain long-term stable operation under limited energy budgets.

4. Core challenges and solutions

4.1. Data privacy and security

BCI can interpret human thought processes, offering immense potential while simultaneously exposing neural data to severe risks of leakage and tampering. The resulting data privacy and security concerns have become widespread public anxieties, with ethical issues such as personal integrity and autonomy also coming to the fore. Therefore, security design must form the cornerstone of chip architecture.

4.1.1. Challenge: Risks of neurodata leakage

The security challenges posed by neural data far exceed those of other data types, manifesting primarily in two following aspects:

(1) Neural data such as EEG directly relate to an individual's thoughts, intentions, and emotional states, constituting the highest dimension of personal privacy. Should raw data leak, it would lead to "thought exposure," placing users in a state of being watched. Such data could be exploited for commercial manipulation or psychological attacks, severely infringing upon cognitive freedom and mental privacy;

(2) Closed-loop BCIs with "writing" capabilities create channels for reverse attacks. Attackers could

manipulate users' perceptions and decisions by fabricating neural feedback signals, potentially inducing mental symptoms and posing direct threats to personal safety and autonomy [12]. Such loss of behavioral control due to technical vulnerabilities or malicious intrusion may not only constitute civil infringement but could even reach criminal offenses under extreme circumstances.

4.1.2. Solution: Privacy-preserving training based on federated learning

To address the aforementioned risks, a distributed training paradigm based on federated learning can be adopted. Within this framework, users' raw neural data is stored and processed entirely within local secure environments without requiring upload to central servers, structurally eliminating the risk of large-scale data leakage during transmission and aggregation. The core of federated learning lies in participants training models locally and uploading only encrypted model parameter updates (e.g., gradients), thereby enabling collaborative learning while protecting data source privacy. To address the risks of collusive attacks and privacy leaks that may exist in traditional server architectures, the latest privacy protection solution effectively enhances system security, verifiability, and practical efficiency by introducing homomorphic proxy re-encryption and a dual-server architecture [13]. To further enhance security, techniques such as homomorphic encryption or secure multi-party computation can be integrated to shield the parameter update process, preventing the reconstruction of raw data from gradient information. Such approaches (e.g., SE-Fed) also optimize communication efficiency through techniques like client grouping and gradient compression. This maintains high model performance while ensuring privacy, making it suitable for real-time-critical applications like BCIs.

4.2. Trade-off between model complexity and hardware resources

The exceptional performance of large AI models relies on massive parameters and complex computations, creating a fundamental conflict with the extremely constrained computational power, memory, and power consumption budgets of BCI chips. This section explores solutions from two perspectives: model optimization and hardware adaptation.

4.2.1. Challenge: The parameter scale of large models far exceeds the computational power of chips

Deploying models with massive parameters on chips primarily faces three major challenges: computational power, memory, and power consumption, as follows:

(1) In terms of computational power, the volume of floating-point operations required for a single inference by large models far exceeds the computational upper limit that embedded processors can provide within real-time requirements (typically < 10ms);

(2) Memory constraints arise because model parameters often span hundreds of megabytes, while on-chip storage typically ranges from kilobytes to megabytes. Frequent access to off-chip, low-speed memory incurs prohibitive latency and power overhead;

(3) Power consumption is further strained by complex computations and frequent data transfers, generating significant heat, a severe challenge for embedded chips operating within milliwatt-level power budgets.

4.2.2. Solution: Model pruning and quantization

Model pruning and quantization techniques aim to directly reduce model size and computational complexity,

which is critical for deploying neural networks on resource-constrained hardware. Model pruning streamlines models by identifying and removing redundant connections or structures within the network. Structured pruning, such as removing entire filters or channels, generates hardware-friendly, regular sparse models that significantly reduce computational load with minimal accuracy loss. Model quantization converts high-precision floating-point parameters and activation values into low-bit fixed-point representations. Key techniques include quantization-aware training (simulating quantization effects during training to enhance final accuracy) and mixed-precision quantization (dynamically allocating bit widths based on layer sensitivity). The synergistic application of pruning and quantization can reduce a model's storage and computational requirements by an order of magnitude, enabling efficient deployment on edge devices. Systematic evaluations, such as those conducted in recent TinyML-oriented research, demonstrate that combining structured pruning with INT8 quantization can achieve substantial model compression (e.g., over 75% size reduction for models like MobileNet) while preserving competitive accuracy, thereby providing a validated pathway for implementing high-efficiency neural networks in real-world, resource-limited applications such as BCI chips [14].

4.3. Long-term stability and biocompatibility

BCI chips, particularly invasive implantable devices, must overcome significant challenges posed by the complex biological environment to achieve long-term clinical application. Long-term stability requires chips to maintain functionality without degradation over periods spanning years to decades, while biocompatibility demands stable coexistence with neural tissue without triggering harmful immune or inflammatory responses.

4.3.1. Challenge: Safety of chronic implantation for invasive chips

The core safety challenge facing invasive BCI lies in the long-term mechanical and biological incompatibility between traditional rigid electrodes and brain tissue. This incompatibility leads to persistent micro-motion friction, triggering chronic inflammatory responses and glial scar encapsulation. Not only can this damage neural tissue, but it also causes progressive degradation in neural signal recording quality and even device failure. Therefore, developing novel materials and structures capable of forming stable, harmonious interfaces with neural tissue is of paramount importance.

4.3.2. Solution: Flexible electronic materials and biocompatible interfaces

The use of flexible electronic materials, such as conductive hydrogels, offers an effective solution to biocompatibility challenges. These materials exhibit a modulus highly compatible with neural tissue, significantly reducing tissue damage caused by mechanical mismatch and effectively suppressing glial scar formation. Functional surface modifications can further guide neuronal synapse adhesion, promoting stable, high-quality electrical coupling and biointegration between the chip and neural circuits. This lays the foundation for acquiring long-term stable neural signals.

4.3.3. Solution: Self-healing circuit

To address potential aging, wear, or localized failure of circuits after implantation, self-healing circuit technology can be introduced. This technology embeds dynamically reversible chemical bonds or microcapsule repair agents within the material, enabling circuits to automatically trigger physical or chemical processes upon minor fractures or performance degradation. This reconfigures conductive pathways and restores functionality. This self-healing

mechanism not only addresses macro-structural damage but has also been validated in critical analog circuit modules such as voltage reference circuits. It effectively counters aging effects like thermal carrier injection, ensuring that key performance parameters, including temperature coefficient and output voltage, remain within acceptable ranges even after prolonged operation [15]. When combined with biodegradable materials serving as temporary scaffolds or repair media, on-demand repair and absorption can be achieved. This self-healing capability significantly enhances the long-term reliability and service life of implantable systems in unattended environments.

5. Future research directions

The key breakthrough direction for future BCI chips lies in integrating the strengths of neuromorphic computing and large artificial intelligence models to construct hybrid architectures that combine biological plausibility with powerful cognitive capabilities, namely, brain-inspired AI fusion chips. Such chips may adopt a layered design: the bottom layer employs ultra-low-power SNN units to directly process neural signals, performing temporal feature extraction and preliminary filtering; the top layer utilizes compressed and optimized modules like Transformers to handle complex semantic decoding and advanced cognitive tasks. Communication and coordination between these layers occurs via efficient cross-paradigm interfaces. Realizing this vision requires developing novel collaborative training methods and hybrid neural architecture search techniques. These approaches must enable hardware-friendly mutual enhancement and co-evolution of SNN perceptual capabilities and ANN cognitive capabilities, ultimately forming high-performance, energy-efficient autonomous intelligent neural processors.

To overcome computational power and power consumption constraints in edge chips, constructing efficient "edge-cloud" collaborative systems represents a core future direction. Within this architecture, edge devices handle lightweight processing ensuring real-time responsiveness and privacy; edge servers host personalized models and perform preliminary aggregation; while the cloud manages large-scale model training and global optimization. 5G/6G networks serve as high-speed connectivity links, guaranteeing real-time synchronization and reliable communication across layers. This collaborative model enables dynamic workload distribution and global resource optimization, allowing endpoints to dynamically access the latest cloud capabilities in real time and driving continuous adaptive evolution of system performance.

To advance technology from laboratory to large-scale application, building an open collaborative ecosystem and unified standardization system is essential. Current fragmentation in hardware interfaces, data formats, communication protocols, and algorithm frameworks severely hinders technological iteration and adoption. Future efforts should focus on advancing standards across all layers, including hardware interfaces, data representation, algorithm deployment, and security protocols. Simultaneously, fostering an open ecosystem involving industry, academia, research institutions, and healthcare providers through open-source hardware design, shared benchmark datasets, and open toolkits will lower R&D barriers, accelerate innovation, and ultimately realize the vision of extending technology benefits to broader populations.

6. Conclusion

Large AI models provide the core algorithmic foundation for achieving efficient and precise adaptive evolution in BCI chips. Their capabilities in online learning and small-sample adaptation enable chips to dynamically respond to changes in users' neural plasticity. Meanwhile, innovations in hardware architecture, particularly bio-inspired

and energy-efficient designs such as compute-in-memory and neuromorphic computing, constitute the critical infrastructure for supporting and unleashing such intelligent algorithms. The synergy between these two elements collectively propels BCIs from static, generic traditional forms toward a new era of dynamic, personalized intelligent interaction. However, technological maturation and widespread adoption still face multiple challenges. Future breakthroughs must address core bottlenecks, including data privacy protection, energy-efficiency optimization, long-term biocompatibility, and the absence of standardized protocols. This requires developing multidisciplinary solutions spanning algorithms, hardware, materials, and security, alongside establishing open industrial ecosystems and collaborative frameworks. Only through persistent technological breakthroughs and collaborative innovation can BCIs transition from laboratory prototypes to safe, reliable, and accessible large-scale applications. This will realize the next generation of human-machine integrated intelligent interaction, unlocking new possibilities for medical rehabilitation, cognitive enhancement, and even the evolution of human existence.

Disclosure statement

The author declares no conflict of interest.

References

[1] He Z, Huang S, Lu Y, et al., 2026, MoTiC: Momentum Tightness and Contrast for Few-Shot Class-Incremental Learning. Pattern Recognition, 2026(173): 112753.

[2] Zhang J, Chang Y, Wang F, 2025, Few-Shot Working Condition Classification within a Meta-Learning Framework based on Multi-Head Attention Autoencoder. Journal of Intelligent Manufacturing, 2025(prepublish): 1–16.

[3] Shi P, Wang H, Liu L, Physiological Signal Emotion Recognition: A Review of Cross-Domain Transfer and Multimodal Fusion. Journal of Frontiers of Computer Science and Technology, 1–23.

[4] Guo J, Lu H, Xu J, 2025, Research on Multimodal Sentiment Analysis Method Based on Cross-Modal Attention Mechanism. Computer Knowledge and Technology, 21(1): 1–4.

[5] Li K, et al., 2025, A Review of Few-Shot Learning Research. Mechanical & Electrical Engineering Technology, 54(6): 160–168.

[6] Wang Y, Chen Y, 2026, Temperature-Driven Category Decoupled Knowledge Distillation with Interpretability for Model Compression. Advanced Engineering Informatics, 69(PC): 104051.

[7] Yang Y, Wei J, Yu Z, et al., 2024, A Trustworthy Neural Architecture Search Framework for Pneumonia Image Classification Utilizing Blockchain Technology. The Journal of Supercomputing, 80(2): 1694–1727.

[8] Boudjadar J, Islam S, Buyya R, 2025, Dynamic FPGA Reconfiguration for Scalable Embedded Artificial Intelligence (AI): A Co-Design Methodology for Convolutional Neural Networks (CNN) Acceleration. Future Generation Computer Systems, 2025(169): 107777.

[9] Ma Z, Wang C, Chen Q, et al., 2025, A High-Precision Hybrid Floating-Point Compute-in-Memory Architecture for Complex Deep Learning. Electronics, 14(22): 4414.

[10] Wei Q, Yang Q, Han L, et al., 2026, Physics-Informed Spiking Neural Networks for Continuous-Time Dynamic Systems. Neurocomputing, 2026(665): 132192.

[11] Chen Y, Huang J, Xiao L, et al., 2025, A DVFS-Weakly Dependent Real-Time Scheduling for Multiple Parallel Applications on Energy-Aware Heterogeneous Systems. Journal of Systems Architecture, 2025(170): 103614.

[12] Brocal F, 2023, Brain-Computer Interfaces in Safety and Security Fields: Risks and Applications. Safety Science,

2023: 160.

[13] Song Z, Huang Z, Cai Y, 2026, A Verifiable Privacy-Preserving Federated Learning Scheme Based on Homomorphic Proxy Re-Encryption. Information Sciences, 2026(730): 122875.

[14] Shabir M, Torta G, Damiani F, 2025, TinyML Model Compression: A Comparative Study of Pruning and Quantization on Selected Standard and Custom Neural Networks. Telecommunication Systems, 88(4): 132.

[15] Zhang J, Hu J, Jiang M, et al., 2023, A HCI-Hardened Self-Healing Operational Amplifier Circuit. Microelectronics Reliability, 2023: 151.