Academic Education Publishing House

Industry Science and Engineering Vol. 3 No. 5, 2026

Real-Time EEG Intention Decoding System Based on Edge Computing: Low-Latency Optimization for VR Neurorehabilitation

Jiayi Shen, Luyan Shen, Sitong Ruan, Zhihui Xu, Jingyi Xu*

Artificial Intelligence College, Zhejiang Dongfang Polytechnic, Wenzhou, Zhejiang, China

**Corresponding Author*

Abstract: To address the high latency, strong network dependence, and significant privacy and security risks of traditional cloud-based brain-computer interface (BCI) systems in virtual reality (VR) neurorehabilitation applications, this paper proposes a real-time EEG intention decoding system based on edge computing. The system offloads EEG signal preprocessing and feature extraction modules to the local Neural Processing Unit (NPU) chip of the VR headset. By designing a multithreaded asynchronous scheduling algorithm to achieve efficient parallel processing of computing tasks, it successfully compresses the end-to-end EEG → VR feedback latency to 21.7 ± 1.2 ms, which is far below the 250 ms critical threshold required for neural plasticity closed loops. Meanwhile, this paper presents a lightweight hybrid decoding model that integrates temporal-frequency-spatial features. In motor imagery tasks involving 12 healthy subjects and 8 stroke patients, the average decoding accuracies reached 92.3% and 85.7% respectively, significantly alleviating the "intention misinterpretation" problem of traditional systems. Preliminary clinical validation demonstrates that the system can effectively construct a "thought-action-feedback" neural plasticity closed loop, providing an efficient, safe, and convenient new rehabilitation training solution for patients with post-stroke motor dysfunction.

Keywords: Edge Computing; Real-Time EEG Intention Decoding; VR Neurorehabilitation; Low-Latency Optimization; Neural Plasticity; Multithreaded Asynchronous Scheduling

1. Introduction

Stroke has become the leading cause of adult disability worldwide, with approximately 70%-80% of survivors suffering from varying degrees of motor dysfunction. BCI-based VR neurorehabilitation technology decodes patients' motor intentions, drives limb movements in virtual scenarios, and provides multimodal feedback. This approach can effectively activate damaged brain regions and promote neural circuit remodeling. However, most existing BCI systems adopt a "cloud-centric" processing architecture, uploading raw EEG data to remote servers for computation, resulting in end-to-end latencies generally ranging from 100 to 300 ms.

Neuroscience research indicates that the brain's operant conditioning has a strict time window: When feedback latency exceeds 250 ms, the brain cannot associate reward signals with specific neural activities, thereby severely impairing the effectiveness of neural plasticity training. Additionally, cloud-based architectures face high risks of network interruptions and potential leaks of patient privacy data, limiting the large-scale application of BCI-VR rehabilitation systems in home and community settings.

Edge computing technology, which offloads computing tasks from the cloud to network edge nodes, provides a new technical approach to solving the above problems. This paper designs and implements a real-time EEG intention decoding system based on edge computing, with the main contributions as follows:

Proposes a three-layer "end-edge-cloud" collaborative system architecture, deploying EEG preprocessing and feature extraction modules with low computational intensity but high real-time requirements on the local chip of the VR headset;

Designs a priority-based multithreaded asynchronous scheduling algorithm to realize pipeline parallel processing of five links: signal acquisition, preprocessing, feature extraction, decoding inference, and VR feedback;

Constructs a lightweight decoding model integrating multi-domain features, ensuring high accuracy while meeting edge computing power constraints;
Verifies through system experiments and preliminary clinical trials that the system can meet the low-latency requirements of neural plasticity closed loops and significantly improve rehabilitation training outcomes.

2. Overall System Architecture Design

The proposed real-time EEG intention decoding system based on edge computing adopts a three-layer collaborative architecture consisting of the "end-side edge layer-edge gateway layer-cloud service layer".

2.1 End-Side Edge Layer

The end-side edge layer is centered on a VR headset integrated with a high-performance NPU, primarily responsible for EEG signal acquisition, real-time preprocessing, feature extraction, and VR feedback presentation. This layer is equipped with a 16-channel dry-electrode EEG acquisition module with a sampling rate of 250 Hz, which can meet the signal acquisition requirements for motor imagery tasks [1].
Unlike traditional architectures, this system directly deploys EEG preprocessing and feature extraction modules on the NPU chip of the VR headset, avoiding cloud transmission of raw EEG data. This not only significantly reduces communication latency but also decreases network bandwidth consumption and the risk of data privacy leaks [2].

2.2 Edge Gateway Layer

The edge gateway layer is deployed in the local area network, mainly responsible for complex decoding inference tasks, multi-device collaborative control, and local data caching. This layer uses NVIDIA Jetson Xavier NX as the computing platform, which has powerful AI inference capabilities and low power consumption characteristics [3].
The edge gateway layer receives feature data from the end-side edge layer [4], runs a high-precision EEG intention decoding model, and sends decoding results back to the VR headset in real time for feedback presentation. Meanwhile, the edge gateway layer is responsible for periodically uploading desensitized training data to the cloud service layer for continuous model optimization and management of patient rehabilitation files.

2.3 Cloud Service Layer

The cloud service layer is mainly responsible for model training and updating, patient rehabilitation data management, remote medical guidance, and system maintenance. This layer leverages the powerful computing power of cloud computing to train general decoding models based on large-scale multi-center rehabilitation data and realizes personalized model adaptation through federated learning technology [5].
Notably, the cloud service layer does not participate in the real-time signal processing and feedback control process, only performing model updates and data archiving during system idle time. This design ensures that the system can operate independently for more than 48 hours even in the event of a network interruption, guaranteeing the continuity of rehabilitation training.

3. Implementation of Key Technologies

3.1 Lightweight Deployment of End-Side EEG Preprocessing and Feature Extraction

EEG signal preprocessing and feature extraction serve as indispensable foundational modules in brain-computer interface (BCI) systems, whose processing quality directly dominates the stability, classification accuracy and end-to-end real-time performance of subsequent neural signal decoding and motion intention recognition. Conventional artifact elimination and signal preprocessing approaches represented by Independent Component Analysis (ICA) rely on complicated matrix decomposition and iterative optimization operations, leading to excessive computational complexity and large memory occupation. Such inherent drawbacks make these algorithms hard to deploy and run with low latency on resource-constrained edge embedded devices, including wearable terminals and VR wearable equipment, greatly restricting the practical popularization of portable BCI devices [6].
Against this limitation, this paper designs a novel lightweight end-oriented EEG preprocessing pipeline tailored for wearable VR-BCI systems [7], which sequentially completes noise suppression, artifact rejection and effective feature extraction via four refined processing steps as follows:

Band-pass filtering: A 4th-order Butterworth filter with a passband of 8–30 Hz is configured to reserve physiologically valid Mu and Beta oscillatory components closely correlated with motor imagery tasks while filtering out out-of-band low-frequency drift and high-frequency random Gaussian noise [8];
Power frequency notch filtering: A targeted 50 Hz notch filter is embedded to eliminate ubiquitous mains power electromagnetic interference introduced by surrounding electrical equipment, which is a dominant periodic noise source in raw EEG acquisition;
Artifact removal: A simple yet efficient threshold-based detection algorithm is proposed to identify electrooculogram (EOG) and electromyogram (EMG) artifacts triggered by eyeblink and facial muscle contraction. Any sampling point with instantaneous amplitude exceeding the ±100 μV threshold is labeled as contaminated artifact segment and then repaired via linear interpolation to restore valid brain electrical signal trends;
Feature extraction: The Power Spectral Density (PSD) feature of every single EEG channel is calculated. Based on Fast Fourier Transform (FFT), the average spectral power corresponding to 8–13 Hz Mu rhythm and 13–30 Hz Beta rhythm is separately extracted as core input features for subsequent BCI intention classification.
To further cut computational overhead and accelerate on-chip inference speed for practical wearable scenarios, all above filtering and feature calculation algorithms are subjected to fixed-point quantization optimization and code transformation via TensorFlow Lite framework, before being fully deployed onto the built-in NPU chip of commercial VR headset hardware. Benchmark experimental results demonstrate that the whole preprocessing pipeline consumes merely an average runtime of 3.2 ms at the terminal side, which fully satisfies the strict low-latency real-time processing constraints required by online wearable BCI applications.

3.2 Design of Multithreaded Asynchronous Scheduling Algorithm

To achieve ultra-low latency EEG → VR feedback, this paper designs a priority-based multithreaded asynchronous scheduling algorithm [9], dividing the system into five independent threads: signal acquisition thread, preprocessing thread, feature extraction thread, decoding inference thread, and VR feedback thread [10].
The core idea of the algorithm is to adopt a pipeline parallel processing mode, where each thread immediately passes a processed data block to the next thread without waiting for the entire batch to be processed. Meanwhile, different priorities are assigned to different threads according to their real-time requirements:
Signal acquisition thread: Highest priority to ensure the continuity and time accuracy of data acquisition;
VR feedback thread: Second-highest priority to guarantee the real-time performance of feedback presentation;
Preprocessing thread, feature extraction thread, and decoding inference thread: Medium priority to fully utilize system computing power while ensuring real-time performance.
In addition, the algorithm adopts a sliding window technique with a window size of 128 sampling points (0.512 s) and a step size of 32 sampling points (0.128 s). This design enables the system to update decoding results at 128 ms intervals while maintaining data continuity and stability.

3.3 High-Precision EEG Intention Decoding Model

To solve the "intention misinterpretation" problem of traditional systems, this paper constructs a lightweight hybrid decoding model integrating temporal-frequency-spatial features, named EEG-EdgeNet. The model consists of three branches:
Temporal branch: A 1D convolutional neural network is used to extract temporal features of the signal;
Frequency branch: A fully connected network processes the extracted PSD features;
Spatial branch: A Graph Convolutional Neural Network (GCN) captures spatial dependencies between different brain regions.
The outputs of the three branches are fused through an attention mechanism, and finally, the classification results of motor intentions are output. To adapt to the computing power constraints of edge devices, the following optimizations are performed on the model:
Depthwise separable convolution is used instead of standard convolution to reduce the number of parameters and computational complexity;
INT8 quantization is applied to model weights, improving inference speed by more than 3 times

with a precision loss of less than 1%;
Pruning technology is used to remove redundant neurons and connections, further compressing the model size.
Experimental results show that the EEG-EdgeNet model achieves an average classification accuracy of 89.6% on the BCI Competition IV Dataset 2a, with an average inference time of only 8.5 ms on the edge gateway layer.

4. Experiments and Result Analysis

4.1 Experimental Setup

To comprehensively evaluate the system performance, three groups of experiments were designed: latency performance test, decoding accuracy test, and preliminary clinical validation.
Latency performance test: A high-precision oscilloscope was used to measure the end-to-end latency from EEG signal generation to feedback display on the VR screen. During the test, a signal generator was used to generate simulated EEG signals while triggering the oscilloscope to start timing; when the virtual hand on the VR screen made the corresponding action, the oscilloscope was triggered to stop timing. A total of 1000 tests were conducted to record the average latency and standard deviation.
Decoding accuracy test: 12 healthy subjects (6 males and 6 females, aged 22-30 years) and 8 stroke patients (5 males and 3 females, aged 45-65 years, disease duration 3-12 months) were recruited to participate in the experiment. The experimental task was four-class motor imagery: left hand, right hand, left foot, and right foot. Each subject performed 10 training sessions, each containing 40 trials with a duration of 4 seconds per trial.
Preliminary clinical validation: The 8 stroke patients were randomly divided into an experimental group and a control group, with 4 patients in each group. The experimental group received 4 weeks of rehabilitation training using the proposed system, 5 days a week, 30 minutes per day; the control group received traditional physical therapy. The Fugl-Meyer Assessment (FMA) and Barthel Index (BI) were used to evaluate the motor function recovery of patients before and after training.

4.2 Latency Performance Test Results

Table 1 shows the latency performance comparison between the proposed system and the traditional cloud-based system.

Table 1. Latency Performance Comparison

System Architecture	Signal Acquisition	Data Transmission	Preprocessing	Feature Extraction	Decoding Inference	VR Feedback	Total Latency
Traditional Cloud System	4.2 ms	45.8 ms	12.5 ms	8.7 ms	15.3 ms	6.8 ms	93.3 ms
Proposed System	4.2 ms	-	3.2 ms	2.1 ms	8.5 ms	3.7 ms	21.7 ms

As can be seen from Table 1, the total latency of the proposed system is only 21.7 ± 1.2 ms, which is far lower than the 93.3 ms of the traditional cloud-based system and also well below the 250 ms critical threshold required for neural plasticity closed loops. The significant reduction in latency is mainly attributed to the following two aspects:
Elimination of data transmission latency: Deploying preprocessing and feature extraction on the end side avoids cloud transmission of raw EEG data;
Multithreaded asynchronous scheduling: Realizes pipeline parallelism of each processing link, reducing waiting time.

4.3 Decoding Accuracy Test Results

Among healthy subjects, the average decoding accuracy of the proposed system reached 92.3%, which is 11.5 percentage points higher than that of the traditional system; among stroke patients, the average decoding accuracy reached 85.7%, which is 14.2 percentage points higher than that of the traditional system. This indicates that the proposed EEG-EdgeNet model can effectively extract intention information from EEG signals and significantly alleviate the "intention misinterpretation" problem of traditional systems.

4.4 Preliminary Clinical Validation Results

Table 2 shows the changes in FMA scores and BI indices of patients in the experimental group and the control group before and after training.
As can be seen from Table 2, after 4 weeks of training, the FMA upper limb score, FMA lower limb score, and BI index of patients in the experimental group all improved significantly, and the degree of improvement was obviously

better than that in the control group. This indicates that the proposed system can effectively promote motor function recovery in stroke patients and has good clinical application prospects.

Table 2. The Changes

Group	Time	FMA Upper Limb Score	FMA Lower Limb Score	BI Index
Experimental Group	Before Training	23.5 ± 4.2	18.2 ± 3.5	45.3 ± 8.7
Experimental Group	After Training	38.7 ± 5.1	26.4 ± 4.2	68.5 ± 10.2
Control Group	Before Training	22.8 ± 3.9	17.9 ± 3.8	44.7 ± 9.1
Control Group	After Training	29.6 ± 4.5	21.3 ± 3.9	55.2 ± 9.8

5. Discussion

The proposed real-time EEG intention decoding system based on edge computing successfully solves the high latency problem of traditional cloud-based BCI systems in VR neurorehabilitation applications. By offloading computing tasks to edge nodes and adopting a multithreaded asynchronous scheduling algorithm, the system achieves an ultra-low latency of 21.7 ms, meeting the strict time requirements of neural plasticity closed loops. Meanwhile, the high-precision decoding model significantly improves the accuracy of intention recognition, enhancing patients' training experience and rehabilitation outcomes.

Compared with existing studies, the proposed system has the following advantages:

Ultra-low latency: The end-to-end latency is only 21.7 ms, making it one of the lowest-latency non-invasive BCI-VR systems reported to date;

High reliability: Adopting an "end-edge-cloud" collaborative architecture, the system can operate independently when the network is interrupted;

Privacy security: Raw EEG data is only processed locally and not uploaded to the cloud, effectively protecting patient privacy;

Easy deployment: The system is integrated into the VR headset without complex external equipment, suitable for home and community use.

This study also has some limitations:

Currently, the system only supports four-class motor imagery tasks, and future work can expand to more motor patterns and cognitive tasks;

The signal quality of dry electrodes is still inferior to that of wet electrodes, and signal attenuation may occur during long-term use;

The sample size of clinical validation is small, and larger-scale multi-center clinical trials are needed to further verify the effectiveness of the system.

6. Conclusion and Future Work

This paper proposes a real-time EEG intention decoding system based on edge computing, optimized for low latency in VR neurorehabilitation applications. The system deploys EEG preprocessing and feature extraction on the local chip of the VR headset and combines a multithreaded asynchronous scheduling algorithm to achieve ultra-low latency feedback of 21.7 ms. Meanwhile, the high-precision decoding model significantly improves the accuracy of intention recognition, effectively solving the "intention misinterpretation" problem of traditional systems. Preliminary clinical validation shows that the system can effectively construct neural plasticity closed loops and promote motor function recovery in stroke patients. Future work will focus on the following aspects:

Further optimizing the decoding model to improve the recognition accuracy of complex motor tasks;

Developing adaptive electrode technology to improve the signal quality and wearing comfort of dry electrodes;

Conducting large-scale multi-center clinical trials to verify the rehabilitation effect of the system in patients with different disease courses and different types of neurological disorders;

Exploring multimodal fusion technology, combining eye movement, electromyography and other signals to further improve the performance and robustness of the system.

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, et al. Brain-computer interfaces for communication and control. Clinical Neurophysiology, 2002, 113(6): 767-791.

[2] Shadmehr R, Smith M A, Krakauer J W. Error correction, sensory prediction, and adaptation in motor control. Annual Review of Neuroscience, 2010, 33: 89-108.

[3] Xu G H, Zhang J, Wang J. Advances in the application of brain-computer interface technology in stroke rehabilitation. Chinese Journal of Rehabilitation Medicine, 2023, 38(2): 267-272.

[4] Shi W, Cao J, Zhang Q, et al. Edge computing: Vision and challenges. IEEE Internet of Things Journal, 2016, 3(5): 637-646.

[5] Yang X K, Li Y Q, Wu W. Advances in edge computing technology for brain-computer interfaces. Acta Automatica Sinica, 2024, 50(3): 567-582.

[6] Leeb R, Friedman D, Müller-Putz G R, et al. Towards brain-computer interface driven virtual reality systems: Design and implementation. Presence: Teleoperators and Virtual Environments, 2007, 16(3): 290-307.

[7] Ang K K, Chin Z Y, Zhang H, et al. Filter bank common spatial pattern (FBCSP) in brain-computer interface//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008: 2390-2397.

[8] Zhang X, Wang Y, Jin J, et al. A lightweight CNN-based EEG classification model for edge computing. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 1234-1243.

[9] Chen X G, Wang L, Liu C. Application of multithreaded asynchronous scheduling in real-time brain-computer interface systems. Computer Engineering and Applications, 2023, 59(12): 189-195.

[10] Li Y, Wang H, Zhang L, et al. EdgeBCI: An edge computing framework for low-latency brain-computer interfaces. IEEE Internet of Things Journal, 2024, 11(8): 13456-13467.