OPTIMIND: Adaptive Neural Variability-Aware Architecture for Real-Time Brain-Computer Interfaces

Jamil Habash
Department of Computer Engineering
Birzeit University
Ramallah, Palestine
1220194

Abdallah Kokash
Department of Computer Engineering
Birzeit University
Ramallah, Palestine
1220116

Hanna Kaibni
Department of Computer Engineering
Birzeit University
Ramallah, Palestine
1220214

***Abstract*—Implantable and edge brain–computer interfaces (BCIs) must operate under strict latency, power, and memory constraints while maintaining robust performance in the presence of long-term neural variability. Emotional state changes introduce additional non-stationarities in neural signals, degrading decoding accuracy and increasing learning overhead if treated as noise. This paper presents OPTIMIND, an adaptive, emotion-aware BCI architecture that exploits emotion-related neural variability as a system-level control signal to improve learning efficiency and memory behavior. Building on a learning-optimized non-volatile memory (NVM) framework, OPTIMIND integrates a lightweight EEG-based emotion recognition pipeline with dedicated feature extraction and kernel computation units, an optimized memory hierarchy with feature and support vector caching, predictive prefetching, and confidence-aware adaptive learning. Emotion-conditioned, margin-based update filtering selectively triggers model updates, significantly reducing unnecessary NVM writes while preserving accuracy. simulation using the EEGEmotions-27 dataset demonstrates that OPTIMIND achieves an average online classification accuracy of 81.0% with a low average end-to-end latency of 31.60 ms in a fully streaming configuration, while maintaining bounded model growth. These results indicate that OPTIMIND enables efficient, real-time, and long-term deployment of emotion-aware, learning-enabled BCIs under tight hardware constraints.**

***Index Terms*—Brain-Computer Interface, Emotion Recognition, Continual Learning, Non-Volatile Memory, Support Vector Machine, Low-Power System, Neural Variability, Affective Computing**

I. Introduction

Brain-computer interfaces (BCIs) connect biological neurons in the brain with computers and machines. BCIs are advancing our understanding of the brain, helping to treat neurological/neuropsychiatric disorders, and helping to restore lost sensorimotor function.

Among the various types of brain-computer interface (BCIs) systems, implantable BCIs are of particular interest because they provide direct interaction with the brain through neural probes implanted beneath the skull. These systems integrate processors that can process neural signals on-site, enabling precise and high-fidelity recording and stimulation of neuronal activity. Implantable BCIs offer superior spatiotemporal resolution compared to non-invasive methods, making them valuable for both research and clinical applications. Prior studies have utilized implantable BCIs to advance our understanding of brain function and to develop treatments for neurological disorders. Clinically approved devices, for example, employ implantable BCIs to detect and prevent seizures or to monitor and slow brain degeneration, demonstrating their potential for improving patient outcomes.

Implantable BCI systems rely on specialized hardware accelerators to perform neural signal processing efficiently and safely. To be effective in medical applications, these systems must operate with sub-second latency, ensuring real-time responsiveness. At the same time, they must stay within a tight power budget of tens of milliwatts to avoid overheating and prevent damage to surrounding brain tissue, making energy-efficient hardware design a critical requirement for implantable BCIs.

Recent studies integrate non-volatile memories (NVM) into BCI Systems for heavy workloads, Implantable BCI technology is scaling with time which increases the number of neurons the system can read at a time, and with complex algorithms like spike sorting and machine learning models, the need of a memory to store these data became inevitable, thats why NVM memory systems are being used for there lower power consumption and smaller area footprints compared to SRAMs.

Despite their benefits, NVM-assisted BCI systems potentially violate the latency and power budgets when supporting learning in BCI applications, Learning is a critical component of BCI applications, as it allows the system to adapt to neural signal variability over time and to improve decoding accuracy for tasks such as motor control or cognitive state detection. In NVM-assisted BCI systems, learning involves frequent reading and writing of neural data and model parameters in memory. While reading from NVM is generally fast and energy-efficient, writing incurs a significant overhead, both in latency and power consumption. This slower write operation can become a bottleneck, potentially violating the strict real-time and energy constraints of implantable BCIs,

thats why in this research we used InfiniMind which is a Learning-Optimized Large-Scale Brain-Computer Interface, used To reduce excessive write operation using four optimization schemes, Update Filtering, Delta Buffering, Out-of-Place Flushing and Waveform Compression.

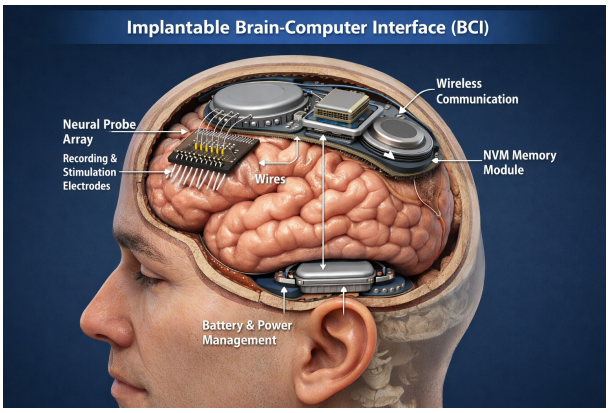

Fig. 1. Implantable BCI

Emotion recognition plays a crucial role in human-computer interaction and serves as the first step in understanding the feelings of a subject. It has broad application prospects, particularly in the diagnosis and treatment of mental disorders, and is considered an important research topic within the field of brain–computer interaction. Traditional emotion recognition methods have primarily relied on information such as questionnaires, facial expressions, and voice signals. However, these approaches face challenges, including strong subjectivity and susceptibility to external interference. Recent advances in brain imaging techniques allow researchers to noninvasively detect neural activity during emotion elicitation, enabling the identification of emotional categories based on neural signatures. Compared with behavioral characteristics, such as facial expressions and voice, neural activity related to emotion is difficult to conceal. Furthermore, compared with physiological signals, such as skin conductivity electromyography, and electrocardiography, neural activity provides higher specificity across different emotional categories and contains more informative features that can effectively distinguish between emotions.

In this work, we propose a low-power, implantable BCI EEG-based emotion recognition, implemented on the InfiniMind in-memory computing platform. The system employs a support vector machine (SVM) classifier and introduces dedicated hardware support for feature extraction, kernel computation, and online inference. the proposed system targets real-time EEG-based emotion recognition while operating under strict power, area, and memory constraints typical of edge and wearable brain–computer interface (BCI) devices.

To support continuous operation on resource-constrained devices, the proposed architecture incorporates an adaptive online learning mechanism. Model updates are selectively triggered based on classification confidence using margin-based filtering, allowing the system to learn from new data while avoiding unnecessary computation and memory writes. This approach enables personalization and long-term deployment without relying on frequent cloud retraining, which is critical for privacy-sensitive IoT applications.

Finally, we introduce an enhanced memory hierarchy optimized for the access patterns of SVM inference and learning. The hierarchy integrates feature and support vector caching, predictive prefetching driven by short-term emotion transitions and deferred update buffering. These techniques significantly reduce memory access latency and nonvolatile memory traffic, enabling efficient, real-time EEG-based emotion recognition suitable for wearable and edge IoT BCI systems.

II. Background

A. Implantable BCI System

Implantable brain-computer interface (BCI) systems utilize neural probes placed directly on brain tissue to establish a direct interface with the brain. These neural probes capture electrical activity by detecting voltage fluctuations generated by nearby neurons. The recorded neural signals are subsequently relayed to a dedicated processor integrated with the neural probes. This processor analyzes and decodes the recorded data to support a wide range of applications, including movement decoding, seizure detection, and neuroscientific research. Based on the decoded outcomes, the BCI system can inject electrical stimuli to treat neurological disorders, such as seizure prevention and prosthetic vision restoration.

An implantable BCI system must satisfy several critical design objectives:

1) Thermal and Power Constraints: The implant must not raise the temperature of surrounding brain tissue by more than 1 °C. Studies have shown that excessive power consumption at the implantation site can lead to overheating and potential damage to brain tissue.

2) Latency Constraints: The processing unit must operate under strict latency requirements to enable real-time medical applications. Millisecond-scale latency is required, such as approximately 50 ms for movement decoding and 10 ms for seizure detection.

3) Workload Coverage: The processor must provide high workload coverage. Modern BCI systems are rapidly scaling to record activity from an increasing number of neurons simultaneously, while application workloads are becoming increasingly complex and data intensive.

4) Device Lifetime: Implantable BCI systems are inserted through burr holes in the skull, which introduces surgical risks such as infection and tissue injury. To minimize the need for surgical replacement, the implanted system must operate reliably over extended periods. Commercial implantable BCI systems are typically designed to guarantee an operational lifetime of at least 10 years.

B. Learning in BCI System

Brain computer interface (BCI) systems operate on neural signals that exhibit significant variability over time. These variations arise from multiple factors, including electrode

degradation, tissue response, long-term neural plasticity, and changes in the user's cognitive or physiological state. As a result, models trained during an initial calibration phase often experience a gradual decline in decoding accuracy if they remain fixed during long-term operation. Learning mechanisms are therefore essential to maintain reliable performance in practical BCI deployments.

Learning in BCI systems enables continuous or periodic adaptation of decoding models to account for changes in neural signal statistics. By updating model parameters during operation, the system can compensate for signal drift and preserve accuracy without requiring complete retraining. This capability is particularly important for implantable BCIs, which are expected to function reliably over long-time spans with minimal user intervention. Studies have shown that adaptive learning significantly improves robustness under conditions such as channel loss, signal amplitude changes, and long-term use.

However, incorporating learning into BCI systems introduces important system-level challenges. Learning requires frequent parameter updates, memory accesses, and intermediate data storage, all of which increase power consumption and computational load. In implantable environments, where strict constraints on energy, heat dissipation, and memory endurance exist, excessive learning activity can negatively impact device safety and lifetime. Repeated memory writes can accelerate wear in non-volatile memory and reduce the long-term reliability of the system.

Despite these challenges, on-device learning remains a key requirement for next-generation BCIs. Local learning reduces dependence on wireless communication, lowers latency, and improves privacy by avoiding continuous data transmission. Consequently, recent research emphasizes the need for learning-aware BCI architectures that intelligently manage when and how learning updates occur. Such approaches aim to balance adaptability and efficiency by aligning learning behavior with system constraints, enabling stable and energy-efficient operation in real-world BCI applications.

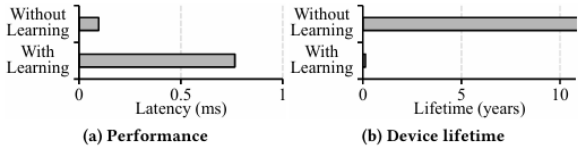

Fig. 2. Limitations of an BCI system when supporting a BCI learning workload

C. Memory in BCI Systems

1) SRAM-Based Memory in BCI Systems: Early BCI systems primarily rely on on-chip static random-access memory (SRAM) tightly coupled with specialized processing units to meet strict latency and power constraints. SRAM provides fast access that is essential for real-time neural signal processing tasks such as spike sorting, feature extraction, and decoding. Architectures such as HALO and its distributed extension SCALO demonstrate how accelerator-rich designs with local SRAM buffers can efficiently support BCI inference workloads under milliwatt-level power budgets. However, SRAM-based designs face limited scalability due to high area and leakage power, making it difficult to store large model parameters or long-term neural histories as BCI channel counts and algorithm complexity continue to increase.

2) NVM-Assisted Memory in BCI Systems: To overcome the scalability limitations of SRAM-only designs, recent BCI architectures integrate non-volatile memory (NVM) into the memory hierarchy. NVM technologies, such as NAND Flash, offer higher density and lower static power consumption, enabling the storage of large neural datasets, model parameters, and historical recordings required for complex and multi-site BCI workloads. Systems like SCALO augment on-chip SRAM with NVM to swap infrequently accessed data while preserving real-time performance. Despite these advantages, NVM-assisted BCI systems face challenges due to high write latency, page-level access granularity, and limited endurance, which can significantly impact performance and device lifetime when write-intensive operations are introduced.

3) NVM-Based Learning in BCI Systems: Long-term BCI deployments require continual on-device learning to adapt to non-stationarities in neural signals caused by probe drift, channel failure, and neuroplasticity. Learning algorithms, however, are inherently write-intensive and expose critical limitations of conventional NVM-assisted BCI systems. InfiniMind introduces a learning-optimized NVM-based BCI architecture that reduces excessive write overhead while preserving learning accuracy. By exploiting properties of neural signals such as sparsity, temporal locality, and redundancy by incorporating techniques such as update filtering, delta buffering, out-of-place flushing, and waveform compression within the memory controller. This design significantly improves both performance and NVM lifetime, enabling practical large-scale, lifelong learning in implantable BCI systems.

D. Emotion recognition in BCI System

Emotion recognition refers to the process of identifying a user's emotional state such as happiness, sadness, stress, or calmness using measurable physiological and neural signals. While traditional approaches infer emotions from external cues such as facial expressions, speech, or body language, these methods rely on voluntary expression and are often unreliable in clinical or assistive settings. Consequently, recent research increasingly focuses on extracting emotional information directly from neural activity, particularly using electroencephalography (EEG). EEG-based emotion recognition is attractive in brain–computer interface (BCI) systems due to its high temporal resolution and suitability for real-time monitoring of neural dynamics associated with emotional processing. However, continuous EEG acquisition and processing impose stringent architectural constraints on latency, memory bandwidth, and energy consumption, as emotion recognition pipelines require sustained preprocessing, feature extraction, and classification operations in low-power environments.

Emotional states significantly modulate cognitive processes such as attention, learning, memory formation, and decision-making, leading to variability in neural signal statistics that can degrade decoding performance if ignored. In adaptive BCI systems, this variability can cause model mismatch, reduced accuracy, and frequent retraining, motivating emotion-aware architectures capable of adapting to neural changes induced by emotional states while minimizing parameter updates, memory writes, and power overhead. Emotion-aware BCIs are particularly valuable in implantable and clinical applications where users may be unable to communicate emotions through motor or speech channels, enabling closed-loop systems that dynamically adjust decoding strategies, stimulation parameters, or learning rates. Supporting such functionality at the architectural level is challenging due to the low amplitude and noise-prone nature of EEG signals, which are susceptible to artifacts from eye blinks, muscle activity, and power-line interference. Mitigating these effects requires additional filtering and artifact removal, increasing computational load and memory traffic, and reinforcing the need for robust, variability-aware, and resource-efficient BCI architectures that meet strict latency, energy, and thermal constraints.

III. Literature

A. InfiniMind: A Learning-Optimized Large-Scale Brain-Computer Interface

The objective of this paper is to address the high energy, latency, and endurance cost of frequent NVM writes in learning-enabled BCI systems. The work introduces a memory-optimized architecture that reduces write frequency and size through four mechanisms: update filtering, delta buffering, out-of-place flushing, and waveform compression. The evaluation on multiple BCI workloads and learning libraries shows significant improvements in power efficiency, system latency, and memory lifetime compared to baseline designs, demonstrating that continual learning can be made practical for long-term BCI deployment.

The main limitation is that the proposed system is evaluated primarily through simulation and modeling, without validation using really neural or implant data, leaving long-term behavior and robustness uncertain. In addition, the added buffering, filtering, and compression logic increases hardware area and design complexity, and the reliance on profiling-based parameter settings may reduce effectiveness as neural signals change over time during real-world use.

B. MINDFUL: Safe, Implantable, Large-Scale Brain-Computer Interfaces from a System-Level Design Perspective

The objective of this paper is to propose an analytical framework to guide the design of implantable BCI architectures under strict power, thermal, and communication constraints. The study evaluates different BCI system types and architectures and quantitatively analyzes the trade-offs between sensing, computation, and wireless communication using realistic SoC models. The findings show that a system-level approach enables more efficient architectural decisions and clarifies the differences between computation-centric and communication-centric implant designs.

A key limitation is that the framework is entirely analytical and simulation-based, with no experimental validation on real implant hardware or in-vivo systems. In addition, application algorithms and hardware architecture are treated largely independently, which limits insight into real-time closed-loop operation and practical deployment of future implantable BCI.

C. EEG-Based BCI Emotion Recognition: A Survey

The objective of this paper was to address all the innovations that was done in the field BCI emotion, by examining methods like data collection, future extraction and classification models to identify current trend and their limitations, the findings reveled that EEG signals are a dependable and challenging-to-fake data source for decoding emotional states, but the complexity of emotions and their neural correlates make standardization difficult. Most research employs machine learning classifiers and features derived from frequency bands or functional connectivity, with many studies focusing on a limited set of emotions, typically mapped within valence-arousal spaces.

The paper's limitations include the lack of standardized protocols across studies, which makes comparisons challenging, and the predominance of small, unbalanced datasets that can bias results. Additionally, most research focuses on binary or limited emotion classes, restricting the generalizability to complex emotional states.

D. Cross-Subject Emotion Recognition Brain-Computer Interface Based on fNIRS and DBJNet

The study aimed to develop a robust cross-subject emotion recognition model based on fNIRS signals utilizing a dual-branch neural network architecture called DBJNet, to accurately distinguish among positive, neutral, and negative emotional states. the model achieved a 74.8% accuracy in distinguishing positive, neutral, and negative emotions. The model performed even better in binary tasks, with accuracies around 89-92%, demonstrating that combining spatial and temporal features greatly improves emotion decoding across individuals.

However, this study still has some limitations. First of all, the positive emotions are only happy, our negative emotions are only sad, and the subdivision of emotions is limited. In addition, the results show that the constructed fNIRS cross-subject emotion recognition framework has poor decoding performance for positive versus negative emotions, due to the problem of brain area coverage, that is, the specific processing brain areas of positive and negative emotions are not taken into account when designing fNIRS channel layout.

E. HALO: Hardware-Software Co-Design for Low-Power Implantable Brain-Computer Interfaces

The paper presents HALO, a hardware-software architecture that decomposes BCI workloads into reusable, configurable processing elements (PEs) connected via low power on-chip network that uses circuit switching. Using realistic 28nm FD-SOI silicon models and in-vivo electrophysiological data from

non-human primates, the paper evaluates trade-offs between flexibility, performance, and power, it compares HALO against monolithic ASICs and software-only RISC-V implementations. The results show that kernel decomposition, PE reuse, and algorithm refactoring enable HALO to support closed-loop tasks such as seizure prediction, movement intent detection, spike detection, compression, and encryption while remaining under a 15mW power budget, achieving 4-57× lower power than software solutions and 2× lower power than monolithic ASICs.

While architecture supports runtime reconfiguration via a RISC-V micro-controller and assumes one task pipeline active at a time, the complexity of multi-task concurrency under strict power budgets remains an open issue.

F. An EEG-Based Brain Computer Interface for Emotion Recognition and Its Application in Patients with Disorder of Consciousness

The objective of this paper is to develop a real-time EEG-based brain–computer interface (BCI) system for recognizing emotional states and to examine whether emotions can be detected in patients with disorders of consciousness (DOC). The system classifies emotions into positive and negative categories using EEG signals recorded while subjects watch emotionally stimulating video clips. It was first validated on ten healthy participants, achieving a high average online accuracy of 91.5%, demonstrating reliable real-time emotion recognition. The system was then applied to eight DOC patients, where three patients achieved statistically significant recognition accuracy, suggesting preserved emotional processing that is not observable through conventional behavioral assessments.

Despite these promising results, the study has several limitations. The small number of DOC patients limits the generalizability of the findings, and the binary emotion classification does not capture more complex emotional states. Additionally, the large variation in accuracy across patients indicates strong individual differences in emotional processing and EEG responses.

G. Improving EEG based brain computer interface emotion detection with EKO ALSTM model

The objective of this paper is to improve the accuracy and efficiency of EEG-based emotion recognition by proposing a novel deep learning model called Enhanced Kookaburra Optimized Adjustable Long Short-Term Memory (EKO-ALSTM). The study uses the DEAP dataset, which contains EEG recordings from participants watching music videos and includes emotion labels based on arousal, valence, and dominance. The methodology involves pre-processing EEG signals using a band-pass filter, extracting features through Discrete Wavelet Transform (DWT), and classifying emotions using the EKO-ALSTM model. The proposed optimization algorithm is used to tune the ALSTM hyperparameters, improve learning performance, and stability. Experimental results show that the proposed model achieves a high classification accuracy of 97.93%, outperforming traditional machine learning methods and existing deep learning models, while also achieving faster training and inference times. These results indicate that the proposed approach is well suited for real-time emotion recognition applications.

A key limitation of this study is that the evaluation is entirely based on an offline public dataset, with no validation on real-time EEG acquisition or wearable BCI systems. Additionally, although the model achieves high accuracy, the complexity of the optimization process may increase computational cost and limit practical deployment on low-power or embedded devices.

H. SCALO: Scalable Low-Power Architecture for Next-Generation Implantable Brain-Computer Interfaces

This paper introduces SCALO, a distributed BCI architecture comprising multiple wirelessly networked implants designed to support multi-site brain interfacing under power and thermal constraints. Building upon the HALO single-site BCI processor, SCALO extends the architecture to enable distributed processing across brain regions to create a critical capability for treating neurological disorders like seizure propagation that manifest as network-level dysfunctions. It augments HALO's processing elements with locality-sensitive hashing, distributed linear algebra, network compression, and non-volatile storage, using optimal scheduling via integer linear programming to map applications across implants. Using post-layout 28 nm silicon models and real neural datasets from human patients, the paper assesses scalability trade-offs across multi-site sensing, feature extraction, signal correlation, and wireless communication, it showed how SCALO achieves 506 Mbps aggregate throughput across 11 implants while maintaining ¡15 mW per implant.

A key limitation is that SCALO's evaluation relies primarily on simulation and synthesized hardware without chronic in-vivo validation or fully integrated implant prototypes. Additionally, the effectiveness of approximate computing and locality-sensitive hashing depends heavily on signal statistics and may be weak under long-term physiological variability.

I. Summary

The table below provides a comparative summary of representative brain–computer interface (BCI) systems and studies reviewed in this work. It highlights their primary objectives, key findings, and main limitations, offering a concise view of design trade-offs and recurring challenges across both algorithmic and hardware-oriented BCI research. Prior work reveals common bottlenecks related to power efficiency, scalability, data availability, and robustness under long-term neural variability. Notably, while existing systems address adaptive learning or emotion recognition in isolation, they do not exploit emotion-induced neural variability as an architectural control signal for learning or memory optimization. This gap motivates the need for more integrated and variability-aware BCI architectures, as pursued in this work.

TABLE I
SUMMARY OF REPRESENTATIVE BCI SYSTEMS AND STUDIES

Paper	Objective	Findings	Limitations
MINDFUL	System-level framework for implantable BCI design	Clarifies power, thermal, and communication trade-offs	Simulation-based; no real implant validation
InfiniMind	Reduce NVM write cost in learning-enabled BCIs	Improves power, latency, and memory lifetime	Higher hardware complexity; no real neural data
EEG-Based BCI Emotion Survey	Review EEG-based emotion recognition methods	Identifies trends in features and classifiers	Small datasets; limited emotion classes
fNIRS + DBiNet Emotion BCI	Cross-subject emotion recognition using fNIRS	Good cross-subject accuracy using spatial–temporal features	Limited emotion diversity; brain coverage issues
HALO	Decompose BCI workloads to reusable elements to maintain low power	Achieved 4–57× less power than software solutions and 2× lower power than monolithic ASICs	Limited support for parallel task execution under strict power constraints
EEG-Based Emotion BCI (DOC)	Real-time EEG emotion detection; assess residual awareness in DOC patients	High online accuracy in healthy subjects; emotion detection in some DOC patients	Small patient cohort; binary emotions only; subject-dependent performance
EKO-ALSTM EEG Emotion	Improve EEG emotion recognition accuracy using optimized DL	Very high accuracy; DWT features; optimized ALSTM outperforms baselines	Offline dataset only; high model complexity; no real-time validation
SCALO	Distributed BCI architecture comprising multiple wirelessly networked implants	506 Mbps aggregate throughput across 11 implants while maintaining < 15 mW per implant	Network's custom protocol lacks long-term durability data for multi-year deployment in RF-noisy brain environment

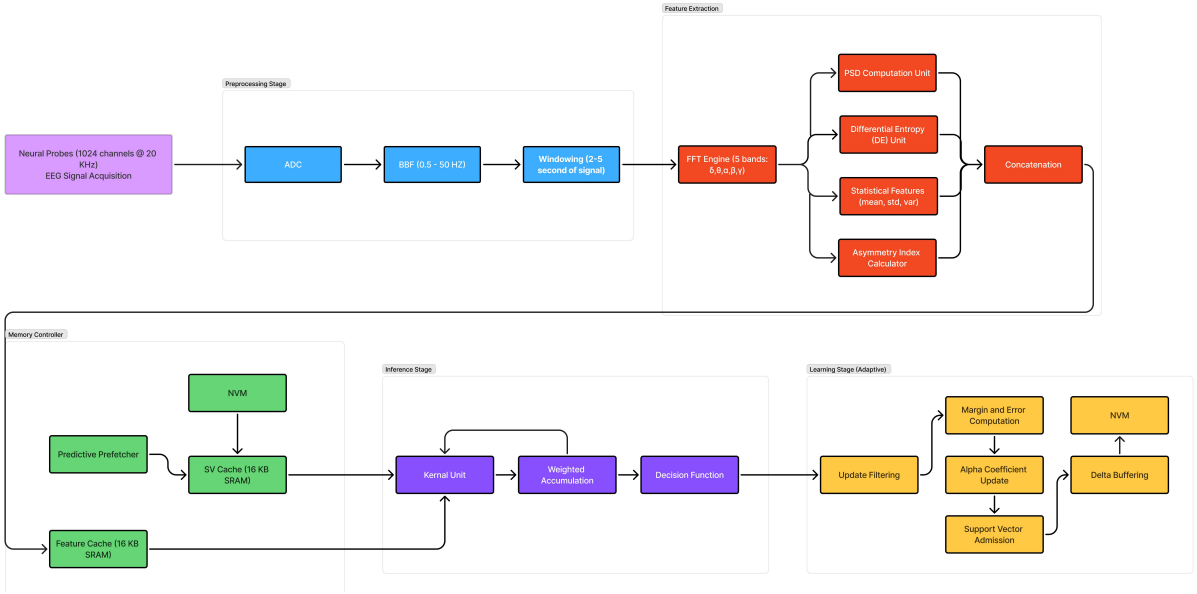

Fig. 3. Full Pipeline

IV. ARCHITECTURE DESIGN

A. System Overview

The proposed system implements a fully pipelined EEG-based emotion recognition architecture that integrates signal acquisition, preprocessing, feature extraction, inference, and adaptive learning into a unified framework. The design enables continuous, real-time operation by overlapping computation across successive EEG windows while maintaining low latency, energy efficiency, and adaptability to non-stationary neural signals. The pipeline processes multichannel EEG signals end-to-end and produces emotion classification outputs, with optional online model adaptation to handle subject variability and temporal drift.

B. Neural Signal Acquisition and Preprocessing Stage

The pipeline begins with neural signal acquisition, where multichannel EEG signals are captured from the scalp using high-density electrode arrays. These analog signals are streamed into the preprocessing stage, where they are digitized through an analog-to-digital converter (ADC). Preprocessing is responsible for conditioning the raw EEG signals to ensure robustness against noise and artifacts. A bandpass filter removes low-frequency baseline drift and high-frequency noise while

preserving physiologically relevant EEG rhythms. The filtered signal is then segmented into fixed-length temporal windows, enabling localized time–frequency analysis and supporting streaming operation. This stage produces clean, windowed EEG segments that serve as standardized inputs to the feature extraction stage.

C. Feature Extraction Stage

In the feature extraction stage, each preprocessed EEG window is transformed into a compact and informative feature representation. The windowed signal is first processed by a shared fast Fourier transform (FFT) engine, which converts time-domain signals into the frequency domain and separates spectral components corresponding to canonical EEG bands. The FFT output is reused by multiple feature computation units that operate in parallel, improving efficiency and reducing redundant computation. These units extract complementary descriptors, including spectral power, entropy-based measures, statistical summaries, and hemispheric asymmetry indicators. The resulting features capture both frequency-specific and spatial characteristics of neural activity. Outputs from all feature units are concatenated to form a fixed-dimensional feature vector that summarizes the neural state within each time window.

D. Feature and Model Data Management

The extracted feature vectors are first stored in a local feature cache to minimize access latency and enable efficient reuse during inference. The memory controller incorporates a predictive prefetching unit that exploits temporal continuity in emotional states by predicting the next likely emotion based on previous classifications. This prediction is used to proactively fetch the corresponding support vectors associated with the anticipated emotion class into a local support vector cache, thereby reducing frequent accesses to non-volatile memory and lowering memory overhead. If the required support vectors are not available in the cache, they are retrieved from non-volatile memory on demand. The cached feature vectors and support vectors are then supplied to the inference stage, enabling low-latency and energy-efficient classification.

E. Inference Stage

The inference stage performs emotion classification using the extracted feature vectors and the stored model parameters. For each feature vector, a kernel computation unit evaluates similarity measures between the input and stored support vectors. The resulting kernel values are combined through weighted accumulation to compute a decision value representing the classifier's confidence. A decision function then maps this value to a discrete emotion class, such as positive or negative affect. The inference stage is fully pipelined, allowing classification of the current window while subsequent windows are simultaneously undergoing preprocessing and feature extraction. The output of this stage is a real-time emotion prediction for each EEG window.

F. Adaptive Learning Stage

To address non-stationarity in EEG signals and subject-specific variability, the system includes an optional adaptive learning stage that operates alongside inference. Rather than updating the model continuously, this stage is selectively activated based on confidence measures derived from the classifier output. Margin-based update filtering determines whether the current sample provides useful information for adaptation. When learning is triggered, lightweight update operations adjust model parameters incrementally, and informative samples may be admitted as new support vectors. Parameter updates are buffered and written to non-volatile memory in a controlled manner to reduce energy consumption and memory wear.

G. End-to-End Pipeline Behavior and Outcome

Prior to real-time deployment, the underlying classification model is first trained offline using labeled EEG data, and its parameters are tuned to optimize performance for the target task and subject population. This initial training phase establishes the support vectors, kernel parameters, and decision thresholds required for reliable inference. Once the model parameters are fixed and loaded into the system memory, the proposed pipeline is activated for continuous, real-time operation.

During runtime, the pipeline processes incoming EEG signals in a streaming, window-based manner, with signal acquisition, preprocessing, feature extraction, inference, and optional adaptive learning operating concurrently across successive windows. The pipelined organization allows each stage to overlap in time, thereby maximizing throughput while maintaining bounded latency and power consumption. The system produces a continuous sequence of emotion classification outputs, and, when enabled, selectively adapts the model to compensate for signal non-stationarity and subject-specific variations without interrupting real-time inference. Overall, the pipeline delivers low-latency emotion recognition with the flexibility to balance stability and adaptability in long-term BCI deployment.

V. Implementation and Experimental Setup

A. Feature Extraction

1) Frequency-Domain Transformation: The windowed EEG samples are first streamed into the FFT engine, which performs a radix-2 fast Fourier transform to convert the time-domain signal into complex frequency-domain coefficients:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1 \qquad (1)$$

The FFT engine outputs real and imaginary components for each frequency bin, which are temporarily stored in an FFT output buffer. This buffer serves as a broadcast point that allows multiple feature units to access the same frequency-domain data in parallel without requiring additional FFT computations.

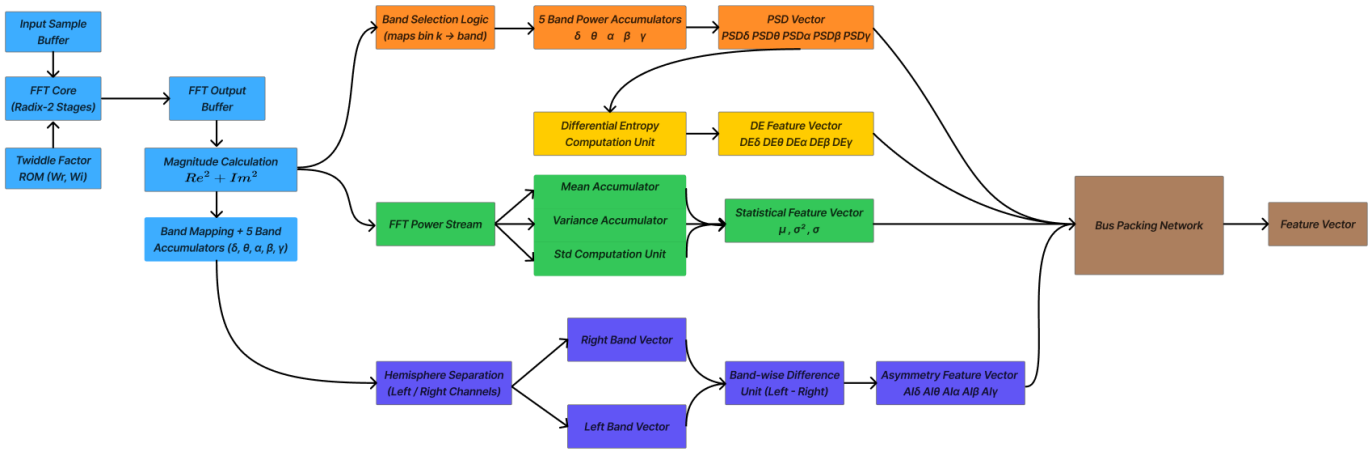

Fig. 4. Feature Extraction Full Diagram

2) Power Spectral Density Computation: Spectral power features are computed by the PSD unit, which consumes the FFT output and calculates the magnitude squared of each frequency bin:

$$P[k] = \Re\{X[k]\}^2 + \Im\{X[k]\}^2 \tag{2}$$

The resulting power values are grouped according to predefined EEG frequency bands (delta, theta, alpha, beta, and gamma) using band mapping logic, and accumulated over the window to produce a band-level PSD feature vector:

$$E_b = \sum_{k \in B_b} P[k] \tag{3}$$

where B_b denotes the set of FFT bins associated with band b. These band energies capture the distribution of signal power across canonical EEG rhythms and form a core component of the feature representation.

3) Differential Entropy Features: To further characterize signal complexity, the differential entropy unit reuses the PSD feature vector as its input. By operating directly on precomputed band energies, the entropy unit avoids repeated access to FFT bins and additional band grouping logic. The unit derives entropy features for each EEG band according to:

$$D_{E_b} = \frac{1}{2} \ln\left(2\pi e \sigma_b^2\right) \tag{4}$$

where σ_b^2represents the variance of the band-limited signal power for band b. This produces a compact representation of spectral variability that complements the power-based features.

4) Statistical Feature Computation: In parallel, statistical features are extracted from the windowed signal or its frequency-derived representations. The statistical feature unit computes first- and second-order statistics, including mean, variance, and standard deviation, over the FFT-derived magnitude or power values. These features summarize the overall distribution and dispersion of spectral components within the EEG window, providing robustness against noise and amplitude variations.

5) Hemispheric Asymmetry Analysis: Hemispheric asymmetry features are computed by the asymmetry index unit, which evaluates differences in spectral activity between left and right hemispheric electrode groups. Using band-level power information, the unit separates band-level spectral power by left and right hemispheric electrode groups and computes band-wise asymmetry values according to:

$$AI_b = \frac{P_b^L - P_b^R}{P_b^L + P_b^R + \epsilon} \tag{5}$$

where P_b^L and P_b^R denote the band energies of the left and right hemispheres, respectively, and ϵ is a small constant added for numerical stability. These features capture spatial characteristics of neural activity that are known to correlate with emotional processing.

6) Feature Vector Formation: Once all feature computation units complete their operation, the resulting PSD, differential entropy, statistical, and asymmetry feature vectors are combined by a concatenation unit. This unit is implemented as a fixed bit-slice packing network that maps each feature field into predefined positions within a fixed-dimensional feature vector. The resulting vector provides a consistent and ordered representation of the neural state for the current EEG window and is forwarded to the memory subsystem for caching and subsequent inference.

B. Memory Controller

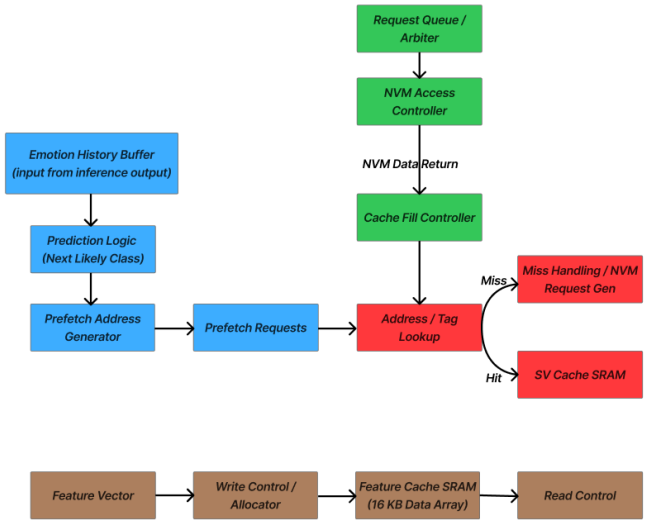

Fig. 5. Memory Controller Full Diagram

1) Feature Cache Management: Once feature extraction is complete, the generated feature vector is forwarded to the memory controller and written into a local feature cache implemented using on-chip SRAM. A write control and allocation unit manage cache placement and ensure that recently generated feature vectors are stored efficiently. During inference, the feature cache provides low-latency access to cached feature vectors through a dedicated read control path, eliminating unnecessary re-computation and avoiding repeated accesses to off-chip or non-volatile storage.

2) Support Vector Cache and Lookup: Support vectors required by the inference engine are managed by a separate support vector cache, also implemented using on-chip SRAM. Incoming support-vector requests are first processed by an address and tag lookup unit, which determines whether the requested vector is available in the cache. On a cache hit, the support vector is directly supplied to the inference engine, ensuring minimal access latency. On a cache miss, the request is forwarded to a miss-handling unit that generates the appropriate NVM access request.

3) Predictive Prefetching Mechanism: To reduce the frequency of cache misses, the memory controller incorporates a predictive prefetcher that exploits temporal locality in emotional state transitions. The prefetcher monitors recent inference outcomes through an emotion history buffer and applies prediction logic to estimate the most likely next emotion class:

$$\hat{L}_{t+1} = \arg\max_{L \in \mathcal{L}} \Pr\left(L \mid L_t, L_{t-1}, \ldots\right) \tag{6}$$

Based on this prediction, a prefetch address generator produces speculative support-vector requests, which are issued ahead of time and routed through the same address and tag lookup path as demand requests. This unified handling ensures consistent cache behavior while allowing speculative data to be staged in the support vector cache before it is explicitly requested.

4) Non-Volatile Memory Access and Cache Fill: All demand and prefetch requests that miss in the support vector cache are queued and arbitrated by a request queue and arbiter. The selected requests are then serviced by the NVM access controller, which manages communication with non-volatile memory and handles variable access latency. Once the requested data is returned from NVM, a cache fill controller inserts the fetched support vectors into the support vector cache, making them immediately available for subsequent inference operations.

5) Architectural Impact: By combining feature caching, support-vector caching, and predictive prefetching within a unified memory controller, the proposed architecture significantly reduces NVM traffic and hides memory access latency. This design enables low-latency and energy-efficient inference while maintaining scalability through non-volatile storage of large model parameters.

C. Inference Stage

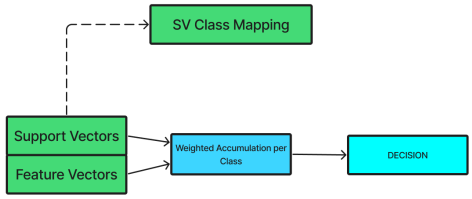

Fig. 6. Inference Engine and Decision Function Architecture

The computation is decomposed into three pipeline stages: element-wise difference computation, accumulation of squared differences to form the Euclidean distance, and Gaussian transformation. These stages are mapped onto existing matrix-acceleration hardware, enabling parallel evaluation across multiple support vectors while minimizing additional hardware overhead.

1) Kernel Computation Unit: The kernel computation unit forms the front-end of the inference stage and is responsible for evaluating the similarity between the incoming feature vector and stored support vectors. For each input feature vector $\mathbf{x}$, the unit retrieves the support vectors $\mathbf{x}_i$ from the support vector cache and computes the Radial Basis Function (RBF) kernel:

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right) \tag{7}$$

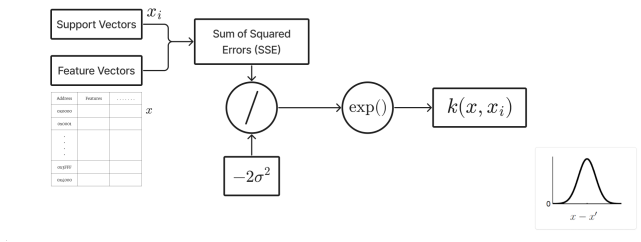

Fig. 7. Mathematical flow of Kernel Function

2) Weighted Accumulation and Decision Value Computation: Kernel outputs are forwarded to the weighted accumulation unit, where each kernel value is multiplied by its corresponding learned coefficient α_i and class label y_i. The unit accumulates these weighted contributions across all support vectors and adds the bias term b to produce the SVM decision value:

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{8}$$

The accumulation proceeds incrementally as kernel values become available, allowing overlap between kernel computation and accumulation. The resulting decision value encodes both the predicted class sign and the classification margin magnitude, which reflects confidence relative to the decision boundary.

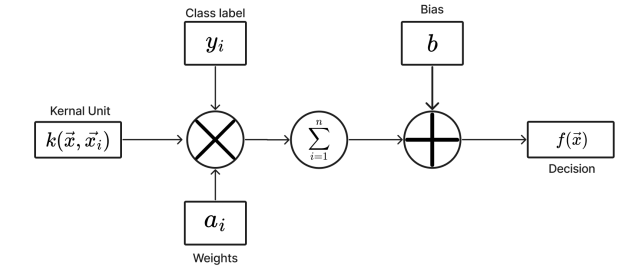

Fig. 8. Mathematical flow of Weighted Accumulation Diagram

3) Classification Logic and Binary Decision: The classification logic converts the continuous decision value into a discrete affective label (positive or negative). A single binary SVM model computes the decision value by accumulating weighted kernel evaluations between the input feature vector and the support vectors. The sign of the resulting decision function determines the final class label.

$$\hat{y} = \text{sign}\left(f(\mathbf{x})\right) \tag{9}$$

4) Pipeline Integration and Margin Output: The inference engine operates as a fully pipelined stage within the OPTIMIND processing chain. While one EEG window is being classified, subsequent windows simultaneously undergo preprocessing and feature extraction. In addition to the predicted emotion label, the inference engine outputs the absolute decision margin $|f(\mathbf{x})|$, which is forwarded to the adaptive learning stage. This margin enables confidence-aware update filtering, allowing high-confidence samples to bypass model updates and reducing unnecessary non-volatile memory writes.

5) Architectural Impact: By tightly coupling kernel computation, weighted accumulation, and classification logic within a unified inference stage, the proposed design achieves low-latency and energy-efficient emotion classification. The use of on-chip caches and hardware reuse minimizes memory access overhead, enabling sub-millisecond inference suitable for real-time and implantable BCI systems.

D. Adaptive Online Learning Mechanism

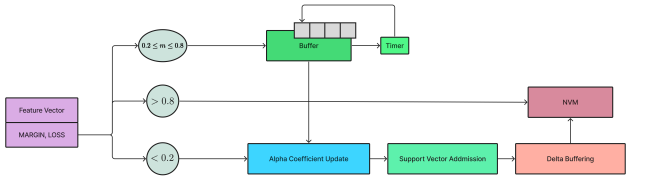

Fig. 9. Adaptive Learning Full Diagram

OPTIMIND incorporates an on-device adaptive learning mechanism that selectively updates model parameters based on classification confidence and novelty, while explicitly constraining non-volatile memory (NVM) write frequency and energy consumption.

1) Margin and Error Computation: Following inference, the decision function output $f(\mathbf{x})$ is combined with the ground-truth label $y \in \{-1, +1\}$ to compute the functional margin:

$$m = y \cdot f(\mathbf{x}) \tag{10}$$

The margin quantifies both prediction correctness and confidence. Positive margins correspond to correct classifications, while negative margins indicate misclassifications. To translate margin into an optimization signal, the system evaluates the hinge loss:

$$\mathcal{L}(\mathbf{x}) = \max(0, 1 - m) \tag{11}$$

This formulation ensures that samples with large positive margins incur zero loss, whereas low-margin or misclassified samples produce non-zero loss values that drive learning updates.

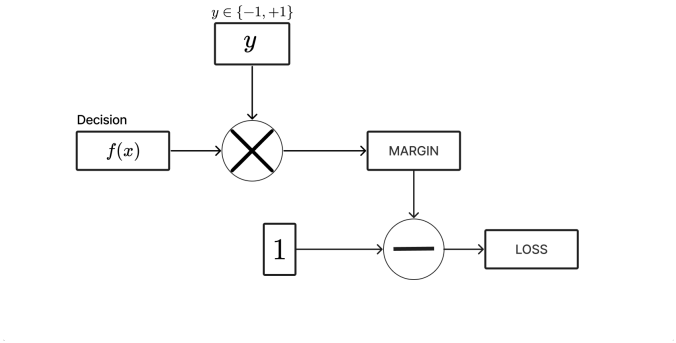

Fig. 10. Margin and Error Computation Diagram

2) Margin-Based Update Filtering: To preserve NVM endurance, OPTIMIND employs a confidence-aware update filtering policy that classifies samples into three operational regimes:

- **Bypass Region ($m > \tau_h$):** High-confidence samples with margins exceeding the upper threshold $\tau_h = 0.8$ produce zero loss and bypass all update logic.
- **Buffered Update Region ($\tau_l \leq m \leq \tau_h$):** Moderately confident samples, where $\tau_l = 0.2$, generate non-zero loss but do not immediately trigger NVM writes. Instead, their update deltas are accumulated in a temporary on-chip buffer.
- **Active Update Region ($m < \tau_l$):** Low-confidence or misclassified samples immediately activate the *Alpha Coefficient Update* unit to correct the model.

3) Alpha Coefficient Update: For samples entering the active update path, the system adjusts the corresponding Lagrange multiplier α_i according to a constrained update rule derived from online SVM optimization:

$$\alpha_{t+1} = \alpha_t + \min\left(C, \frac{\max(0, Loss(m))}{\|x\|^2}\right) \tag{12}$$

where C is the regularization bound and η is the learning rate (0.05). The Normalization by $\|\mathbf{x}_i\|^2$, shown in Fig. 11, ensures scale-invariant updates and prevents instability under varying signal amplitudes.

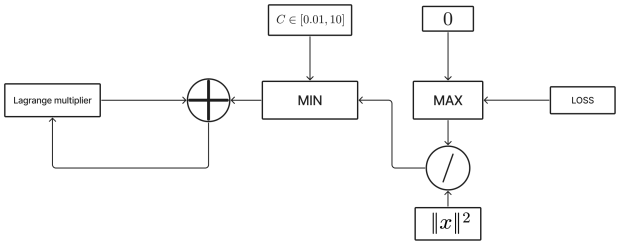

Fig. 11. Mathematical Flow of Alpha Coefficient Update Diagram

4) Buffered Update Aggregation and Bulk Flushing: Samples in the buffered update region are not discarded. Instead, their parameter deltas $\Delta\alpha_i$ are accumulated in a bounded on-chip buffer:

$$\Delta\alpha_i^{(k)} = \sum_{j=1}^{k} \eta \cdot \mathcal{L}(\mathbf{x}_j) \tag{13}$$

Buffered updates are committed to NVM only under one of two conditions:

- **Timer Expiry:** A programmable timer expires, enforcing a maximum update rate independent of sample arrival frequency.
- **Buffer Saturation:** The buffer reaches capacity, triggering a bulk flush to prevent overflow.

This mechanism, illustrated in Fig. 11, prevents repeated small writes while preserving learning information, effectively trading temporal precision for NVM endurance.

5) Support Vector Admission and Novelty Control: Following coefficient updates, the system evaluates whether the current sample represents a novel support vector. Novelty is determined by comparing the squared norm of the input feature vector against existing support vectors:

$$\|\mathbf{x}_{new} - \mathbf{x}_i\|^2 > \delta \tag{14}$$

If admitted, the least significant support vector is evicted to maintain the fixed 16 KB support vector cache size. All structural updates are staged through delta buffering and written to NVM using the same bulk-flush mechanism.

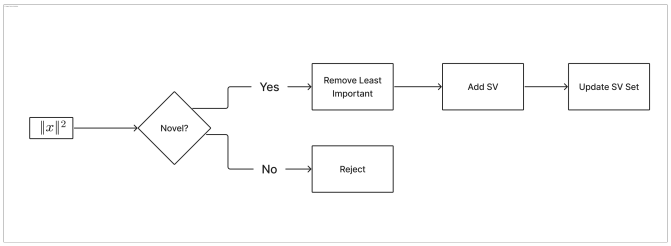

Fig. 12. Support Vector Admission Diagram

VI. Evaluation

A. Experimental Methodology

We evaluate the proposed architecture using an accurate Python-based simulator that models the full EEG emotion recognition pipeline shown in Fig. 3, including preprocessing, feature extraction, inference, memory access behavior, and adaptive learning. The evaluation is conducted using the EEGEmotions-27 dataset, which contains EEG recordings labeled across 27 fine-grained emotional states where each emotion is mapped to a binary affective label, ranging from low-arousal affect (e.g., calmness, boredom) to high-arousal affect (e.g., excitement, fear, anger). This dataset allows us to assess both classification accuracy and architectural behavior under diverse emotional dynamics.

Each EEG recording is processed in an online streaming mode, where signals are segmented into temporal windows (5

s), and each window is classified independently. Unlike offline evaluation, this setup reflects real-world BCI operation, where latency, memory accesses, and learning updates occur continuously during inference. Prior to online execution, the SVM classifier is initialized using offline-trained support vectors. During runtime, adaptive learning is enabled using margin-based update filtering, allowing the system to selectively update the model when classification confidence is low.

B. Emotion Recognition Evaluation

TABLE II
ONLINE EVALUATION RESULTS ACROSS REPRESENTATIVE SUBJECTS AND EMOTIONS

Subject	Emotion ID	Windows	Accuracy	End-to-End Latency (ms)
P69	16 (Excitement)	13	1.00	44.67
P23	19 (Interest)	17	1.00	58.41
P12	4 (Amusement)	8	1.00	27.93
P31	15 (Entrancement)	4	1.00	14.21
P89	1 (Admiration)	9	1.00	40.49
P72	7 (Awe)	6	1.00	25.94
P35	21 (Nostalgia)	2	1.00	7.35
P56	22 (Relief)	9	1.00	31.55
P66	9 (Boredom)	7	0.00	24.60
P40	12 (Craving)	10	0.10	40.90
Average	–	–	**0.81**	**31.60**

Table II summarizes representative online classification results across multiple subjects and emotion classes. The proposed system achieves high online accuracy across most tested emotions, with several files reaching 100% window-level accuracy, including excitement (emotion 16), interest (emotion 19), amusement (emotion 4), admiration (emotion 1), and awe (emotion 7). These results demonstrate that the selected frequency-domain and statistical features, combined with the SVM classifier, are sufficiently expressive to distinguish complex emotional states in an online setting.

More challenging emotional classes, such as boredom (emotion 9) and craving (emotion 12), exhibit lower accuracy. These emotions are known to have weaker and more diffuse neural signatures, which increases overlap in feature space and leads to classification ambiguity. Notably, even in these cases, the adaptive learning mechanism is triggered more frequently, indicating that the system correctly identifies uncertainty and attempts to compensate through model updates rather than blindly reinforcing incorrect predictions.

Across all evaluated test files, the proposed system achieves an average online classification accuracy of 81.0% while operating in a fully streaming configuration with adaptive learning enabled. On average, feature extraction requires 12.57 ms per EEG window and inference consumes 12.67 ms, reflecting the computational cost of FFT-based spectral analysis and kernel-based SVM evaluation, respectively. Memory overhead remains well controlled, with an average memory access latency of 4.42 ms, despite occasional cold-cache effects that introduce higher latency during early execution. The adaptive learning stage incurs only 0.98 ms on average, confirming that margin-based update filtering effectively limits unnecessary learning activity. Overall, the system achieves an average end-to-end latency of 31.60 ms per window, satisfying real-time BCI constraints while supporting continual online adaptation, and the model stabilizes at 224 support vectors by the end of execution, demonstrating bounded model growth under adaptive learning.

In contrast, the baseline system implemented without prefetching, margin-based update filtering, or delta-buffered, etc... updates incurs a substantially higher average end-to-end latency of approximately 250 ms per window, primarily due to frequent NVM accesses and immediate support vector updates during online learning. This comparison highlights that the proposed optimizations reduce end-to-end latency while maintaining classification accuracy and stable model size.

C. Impact of Adaptive Learning

Adaptive learning plays a crucial role in maintaining accuracy under non-stationary EEG signals. By selectively admitting new support vectors only when the classification margin is low, the system avoids excessive model growth and unnecessary NVM writes. The low average learning latency and controlled support vector count demonstrate that the learning mechanism is both effective and resource-efficient.

Importantly, learning operates transparently alongside inference, without interrupting real-time processing. This behavior is essential for long-term deployment of emotion-aware BCIs, where neural statistics evolve due to fatigue, emotional drift, and electrode variability.

VII. CONCLUSION AND FUTURE WORK

This paper presented OPTIMIND, an adaptive, emotion-aware BCI architecture designed to address emotion-induced neural variability while operating under the strict latency, power, and memory constraints of implantable and edge brain–computer interfaces. By integrating a lightweight EEG-based emotion recognition pipeline with a learning-optimized NVM assisted architecture, OPTIMIND transforms emotional variability from a source of noise into actionable system-level information. The proposed design introduces dedicated feature extraction and kernel computation units, an optimized memory hierarchy with feature and support vector caching, predictive prefetching, and a confidence-aware adaptive learning mechanism that selectively updates the model while minimizing non-volatile memory writes.

Comprehensive simulation using the EEGEmotions-27 dataset demonstrates that OPTIMIND achieves an average online classification accuracy of 81.0% while maintaining a low average end-to-end latency of 31.60 ms in a fully streaming configuration. The evaluation confirms that margin-based update filtering effectively bounds model growth, limits unnecessary learning activity, and preserves NVM endurance without interrupting real-time inference. These results indicate that OPTIMIND is well suited for long-term, on-device deployment in emotion-aware and learning-enabled BCI systems, particularly in implantable and privacy-sensitive applications.

Future work will focus on several directions to further extend the proposed architecture. First, validating OPTIMIND

using real-time EEG acquisition and wearable BCI platforms will be critical to assess robustness under real physiological noise, motion artifacts, and long-term signal drift. Second, support for multi-task BCI workloads, where emotion recognition dynamically modulates primary decoding tasks such as motor intention or seizure detection, represents an important extension toward fully adaptive closed-loop systems. Finally, exploring richer emotion models, including continuous valence–arousal representations and cross-subject generalization, could improve scalability and clinical relevance.Together, these directions aim to advance OPTIMIND toward practical, lifelong, emotion-aware brain–computer interfaces.

VIII. References

[1] G. Eichler, Y. Gilhotra, N. Zeng, M. Kim, K. Shepard, and L. Carloni, "MINDFUL: Safe, Implantable, Large-Scale Brain-Computer Interfaces from a System-Level Design Perspective," 58th IEEE/ACM International Symposium on Microarchitecture (MICRO), Seoul, South Korea, pp. 1–18, Oct. 2025.

[2] Karthik Sriram, Raghavendra Pradyumna Pothukuchi, Michał Gerasimiuk, Muhammed Ugur, Oliver Ye, Rajit Manohar, Anurag Khandelwal, and Abhishek Bhattacharjee, "SCALO: An Accelerator-Rich Distributed System for Scalable Brain-Computer Interfacing," in Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA '23), Orlando, FL, USA, pp. 1–20, June 2023.

[3] Y. Jang, D. Jung, S. Song, H. Lee, and J. Kim, "InfiniMind: A Learning-Optimized Large-Scale Brain-Computer Interface," Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA), Tokyo, Japan, pp. 1–17, Jun. 2025.

[4] Si, X., He, H., Yu, J., & Ming, D. (2023). Cross-Subject Emotion Recognition Brain–Computer Interface Based on fNIRS and DBJNet. Cognitive Systems Research.

[5] Torres, E.P.; Torres, E.A.; Hernández-Álvarez, M.; Yoo, S.G. EEG-Based BCI Emotion Recognition: A Survey. Sensors 2020, 20(17), 5083.

[6] I. Karageorgos, K. Sriram, J. Veselý, M. Powell, M. Wu, D. Borton, R. Manohar, and A. Bhattacharjee, "Hardware-Software Co-Design for Brain-Computer Interfaces," in 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), IEEE, pp. 1–18, 2020.

[7] H. Huang, Q. Xie, J. Pan, Y. He, Z. Wen, R. Yu, and Y. Li, "An EEG-Based Brain-Computer Interface for Emotion Recognition and Its Application in Patients with Disorder of Consciousness," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 832–846, Oct.–Dec. 2021.

[8] R. Kishore Kanna, P. Shoran, M. Yadav, M. N. Ahmed, S. Burje, G. Shukla, A. Sinha, M. R. Hussain, and S. Khalid, "Improving EEG-Based Brain–Computer Interface Emotion Detection with EKO-ALSTM Model," Scientific Reports, vol. 15, Art. no. 20727, 2025.