

Toward Real-Time Cognitive State Sensing in Everyday Work: A Lightweight CNN–Transformer for 4-Channel Wearable EEG

Ngo Tran Tuan Phong¹ [0009-0001-3893-4744], Nguyen Thi Hoang Trang² [0000-0002-9638-4672]
and Vi Chi Thanh³ [0000-0002-1686-5549]

¹ International University, Vietnam National University – Ho Chi Minh City
phongntt32@mp.hcmiu.edu.vn

² University of Liège, Belgium; Brain-Life Link Technology JSC, HCMC, Vietnam
hoangtrangnguyen181@gmail.com

³ Brain-Life Link Technology JSC, HCMC, Vietnam; Institute of Engineering Technology, Thu
Dau Mot University, Vietnam
thanh.vi@brainlife.ai

Abstract. Knowledge work involves frequent transitions between cognitive modes such as focused effort, engagement, relaxation, and sustained immersion. Detecting these shifts in real time could support reflective productivity tools, adaptive work interfaces, and early intervention for cognitive overload. However, most EEG-based cognitive-state classifiers rely on research-grade hardware, handcrafted features, or narrowly defined binary states, limiting their suitability for everyday interactive systems. We present a real-time cognitive-state sensing pipeline for consumer-grade 4-channel wearable EEG that combines Short-Time Fourier Transform (STFT) spectrograms with a lightweight CNN–Transformer architecture. Using four channels corresponding to a 10-20 international system (T3, T4, O1, O2), our approach classifies four target states: Focus, Relaxation, Engagement, and Flow, without manual feature engineering. Using a subject-stratified dataset of 6,220 four-second segments, the model achieved 94.69% test accuracy, while operating with an initial latency of approximately 4.2 seconds and 0.5-second update intervals. Beyond model performance, we contribute a design-oriented analysis of what minimal-channel EEG can and cannot support in interactive workplace settings. Our results suggest that end-to-end spectrogram learning can approach the performance of handcrafted-feature pipelines while reducing engineering overhead and enabling faster adaptation to new devices or cognitive-state definitions. At the same time, we show that high performance under controlled task-derived labels should not be conflated with validated inference of lived psychological states, especially for Flow. We discuss implications for human-centred cognitive sensing, ecological validity, calibration, interpretability, and future deployment in adaptive and reflective work systems.

Keywords: EEG, cognitive state classification, STFT CNN-Transformer, occupational health, wearable systems

1 Introduction

An individual’s mental state naturally shifts throughout the workday, ranging from deep analytical focus to collaborative engagement with colleagues to more open, flexible creative thinking. Typically, these different working modes are most effectively carried out in a FLOW state, where tasks feel more seamless, and progress comes with greater ease [1]. If such states could be captured in real time, they could provide valuable insights for more productive task planning, earlier identification of burnout risks, and greater adaptability across varied work environments [2, 3]. As concerns surrounding occupational health and safety continue to grow, the need for objective approaches to assessing cognitive state becomes increasingly important, rather than relying solely on attendance records or subjective self-reports.

For decades, EEG has served as a powerful method for analysing brain states within controlled laboratory settings [4, 5], offering millisecond-level resolution in detecting rapid cognitive changes. However, its use in everyday work environments is limited as research-grade systems require 32–64 electrodes, conductive gel, and trained technicians, making them impractical for daily use. More accessible wearable devices, such as BrainBit (4 channels), Emotiv EPOC+ (14 channels), and Muse (4 channels), use fewer channels and dry electrodes, making setup easier. Still, their reliability remains uncertain, particularly whether a small number of channels can capture sufficient information to accurately distinguish between a wide range of cognitive states.

Recent advanced studies using consumer-grade EEG devices suggest that speaking cognitive information can still be identified from minimal electrode configurations [6–8]. Specifically, lightweight machine learning models such as LightGBM and linear SVM have achieved over 95% accuracy in binary cognitive classification tasks using such data [9]. Despite this progress, most existing approaches focus on only one or two states (e.g., Relaxation vs Attention). A more comprehensive framework covering key workplace-relevant states: Relaxation, Focus, Engagement, and Flow, which has not been researched. Among these, the Flow state has only recently begun to receive attention in studies using wearable EEG in real-world settings [10].

Traditional EEG pipeline classification methods often rely on manually designed spectral features, such as bandwidth power, entropy measures, and spectral ratios, which require domain expertise and depend on established biomarkers. For instance, Jain et al. [9] combined VMD with LightGBM and 80 manually designed features to identify binary cognitive payloads, achieving 97.2% on 14–36 channels. Atkinson and Campos [14] used selected statistical features with SVM for emotion classification. In practice, these methods all require field expertise to decide which features to extract, and the results depend heavily on the choices made. Not to mention that manual feature extraction often simplifies the signal into overall summary values for a timeframe, which can miss how the frequency spectrum changes within the segment, even though these subtle dynamics may be important for distinguishing states with similar spectral signatures.

Notably, end-to-end deep learning methods that preserve both temporal and frequency information have shown strong execution and often outperform traditional approaches on research-grade datasets [11]. For instance, Xu et al. [11] proposed

AMDET, a multidimensional Transformer that applies self-attention across spectral, spatial, and temporal domains, achieving 97.48% accuracy on the DEAP dataset using 32 channels without manual feature engineering. However, applying these deep learning models to minimal-channel consumer EEG devices remains underexplored. In reality, there is still a noticeable performance gap between high-density systems (32–64 channels) and low-channel devices (around 4 channels), often around 10–15% in accuracy [6, 15], which remains an open challenge.

In this paper, we propose a lightweight CNN–Transformer hybrid model that operates directly on STFT spectrograms derived from 4-channel EEG signals, removing the need for manual feature engineering. We make five contributions.

1. A low-channel EEG framework for cognitive-state classification under wearable and consumer-grade sensing constraints, covering four workplace-relevant states: Relaxation, Focus, Engagement, and Flow.
2. A lightweight STFT-based CNN–Transformer model that reduces reliance on handcrafted EEG features by learning directly from time–frequency spectrograms, achieving 94.69% accuracy without any manual feature engineering.
3. A systematic comparison of handcrafted-feature, attention-based, and spectrogram-based learning approaches on the same four-channel EEG dataset under identical preprocessing and evaluation conditions.
4. Evidence that spectrogram-based learning provides a practical trade-off between accuracy, model complexity, and deployability for real-time cognitive-state monitoring, matching the best handcrafted pipeline (94.69% vs. 94.86%) while eliminating manual feature design.
5. Deployment-oriented observations on latency, calibration, and classification errors to guide future wearable EEG and neuroadaptive HCI applications.

2 Overview of related research

2.1 Detection of Cognitive States based on EEG

EEG-based cognitive state monitoring has increasingly moved from controlled laboratory settings using high-density electrode arrays towards more deployable systems based on low-density and consumer-grade wearable devices. This transition is central to human-centred research because real-world cognitive monitoring systems must balance signal quality, user comfort, calibration burden, computational cost, and ecological validity. Table 1 summarises representative EEG-based approaches relevant to cognitive-state and affective-state classification.

Table 1. Summary of representative EEG-based cognitive and affective state classification studies.

Study	Year	Target State	Channels	Best Accuracy	Method
Jain et al. [8]	2024	Cognitive load (2 classes)	14–36	97.2%	VMD + LightGBM
AMDET [10]	2023	Emotion (V/A)	32	97.5%	STFT + Transformer
Dual Trans-former [11]	2024	Emotion(3 classes)	62	95.7%	Dual Transformer
SSAA [12]	2024	Emotion (cross-subject)	62	90.4%	Spectral-Spatial Alignment
SVM+FFS [13]	2024	Emotion (V/A)	32	81.8%	SVM + Forward Selection
Engagement [14]	2024	Engagement (2 classes)	4–14	~80%	RF / SVM
UX-EEG [15]	2024	Emotional UX	4–14	~75%	Consumer-grade EEG

The reviewed literature suggests three important trends for the design of practical EEG-based cognitive monitoring systems.

First, recent deep learning approaches increasingly avoid manually engineered EEG features by learning directly from richer time–frequency representations. For example, AMDT combines STFT-based representations with Transformer modelling and reports high classification performance on 32-channel EEG data [10]. Similarly, Xu et al. [11] show that attention mechanisms operating across spectral, spatial, and temporal dimensions can learn complementary EEG representations without relying entirely on handcrafted feature design. This is important for EEG systems because manual feature engineering often depends on task-specific assumptions, expert tuning, and dataset-specific optimisation, which can limit generalisability in real-world deployments.

Second, performance appears to depend not only on model complexity but also on the information preserved in the input representation. Attention-based architectures such as SSAA [12] provide sophisticated mechanisms for modelling spectral and spatial relationships, yet their performance is still constrained by the quality and expressiveness of the input features. Methods based on compact handcrafted descriptors, such as differential entropy or selected band-power features, may be efficient, but they can dis-

card temporal and cross-frequency patterns that are potentially informative for cognitive-state inference. In contrast, time–frequency representations such as STFT spectrograms preserve a denser description of EEG dynamics and may therefore provide a stronger basis for lightweight neural architectures.

Third, there remains a substantial gap between high-accuracy results obtained with research-grade EEG systems and the performance typically observed with low-density consumer-grade headsets. Studies using 32–64 channels often report accuracies above 90%, whereas studies using 4–14 channels usually report lower performance, often around 75–80% for engagement or emotional UX classification [14,15]. This drop is expected: fewer channels reduce spatial resolution, and consumer-grade devices are more susceptible to noise, motion artefacts, electrode placement variability, and user-specific differences. However, this gap does not necessarily imply that low-channel EEG is unsuitable for real-time HCI applications. Rather, it suggests that practical systems must compensate for reduced spatial coverage by using better preprocessing, richer input representations, model architectures that efficiently exploit limited signals, and lightweight calibration procedures.

Several recent models further demonstrate the potential of attention-based and high-dimensional neural architectures for EEG classification. For example, SST-Emotion Net [28] and the four-dimensional neural network proposed by Xiao et al. [29] report strong performance on benchmark datasets such as SEED. However, these methods are primarily evaluated on research-grade EEG datasets with relatively high channel counts. Their suitability for minimal-channel consumer EEG remains unclear, particularly under constraints relevant to real-world HCI deployment: short calibration time, low computational latency, noisy workplace environments, and the need for unobtrusive wearable use.

Taken together, prior work indicates that high-performing EEG classification increasingly benefits from preserving rich time–frequency information and from using attention-based architectures. At the same time, there is still limited evidence on whether such approaches remain effective when applied to noisy, low-channel EEG collected from consumer-grade devices. This motivates our investigation of a lightweight STFT + CNN–Transformer architecture for 4-channel EEG cognitive-state classification. Rather than maximising performance under ideal laboratory conditions, our aim is to examine whether a compact model can achieve competitive accuracy while remaining compatible with real-time, low-burden deployment in workplace monitoring scenarios.

2.2 Consumer-grade EEG for Workplace Monitoring

Consumer-grade EEG has become increasingly relevant for workplace monitoring, offering a practical alternative to laboratory-grade systems. Conventional EEG studies often rely on 32–64 electrodes, gel-based preparation, and controlled experimental environments. In contrast, workplace-oriented EEG systems must support rapid setup, comfort, wireless operation, and repeated use by non-expert users. These requirements have motivated growing interest in low-density dry-electrode headsets for monitoring cognitive states in naturalistic settings.

The BrainBit headset [30] provides four dry electrodes at T3, T4, O1, and O2 with 250 Hz sampling and BLE connectivity. Compared to 32–64-electrode research systems, it trades spatial resolution for practicality: lightweight, gel-free, full-day battery, and self-applicable. Key advantages of this configuration are: O1/O2 sit directly over the occipital alpha source, the strongest relaxation biomarker, first described by Berger in 1929; T3/T4 lie near the amygdala and capture temporal theta via volume conduction from frontal sources [10], with temporal-frontal alpha asymmetry correlating at $r \approx 0.65$ [6]; and occipital ERD at O1/O2 tracks attentional engagement reliably. Limitations include the absence of frontal electrodes (F3, F4, Fz), precluding direct frontal alpha asymmetry and midline theta measurement, and susceptibility of T3/T4 to temporalis EMG. Xu et al. [11] report that Transformer-based models retain $\sim 90\%$ accuracy even when reduced to 8 channels, suggesting attention mechanisms can partially compensate for sparse coverage.

At the same time, prior work also highlights the limitations of low-density consumer-grade EEG. Compared with research-grade systems, four-channel headsets have reduced spatial resolution and limited access to frontal and central cortical regions. In particular, the absence of frontal electrodes such as F3, F4, and Fz restricts the direct measurement of frontal alpha asymmetry and frontal midline theta, both of which are commonly used in studies of workload, executive control, and affective processing. Temporal dry electrodes are also susceptible to artefacts from jaw movement, facial tension, and temporalis muscle activity. These issues make preprocessing and robust feature extraction especially important in consumer-grade EEG studies.

Recent modelling work indicates that sparse-channel EEG can still support reliable classification when combined with suitable signal representations and learning architectures. For example, Transformer-based approaches have been reported to maintain strong performance under channel-reduction settings, suggesting that attention mechanisms can exploit temporal and spectral regularities even when spatial coverage is limited. However, most high-performing models have been developed and evaluated on research-grade datasets, often with substantially more channels than consumer headsets provide. There remains limited evidence on how such architectures perform in practical workplace-oriented scenarios with four-channel dry-electrode EEG.

This body of work motivates the present study’s focus on low-channel EEG classification using a deployable consumer-grade headset. Rather than treating consumer-grade EEG as a direct substitute for high-density systems, existing literature positions it as a complementary sensing approach for contexts where comfort, setup time, mobility, and ecological validity are central design constraints. Our work builds on this direction by examining whether STFT-based representations and a lightweight CNN–Transformer model can support cognitive-state classification from four-channel BrainBit EEG data.

2.3 Definition of four target cognitive states and neurophysiological markers

Distinct EEG Patterns Across Cognitive States. Based on a synthesis of 13 recent studies [4, 6, 9, 10, 15–21], we aim to define spectral markers for each cognitive state

relative to an individual's resting baseline. Table 2 summarises the full biomarker matrix.

Table 2. EEG spectral signatures for four cognitive states on BrainBit channels

Band	Relaxation	Focus	Engagement	Flow
Delta (1–4 Hz)	Low	Normal	Low	Stable
Theta (4–8 Hz)	Low	High	Normal	Moderate, stable
Alpha (8–13 Hz)	Low	Suppressed	Strongly suppressed	Moderate
Beta (13–30 Hz)	Low	High	Very high	Moderate, stable
Gamma (30–45 Hz)	Low	Normal	High	Moderate
Main channels	O1, O2	T3, T4	All channels	All channels
Key index	$RI = \alpha/(\beta+\gamma)$	$TBR = \theta/\beta$	$EI = \beta/(\alpha+\theta)$	Composite

Relaxation. Relaxation state refers to a state of brain functioning in which thalamocortical communication occurs less actively [4, 17]. Alpha activity is often observed in the occipital area. Because the BrainBit electrodes O1 and O2 are located precisely in this part of the brain, it becomes easier to recognise relaxation states. To quantify relaxation, parameters such as RI ($P\alpha/(P\beta+P\gamma)$) and TAR may be used. When TAR is below 1.0, the individual is experiencing relaxation, while TAR above 1.2 suggests drowsiness [4].

Focus. When someone is focusing hard on a task, the brain shifts out of its relaxed state. This shows up as an increase in theta waves (4–8 Hz) in the frontal area and a drop-in alpha activity (ERD), a pattern often described as the brain “switching off” its resting mode [4, 18]. Although BrainBit does not include frontal electrodes, theta activity recorded at T3 and T4 can still reflect frontal processes via brain tissue [10]. A common way to measure this is the Theta/Beta Ratio ($TBR = P\theta / P\beta$), which has been linked to attention levels in studies, including ADHD research ($r \approx -0.48$, as reported in [15, 19] though this correlation value varies across studies and should not be taken as universal)

Engagement. The state of engagement is characterised by the brain being truly captivated, which activates the pervasive cortical network and leads to increased beta and gamma activity, along with a strong reduction in alpha waves across multiple regions [15, 19, 20]. A common measure is the Engagement Index ($EI = P\beta / (P\alpha + P\theta)$), originally developed for monitoring pilot workload, which has shown a moderate correlation ($r \approx 0.55$, per [15, 19] noting that this figure comes from specific experimental paradigms and may not generalize to all populations) with self-reported engagement [15, 19]. Compared to Focus, which mainly affects specific areas such as reduced alpha in occipital regions, Engagement tends to involve a more widespread suppression of alpha activity across the brain.

Flow. Flow, often referred to as the "meditative" experience of working, involves being totally absorbed, with activities feeling seamless and almost effortless [1]. Unlike the pure states of Focus and Relaxation, Flow does not rely on any specific brainwave. Instead, it reflects a balanced combination of moderate theta and alpha activity, plus stable beta activity across both hemispheres. What matters most in identifying Flow is not the strength of the signals, but their consistency over time. In this state, brain activity remains relatively stable across frequency bands, often reflected by low variation ($CV < 0.25$) [10, 20, 21]. Studies by Knierim et al. [10] using wearable ear-EEG in real-world work settings have supported these patterns.

3 Method

3.1 System Overview

The proposed system combines on-device signal acquisition with cloud-based deep learning inference in a four-layer architecture. Raw EEG signals are captured by a 4-channel headset, preprocessed to remove noise and artefacts, then transformed into time–frequency spectrograms via STFT. The complete pipeline includes data acquisition and labelling (Section 3.2), preprocessing (Section 3.3), STFT feature extraction (Section 3.4), CNN–Transformer model architecture (Section 3.5), training and evaluation protocol (Section 3.6), post-processing and calibration (Section 3.7), and real-time latency analysis (Section 3.8).

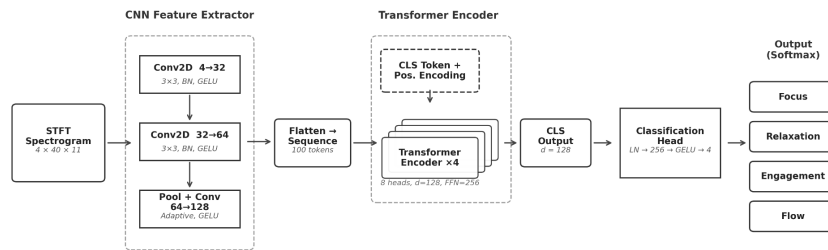


Fig. 1. Overview of the system architecture. On the left is the STFT spectrogram input ($4 \times 40 \times 11$). Middle left: CNN feature extraction (3 Conv2D layers).

Middle: Transformer encoder (4 layers, 8 heads).

Right: Classification head outputting four cognitive states. Total: 674,852 parameters.

The choice of combining STFT with a CNN–Transformer architecture hinges on three key advantages. First, information preservation: traditional methods discard within-window temporal dynamics, averaging statistics across the entire spectrum over a fixed 4-second interval. STFT retains both frequency content and its progression, making it easier to capture patterns linked to cognitive-state transitions. Second, automatic feature discovery: the CNN–Transformer identifies useful patterns directly from the spectrogram, including relationships across time, frequency, and channels. This eliminates the need for handcrafted features while still achieving strong performance. Third, rapid scalability: new cognitive states or different devices can be incorporated by retraining the model without redesigning the feature-extraction pipeline.

3.2 Dataset and Labelling Strategy

The dataset was compiled from 64-channel PhysioNet recordings of 20 participants, reduced to four channels (T3, T4, O1, and O2), and resampled to 250 Hz to match BrainBit parameters. Cognitive-state labels were assigned based on task conditions: Focus for imagery-based motor tasks, Relaxation for eyes-closed rest, Engagement for actual motor execution, and Flow for eyes-open rest during task engagement. The dataset consisted of 6,220 4-second segments, evenly divided among four classes (1,555 segments per class).

A key limitation: these labels reflect actions, not self-reported feelings. Motor imagery maps to Focus; eyes-closed rest to Relaxation; motor execution to Engagement; and eyes-open rest during task blocks to Flow. The first three mappings are grounded in well-documented neurophysiological signatures (Section 2.3). The Flow label is a rough proxy; eyes-open rest during a task block doesn't capture deep immersion as described by Csikszentmihalyi. The near-perfect F1 scores for Relaxation and Flow (Section 4.1) likely result from clear-cut task conditions rather than the genuine ability to separate these states in real life. Proper validation with experience sampling and concurrent self-reports from at least 30 participants is needed before these labels carry weight.

3.3 Preprocessing Pipeline

A sequential preprocessing algorithm comprising four stages is applied to the raw EEG data, with each stage targeting the minimisation of noise from a specific source, thereby yielding a signal that better represents brain activity. The whole process is embedded

in the provided web application demonstration tool and allows visualising the input/output of each processing stage, along with the resulting PSD, wavelet analysis results, and excluded channels.

A 50 Hz IIR Notch Filter ($Q = 30$) removes power-line interference. On shielded PhysioNet recordings, this step reduces RMS by only $\sim 0.07\%$, but in open office environments with BrainBit, the contribution is substantially larger.

Fifth-order Butterworth bandpass filter (0.5–45 Hz) applied via zero-phase to retain all five neurophysiologically relevant bands: delta, theta, alpha, beta, gamma, while removing DC drift below 0.5 Hz and EMG noise above 45 Hz.

Wavelet denoising (Daubechies-4, level 4, soft thresholding), the signal is decomposed into five levels; approximation coefficients (cA4) are preserved while detail coefficients (cD1–cD4) undergo VisuShrink soft thresholding [26]: $\sigma = \text{median}(|cDi|)/0.6745$, $\text{threshold} = 2\sigma\sqrt{2 \ln N}$. This suppresses EOG and EMG artifacts while preserving brainwave morphology [4].

Baseline removal (zero-mean per channel) subtracts the channel mean to center the signal at 0 μV , removing residual DC offset without altering the absolute microvolt scale needed for band-power analysis.

Sliding window segmentation: A 4-second window (1,000 samples at 250 Hz) with a 0.5-second hop feeds the classification pipeline continuously. The accompanying web demo visualises the full pipeline using real data from 6,220 PhysioNet segments: overlaid before/after signal plots, PSD curves showing energy removed at each stage, wavelet coefficient maps before and after thresholding, and relative band-power charts across all four channels for each cognitive state.

3.4 STFT Feature Extraction

Each pre-processed 4-second EEG segment (4 channels $\times$ 1000 samples at 250 Hz) was transformed into a time-frequency spectrogram using the Short-Time Fourier Transform [4].

STFT Formula:

$$\text{STFT}\{x\}(\tau, \omega) = \sum x[n] \cdot w[n - \tau] \cdot e^{-j2\pi\omega n} \quad (1)$$

with **Hann Window**

$$W[n] = 0.5 \cdot (1 - \cos(2\pi n / (N - 1))) \quad (2)$$

The spectral amplitude is converted to a logarithmic scale to compress the dynamic range [11]:

$$X_{\text{log}} = \log(|\text{STFT}| + 10^{-8}) \quad (3)$$

The EEG power spectrum spans a wide range of values, from strong low-frequency activity (delta) to much weaker high-frequency activity (gamma). Applying a logarithmic scale compresses this range, making both strong and weak signals more visible to the CNN. A small constant (10^{-8}) is added to avoid taking the log of zero.

STFT parameters:

- Hann window, 256 samples (≈ 1.02 s at 250 Hz)
- Hop size 64 samples, giving 75% overlap
- Frequency range 0–45 Hz, mapped into 40 bins (≈ 1.1 Hz resolution)
- 12 time frames per segment

Each segment is represented as a 3D tensor with shape $(C, F, T) = (4, 40, 11)$, where $C = 4$ channels, $F = 40$ frequency bins, and $T = 12$ time frames. In simple terms, this can be viewed as a stack of four “image-like” representations, where each layer corresponds to the spectrogram of one EEG channel.

The 40 frequency bins give a resolution of about 1.1 Hz across the 0–45 Hz range. This level of detail is sufficient to clearly capture the main EEG rhythm bands, including delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz), which are commonly associated with different brain activities.

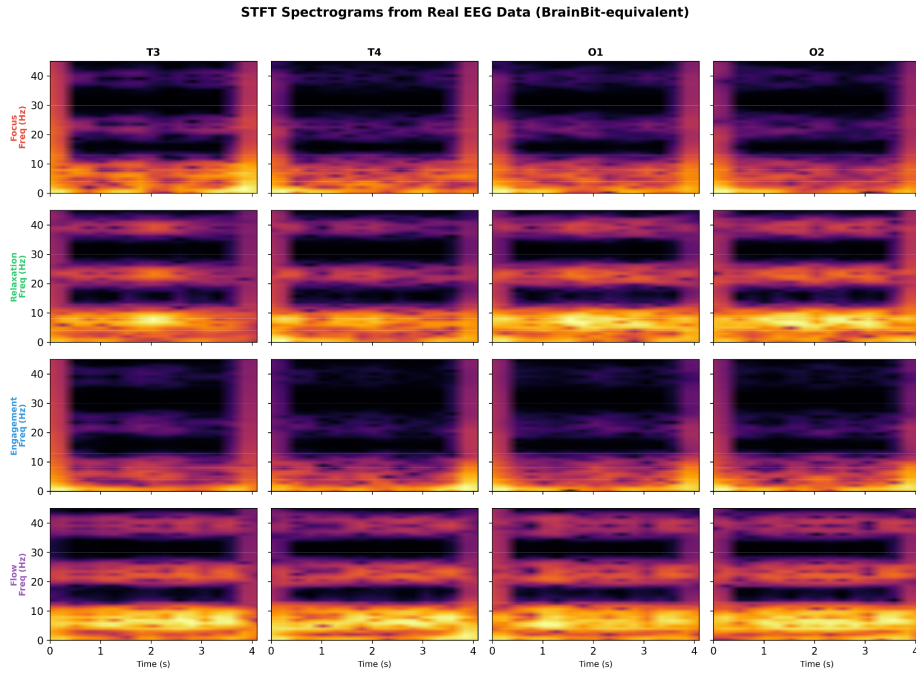


Fig. 2. Example STFT spectrograms for different cognitive states across four BrainBit channels. In the relaxation state, alpha activity (8–13 Hz) is clearly dominant, especially in the O1 and O2 channels. During focused attention, beta activity (13–30 Hz) becomes more prominent, along with noticeable theta presence at T3 and T4. The Engagement state shows broader spectral activation, including strong gamma activity (30–45 Hz), suggesting greater cognitive integration. In contrast, the flow state shows a more balanced, stable distribution across frequency bands. The dashed lines mark the conventional EEG band boundaries.

3.5 CNN–Transformer Architecture

CNN-Based Feature Extraction. The spectrogram is fed into a three-layer CNN to capture local patterns in both time and frequency domains.

Layer 1: Conv2D ($4 \rightarrow 32$, kernel 3×3 , padding 1) with BatchNorm and GELU. Input size is (4, 40, 11). This layer learns basic spectral features, such as band boundaries, power changes, and channel relationships.

Layer 2: Conv2D ($32 \rightarrow 64$, kernel 3×3 , padding 1) with BatchNorm and GELU. Here, the model combines simpler patterns into more meaningful ones, such as changes across regions and frequency bands.

Layer 3: Adaptive average pooling to (20, 5), followed by Conv2D ($64 \rightarrow 128$, kernel 3×3 , padding 1) with BatchNorm and GELU. Pooling reduces spatial size while keeping key information, and the 128 channels provide a richer representation for the next stage.

The final CNN output is reshaped into a sequence of $20 \times 5 = 100$ patches, each represented by a 128-dimensional embedding, similar to the patch-based approach used in Vision Transformers (ViT) [22].

Transformer Encoder. The patch sequence processed by the Transformer encoder captures global dependencies across all time positions as well as frequencies:

CLS Token: A classification token (CLS) is added to the beginning of the sequence, serving as a global information aggregation point and summarising the most important information. After processing through the Transformer layers, the output vector of the encoded CLS token represents the overall representation of the entire spectrogram.

Position Coding: Sinusoidal position coding by Vaswani and cs is added to preserve the spatial order of the patch:

$$PE(pos, 2i) = \sin(pos / 10000^{(2i/d_model)}) \quad (4)$$

$$PE(pos, 2i+1) = \cos(pos / 10000^{(2i/d_model)}) \quad (5)$$

This coding allows the model to distinguish between low-frequency (theta/alpha) and high-frequency (beta/gamma) regions.

Self-Attention Mechanism: each Transformer layer applies multi-head self-attention with 8 heads, as described by Vaswani et al. [25].

$$\text{Attention}(Q, K, V) = \text{softmax}(Q \cdot K^T / \sqrt{d_k}) \cdot V \quad (6)$$

where Q, K, V are linear projections of the input, $d_k = d_model / h = 16$ is the ff each end. After attention, each layer applies a feed-forward block with redundant connections and layer normalisation [25]:

$$FFN(x) = \text{GELU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (7)$$

$$\text{output} = \text{LayerNorm}(x + \text{MHA}(x)) \quad (8)$$

$$\text{output} = \text{LayerNorm}(\text{output} + \text{FFN}(\text{output})) \quad (9)$$

The architecture consists of 4 Transformer layers, totalling approximately 470,000 parameters.

This mechanism allows every patch to pay attention to every other patch, enabling the learning model to:

- **Frequency interaction:** alpha inhibition at 8–13 Hz correlates with beta increase at 13–30 Hz (Focus sign).
- **Channel interaction:** hemispherical asymmetry pattern between T3/T4 and O1/O2.
- **Time interaction:** spectral stability over the 4-second window (Flow sign).
- **Architecture Summary:** 4-layer Transformer encoder, feed-forward dimension 256, GELU enabled, dropout 0.3. Total parameters: 674,852.

Classification and Training Head. The CLS token output goes through: LayerNorm $\rightarrow$ Linear(128, 256) $\rightarrow$ GELU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(256, 4). The output is a 4-class logit vector converted into probabilities via softmax. The entire network is optimized using a frequency-weighted cross-entropy loss function to balance the four classes.

Training configuration:

- **Data split:** 68% training, 12% validation, 20% testing (stratified)
- **Optimization:** AdamW ($\text{lr} = 1 \times 10^{-3}$, $\text{weight_decay} = 1 \times 10^{-4}$)
- **Schedule:** Cosine annealing
- **Regularization:** Dropout 0.3, gradient cut (max norm = 1.0), early stop (patience = 20)
- **Batch size:** 64

3.6 Experimental Setup

The dataset was created from 64-channel recordings of 20 participants, reduced to four channels (T3, T4, O1, and O2) and resampled to 250 Hz to match BrainBit parameters. To define cognitive states, labels were assigned based on the experimental task conditions. Labels were assigned based on task conditions: Focus for imagery-based motor tasks, Relaxation for eyes-closed resting periods, Engagement for actual motor execution tasks, and Flow for eyes-open resting periods during task engagement. The dataset consisted of 6,220 4-second segments, balanced between four classes (approximately 1,555 segments per class).

A key limitation: these labels reflect what participants were doing, not what they reported feeling. Motor imagery maps to Focus; eyes-closed rest to Relaxation; motor execution to Engagement; and eyes-open rest during task blocks to Flow. The first three mappings are grounded in well-documented neurophysiological signatures (Section 3). The Flow label is a rougher proxy; eyes-open rest during a task block does not capture the deep immersion described by Csikszentmihalyi. The near-perfect F1 scores for Relaxation and Flow (Section 6.2) likely reflect the clear-cut nature of the task conditions more than a genuine ability to separate these states in the real world. Proper validation

with experience sampling and concurrent self-reports from at least 30 participants is needed before these labels carry weight.

We used a stratified 68/12/20% train/validation/test partition to split the data. To avoid within-subject data leakage, we ensured that all segments from a given participant appeared in only one partition. A more rigorous approach, such as Leave-One-Subject-Out cross-validation, would provide stronger evidence of cross-subject generalisation, and we consider this an important direction for follow-up work.

For a fair assessment, all five classification methods were tested using the same pre-processed dataset, identical splits, and evaluation metrics.

The 4-minute calibration protocol described in Section 3.4 adapts the system to each individual's spectral baseline after calibration, leading to more stable predictions during real-time monitoring. However, the current evaluation does not include a controlled before-and-after calibration comparison; quantifying the accuracy gain from calibration remains a direction for future work.

3.7 Post-Processing and Calibration

Time smoothing is achieved using the Exponential Moving Average (EMA) with $\alpha = 0.3$, yielding a time constant of about 5 seconds. This prevents frequent switching between classes within a short time frame, resulting in more reliable output.

For improved reliability, the model outputs the prediction based on the state with a probability exceeding 0.6 according to the softmax function. Lower-probability predictions are considered ambiguous and not used for classification.

The Calibration Protocol lasts 4 minutes. Although STFT + CNN-Transformer is less dependent on calibration than rule-based methods, a brief calibration session still boosts cross-subject performance. The protocol consists of three stages:

The Resting Baseline lasts 2 minutes: eyes open, establishing a spectral baseline for each channel in its normal state.

The Relaxation Reference lasts 1 minute: eyes closed, confirming the strongest alpha rhythm at O1/O2.

Activity Reference (1 minute): Simple cognitive exercise, confirm beta activation at T3/T4 so the device can record how the brain "turns on" to work mode.

The 4-minute calibration protocol adapts the system to each individual's spectral baseline after calibration, leading to more stable predictions during real-time monitoring. However, the current evaluation lacks a controlled before-and-after calibration comparison; quantifying the accuracy gain from calibration remains a direction for future work.

3.8 Latency Analysis

Table 3. End-to-end latency budget for real-time cognitive state detection.

Component	Latency	Notes
-----------	---------	-------

EEG sampling + buffering	4000 ms	Requires a 4-second window
BLE transmission	15–25 ms	BrainBit BLE 4.0
On-device preprocessing	5–10 ms	Filtering + artifact checking
Network upload	50–100 ms	4 channels $\times$ 1000 samples $\approx$ 32 KB
STFT computation	2–5 ms	Server-side, vectorized
CNN–Transformer inference	10–20 ms	GPU (T4/A10), batch = 1
Network download	20–40 ms	4 probabilities $\approx$ 64 bytes
Total per update	$\sim$ 4.2 seconds	Dominated by window length
Update frequency	0.5 seconds	Sliding window step

The initial 4-second latency stems from the STFT window required to capture enough signal for reliable analysis, since cognitive states typically evolve over seconds rather than instantaneously. After this primary step, the system updates every 0.5 seconds, with an additional processing delay of about 200 ms.

4 Experimental Results and Discussion

This section presents the experimental evaluation. Section 5.1 describes the dataset and labelling strategy. Section 5.2 reports classification results, method comparison, and feature importance analysis. Section 5.3 discusses findings, organised around six themes: main findings, why the proposed model works, comparison with the Dual Transformer and feature-based models, error patterns and their neurophysiological interpretation, cautious interpretation of Relaxation and Flow scores, and deployment implications.

4.1 Experimental Results

Table 4 presents the per-class classification results of the proposed STFT + CNN–Transformer. The model achieved 94.69% overall accuracy with a weighted F1-score of 0.947 and Cohen’s κ of 0.929. Relaxation and Flow were classified with F1 = 1.00, while Focus and Engagement showed F1 $\approx$ 0.90 with approximately 10% mutual confusion.

Table 4. Classification results by class of the proposed STFT + CNN–Transformer.

State	Precision	Recall	F1-Score	Samples
Focus	0.91	0.90	0.91	311
Relaxation	1.00	1.00	1.00	311
Engagement	0.90	0.91	0.90	311
Flow	1.00	1.00	1.00	311
Overall	0.95	0.95	0.95	1,244

Table 5. Summary metrics of the proposed model.

Metric	Value
Test Accuracy	94.69%
Best Validation Accuracy	92.90%
Weighted F1-score	0.947
Cohen’s Kappa (κ)	0.929
Parameters	674,852
Training Epochs	57 (early stopping from max 150)

Fig. 3 shows the confusion matrix. The dominant error pattern is bidirectional confusion between Focus and Engagement, while Relaxation and Flow are separated with near-zero error.

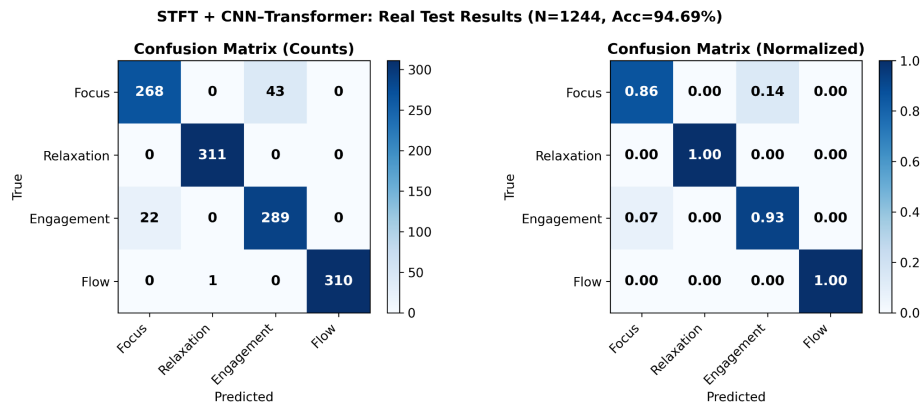


Fig. 3. Confusion matrix of STFT + CNN–Transformer on test set ($N = 1,244$). Left: raw number. Right: normalized by row. The main source of error is confusion between Focus and Engagement ($\sim 10\%$), while Relaxation and Flow are classified with near-zero error.

Table 6 compares the five classification methods evaluated on the same dataset. The proposed model achieves the highest accuracy among feature-engineering-free methods. The Dual Transformer reaches 94.86% but requires 39 handcrafted features per channel.

Table 6. Comparison of five classification methods on the equivalent BrainBit dataset (6,220 segments, 4 classes, 4 channels).

State	Precision	Recall	F1-Score	Samples	Feature Engineering
SSAA	89.71%	0.896	0.863	22,824	Manual: 10/ch
SVM + FFS	92.12%	0.921	0.895	-	Manual: 162 $\rightarrow$ 15
VMD + LightGBM	93.41%	0.934	0.912	-	Manual: 80
STFT + CNN–Transformer	94.69%	0.947	0.929	674,852	None
Dual Transformer	94.86%	0.949	0.931	872,326	Manual: 39/ch

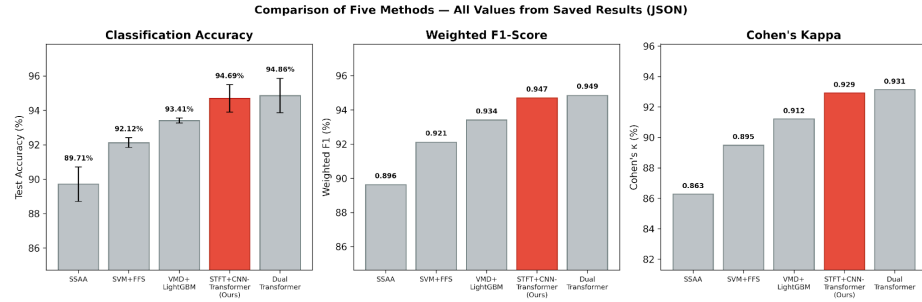


Fig. 4. Comparison of five classification methods. Left: accuracy tested with error bars (± 1 SD from cross-validation). Middle: Weighted F1. Right: Cohen's κ . STFT + CNN–Transformer proposed (red) achieves the highest accuracy among methods that do not require feature design.

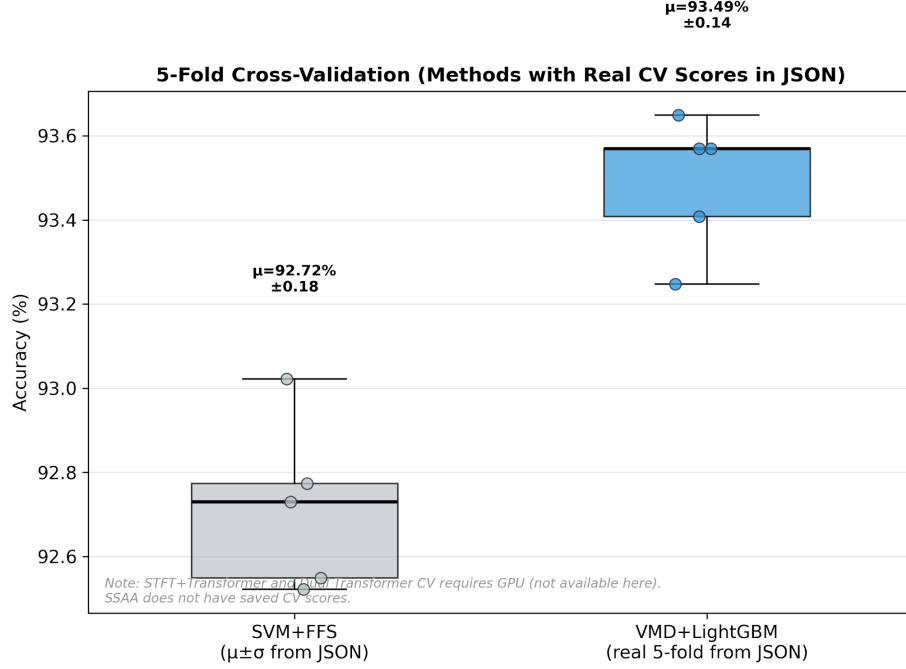


Fig. 5. Distribution of 5-fold cross-validation accuracy for methods with stored CV scores. VMD + LightGBM has true 5-fold scores from training; SVM + FFS is reconstructed from stored mean $\pm$ SD. STFT + CNN–Transformer and Dual Transformer require GPU to run CV.

The SVM + FFS pipeline also provides insight into feature importance. The top five selected features, in order, were: Hjorth mobility at T3 (66.2% single-feature accuracy), relative theta energy at T3 (cumulative 78.7%), theta asymmetry between T4 and O2 (83.2%), Hjorth mobility at T4 (85.4%), and relative alpha energy at T3 (87.7%). Hjorth mobility at T3 alone reached 2.6 times above the 25% chance level, indicating that even sparse-channel EEG carries meaningful cognitive-state information.

4.2 Discussion

Main Findings. The proposed STFT + CNN–Transformer classifies four cognitive states from a 4-channel EEG with accuracy matching the best handcrafted-feature pipeline, eliminating manual feature engineering entirely. Relaxation and Flow are consistently well separated, while Focus and Engagement show persistent mutual confusion. This pattern holds regardless of the classification approach, pointing to hardware-level and neurophysiological constraints rather than model-specific weaknesses.

Why STFT + CNN–Transformer Works. The model operates on STFT spectrograms that preserve both frequency content and its temporal evolution within each 4-second segment. Cognitive states are not static: alpha suppression during Focus builds over

time, and Flow is characterized by spectral stability rather than a single dominant band. The CNN extracts local spectral patterns, while the Transformer captures longer-range dependencies across time, frequency, and channels. The model learns discriminative representations, such as the correlation between alpha suppression at O1/O2 and theta elevation at T3/T4 during Focus, without manual specification.

Comparison with the Dual Transformer and Feature-Based Models. The Dual Transformer achieved 94.86%, just 0.17 percentage points above the proposed model. This gap corresponds to roughly two segments on 1,244 test samples and has not been tested for statistical significance. The Dual Transformer's performance appears driven by feature richness rather than architecture: versions with 5 features per channel reached 69.6–71.5%, while the version with 39 features per channel jumped to 94.86%. The architecture itself contributed less than 2 percentage points. The Dual Transformer was designed for 62-channel data; its spatial encoder operates below capacity on four channels. In contrast, the STFT pipeline requires only a standard FFT per channel and is portable to browsers and embedded devices, while the Dual Transformer depends on Butterworth filtering, Welch PSD, numerical integration, and Hjorth parameter computation.

Among feature-based methods, input representation quality proved at least as important as classifier complexity. SVM with 38 features per channel outperformed SSAA with its attention mechanism, but only 10 features per channel. VMD + LightGBM with 80 features reached 93.41%, confirming that richer representations consistently improve performance regardless of the downstream model.

Error Pattern and Neurophysiological Interpretation. The 10% bidirectional confusion between Focus and Engagement is neurophysiologically plausible. Both states involve increased beta activity and alpha suppression, producing similar spectral profiles. Focus tends to activate localized cortical areas, while Engagement recruits broader regions. The spatial resolution of four electrodes is insufficient to capture this difference reliably. The consistency of this error pattern across all five methods confirms it's a hardware constraint of the 4-channel montage, not a model limitation. Adding frontal electrodes or combining EEG with heart rate variability data could reduce this confusion.

Cautious Interpretation of Relaxation and Flow. Relaxation and Flow were classified with perfect accuracy, but this shouldn't be overgeneralized. These states have distinct spectral signatures: eyes-closed rest produces strong occipital alpha, while eyes-open rest during task blocks shows balanced theta–alpha with low temporal variability. The model likely captures task-level contrasts as much as genuine cognitive-state boundaries. In naturalistic settings, where a person may be relaxed with eyes open or in a flow-like state, the separation would decrease. These scores are an upper bound under controlled conditions, not a guarantee of real-world performance.

Deployment Implications. The pipeline runs within real-time constraints: after a 4-second buffer, inference takes 200 ms with updates every 0.5 seconds. The 4-minute calibration protocol is shorter than typical rule-based systems, since the model has learned to handle most between-subject variation from training data. However, the current implementation depends on cloud-based GPU inference, adding network latency and requiring a stable connection. Model quantisation (INT8) or knowledge distillation into a smaller network are viable paths toward edge deployment. The STFT-based pipeline is well-suited for this transition, as FFT is hardware-accelerated on virtually every platform.

5 Limitations and Future Development

5.1 Limitations

The current CNN–Transformer implementation requires cloud-based GPU inference rather than fully on-device processing. This introduces network latency of approximately 100–200 ms and assumes a stable internet connection. Although this latency remains acceptable for non-critical workplace monitoring, it limits deployment in off-line or privacy-sensitive settings. Model quantisation, knowledge distillation, and lightweight CNN variants are promising routes towards mobile or edge inference.

Synthetic labels and ecological validity. The most important limitation concerns the construction of the label. The cognitive-state labels are derived from PhysioNet task conditions rather than participants’ self-reported experiences. The Flow label is therefore an approximation rather than a direct measure of subjective flow. In particular, eyes-open rest during a task block should not be equated with the deep absorption described in flow theory. The near-perfect F1-scores for Relaxation and Flow are likely influenced by the separability of the underlying task conditions, especially eyes-closed versus eyes-open states. These results should therefore be interpreted as an upper-bound estimate under controlled conditions. Before workplace deployment, the system should be validated with naturalistic data, experience sampling, and concurrent self-reports from a larger participant sample.

Channel limitations. The four-channel BrainBit montage does not include frontal electrodes such as F3, F4, or Fz. As a result, the system cannot directly measure frontal alpha asymmetry or frontal midline theta, both of which are commonly used in studies of focus, workload, and executive control. The model may partially compensate by learning temporal theta patterns from T3/T4, but these channels remain indirect proxies. Additional frontal coverage would likely improve the separation between Focus and Engagement.

Focus–Engagement ambiguity. The 9–10% confusion between Focus and Engagement appears consistently across all evaluated methods. This suggests that the ambiguity reflects genuine overlap in the EEG signatures of these states, combined with limited spatial resolution, rather than a failure of the proposed model. Multimodal sensing, for example, combining EEG with heart-rate variability, interaction logs, or behavioural markers, may improve discrimination, although this would increase system complexity.

Individual differences and cross-subject robustness. EEG patterns vary substantially across individuals. Although our evaluation uses a subject-stratified split to ensure that segments from the same participant do not appear in both the training and test sets, we did not conduct Leave-One-Subject-Out cross-validation. LOSO evaluation is a stronger test of cross-subject generalisation and should be included in future work. Until then, claims about subject-independent deployment remain preliminary.

Lack of statistical significance testing. We did not conduct paired statistical comparisons between methods, such as McNemar’s test, paired bootstrap testing, or paired t-tests over repeated runs. Therefore, small differences between models, including the 0.17 percentage-point gap between the CNN–Transformer and Dual Transformer, should not be interpreted as statistically meaningful. Future evaluations should report confidence intervals and paired significance tests.

5.2 Future Work

First, device-based inference should be explored through INT8 quantisation, pruning, or knowledge distillation into a smaller model. This would reduce reliance on cloud inference and enable real-time deployment on mobile phones or wearable devices.

Second, multimodal integration could help resolve the Focus–Engagement ambiguity. Combining EEG spectrograms with heart-rate variability from the BrainBit PPG sensor, behavioural interaction data, or task-performance indicators may provide complementary evidence about cognitive state.

Third, transfer learning should be investigated to improve generalisation across users. Pre-training on larger public EEG datasets such as DEAP, DREAMER, or STEW, followed by fine-tuning on BrainBit data, may reduce the calibration burden and improve robustness to individual differences.

Fourth, longitudinal workplace deployment is needed to assess whether the model remains reliable over extended periods. Such data would allow the system to model daily cognitive rhythms, productivity cycles, fatigue accumulation, and personalised work–rest patterns.

Fifth, attention-map analysis should be used to improve interpretability. Visualising the Transformer’s attention weights over time–frequency regions may help identify which spectral patterns drive classification decisions and whether they align with established EEG markers.

Finally, Leave-One-Subject-Out cross-validation should be conducted to quantify subject-independent performance more rigorously. This is an important next step before making strong claims about generalisability beyond the current dataset.

6 Conclusion

We proposed a four-channel consumer EEG pipeline combining STFT spectrograms with a lightweight CNN–Transformer model. The system achieved 94.69% accuracy across four cognitive states without manual EEG feature engineering. Across five evaluated pipelines, three findings emerged. First, richer input representations were more

important than model complexity alone. Second, preserving within-window temporal dynamics improved performance over static summary features. Third, spectrogram-based end-to-end learning achieved performance comparable to handcrafted feature pipelines, while reducing the need for expert-designed EEG features. The results also show clear limitations. Relaxation and Flow were classified almost perfectly, but this likely reflects the separability of the controlled task conditions rather than guaranteed real-world performance. The main remaining challenge is the confusion between Focus and Engagement, which appears to stem from overlapping EEG signatures and the limited spatial coverage of four electrodes. Overall, the findings suggest that low-cost wearable EEG, combined with lightweight deep learning, can support practical cognitive-state monitoring. However, real-world validation with longitudinal workplace data and self-reported cognitive states is needed before deployment claims can be made.

References

1. M. Csikszentmihalyi, **Beyond Boredom and Anxiety**. San Francisco: Jossey-Bass, 1975.
2. X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Martinen, "EEG-based emotion recognition: A tutorial and review," **ACM Computing Surveys**, vol. 55, no. 14s, pp. 1–57, 2023.
3. C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski, "Estimating cognitive workload in an interactive virtual reality environment using EEG," **Frontiers in Human Neuroscience**, vol. 13, p. 401, 2019.
4. S. Sanei and J. A. Chambers, **EEG Signal Processing**. John Wiley & Sons, 2013.
5. Biasiucci, B. Franceschiello, and M. M. Murray, "Electroencephalography," **Current Biology**, vol. 29, no. 3, pp. R80–R85, 2019.
6. W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," **IEEE Transactions on Cybernetics**, vol. 49, no. 3, pp. 1110–1122, 2019.
7. R. Zafar, S. C. Dass, and A. S. Malik, "Electroencephalogram-based decoding cognitive states using convolutional neural network and likelihood ratio based score fusion," **PLOS ONE**, vol. 12, no. 5, p. e0178410, 2017.
8. M. Abo-Zahhad, S. N. Ahmed, and S. N. Abbas, "A new EEG acquisition protocol for biometric identification using eye blinking signals," **International Journal of Intelligent Systems and Applications**, vol. 7, no. 6, pp. 48–54, 2015.
9. P. Jain, J. Yedukondalu, H. Chhabra, U. Chauhan, and L. D. Sharma, "EEG-based detection of cognitive load using VMD and LightGBM classifier," **International Journal of Machine Learning and Cybernetics**, vol. 15, pp. 4193–4210, 2024.
10. M. T. Knierim, M. Berger, and C. Weinhardt, "Exploring flow in real-world knowledge work using discreet cEEGrid sensors," in **Proc. CHI Conference on Human Factors in Computing Systems (CHI '25)**, Yokohama, Japan, 2025.
11. Y. Xu, Y. Du, J. Zou, T. Zhou, L. Xiao, L. Liu, and P. Ma, "AMDET: Attention-based multiple dimensions EEG Transformer for emotion recognition," **IEEE Transactions on Affective Computing**, vol. 15, no. 3, pp. 1392–1404, 2024.
12. Y. Sun, J. Wang, S. Fang, and Z. Cao, "Dual Attentive Transformer in long-term continuous EEG emotion analysis," **arXiv preprint arXiv:2407.20519**, 2024.

13. Z. Li, Y. Zheng, and W. Zheng, "Spectral-spatial attention alignment for multi-source domain adaptation in EEG-based emotion recognition," **IEEE Transactions on Affective Computing**, vol. 15, no. 4, pp. 1894–1907, 2024.
14. J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," **Expert Systems with Applications**, vol. 47, pp. 35–41, 2016.
15. K. Sharma, Z. Papamitsiou, and M. Giannakos, "Measuring affective and motivational states as conditions for cognitive processing," in **Proc. 14th International Learning Analytics and Knowledge Conference (LAK 2024)**, pp. 234–244, 2024.
16. W. Zheng and B. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," **IEEE Transactions on Autonomous Mental Development**, vol. 7, no. 3, pp. 162–175, 2015.
17. K. Katahira, Y. Yamazaki, C. Yamaoka, H. Ozaki, S. Nakagawa, and N. Nagata, "EEG correlates of the flow state: A combination of increased frontal theta and moderate fronto-central alpha rhythm in the mental arithmetic task," **Frontiers in Psychology**, vol. 9, p. 300, 2018.
18. M. Arns, C. K. Conners, and H. C. Kraemer, "A decade of EEG theta/beta ratio research in ADHD: A meta-analysis," **Journal of Attention Disorders**, vol. 17, no. 5, pp. 374–383, 2013.
19. T. Pope, E. H. Bogart, and D. S. Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," **Biological Psychology**, vol. 40, pp. 187–195, 1995.
20. Tremmel, C. Herff, and D. J. Krusienski, "EEG-based brain-computer interactions in immersive virtual and augmented reality: A systematic review," **ACM Computing Surveys**, vol. 55, no. 14s, 2023.
21. J. Harris, S. J. Vine, and M. R. Wilson, "Neurocognitive mechanisms of the flow state," **Progress in Brain Research**, vol. 234, pp. 221–243, 2017.
22. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in **Proc. ICLR**, 2021.
23. G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," **IEEE Transactions on Bio-medical Engineering**, vol. 51, no. 6, pp. 1034–1043, 2004.
24. Y. Ding, J. H. Lee, S. Zhang, T. Luo, and C. Guan, "Decoding human attentive states from spatial-temporal EEG patches using Transformers," **arXiv preprint arXiv:2502.03736**, 2025.
25. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in **Advances in Neural Information Processing Systems (NeurIPS)**, vol. 30, 2017.
26. D. L. Donoho, "De-noising by soft-thresholding," **IEEE Transactions on Information Theory**, vol. 41, no. 3, pp. 613–627, 1995.
27. BrainBit, "BrainBit EEG Headband — Technical Specifications," BrainBit LLC, 2024. [Online]. Available: <https://brainbit.com/system-eeeg-software/specifications/>