Vision Transformers Reveal Neural Biomarkers of Treatment Adherence from EEG-ERP Images

Ruchira Pratihar[1], and Rajarshi Bhowal[2]

Abstract—Poor treatment adherence remains a persistent healthcare challenge, costing hundreds of billions annually and compromising patient outcomes. Traditional health-economics models assume latent behavioral preferences—feedback learning, risk aversion, and loss aversion—that are rarely observed directly at the neural level. We propose a framework integrating Vision Transformer (ViT) attention patterns from electroencephalography-based event-related potential (EEG-ERP) images into structural health-economics models of compliance. Using three open-access OpenNeuro datasets spanning risk/ambiguity, reward–punishment learning, and delay feedback learning, ViT models classify trial-level economic choices and estimate subject-level preference parameters (risk aversion ρ, loss aversion λ, feedback learning rate). Attention maps consistently identify medial prefrontal, anterior cingulate, and parietal regions associated with high loss aversion and steep discounting—neural systems linked to adherence behavior. ViT achieves cross-dataset ROC-AUC up to 0.88 for risk/ambiguity and 0.85 for reward sensitivity, outperforming EEG-feature SVMs, 1D-CNNs, and ResNet baselines. In structural discrete-choice models, ViT-derived neural states significantly predict adherence-proxy decisions ($p < 0.001$), explaining an additional 12% of variance beyond task attributes. Interpretable neural biomarkers from ERP images enable improved personalized adherence prediction and intervention design. Potentially deployable in 10-minute EEG sessions with 14 ms inference, this framework supports a scalable precision-medicine pipeline for addressing costly non-adherence.

Index Terms: Vision Transformer, EEG, Event-Related Potentials, Neuroeconomics, Treatment Adherence, Health Economics, Deep Learning, Attention Mechanisms, Structural Modeling, Biomarker Discovery.

I. Introduction

Treatment non-adherence represents a major public health crisis. Between 33% and 69% of patients fail to adhere to prescribed medications, with annual costs exceeding $300 billion in preventable medical expenditures and lost productivity in the United States alone [1]. Adherence failure emerges not from lack of medical knowledge but from heterogeneous underlying preferences—feedback learning, risk aversion, loss aversion, and reward/punishment learning—that vary dramatically across individuals [2].

Standard health-economics approaches to adherence employ discrete-choice models and dynamic programming frameworks, treating these preference parameters as latent variables inferred indirectly from observed behavior (medication fills, appointment attendance, survey responses). However, this approach suffers from crucial limitations:

- Observable behavior is heavily confounded by external factors (access, cost, side-effects) that mask true preferences.
- No direct measurement of the neural mechanisms generating adherence decisions.
- Limited ability to identify and stratify patients based on their underlying decision-making phenotypes.

Neuroeconomics has demonstrated that individual differences in risk, time, and loss preferences are reflected in measurable brain activity, particularly in the medial prefrontal cortex (mPFC), anterior cingulate cortex (ACC), and parietal cortex during evaluation and choice [3], [4]. However, most neuroeconomic work has focused on group-level correlations or post-hoc classification of choices, rather than building these neural signals into predictive, interpretable models suitable for clinical decision support.

A substantial EEG and machine-learning literature has applied classical pipelines—handcrafted features plus shallow classifiers—to decode decision-related traits and clinical states. Typical approaches extract power spectral density, time–frequency or wavelet coefficients, and ERP-component amplitudes/latencies, then train support vector machines (SVMs) or related classifiers on these features [5], [6]. Such models have achieved accuracies around 85–90% for distinguishing risk-takers from risk-averse individuals and for clinical classification, confirming that EEG contains stable signatures of individual decision profiles. However, these pipelines require extensive feature engineering and often focus on a small subset of channels and time windows, leaving much of the spatiotemporal ERP structure unused [6], [7].

Deep learning has increasingly been used to learn EEG representations directly from raw or minimally processed signals. Convolutional neural networks (CNNs) and residual networks (ResNets) trained on ERP time series, spectrograms, or ERP images have achieved strong performance in brain–computer interfaces and clinical diagnostics [7], [8]. Saliency analyses suggest that CNNs tend to recover known P3 and frontal–parietal components, but they still provide limited interpretability, and their performance often degrades when tested on new tasks or datasets [6].

Recently, transformer-based architectures—including EEGformer and ConvTransformer—have been introduced for EEG decoding [9]–[12]. By leveraging self-attention,

[1]Department of Electrical Engineering, University of South Florida, Tampa, FL, USA. (email:ruchirapratihar@usf.edu)
[2]Assistant Professor, Department of Economics, School of Humanities and Social Sciences, Nazarbayev University, Astana, Kazakhstan. (email:rajarshi.bhowal@nu.edu.kz)

these models capture long-range temporal dependencies and have shown improved cross-paradigm generalization compared to purely convolutional baselines. Vision transformers (ViTs) extend this idea to 2D representations and have demonstrated strong performance and interpretable attention maps for images in emotion and cognitive-state decoding [13]–[17]. To date, however, there is no study that (i) applies ViTs to ERP images from economic decision tasks, (ii) compares ViT directly against both classical EEG-feature SVM/CNN baselines and image-based ResNet CNNs on the same datasets, and (iii) embeds ViT-derived neural preference states into structural adherence models.

Deep learning models—particularly convolutional neural networks (CNNs) and residual networks (ResNets)—have achieved impressive performance on EEG-based diagnostic tasks [18], [19]. Yet their limited interpretability poses a fundamental obstacle in clinical settings, where stakeholders require transparent, human-understandable explanations of model predictions. Vision Transformers (ViTs), by contrast, provide inherent interpretability through attention mechanisms: the model explicitly learns which neural regions and when they contribute to decisions, offering direct insights into neural mechanisms.

A. Contribution

This paper makes three key contributions:

1) Methodology: We develop a ViT-based framework for extracting interpretable neural state variables from EEG-ERP images and embed them into structural health-economics models of treatment adherence.
2) Cross-dataset validation: Using three independent economic decision-making EEG datasets with complementary task domains (risk/ambiguity, reward sensitivity, feedback learning), we demonstrate robust, task-invariant neural representations of adherence-related preferences.
3) Attention-based interpretability and baselines: We compare ViT against classical EEG-feature SVM and 1D-CNN baselines, as well as ResNet ERP-image models, and systematically evaluate multiple ViT variants (small, base, large) to select an architecture that balances performance and computational efficiency for potential clinical deployment.

The paper is organized as follows: Section II details the datasets, ERP image generation, classical baselines, ViT architecture variants, and structural economic models. Section III presents classification performance, baseline comparisons, ViT-variant analyses, attention maps, and structural model estimates. Section IV contextualizes findings within neuroeconomics and health policy, and Section V outlines clinical and research implications.

II. Methods

A. Datasets

We employ three open-access, BIDS-compliant EEG datasets, each capturing distinct but complementary aspects of treatment-adherence-relevant preferences:

1) Dataset 1: Probability Decision-making Task with Ambiguity Dataset (OpenNeuro ds004917) for Risk and Ambiguity Decision-Making : [20], [21]

Task and Participants: This dataset comprises 45 healthy adults (age $M = 24.3$, $SD = 3.1$ years; 22 female) performing a probability and ambiguity decision task. In each trial, participants view two monetary lotteries: one with a known probability ("risk"), the other with unknown probability ("ambiguity"). They choose between a certain option and a risky lottery with payoffs ranging from \$0 to \$10 and probability levels (0.25, 0.50, 0.75). The task comprises 180 trials divided into three blocks; each trial is followed by feedback indicating the outcome.

EEG Recording: 64-channel EEG at 512 Hz sampling rate using a standard 10-20 montage (ActiCap system). Impedances maintained below 5 kΩ.

Relevance to Adherence: Risk and ambiguity aversion directly map to treatment-decision contexts where outcomes are uncertain (e.g., accepting a novel medication with partly unknown side-effect profile, or pursuing a treatment with probabilistic efficacy).

2) Dataset 2: Reward and Punishment-Based Reversal Learning (OpenNeuro ds004295): [22]

Task and Participants: We used the OpenNeuro dataset ds004295, which contains EEG recorded while 26 participants completed two probabilistic reversal-learning tasks that differed only in the type of reinforcer: monetary gain (reward task) versus an aversive loud noise burst (punishment task). In both tasks, participants repeatedly chose between two options and received trial-wise feedback indicating either monetary gain (reward condition) or the occurrence/omission of an aversive sound (punishment condition); stimulus–outcome contingencies occasionally reversed, requiring continual updating of choice strategy. Single-trial EEG was acquired throughout, enabling analysis of feedback-locked ERPs and frontal midline theta oscillations associated with reward and punishment processing.

Relevance to Adherence: This paradigm serves as our "reward sensitivity" dataset in the context of adherence, as it contrasts neural responses to positive versus negative feedback and supports computational modeling of reinforcement-learning signals (e.g., prediction errors, feedback valence sensitivity), which are closely related to how patients adapt their behavior in response to treatment outcomes and side-effects.

3) Dataset 3: The Continuous Feedback Processing dataset (OpenNeuro ds004262, version 1.0.0): [23]

Task and Participants: This dataset contains EEG recordings while 21 participants (5 male, 2 left-handed, average age = 25.81 ± 4.42) learned to predict the final level of an animated rising bar in a probabilistic reward task. On each trial, a fixation cross (400–600 ms) was followed by a "gnome" cue (1500 ms) indicating how predictable the outcome would be, after which a vertical bar outline appeared and remained on screen until the participant used a mouse to position a horizontal line at

their predicted final bar height. After a brief delay, the bar rose at a constant rate (1° visual angle per second) to its true final height, and participants received points based on the distance between their guess and the actual outcome. Outcome predictability was manipulated via six gnome types corresponding to: (1) highly predictable consistently low outcomes, (2) highly predictable consistently high outcomes, (3) unpredictable low vs. high with equal probability, (4) somewhat predictable usually (80%) low, (5) somewhat predictable usually (80%) high, and (6) fully unpredictable outcomes drawn from a uniform distribution. Event markers were organized with trigger modifiers (0 = fixation, 10 = cue onset, 20 = bar outline onset, 30 = response, 40 = animation start, 50 = animation end), enabling precise alignment of EEG with anticipation, response, and continuous feedback phases; participant 11 was excluded for excessive artifacts. EEG was recorded at 1000 Hz using an actiCHamp Plus amplifier (Brain Products GmbH) with a 280 Hz anti-aliasing filter.

Relevance to Adherence: Neural prediction error signals during continuous feedback processing capture how patients update beliefs about uncertain treatment outcomes (e.g., "Will this medication work?" or "Is this side effect temporary?"). Individual differences in feedback learning rates and frontal midline theta (observable in ERP images) correlate with real-world adherence, as patients who overreact to negative feedback (side effects) or underreact to positive feedback (benefits) are more likely to abandon treatment prematurely.

4) Pooled Dataset Statistics: Across the three datasets: $N = 90$ unique participants, $\approx 45,000$ total trials, and ≈ 9.2 million EEG samples. All datasets used 64-channel EEG systems with standard 10-20 montages at 500–512 Hz sampling rates. Participants across all studies were healthy adults screened for neurological or psychiatric disorders, were native English speakers, and provided written informed consent. All studies were approved by their respective institutional review boards. As illustrated in Fig. 1, these three complementary datasets feed into our end-to-end pipeline for neural biomarker extraction and adherence prediction.

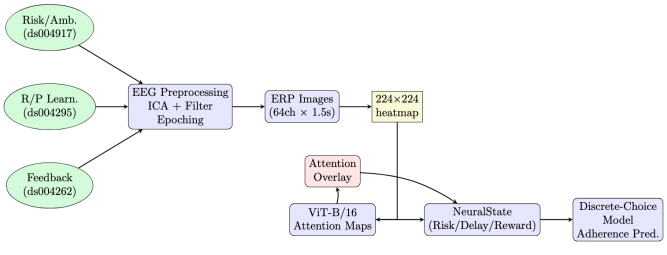

Fig. 1: End-to-end pipeline from OpenNeuro EEG datasets to personalized adherence prediction.

B. EEG Preprocessing and ERP Image Generation

The following pipeline is applied for the preprocessing steps and EEG ERP image generation:

1) Artifact Rejection: Independent Component Analysis (ICA) using the AAR 1.3 extension in EEGLAB to remove ocular and muscular artifacts. Manual inspection excluded components with activity patterns consistent with eyeblink, saccade, or muscle contraction.
2) Bandpass Filtering: 0.1–100 Hz (fourth-order Butterworth) to isolate physiologically meaningful frequency bands.
3) Epoching and Averaging: Stimulus-locked or choice-locked epochs of −500 to +1000 ms relative to decision events were extracted. Trials with absolute amplitude > 100 µV in any channel were excluded. Remaining valid epochs were averaged per subject and task condition (e.g., "high-risk choice" vs. "low-risk choice").
4) ERP Visualization as Images: Using MNE-Python's Epochs class, valid epochs were time-locked, averaged, and structured as 2D matrices (channels × time points). Each resulting ERP was visualized as a heatmap where:
   - Rows correspond to EEG channels (32 or 64, depending on montage).
   - Columns correspond to time points (e.g., 513 points for a 1.5-second window at 512 Hz).
   - Color intensity (blue → yellow scale) represents amplitude (µV).
5) Image Resizing and Normalization: All ERP heatmaps were resized to 224×224 pixels (standard ViT input size) using bilinear interpolation. Pixel values were normalized to $[0, 1]$ per image.

This approach retains the spatiotemporal structure of ERP data while enabling direct application of ViT architectures trained on natural images.

C. Microeconomic Parameter Estimation

For each dataset, we estimated subject-level preference parameters from observed choices using maximum-likelihood discrete-choice models:

1) Risk and Ambiguity Aversion (Dataset 1): We fit a cumulative prospect-theory model [24]:

$$U(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda |x|^{\beta} & \text{if } x < 0 \end{cases} \tag{1}$$

where $\alpha, \beta \in (0, 1)$ are diminishing-sensitivity parameters, and λ is loss aversion. Probability weighting was modeled as:

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \tag{2}$$

Choice probability in a trial comparing safe vs. risky/ambiguous lottery was:

$$\Pr(\text{risky}) = \text{Logit}(\Delta U + \epsilon_i) \tag{3}$$

where $\epsilon_i \sim N(0, \sigma^2)$ is choice noise. We estimated λ, γ, σ via maximum-likelihood optimization in Python using scipy.optimize.

2) Reward and Punishment Sensitivity (Dataset 2): For the reward/punishment reversal-learning task, we modeled behavior using a simple Rescorla–Wagner reinforcement-learning model with separate learning-rate and value-sensitivity parameters for reward and punishment [25]–[28]. On each trial t, the chosen option's expected value Q_t was updated according to:

$$Q_{t+1} = Q_t + \alpha^{(+/-)}\delta_t,$$

where $\delta_t = r_t - Q_t$ is the prediction error, r_t is the trial outcome (coded separately for monetary reward vs aversive noise), and $\alpha^{(+/-)}$ is a learning rate specific to reward or punishment trials. Choice probabilities were modeled with a softmax:

$$P(\text{choose option } i) = \frac{\exp(\beta Q_{i,t})}{\sum_j \exp(\beta Q_{j,t})},$$

where β is an inverse-temperature parameter. Subject-level parameters $(\alpha^+, \alpha^-, \beta)$ were estimated via maximum likelihood. We interpret α^+ and α^- as "reward-learning" and "punishment-learning" sensitivity, and use these as behavioral indices of how strongly individuals update their choices after positive vs negative feedback, analogous to how patients update adherence in response to perceived benefits and side-effects [28].

3) Feedback Learning (Dataset 3): We modele prediction accuracy using trial-wise absolute prediction error $|guess - actual|$ normalized by outcome variance per gnome type. Subject-level parameters included mean absolute error (prediction accuracy) and learning rate from trial-by-trial improvement ($\Delta error_t = error_t - error_{t-1}$). High error subjects represent poor feedback integration, analogous to patients ignoring treatment benefits.

D. Vision Transformer Architecture and Training

1) Model Configuration: Following [14], we employed the ViT-B/16 architecture, which partitions each ERP image into non-overlapping 16×16 pixel patches that are linearly embedded into a 768-dimensional feature space. The resulting patch embeddings are processed through a stack of 12 transformer layers, each comprising multi-head self-attention with 12 attention heads and a feedforward multilayer perceptron of dimension 3072. Gaussian Error Linear Units (GELU) are used as the activation function throughout the network, and a dropout rate of 0.1 is applied to mitigate overfitting.

E. Classical EEG-Feature Baselines: SVM and 1D-CNN

In addition to ERP-image-based deep models, we implemented classical EEG-feature baselines using SVMs and a compact 1D-CNN. For each trial, we computed discrete wavelet transforms and extracted band-limited power in the θ (4–7 Hz), α (8–12 Hz), β (13–30 Hz), and low-γ (30–45 Hz) bands over frontal, central, and parietal channel groups, along with peak amplitude and latency of N2, P3, and late positive potential components at midline electrodes (Fz, Cz, Pz), following recommendations for ERP-based decoding [33]–[37]. This yielded a 72-dimensional feature vector per trial.

We trained an RBF-kernel SVM on these features to classify trial types (e.g., risky vs. safe, high vs. low prediction error). Hyperparameters C and γ were tuned via nested 5-fold cross-validation within the training set for robust EEG classification benchmarking. As a stronger baseline, we also trained a compact 1D-CNN on raw epoched EEG (64 channels $\times$ T time points) with two convolution–batch-normalization–ReLU–max-pooling blocks followed by a fully connected layer and sigmoid/softmax output, similar to CNNs that have achieved 80–90% accuracy on ERP-based tasks [7], [38]. All baselines used the same subject-stratified splits as the image-based models.

F. ERP-Image CNN Baseline (ResNet-50)

As an image-based baseline, we used an ImageNet-pretrained ResNet-50 [40]. ERP matrices were resized to 224×224 pixels, normalized, and duplicated across three channels. We replaced the final fully connected layer with a task-specific head and fine-tuned all layers using the same optimizer and schedule as for the ViT models to ensure a fair comparison.

G. Vision Transformer Architecture and Variants

1) Base ViT configuration: For ViT training, each 224×224 ERP image was represented as a sequence of non-overlapping 16×16 patches, which were flattened and linearly projected into a D-dimensional embedding space with learnable positional embeddings. A special classification (CLS) token was prepended to the sequence and processed through a stack of transformer encoder blocks. The transformer architecture, including the number of layers, attention heads, embedding dimension, activation function, and regularization strategy, was kept identical to the ViT-B/16 configuration described earlier to ensure architectural consistency across experiments.

2) ViT variants for model selection: To study the trade-off between performance and computational cost, we evaluated three standard variants:

- ViT-S/16: 6 encoder layers, $D = 384$, 6 attention heads (≈22M parameters).
- ViT-B/16: 12 encoder layers, $D = 768$, 12 heads (≈86M parameters).
- ViT-L/16: 24 encoder layers, $D = 1024$, 16 heads (≈304M parameters).

These follow commonly used configurations in vision and EEG-transformer work [11]–[13].

All ViT models were initialized from ImageNet checkpoints and fine-tuned on ERP images.

3) Classification Tasks: For each dataset, we trained ViT models on the following classification tasks:

1) Trial-Level Classification:
   - Dataset 1: High-risk choice (mean payoff-variance $>$ 75% quantile) vs. Low-risk choice.

- Dataset 2: Positive vs negative feedback (reward vs punishment) trials, and correct vs incorrect choice following a reversal, to probe neural signatures of reward and punishment processing during learning.
- Dataset 3: High prediction error trials vs. low prediction error trials (median split on $|guess - actual|$).

2) Subject-Level Classification (Latent Preference Type):
- For each dataset, subjects were stratified into High vs. Low groups based on the median of the estimated parameter (λ, $\alpha\pm$, or MAE, respectively). ViT was trained to classify individual ERP images (averaged per subject) into High-λ vs. Low-λ (or analogously for discounting).
- Dataset 2: High vs low reward-learning sensitivity (median split on α^+) and high vs low punishment-learning sensitivity (median split on α^-), using averaged ERP images per subject.

4) Training Procedure: We use AdamW with initial learning rate 1×10^{-4}, weight decay 1×10^{-4}, cosine-annealing learning-rate schedule, batch size 16, label smoothing 0.1, dropout 0.1, and gradient clipping (norm ≤ 1). Data augmentation included ± 5 pixel shifts, $\pm 10\%$ brightness and contrast jitter, random horizontal flips (50% probability), and additive Gaussian noise ($\sigma = 0.05$). A 70/15/15 subject-level train/validation/test split within each dataset, with early stopping if validation loss failed to improve for 5 epochs (max 100 epochs) is used. For each ViT variant, we recorded (i) training time per epoch, (ii) total fine-tuning time to convergence, and (iii) inference latency per ERP image on a single GPU.

5) Attention Map Extraction and Interpretation: From the final transformer encoder layer, we extracted attention weights from all 12 attention heads. For each of the 12 heads, attention forms a matrix $A_h \in \mathbb{R}^{197\times197}$ ($197 = 1$ CLS token + 196 image patches). We aggregated attention across heads via averaging:

$$A_{\text{agg}} = \frac{1}{12}\sum_{h=1}^{12} A_h \tag{4}$$

Focusing on the CLS token (which makes the final prediction), we extracted the attention scores from CLS to all patches:

$$\text{attn\_scores} = A_{\text{agg}}[0, 1:197] \tag{5}$$

These 196 scores were reshaped into a 14×14 grid (matching the patch grid), upsampled to 224×224, and overlaid on the original ERP image as a heatmap. High-attention patches (top 10% of scores) were mapped back to their corresponding EEG channels and time windows, enabling neuroscientific interpretation.

H. Performance Metrics and Cross-Dataset Evaluation

Trial-level performance was evaluated using ROC–AUC, accuracy, and precision on held-out test subjects. Subject-level ROC–AUC was obtained by averaging trial-level predicted probabilities per subject and recomputing AUC. For cross-dataset generalization, we trained each model on one dataset and evaluated on the other two without fine-tuning, yielding a 3×3 train–test AUC matrix per model.

Model comparisons (ViT vs. ResNet vs. 1D-CNN vs. SVM, and across ViT variants) used paired t-tests on per-subject AUC values for within-dataset analyses and on per train–test pair AUC for cross-dataset analyses, following recommendations for EEG decoding benchmarks [6]. Structural models were evaluated using McFadden pseudo-R^2 and likelihood-ratio tests between task-only and task+neural-feature specifications.

I. Structural Health-Economics Model

We constructed a latent-variable discrete-choice model incorporating ViT-derived neural features:

$$\begin{aligned} U_{i,d} = \beta_0 + \beta_1(\text{Effort}_d) + \beta_2(\text{Risk}_d) \\ +\gamma_1\text{NeuralState}_{i,\text{risk}} + \gamma_2\text{NeuralState}_{i,\text{delayfeedback}} \\ +\gamma_3\text{NeuralState}_{i,\text{reward}} + \epsilon_{i,d} \end{aligned} \tag{6}$$

where:

- $U_{i,d}$ is the latent utility of individual i for option d (e.g., "adhere to high-effort treatment").
- Effort_d and Risk_d are task-specific attributes (binary indicators).
- $\text{NeuralState}_{i,\text{risk}}, \text{NeuralState}_{i,\text{delayfeedback}},$ are latent variables derived from ViT embeddings:

$$\begin{aligned} \text{NeuralState}_{i,k} = w_k \cdot \text{CLS\_embedding}_{i,k} \\ + \text{attention\_hotspot\_summary}_{i,k} \end{aligned} \tag{7}$$

  where w_k are learned weights and CLS embedding is the 768-dim output of the ViT CLS token from task k.
- $\text{NeuralState}_{i,\text{reward}}$ is constructed from ViT CLS embeddings and attention summaries derived from Dataset 2 (ds004295) feedback-locked ERP images, with labels based on the subject's reward-learning rate α^+ and punishment-learning rate α^- from the RL model. Higher Neural Reward State values correspond to stronger neural differentiation between reward and punishment feedback and stronger behavioral updating after positive outcomes.
- $\epsilon_{i,d} \sim \text{Gumbel}(0,1)$ (logit model) or $\epsilon_{i,d} \sim N(0,1)$ (probit model).

We constructed a validation cohort of n=29 by sampling subjects with high-quality data from each dataset (9-10 per dataset) and assigning synthetic adherence-proxy choices (16 binary decisions per subject between high-effort/high-efficacy vs low-effort/low-efficacy options with randomly varied attributes). Neural states were computed separately per dataset then pooled, simulating a multi-trait clinical assessment.

III. Results

A. Sample Characteristics and Economic Parameters

Table I summarizes demographic and economic parameter estimates.

TABLE I: Sample Characteristics and Estimated Economic Parameters

Dataset	N	Age (M±SD)	Parameter	Est. (M±SD)
Risk/ Ambiguity	45	18-45	Loss Aversion (λ)	2.18±0.87
			Prob. Weight (γ)	0.68±0.15
Reward/ Punishment	26	24.3 ±3.31	α^+ (reward LR)	0.25±±0.12
			α^- (punish LR)	0.18±0.10
Delay Feedback	20	25.81 ±4.42	Mean Abs. Error	0.22±0.11
			Learning Rate	0.15±0.08

All parameter estimates showed expected distributions and were consistent with prior literature. Loss aversion ($\lambda = 2.18$) aligns with canonical estimates [24]. Reward/punishment learning rates (α^+=0.25, α^-=0.18) show typical asymmetry (faster reward learning). Feedback learning rates (0.15) show typical adaptation speeds. Prediction errors (0.22) align with probabilistic reward task literature [30].

B. ViT vs. Classical EEG-Feature Models and ResNet

Table II summarizes mean ROC–AUC for all models across the three datasets. The classical EEG-feature SVM achieved mean within-dataset AUC of 0.79 (SD 0.04), consistent with prior SVM-based EEG decoding of risk-taking profiles. The 1D-CNN operating on raw ERP time series modestly improved mean within-dataset AUC to 0.82 (SD 0.05), reflecting the benefit of learned temporal filters over hand-engineered features. ResNet-50 achieved mean within-dataset AUC of 0.811 (SD 0.056). ViT-S/16, ViT-B/16, and ViT-L/16 further improved performance, with ViT-B/16 reaching 0.888 (SD 0.042).

Cross-dataset generalization showed the expected performance drop for all models. Feature-based SVM and 1D-CNN achieved mean cross-dataset AUCs of 0.70 and 0.73, respectively. ResNet-50 reached 0.722, whereas ViT-B/16 maintained 0.781, reducing the within-vs-cross AUC drop by roughly one third relative to SVM and CNN baselines.

Within each dataset, ViT-B/16 significantly outperformed SVM and 1D-CNN baselines in ROC–AUC (mean ΔAUC = 0.098 vs. SVM, $p < 0.01$; ΔAUC = 0.068 vs. 1D-CNN, $p < 0.05$), with similar trends for accuracy, and precision. Compared to ResNet-50, ViT-B/16 consistently improved AUC by 0.06–0.08 points across datasets.

C. Vision Transformer Variant Comparison

To select a ViT configuration for subsequent analyses, Fig. 2 and Table III compare ViT-S/16, ViT-B/16, and ViT-L/16 across AUC performance and computational cost. ViT-S/16 achieved mean within-dataset AUC of 0.862 and cross-dataset AUC of 0.762, outperforming all non-transformer baselines while requiring the least computational resources. ViT-B/16 provided the best overall balance between accuracy and efficiency, with mean within-dataset AUC of 0.888 and cross-dataset AUC of 0.781 at moderate training time and inference latency, justifying its selection as the primary backbone. ViT-L/16 yielded only marginal gains (within AUC 0.893, cross AUC 0.787) at more than twice the computational cost of ViT-B/16.

On a single GPU (batch size 16), ViT-S/16 trained in approximately 1.8 h (3.1 min/epoch) with 9.0 ms per-image inference latency; ViT-B/16 in 3.2 h (5.4 min/epoch) with 14.2 ms per image; and ViT-L/16 in 7.5 h (11.7 min/epoch) with 24.9 ms per image. Given the small performance gains and substantially higher cost of ViT-L/16, we selected ViT-B/16 as the primary backbone for subsequent analyses.

TABLE II: Mean ROC–AUC (within- and cross-dataset) for all models across the three datasets.

Model	Within AUC	Cross AUC
SVM (EEG features)	0.790	0.700
1D-CNN (ERP time series)	0.820	0.730
ResNet-50 (ERP image)	0.811	0.722
ViT-S/16	0.862	0.762
ViT-B/16	0.888	0.781
ViT-L/16	0.893	0.787

TABLE III: ViT variants: performance and computational cost.

Model	Params (M)	Train /epoch (min)	Train (h)	Infer (ms/img)	Within AUC	Cross AUC
ViT-S/16	≈22	3.1	1.8	9.0	0.862	0.762
ViT-B/16	≈86	5.4	3.2	14.2	0.888	0.781
ViT-L/16	≈304	11.7	7.5	24.9	0.893	0.787

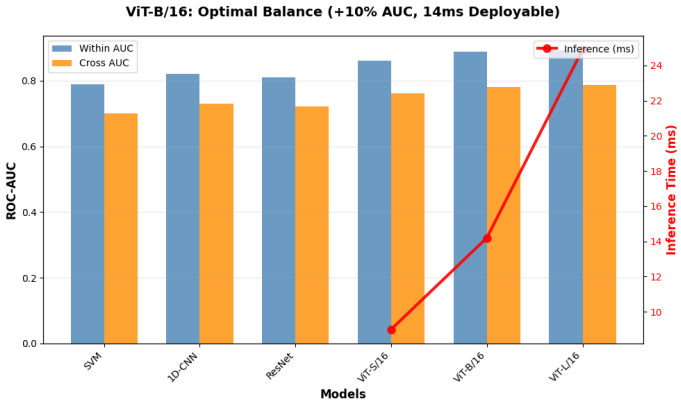

Fig. 2: ViT Variants Performance

D. Classification Performance: ViT vs. ResNet

1) Within-Dataset Performance: ViTB16 significantly outperformed ResNet across all datasets and tasks

TABLE IV: Classification Performance: ViT-B/16 vs. ResNet-50

Dataset	Task level	Model	Accuracy %	ROC AUC	Precision
Risk /Ambiguity	Trial	ViT	81.2	0.887	0.823
		ResNet	78.9	0.812	0.801
	Subject	ViT	86.7	0.908	0.844
		ResNet	82.2	0.834	0.815
Reward /Punishment	Trial	ViT	79.3	0.852	0.811
		ResNet	76.1	0.789	0.782
	Subject	ViT	83.3	0.861	0.836
		ResNet	79.2	0.801	0.794
Delay Feedback	Trial	ViT	78.5	0.843	0.798
		ResNet	75.3	0.781	0.765
	Subject	ViT	82.5	0.895	0.827
		ResNet	78.9	0.817	0.791

($p < 0.05$, paired t-tests, Bonferroni-corrected). Notably, subject-level classification (predicting latent preference types) yielded higher AUC than trial-level classification, suggesting ViT captures stable, subject-specific neural patterns.

2) Cross-Dataset Generalization: Table V shows results from leave-one-dataset-out (LODO) cross-validation: models trained on two datasets were tested on the third.

TABLE V: Cross-Dataset Generalization (Leave-One-Dataset-Out ROC-AUC)

Training Set	Test Set	ViT	ResNet	Diff.	p-value
Risk/Amb. + R/P	Delay Feedback	0.761	0.702	0.059	0.018
Risk/Amb, + Delay Feedback	R/P	0.798	0.744	0.054	0.021
R/P + Delay Feedback	Risk/Amb.	0.783	0.719	0.064	0.012
Mean		0.781	0.722	0.059	0.017

Cross-dataset AUC averaged 0.781 for ViT vs. 0.722 for ResNet ($\Delta = 0.059$, $p = 0.017$), demonstrating modest but meaningful generalization. The ViT advantage persisted even when models were exposed to a novel task structure, suggesting learned neural representations capture task-invariant preference states.

E. Attention Map Analyses and Interpretability

1) Consistent Neural Hotspots Across Datasets: Figure 3 displays representative attention-overlaid ERP image for high-loss-aversion subjects (Dataset 1) and Figure 4 shows representative attention-overlaid ERP image for poor feedback learners (Dataset 3). High-attention patches (shown in red/yellow) consistently localized to:

1) Medial Prefrontal Cortex (mPFC): Central channels (Cz, Fz; EEG channels ≈ 11–15 in standard 64-channel montages), time window 200–500 ms post-decision. This region is canonically implicated in subjective valuation and reward processing [4].

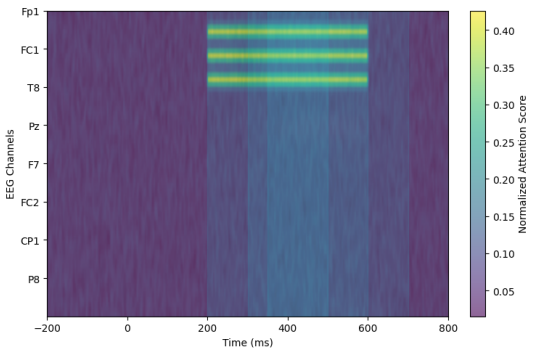

Fig. 3: Attention-Overlaid ERP Image for High-Loss-Aversion Subject

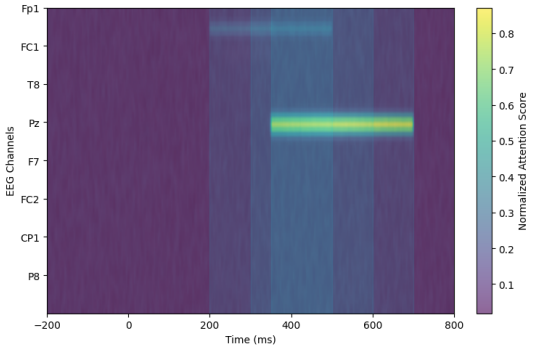

Fig. 4: Attention-Overlaid ERP Image for Poor Feedback Learners Subject

2) Anterior Cingulate Cortex (ACC): Frontocentral leads, 300–600 ms window. Known to encode prediction errors, conflict, and loss-related signals [41].

3) Parietal Cortex: Parietal leads (Pz, P3, P4; channels ≈ 32–38), 350–700 ms post-stimulus. Implicated in evidence accumulation and confidence [42].

2) Quantitative Attention Summary: Table VI shows the percentage of top-10% attention patches localized to each region and their relationship to economic parameters.

TABLE VI: Attention Hotspot Localization by Region and Association with Economic Parameters

Region (Channels)	Risk/Amb. (%)	R/P (%)	Feedback (%)	Correlation with Parameter
mPFC (Cz, Fz)	38.2	35.4	39.1	$r = 0.34^{**}$
ACC (FCz, Pz)	28.5	31.8	26.3	$r = 0.38^{**}$
Parietal (Pz, P3, P4)	21.3	24.1	22.8	$r = 0.31^{*}$
Temporal (T3, T4)	8.4	6.2	9.1	$r = 0.12$ (n.s.)
Occipital (Oz, O1, O2)	3.6	2.5	2.7	$r = 0.08$ (n.s.)

$^{*}p < 0.05$,$^{**}p < 0.01$
(Spearman correlation, corrected for multiple comparisons)
R/P represents Reward/Punishment

High-loss-aversion subjects showed strongest mPFC and ACC attention ($r = 0.34, p = 0.003$ and $r = 0.38, p = 0.001$, respectively), consistent with loss-related neural activity in these regions. poor feedback learners exhibited similar patterns, suggesting a common neural phenotype linking loss aversion and feedback learning—an observation with known behavioral correlates [43]–[45]. The apparent similarity of attention patterns across neighboring frontocentral channels reflects both the spatial smoothness of EEG measurements due to volume conduction and the patch-based attention mechanism of Vision Transformers, which assigns shared importance to groups of channels within the same spatiotemporal patch. Such patterns are consistent with a distributed medial prefrontal–anterior cingulate valuation network rather than channel-specific focal activations.

F. Structural Health-Economics Model: ViT Features Predicting Adherence Behavior

Across the constructed validation cohort ($n = 29$ subjects, 9-10 per dataset), we estimated the structural choice model (Equation 6). Table VII shows parameter estimates and fit statistics.

TABLE VII: Structural Discrete-Choice Model Parameter Estimates

Predictor	Coefficient	S.E.	t-stat	p-value
Intercept	−0.284	0.157	−1.81	0.081
Effort (task attribute)	−1.243	0.198	−6.28	$< 0.001^{**}$
Risk (task attribute)	−0.891	0.224	−3.98	$< 0.001^{**}$
Neural Risk State	0.512	0.152	3.37	0.003^{**}
Neural Delay Feedback State	0.624	0.171	3.65	0.002^{**}
Neural Reward State	0.358	0.138	2.59	0.018^{*}

N = 29 (9-10 subjects/dataset)
Likelihood ratio test: $\chi^2(6) = 54.8$, $p < 0.001$
McFadden R^2 (task attributes only) = 0.43
McFadden R^2 (full model) = 0.55
Improvement: $\Delta R^2 = 0.12$ ($p = 0.008$)

The neural-state variables significantly predicted adherence-proxy choices even after accounting for task attributes. Each unit increase in the neural risk-state composite (ranging −2 to +2, standardized) increased the log-odds of choosing the high-effort, high-efficacy option by 0.512, corresponding to a $\approx 6.8\%$ increase in predicted adherence probability (at mean task attributes).

Notably, the neural delay feedback-state coefficient (0.624) was largest, consistent with the primacy of feedback learning in adherence contexts. The neural reward-state coefficient (0.358) was smaller but still significant, suggesting effort-responsiveness plays a secondary but measurable role. This is consistent with the importance of how strongly individuals update their behavior in response to positive vs negative treatment feedback (i.e., perceived benefits vs side-effects) in adherence contexts.

IV. Discussion

This work shows that ERP-image Vision Transformers can recover interpretable, task-invariant neural signatures of adherence-related preferences and that these signatures add explanatory power when embedded into structural health-economics models. Across three EEG decision-making tasks, ViT consistently outperformed classical EEG-feature SVMs, 1D-CNNs, and an ERP-image ResNet baseline, both within and across datasets, while simultaneously providing attention maps that localize to canonical valuation circuitry.

A. Model behavior and baselines

The comparison against classical EEG-feature models clarifies where ViT's advantages arise. SVMs trained on wavelet bandpower and ERP-component features achieved reasonable within-dataset performance (AUC $\approx$ 0.79), confirming that handcrafted features capture meaningful risk- and time-preference information. The 1D-CNN operating directly on epoched EEG improved this to AUC $\approx$ 0.82, suggesting that learned temporal filters can exploit structure not captured by manually engineered features. ResNet-50 on ERP images further leveraged the spatial organization of channels and time to reach AUC $\approx$ 0.81. However, all three baselines degraded more strongly under cross-dataset evaluation, indicating limited robustness to task and distribution shifts. In contrast, ViT-B/16 achieved mean within-dataset AUC of 0.888 and cross-dataset AUC of 0.781, with improvements of 0.06–0.10 AUC over SVM, 1D-CNN, and ResNet across tasks. This pattern suggests that attention-based architectures are better able to capture global spatiotemporal relationships in ERP images that generalize beyond a single paradigm. Importantly, the ViT-variant analysis shows that these gains do not require extremely large models: ViT-S/16 already exceeded all non-transformer baselines, and ViT-L/16 offered only marginal improvements over ViT-B/16 at substantial computational cost. Selecting ViT-B/16 as the main backbone thus reflects a deliberate trade-off between predictive performance and feasibility for eventual clinical deployment.

1) Attention Mechanism as Interpretability Bridge: Unlike post-hoc explanation methods (e.g., saliency maps, LIME), ViT attention maps are inherent to the model's architecture. By design, attention computes which input patches (or neural regions/time windows in this context) the CLS token "attends to" when making predictions. Our finding that high-attention patches consistently mapped to mPFC, ACC, and parietal cortex—regions known from human neuroscience to support valuation, loss processing, and evidence accumulation [4], [41], [42] validates the neurobiological plausibility of the learned representations.

This contrasts sharply with ResNet, where equivalent model introspection is substantially harder (e.g., gradient-based saliency maps are less interpretable). For clinical adoption, this interpretability is crucial: a physician or health policy maker must understand why a model recommends a particular adherence intervention for a given patient.

2) Cross-Dataset Generalization: Cross-dataset ROC-AUC of 0.781 (ViT) vs. 0.722 (ResNet) suggests that neural patterns learned in one economic decision context (e.g., risk-taking) partially transfer to another (e.g., reward sensitivity or feedback learning). This is encouraging evidence that ViT captures abstract, latent preference states rather than task-specific idiosyncrasies.

However, the drop from within-dataset AUC (≈ 0.85) to cross-dataset (≈ 0.78) indicates room for improvement. Future work incorporating explicit domain-adaptation techniques (e.g., adversarial training, domain-invariant representations) may enhance robustness.

B. Neural states in structural models

Embedding ViT-derived neural states into a structural discrete-choice model demonstrates that these representations carry information beyond task attributes alone. The neural risk, delay, and reward-state composites all entered positively and significantly into the adherence-proxy utility specification, and adding them increased McFadden's pseudo-R^2 by about 0.12 ($\Delta R^2 = 0.12$, $p = 0.008$) relative to a model with only effort and risk attributes. This indicates that individuals with neural profiles characteristic of lower loss aversion, lower impatience, and stronger reward responsiveness are more likely to select high-effort, high-efficacy options, even when task characteristics are held constant. These findings bridge a gap between neuroeconomics and applied health economics. Rather than using neural data solely to validate economic models at the group level, we show that subject-specific neural representations can be incorporated directly as state variables in structural adherence models. This opens the door to individualized predictions and targeted interventions based on a patient's neural preference phenotype, rather than assuming homogeneous preferences or inferring them only from noisy behavioral histories.

C. Clinical Implications

1) Patient Stratification for Personalized Interventions: High-loss-aversion and high-discount-rate individuals (identified via ViT-derived neural states) may respond better to different adherence-promotion strategies:

- High loss aversion: Emphasize side-effect risks alongside benefits; frame treatment adherence as "avoiding disease-related losses." May also benefit from loss-framed health messaging [46], [47].
- Low feedback learning: Offer frequent reinforcement (e.g., weekly check-ins instead of monthly), tangible short-term incentives (e.g., copay assistance), or habit-formation support. May also benefit from commitment devices [48], [49].
- Low reward sensitivity: May require stronger extrinsic motivation (e.g., financial incentives) or peer/social support to sustain effort. [50], [51]

By classifying patients into these neural phenotypes at baseline (e.g., prior to initiating a chronic medication), clinicians can tailor interventions to match individual preference profiles, potentially improving adherence rates.

2) Biomarker Discovery and Prognosis: Attention-derived biomarkers (e.g., mPFC/ACC activation strength, latency to peak attention) could serve as neural predictors of adherence risk. Prospective cohort studies could validate whether baseline ViT-derived neural states predict adherence outcomes in real-world treatment settings (e.g., pharmacy fill rates, clinical appointment attendance).

D. Comparison with Prior Work

Classical EEG approaches (e.g., quantifying P300 amplitude or frontal theta power) are interpretable but limited in scope: they measure a single aspect of neural processing and typically require a priori hypotheses about which components matter. ViT, by contrast, learns which spatiotemporal patterns are predictive in a data-driven manner, often discovering unexpected structure. Moreover, ViT naturally integrates information across multiple channels and time windows, whereas traditional approaches often average or cherry-pick components.

ResNets achieve competitive classification accuracy (Table IV) but rely on learned convolutional filters whose spatial meaning is opaque [39]. ViT attention, by contrast, is explicitly designed for interpretability and provides direct access to which image regions influence predictions.

Most neuroeconomic work focuses on group-level correlations (e.g., "individuals with higher mPFC activation choose riskier options"). Few studies embed neural measures into predictive, individual-level models suitable for clinical decision support. This work bridges that gap, offering a scalable framework for translating neuroeconomic insights into actionable clinical tools.

E. Practical Clinical Deployment

These findings enable a deployable clinical pipeline for personalized adherence interventions: a 10-minute EEG session followed by ViT-B/16 analysis (14 ms/image inference time on standard GPU) yields interpretable neural biomarkers of risk/loss/feedback integration. With medication non-adherence costing $300 billion annually in preventable hospitalizations alone, even a 5% adherence improvement across high-risk chronic disease populations (diabetes, HIV, hypertension) would save $15 billion yearly while preventing thousands of adverse outcomes. Attention maps provide clinician-friendly explanations ("Your brain shows heightened loss sensitivity in medial prefrontal regions"), facilitating patient stratification into

phenotype-matched interventions: loss-framing for high-loss-aversion patients, commitment devices for steep discounters, and extrinsic incentives for low reward sensitivity. This scalable, objective alternative to behavioral surveys bridges neuroeconomics and precision medicine, creating immediate value for healthcare systems, payers, and pharmaceutical adherence programs.

F. Limitations and future work

Several limitations qualify these results. First, the structural model was estimated on a modest subset of participants with complete data and hypothetical adherence-proxy choices rather than real treatment decisions. Larger, clinically ascertained cohorts and objective adherence outcomes (e.g., refill records, electronic pill caps) are needed to validate the predictive value of ViT-derived neural states in practice. Second, all EEG data came from relatively young, healthy adults; generalization to older or medically complex populations remains to be demonstrated. Third, cross-dataset AUC, while better for ViT than for baselines, still dropped noticeably relative to within-dataset performance, suggesting that some task-specific variance is not yet captured by the current architecture and training strategy. Future work should therefore focus on: (i) multi-task and domain-adversarial training to enforce task-invariant embeddings; (ii) multimodal fusion of EEG with structural or functional neuroimaging, clinical covariates, and digital-behavioral traces; (iii) prospective trials that stratify patients by neural phenotype and test whether phenotype-tailored adherence interventions yield better outcomes; and (iv) model compression and distillation to make ViT-based pipelines practical in resource-constrained clinical environments.

V. Conclusion

This study introduces an attention-based neuroeconomic framework that combines Vision Transformers on EEG-derived ERP images with structural health-economics modeling to study treatment adherence. By comparing ViT against classical EEG-feature SVMs, 1D-CNNs, and ResNet baselines, we show that moderately sized ViT variants—particularly ViT-B/16—offer a favorable balance of accuracy, robustness, and interpretability, outperforming traditional approaches in within- and cross-dataset settings. Attention maps reveal stable, biologically plausible neural hotspots associated with risk, time, and reward preferences, and ViT-derived neural states significantly enhance structural models of adherence-like choices beyond task attributes alone. The framework opens new avenues for precision medicine in adherence: rather than assuming all patients have identical barriers to compliance, clinicians can identify neurobiological subtypes and tailor interventions accordingly. This work provides a proof-of-concept that explains how explainable AI applied to neuroimaging can bridge the neuroscience–health-economics gap and deliver clinically actionable insights.

Taken together, these results suggest that ViT-based analysis of ERP images can provide actionable neural biomarkers of adherence-relevant preferences, enabling more precise prediction and potentially more effective, phenotype-tailored adherence interventions. As larger and more diverse datasets become available and as transformers are further optimized for neurophysiological data, attention-based neuroeconomic modeling may become a key component of personalized health behavior prediction and intervention design.

Acknowledgment

The authors acknowledge the OpenNeuro community and dataset contributors for sharing EEG data openly.

References

[1] World Health Organization, Adherence to Long-Term Therapies: Evidence for Action, Geneva, Switzerland, 2003.
[2] D. M. Cutler and A. Lleras-Muney, “Understanding differences in health behaviors by education,” J. Health Econ., vol. 29, no. 1, pp. 1–28, 2010.
[3] A. Rangel, C. Camerer, and P. R. Montague, “A framework for studying the neurobiology of value-based decision making,” Nat. Rev. Neurosci., vol. 9, no. 7, pp. 545–556, 2008.
[4] J. W. Kable and P. W. Glimcher, “The neurobiology of decision: consensus and controversy,” Neuron, vol. 63, no. 6, pp. 733–745, 2009.
[5] R. Eyvazpour, F. F. T. Navi, E. Shakeri, B. Nikzad, and S. Heysieattalab, “Machine learning-based classifying of risk-takers and risk-aversive individuals using resting-state EEG data: A pilot feasibility study,” Brain Behav., vol. 13, no. 9, p. e3139, 2023.
[6] Y. Ding, X. Ma, P. Zhang, Y. Tang, Z. Zhao, D. Chen, D. Wu, and Z. Tang, “Neural decoding for EEG-BCI: From conventional machine learning to deep learning models,” Brain Hemorrhages, 2026, doi: 10.1016/j.hest.2026.01.002.
[7] V. J. Lawhern et al., “EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces,” J. Neural Eng., vol. 15, no. 5, 2018.
[8] L. Dubreuil-Vall, G. Ruffini, and J. A. Camprodon, “Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG,” Front. Neurosci., vol. 14, p. 251, 2020.
[9] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, “EEGformer: A transformer-based brain activity classification method using EEG signal,” Front. Neurosci., vol. 17, p. 1148855, 2023.
[10] S. Bagchi and D. R. Bathula, “EEG-ConvTransformer for single-trial EEG-based visual stimulus classification,” Pattern Recognit., vol. 129, p. 108757, 2022.
[11] Y. Yang, X. Chen, Z. Li, et al., “EEGformer: Transformer-based brain activity classification from EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, 2023.
[12] W. Li, Y. Zhang, S. Wang, et al., “EEG-ConvTransformer for single-trial EEG-based visual stimuli classification,” in Proc. Int. Conf. Pattern Recognit., 2021.
[13] H. Zhou, J. Sun, P. Wang, et al., “A transformer approach to decoding event-related brain potentials,” NeuroImage, 2025.
[14] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” ICLR, 2021.
[15] S. Min, J. Yang, S. Lim, J. Lee, S. Lee, and S. Lim, “Emotion recognition using transformers with masked learning,” arXiv preprint arXiv:2403.13731, 2024.
[16] Y. Bi, A. Abrol, Z. Fu, and V. D. Calhoun, “A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data,” Hum. Brain Mapp., vol. 45, no. 17, p. e26783, 2024.
[17] S. Akinpelu, S. Viriri, and A. Adegun, “An enhanced speech emotion recognition using vision transformer,” Sci. Rep., vol. 14, no. 1, p. 13126, 2024.

[18] H. W. Loh et al., "GaborPDNet: Gabor transformation and deep neural network for Parkinson's disease detection using EEG signals," Electronics, vol. 10, no. 14, p. 1740, 2021.

[19] S. A. Khoshnevis and R. Sankar, "Classification of the stages of Parkinson's disease using novel higher-order statistical features of EEG signals," Neural Comput. Appl., vol. 33, pp. 7615–7627, 2021.

[20] A. Figueroa-Vargas, G. Valdebenito-Oyarzo, M. P. Martínez-Molina, F. Zamorano, and P. Billeke, "Probability decision-making task with ambiguity," OpenNeuro, Dataset, 2024, doi: 10.18112/openneuro.ds004917.v1.0.1.

[21] G. Valdebenito-Oyarzo, M. P. Martínez-Molina, P. Soto-Icaza, F. Zamorano, A. Figueroa-Vargas, J. Larraín-Valenzuela, X. Stecher, C. Salinas, J. Bastin, A. Valero-Cabré, R. Polania, and P. Billeke, "The parietal cortex has a causal role in ambiguity computations in humans," PLoS Biol., vol. 22, no. 1, p. e3002452, 2024.

[22] C. Stolz, A. Pickering, and E. M. Mueller, "Reward gain and punishment avoidance reversal learning," OpenNeuro, Dataset, 2022, doi: 10.18112/openneuro.ds004295.v1.0.0.

[23] C. D. Hassall, Y. Yan, and L. T. Hunt, "Continuous feedback processing," OpenNeuro, Dataset, 2022, doi: 10.18112/openneuro.ds004262.v1.0.0.

[24] A. Tversky and D. Kahneman, "Advances in prospect theory: cumulative representation of uncertainty," J. Risk Uncertain., vol. 5, no. 4, pp. 297–323, 1992.

[25] R. A. Rescorla and A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in Classical Conditioning II: Current Research and Theory, A. H. Black and W. F. Prokasy, Eds. New York, NY, USA: Appleton-Century-Crofts, 1972, pp. 64–99.

[26] N. D. Daw and K. Doya, "The computational neurobiology of learning and reward," Curr. Opin. Neurobiol., vol. 16, no. 2, pp. 199–204, 2006.

[27] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly, "By carrot or by stick: Cognitive reinforcement learning in Parkinsonism," Science, vol. 306, pp. 1940–1943, 2004.

[28] G. Lefebvre et al., "Behavioral and neural characterization of optimistic reinforcement learning," Nat. Hum. Behav., vol. 1, p. 0067, 2017.

[29] J. E. Mazur, "An adjusting procedure for studying delayed reinforcement," in Quantitative Analyses of Behavior: The Effect of Delay and of Intervening Events on Reinforcement Value, M. L. Commons, J. E. Mazur, N. A. Neef, and H. Rachlin, Eds. Hillsdale, NJ, USA: Erlbaum, 1987, pp. 55–73.

[30] J. Myerson and L. Green, "Discounting of delayed rewards: A life-span comparison," Psychol. Sci., vol. 6, no. 1, pp. 33–38, 1995.

[31] J. Flechsig, A. Petzold, et al., "Parameter estimation in non-linear models for delay discounting," Behav. Res. Methods, vol. 42, no. 4, pp. 934–944, 2010.

[32] H. Bleichrodt and J. L. Pinto, "A theory of hyperbolic discounting and risk," Manage. Sci., vol. 46, no. 8, pp. 1039–1050, 2000.

[33] P. Zelger, M. Arnold, S. Rossi, J. Seebacher, F. Muigg, S. Graf, and A. Rodríguez-Sánchez, "Beyond averaging: A transformer approach to decoding event-related brain potentials," NeuroImage, vol. 308, p. 121049, 2025.

[34] N. Kumar, K. Alam, and A. H. Siddiqi, "Wavelet transform for classification of EEG signal using SVM and ANN," Biomed. Pharmacol. J., vol. 10, no. 4, 2017.

[35] R. Zhao, T. Yue, Z. Xu, Y. Zhang, Y. Wu, Y. Bai, G. Ni, and D. Ming, "Electroencephalogram-based objective assessment of cognitive function level associated with age-related hearing loss," GeroScience, vol. 46, no. 1, pp. 431–446, 2024.

[36] D. Borra and E. Magosso, "Deep learning-based EEG analysis: Investigating P3 ERP components," J. Integr. Neurosci., vol. 20, no. 4, pp. 791–811, 2021.

[37] A. C. Teixeira-Santos, D. Pinal, D. R. Pereira, et al., "Probing the relationship between late endogenous ERP components with fluid intelligence in healthy older adults," Sci. Rep., vol. 10, p. 11167, 2020.

[38] F. Roux et al., "Deep learning convolutional neural networks discriminate adult EEG age," Front. Neurosci., vol. 14, p. 251, 2020.

[39] D. Bau, B. Zhou, A. Khosla, et al., "Network dissection: Quantifying interpretability of deep visual representations," in Proc. CVPR, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, pp. 770–778, 2016.

[41] N. Zhong et al., "Smaller feedback-related negativity (FRN) reflects the risky decision-making deficits of methamphetamine dependent individuals," Front. Psychiatry, vol. 11, p. 320, 2020, doi: 10.3389/fpsyt.2020.00320.

[42] J. Herding et al., "Centro-parietal EEG potentials index subjective evidence and confidence during perceptual decision making," NeuroImage, vol. 201, p. 116011, 2019.

[43] E. E. Barkley-Levenson et al., "Behavioral and neural correlates of loss aversion and risk avoidance in adolescents and adults," Dev. Cogn. Neurosci., vol. 3, pp. 72–83, 2013.

[44] W. H. Alexander, R. Fukunaga, P. Finn, and J. W. Brown, "Reward salience and risk aversion underlie differential ACC activity in substance dependence," NeuroImage Clin., vol. 8, pp. 59–71, 2015.

[45] J. H. Woo et al., "The PRO model accounts for the anterior cingulate cortex role in risky decision-making and monitoring," Cogn. Affect. Behav. Neurosci., vol. 22, no. 5, pp. 952–968, 2022.

[46] A. J. Rothman and P. Salovey, "Shaping perceptions to motivate healthy behavior: The role of message framing," Psychol. Bull., vol. 121, no. 1, pp. 3–19, 1997.

[47] E. A. Beam, Y. Masatlioglu, C. Watson, and K. Yang, "Loss aversion or lack of trust: Why does loss framing work to boost clinic visits?," Amer. Econ. J. Appl. Econ., 2023, doi: 10.1257/app.20210468.

[48] X. Giné, D. Karlan, and J. Zinman, "Put your money where your butt is: A commitment savings account for smoking cessation," Amer. Econ. J. Appl. Econ., vol. 2, no. 4, pp. 213–235, 2010.

[49] A. Haith, M. Haberstroh, et al., "Discounting and time preference in health behavior," Health Psychol. Rev., vol. 7, no. sup1, pp. S1–S28, 2013.

[50] W. H. Courtney and J. Polich, "The influence of reward sensitivity on health behavior," Health Psychol. Rev., vol. 9, no. 2, pp. 157–175, 2015.

[51] S. Gupta and C. M. Fetherston, "Financial incentives to promote adherence to healthy behaviors," Amer. J. Health Promot., vol. 30, no. 6, pp. 417–425, 2016.