

ERP-XTTN: Interpretable Cross-Subject Error-Related Potential Classification via Cross-Attention to Data-Driven ERP Prototypes

Charlotte Genevieve Wyman¹, Leanne Hirshfield¹

¹University of Colorado Boulder, Boulder, CO, United States

E-mail: charlotte.wyman@colorado.edu

ABSTRACT: Error-related potential (ErrP) classification methods for brain-computer interface deployment have separately addressed causal filtering, calibration-free generalization, and interpretability, but no existing approach combines all three. We propose ERP-XTTN, a cross-attention architecture using fixed difference-wave prototypes with temporal windows derived from training data. A peak-detection algorithm identifies ErrP components from the grand-average difference wave and sets window boundaries at smoothed zero-crossings. The prototypes serve as keys in a cross-attention mechanism with no value projection, so classification depends entirely on interpretable attention routing. We evaluate via leave-one-subject-out (LOSO) cross-validation on two datasets (BNCI Horizon 2020, 6 subjects; HRI Cursor, 11 subjects) with causal filtering and no hyperparameter tuning. ERP-XTTN achieves 0.782/0.836 AUROC (BNCI/HRI) using three midline channels and remains competitive with two. Attention maps reveal class-discriminative routing aligned with event-related potential (ERP) component timing on both datasets. These results show that data-driven prototypes can approach black-box performance while preserving interpretability grounded in ERP component timing, without sacrificing causal filtering or calibration-free generalization.

INTRODUCTION

Error-related potentials (ErrPs) are neural signals measurable via electroencephalography (EEG) that arise when a person observes or commits an erroneous action [1]. These potentials can provide brain-computer interfaces (BCIs) with implicit error-detection capability, enabling system correction without explicit user input.

Deployment-ready ErrP classification requires meeting several constraints at once. For online compatibility, filtering must be causal (forward-only), which can reduce signal quality relative to acausal filtering. We emphasize causal filtering because it is a necessary prerequisite that most offline evaluations omit: acausal filtering incorporates future samples, producing inflated accuracy estimates that cannot be reproduced in any forward-processing pipeline. Practical deployment also limits channel count (many commercial devices lack research-grade coverage [2]) and cannot depend on subject-specific calibration, which delays setup and limits scal-

ability. Finally, clinical and safety-critical applications demand that classification decisions be interpretable, not merely accurate.

Prior works have addressed each of these issues separately. Cross-subject causal ErrP classification has been demonstrated online using PCA-based dimensionality reduction and shrinkage linear discriminant analysis (LDA) with personalized decision thresholds under a full 61-channel montage [3]. This pipeline requires dedicated electrooculographic (EOG) channels with a calibration recording for artifact removal, and the PCA feature space is defined over the full electrode set, preventing direct application to reduced montages. Calibration-free cross-subject ErrP decoders have also been demonstrated using ensemble classifiers in a simulated online setting [4] and using generalized LDA models on one of the datasets used in this work [5], though without inherent interpretability or minimal-channel constraints. At reduced channel counts, Ancau et al. [6] demonstrated online single-channel ErrP decoding from a consumer-grade headset using deep learning, though without cross-subject evaluation or interpretability. Wimmer et al. [7] showed that combining minimal EEG montages with pupillometric signals can partially recover accuracy lost from channel reduction, though under acausal preprocessing and without interpretability. LOSO evaluation on BNCI has been demonstrated using convolutional neural network (CNN)-transformer architectures [8, 9], reaching 0.755 area under the receiver operating characteristic curve (AUROC) with two channels (FCz, Cz), but without inherent interpretability. Gradient-weighted Class Activation Mapping (Grad-CAM) has been applied to a neuroscience-informed CNN architecture for interpretable cross-subject error-related negativity (ERN) classification [10], demonstrating alignment between learned features and known ERP component timing, though under offline preprocessing and with post-hoc rather than inherent interpretability. Chan et al. [11] used post-hoc explainability methods to systematically reduce the EEG setup for asynchronous ErrP decoding under LOSO evaluation, retaining performance with as few as 10 channels, though the interpretability informs setup design rather than individual classification decisions. Prototype-based ERP classification has been explored for P300 in rapid serial visual presentation (RSVP) tasks [12], using learned prototypes matched via

Euclidean distance in a CNN feature space, but without attention-based routing or application to ErrPs. To the best of our knowledge, no prior work has met all these deployment constraints at once: causal filtering, cross-subject generalization without calibration, minimal channels, and inherent interpretability.

We propose a cross-attention architecture designed to satisfy all these constraints simultaneously. Fixed difference-wave prototypes are extracted automatically from training data via a constrained peak-detection procedure applied to the grand-average difference wave of the training subjects. Classification is based directly on attention weights, ensuring that all discriminative information flows through interpretable prototype-to-signal routing. We evaluate the model under LOSO cross-validation with causal infinite impulse response (IIR) filtering on two public benchmarks and show competitive performance on 2- and 3-channel montages.

To our knowledge, this is the first inherently interpretable ErrP classifier evaluated under LOSO with causal filtering on standard benchmarks. We show that this interpretability comes at an average cost of only 0.026 AU-ROC relative to EEGNet [13] in the 3-channel configuration, and that attention routing patterns suggest subject-specific differences in component-level routing across two datasets.

METHODS

Datasets: Two publicly available datasets (BNCI Horizon 2020 and HRI Cursor) were used to evaluate the proposed model.

BNCI Horizon 2020 is a 64-channel EEG dataset collected from 6 subjects in an observation-error paradigm [14]. Subjects were asked to monitor the performance of an agent moving a cursor toward a target; errors occurred on roughly 20% of trials. The dataset is publicly available at <https://bnci-horizon-2020.eu/database/data-sets>.

HRI Cursor is a 27-channel EEG dataset collected from 11 subjects during a cursor-control choice-reaction-time task [15]. Participants responded to one of three target stimuli via keypress, and feedback was presented as a cursor moving toward (congruent) or away from (incongruent) the target; errors occurred on roughly 35% of trials. Only the cursor scenario was used in this work; a companion robot-head-turn scenario from the same study was excluded. The dataset is publicly available at <https://github.com/stefan-ehrich/dataset-ErrP-HRI/>.

Preprocessing: A causal IIR 4th-order Butterworth 1-10Hz bandpass filter was applied to both datasets before epoching to 0-800ms after event onset. BNCI was downsampled from 512Hz to 256Hz after filtering. HRI was sampled at 256Hz and required no downsampling. Baseline correction, re-referencing, and spatial filtering were omitted to evaluate performance under constraints compatible with online, zero-calibration deployment.

Three channel sets were used: a 2-channel midline set (Fz, Cz), a 3-channel midline set (Fz, Cz, Pz), and the

full montage (64 for BNCI and 27 for HRI). The 2- and 3-channel configurations were selected based on the known frontocentral topography of ErrP components [1]. While FCz is often reported as the highest-amplitude ErrP channel [16], it is absent from many commercial and research-grade montages, including HRI. Selecting a channel set available across both datasets tests robustness to realistic montage variability and supports deployment across heterogeneous recording setups.

Architecture (ERP-XTTN): The proposed model, ERP-XTTN, classifies trial epochs by routing input EEG patches to a set of fixed ERP prototypes via query-key (QK)-only cross-attention, then classifying directly from the resulting attention-weight distribution (Figure 1).

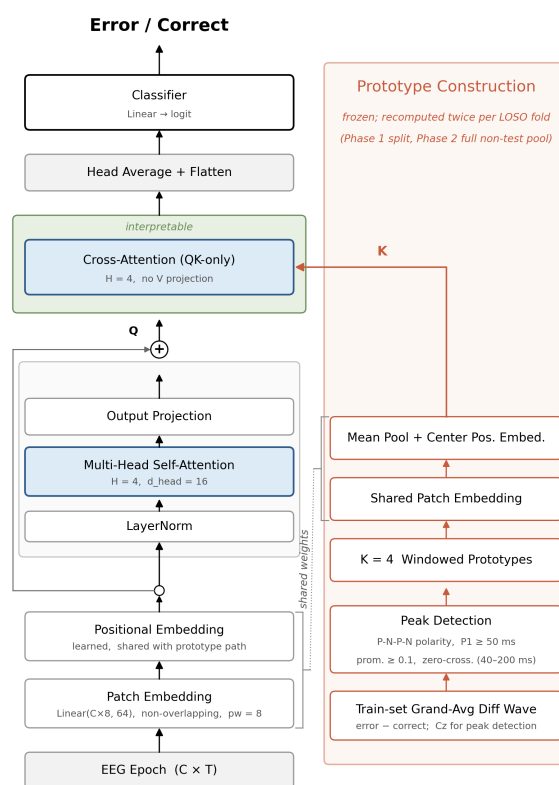


Figure 1: ERP-XTTN architecture. Input patches serve as queries (Q) into a QK-only cross-attention module against K=4 frozen ERP prototypes derived from the training set's grand-average difference wave via P-N-P-N peak detection. The attention-weight distribution over prototypes (the interpretable representation) is directly classified. Patch and positional embeddings (dotted line) are shared between the input and prototype paths. Prototypes are recomputed per LOSO fold.

Prototypes are derived from the grand-average training-set difference wave (mean error - mean correct) and, within each LOSO fold, are recomputed twice: first on the phase-1 training split used for validation-based epoch selection, and then on the full non-test training pool for phase-2 retraining at the selected epoch count. On the Cz channel of the smoothed difference wave (Gaussian kernel, $\sigma = 2$ samples), peaks are detected subject to a P-N-P-N polarity constraint reflecting the canonical ErrP morphology: an initial positivity (PI), the error negativ-

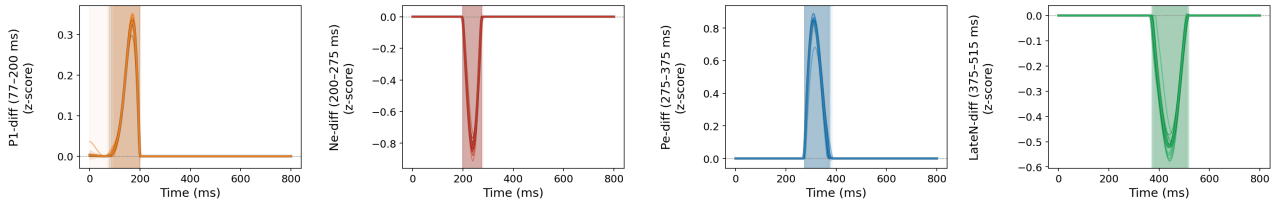


Figure 2: Difference-wave prototypes (Cz shown; HRI dataset; 3-channel configuration) derived from the grand-average error-minus-correct waveform. Each column shows one prototype: *P1* (77–200ms), *Ne* (200–275ms), *Pe* (275–375ms), and *LateN* (375–515ms). Shaded regions indicate prototype windows. Thin lines show individual LOSO fold prototypes ($n = 11$); thick lines show their mean.

ity (*Ne*), the error positivity (*Pe*), and a late negativity (*LateN*) [16]. All local extrema exceeding a prominence threshold of 0.1 are identified; *P1* is anchored as the earliest positive peak at least 50ms post-stimulus, and the remaining three components are selected as the N-P-N chain maximizing total prominence. Each peak is expanded to a window bounded by zero-crossings of the smoothed signal and clamped to 40–200ms width. The multichannel difference wave within each window forms one prototype; values outside the window are zeroed, yielding a prototype tensor of shape ($K=4$, C , T). Prototypes are frozen during training. An example set of resulting prototypes for the HRI 3-channel configuration is shown in Figure 2.

Input epochs (shape $C \text{ channels} \times T = 206$ samples) are divided into non-overlapping patches of width 8 samples, linearly projected to $d_{\text{model}} = 64$, and summed with a learned positional encoding ($N = 25$ patches). Dropout of 0.3 is applied after embedding. Prototypes pass through the same shared patch embedding, are mean-pooled across patch positions, and receive a positional encoding indexed at each prototype’s temporal center. Input patches are first passed through a single transformer-style self-attention layer. This layer applies pre-norm LayerNorm, multi-head self-attention ($H = 4$ heads, $d_{\text{head}} = 16$), output projection, and a residual connection, with dropout of 0.3 on the attention weights. Cross-attention is then computed between input patch queries and prototype keys using separate LayerNorms, linear projections to d_{model} , and multi-head scaled dot-product attention ($H = 4$), with logits clamped to $[-10, 10]$. Crucially, no value projection is used; instead, the softmax attention weights over the $K = 4$ prototypes are themselves the learned representation. This design constrains the model’s discriminative representation to prototype routing patterns, providing inherent interpretability without post-hoc analysis.

Finally, attention weights are averaged across heads, yielding a (B, N, K) tensor, flattened to $(B, 100)$, and passed through a single linear layer to produce a scalar logit. The full model has approximately 28,645 trainable parameters in the 3-channel configuration. Single-trial inference time is 0.94 ms (median) on CPU (AMD Ryzen 9 8945HS).

The model was trained with AdamW ($lr = 1 \times 10^{-3}$, weight decay $= 1 \times 10^{-4}$) using linear warmup for 5 epochs followed by cosine annealing over 100 epochs;

beyond 100 epochs, the learning rate was held at 10^{-5} . Gradient norms were clipped at 1.0. The loss was binary cross-entropy with a positive class weight equal to the inverse class frequency ratio ($n_{\text{correct}}/n_{\text{error}}$). Batch size was 32. Data augmentation consisted of temporal jitter (uniform ± 10 samples) and additive Gaussian noise ($x \leftarrow x + 0.1 \cdot N(0, 1)$ in normalized units), applied during training only.

Baseline (EEGNet): As a deep learning baseline for comparison, we used EEGNet with $F1=8$, $D=2$, and a kernel length of 128 samples, yielding 2,017 trainable parameters for the 3-channel configuration. We used the same preprocessing, training pipeline, hyperparameters, and evaluation protocol as for ERP-XTTN, including matched augmentation (temporal jitter and amplitude noise).

Evaluation: We evaluated both models using leave-one-subject-out (LOSO) cross-validation. Within each dataset, each subject served as the test set exactly once, with all remaining subjects forming the training set. To simulate zero-calibration deployment, no subject data from the held-out test subject was used during training. Within each LOSO fold, early stopping on a stratified validation split determined the optimal epoch count (patience 15, max 250 epochs, AUROC criterion). The final model was then retrained from scratch on all training and validation data for that epoch count.

The primary evaluation metric was AUROC, which is threshold-independent and robust to class imbalance. Balanced accuracy was computed at a fixed decision threshold of 0.5 on the sigmoid output, applied uniformly across all subjects and configurations. All reported results were obtained without hyperparameter optimization. Identical architectures, hyperparameters, and training procedures were used across both datasets and all channel configurations.

RESULTS

Table 1 summarizes AUROC for both architectures across all channel configurations, while Table 2 reports balanced accuracy (BA). At the 3-channel midline configuration, ERP-XTTN achieves 0.782 ± 0.058 AUROC on BNCI and 0.836 ± 0.107 on HRI, compared to 0.806 ± 0.077 and 0.864 ± 0.091 for EEGNet. At the 2-channel midline configuration, ERP-XTTN scores 0.775 ± 0.076 and 0.817 ± 0.115 , versus 0.777 ± 0.082 and

Table 1: AUROC (mean $\pm$ std, LOSO)

Dataset	Channels	ERP-XTTN	EEGNet
BNCI	2	.775 $\pm$.076	.777 $\pm$.082
BNCI	3	.782 $\pm$.058	.806 $\pm$.077
BNCI	64	.805 $\pm$.060	.842 $\pm$.071
HRI	2	.817 $\pm$.115	.839 $\pm$.105
HRI	3	.836 $\pm$.107	.864 $\pm$.091
HRI	27	.834 $\pm$.100	.884 $\pm$.073

Table 2: Balanced accuracy (mean $\pm$ std, LOSO)

Dataset	Channels	ERP-XTTN	EEGNet
BNCI	2	.698 $\pm$.062	.703 $\pm$.065
BNCI	3	.687 $\pm$.058	.722 $\pm$.064
BNCI	64	.719 $\pm$.056	.756 $\pm$.084
HRI	2	.739 $\pm$.102	.762 $\pm$.104
HRI	3	.759 $\pm$.100	.779 $\pm$.091
HRI	27	.747 $\pm$.089	.784 $\pm$.082

0.839 $\pm$ 0.105 for EEGNet. At full montage (64-channel BNCI, 27-channel HRI), ERP-XTTN reaches 0.805 $\pm$ 0.060 and 0.834 $\pm$ 0.100, compared to 0.842 $\pm$ 0.071 and 0.884 $\pm$ 0.073. BA patterns are consistent with AUROC: EEGNet leads across all configurations, and the gap between models follows the same channel-scaling trend.

The 3-channel midline configuration retained nearly all of the full-montage AUROC for ERP-XTTN. On HRI, 3-channel AUROC slightly exceeds full-channel performance (0.836 vs 0.834).

For the 3-channel midline configuration, ERP-XTTN outperformed EEGNet on 2 of 6 BNCI subjects, including subject 6 (0.702 vs 0.672 AUROC). On HRI, subject 4 showed near-ceiling performance for both models (0.980 and 0.986), while subject 10 had the lowest AUROC for both (0.546 and 0.636). Cross-subject AUROC standard deviation on BNCI was lower for ERP-XTTN than EEGNet (0.058 vs 0.077). Per-subject AUROC rank-order agreement between models was high (Spearman $\rho = 0.94$ on BNCI and $\rho = 0.93$ on HRI).

At the 3-channel midline configuration, attention routing differed systematically between error and correct trials. The most consistent pattern was reduced attention to the *Ne* prototype on error trials: mean *Ne* attention dropped from 0.466 to 0.340 on BNCI and from 0.503 to 0.337 on HRI, with compensatory increases distributed across the remaining prototypes. The strongest temporal contrast on BNCI occurred at ~ 359 ms, where *Ne* routing decreased (-0.190) and *Pe* routing increased ($+0.221$). On HRI, positive routing effects were distributed primarily between *PI* ($+0.234$ at 266ms) and *LateN* ($+0.234$ at 422ms), with a smaller *Pe* contribution.

Per-subject routing was heterogeneous but structured. On BNCI, *Pe* carried the dominant positive routing effect in 5/6 subjects, with subject 6 dominated by *PI* instead. On HRI, subjects split between *PI*-dominant (5/11) and *LateN*-dominant (5/11) routing, with one *Pe*-dominant subject. *Ne* routing remained negative (stronger for correct than error) in 5/6 BNCI and 10/11 HRI subjects. Routing contrast magnitude tracked with perfor-

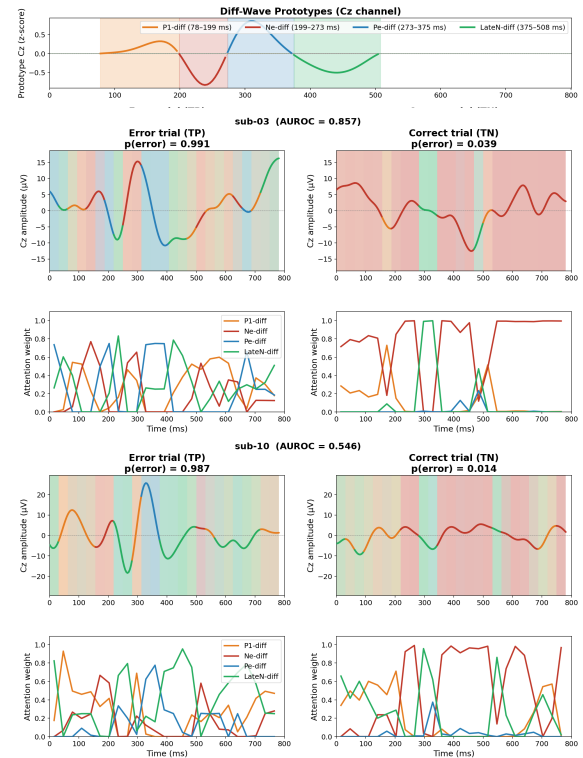


Figure 3: Difference-wave prototypes (top) and attention routing for high-confidence, true positive (error) and true negative (correct) trials from two HRI subjects in the 3-channel midline configuration. Both subjects show class-dependent routing, with *Ne*-dominant attention on correct trials and increased *Pe* and *LateN* engagement on error trials. Per-subject differences are visible in which prototypes are attended: for example, subject 3 shows *Ne*-dominant routing in the 250–300ms range on error trials, while subject 10 routes more strongly through *LateN* and *PI* in the same window. These routing profiles are read directly from the model’s attention weights without post-hoc analysis.

mance: the lowest-performing subjects showed the weakest error-vs-correct separation (BNCI subject 6: AUROC 0.702, max $|\Delta| = 0.208$; HRI subject 10: AUROC 0.546, max $|\Delta| = 0.129$), while the highest-performing subject showed the largest (HRI subject 4: AUROC 0.980, max $|\Delta| = 0.722$). The same patterns appeared at 2 channels and full montage, though *Ne*-dominant contrast was less stable. Figure 3 shows representative high-confidence error and correct trials from two HRI subjects at the 3-channel midline configuration, illustrating the class-dependent and subject-specific routing patterns described above.

DISCUSSION

The proposed ERP-XTTN classifier provides an inherently interpretable approach to deployment-compatible, cross-subject ErrP classification. With sub-millisecond CPU inference, it is compatible with deployment on modest hardware. Classification is based entirely on cross-attention routing to data-driven prototypes derived from canonical ErrP morphology, making the decision process

directly inspectable.

Figure 3 shows trial-level routing for two HRI subjects spanning a wide AUROC range. The *Ne*-dominant routing on correct trials, present at the group level on both datasets, may reflect the absence of the distinctive ErrP deflection sequence: without the *Pe* and *LateN* components that characterize error processing, the signal in the *Ne* window becomes the most informative template for identifying correct trials. This is consistent with the model exploiting known ERP component timing, though the alignment is not guaranteed by the architecture; the model could in principle learn to ignore the prototypes or route arbitrarily. Regardless of the underlying neurophysiological explanation, the routing patterns provide a concrete, trial-level account of each classification decision. This degree of transparency is difficult to obtain with black-box alternatives.

As Table 1 shows, the cost of this interpretability is modest: the mean gap between ERP-XTTN and EEGNet is 0.026 AUROC at the 3-channel midline configuration, similar across both datasets (0.024 BNCI, 0.028 HRI). The gap widens with channel count, from 0.002–0.022 at 2 channels to 0.024–0.027 at 3 channels and 0.037–0.050 at full montage. This pattern is expected: EEGNet’s depthwise-separable convolution learns spatial filters that benefit from additional channels, while ERP-XTTN’s fixed prototypes constrain spatial learning by design. The interpretability cost is therefore smallest precisely where deployment constraints are tightest. At 3 midline channels, ERP-XTTN largely matches its own full-channel AUROC, suggesting that additional channels may introduce noise without improving discrimination for this architecture.

BNCI subject 6 is the sole case where ERP-XTTN outperforms EEGNet (0.702 vs 0.672). This exception is notable given the otherwise high rank-order agreement between models on BNCI (Spearman $\rho = 0.94$). Subject 6 has the fewest error trials and the highest noise floor in the dataset, and Ren et al. [9] independently noted that this subject’s ERP morphology differed substantially from that of the remaining five participants. For such atypical subjects, fixed prototypes may provide a stronger inductive prior: when error trials are scarce, they regularize against overfitting, and when individual ERP morphology deviates from the population, they anchor classification to canonical component timing rather than to idiosyncratic patterns. ERP-XTTN also shows lower cross-subject variance on BNCI (0.058 vs 0.077), though $N=6$ limits the strength of this observation. If this pattern holds in larger samples, ERP-XTTN may offer an advantage precisely where it matters most: atypical or low-data subjects who are hardest to serve in zero-calibration settings.

The correlation between routing contrast magnitude and per-subject AUROC suggests that attention weights carry information beyond classification. In an online BCI context, routing profiles could help flag subjects whose ERP morphology deviates from the training population, in-

forming whether to trust the classifier output or initiate recalibration. This dual-use potential (classification and diagnostics from the same forward pass) is unavailable from black-box classifiers, even when augmented with post-hoc analysis. Jalilpour and Müller-Putz [10] achieve interpretability only through post-hoc Grad-CAM applied to a learned feature space, and Wei et al. [12] learn prototypes for P300 classification in a learned feature space using Euclidean distance with no attention transparency over the prototypes. In ERP-XTTN, the attention weights are themselves the explanation.

On the classification pipeline side, several prior works have addressed subsets of the constraints targeted here. Lopes-Dias et al. [3] demonstrated online cross-subject causal ErrP classification with causal filtering, PCA-based dimensionality reduction, and shrinkage LDA over a full 61-channel montage, with subspace subtraction requiring dedicated EOG channels and a pre-experiment calibration recording. Personalized decision thresholds were additionally updated during the first three experimental blocks. These montage and calibration requirements limit direct transfer to reduced or heterogeneous channel configurations. Ren et al. [8] evaluated a CNN-transformer architecture under LOSO on BNCI but used common average reference (CAR) re-referencing and non-causal finite impulse response (FIR) filtering across all 64 channels and included no interpretability mechanism. A follow-up study [9] reduced the input to 2 channels (FCz, Cz) with Butterworth filtering, achieving 0.755 and 0.737 AUROC on BNCI sessions 1 and 2 respectively under LOSO. ERP-XTTN achieves 0.775 AUROC at 2 channels (Fz, Cz), exceeding both per-session results despite stricter, deployment-oriented preprocessing constraints and the substitution of Fz for the typically higher-amplitude FCz. On HRI, Schönleitner et al. [5] reported 72.7% balanced accuracy with a generalized LDA using all 27 channels; ERP-XTTN achieves 75.9% balanced accuracy with three midline channels and inherent interpretability. Taken together with the interpretability distinctions from Jalilpour and Müller-Putz [10] and Wei et al. [12] discussed above, ERP-XTTN is, to our knowledge, the first ErrP classifier to combine LOSO cross-validation, causal filtering, no spatial filtering, and inherent interpretability on standard benchmarks, with competitive AUROC scores.

Limitations: Several limitations should be noted. The evaluation is limited to two datasets totaling 17 subjects, which limits the strength of the generalization claims. All results are based on offline replay; no online or real-time validation was performed. Causal filtering addresses only one prerequisite of online deployment; asynchronous detection without time-locking markers remains an open challenge. No hyperparameter tuning was conducted for either model, so reported performance may not reflect the full capacity of either architecture. Fixed prototypes are derived from grand-average difference waves and may not optimally capture individual variation in ERP morphology. No direct comparison with classical pipelines

(xDAWN+LDA) was included. Finally, balanced accuracy lags AUROC more for ERP-XTTN than for EEG-Net, suggesting that the fixed 0.5 decision threshold is suboptimal in the LOSO setting. In deployment, where a fixed threshold must be chosen without per-subject calibration, this gap warrants further investigation.

CONCLUSION

This work introduced ERP-XTTN, a cross-attention ErrP classifier that routes input EEG through fixed, data-derived prototypes grounded in canonical ErrP morphology. We evaluated it under strict deployment-relevant constraints across two public datasets and three channel configurations, demonstrating competitive performance with inherent per-trial interpretability.

Several directions remain for future work. Expanded evaluation across additional datasets (including different ERPs such as P300, N200, and feedback-related negativity), FIR filter comparison, and classical pipeline benchmarking (xDAWN+LDA) are planned for a journal extension. Online real-time validation is the most important next step for assessing deployment viability. Minimal subject-specific fine-tuning with a brief calibration session could further close the performance gap while retaining the interpretability benefits of the fixed-prototype architecture.

ERP-XTTN demonstrates that inherent interpretability in ErrP classification need not come at a substantial performance cost. With 3 midline channels, it achieves AUROCs of 0.782 on BNCI and 0.836 on HRI, within 0.026 of EEGNet on average, while providing direct, per-trial transparency through interpretable routing over data-driven ERP prototypes. These results hold under causal filtering, LOSO cross-validation, and no spatial filtering, satisfying the constraints that online, calibration-free BCI deployment demands.

Code is publicly available at <https://github.com/cgenevier/ERP-XTTN>; the version used to produce these results is archived on Zenodo [17].

REFERENCES

- [1] Chavarriaga R, Sobolewski A, Millán JDR. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*. 2014;8:208.
- [2] Sabio J, Williams NS, McArthur GM, Badcock NA. A scoping review on the use of consumer-grade EEG devices for research. *PLOS ONE*. 2024;19(3):e0291186.
- [3] Lopes-Dias C et al. Online asynchronous detection of error-related potentials in participants with a spinal cord injury using a generic classifier. *Journal of Neural Engineering*. 2021;18(4):046022.
- [4] Bhattacharyya S, Konar A, Tibarewala DN, Hayashibe M. A Generic Transferable EEG Decoder for Online Detection of Error Potential in Target Selection. *Frontiers in Neuroscience*. 2017;11:226.
- [5] Schönleitner FM, Otter L, Ehrlich SK, Cheng G. Calibration-Free Error-Related Potential Decoding With Adaptive Subject-Independent Models: A Comparative Study. *IEEE Transactions on Medical Robotics and Bionics*. 2020;2(3):399–409.
- [6] Ancau DM, Ancau M, Ancau M. Deep-learning online EEG decoding brain-computer interface using error-related potentials recorded with a consumer-grade headset. *Biomedical Physics & Engineering Express*. 2022;8(2):025006.
- [7] Wimmer M, Weidinger N, Veas E, Müller-Putz GR. Multimodal decoding of error processing in a virtual reality flight simulation. *Scientific Reports*. 2024;14(1):9221.
- [8] Ren G, Mahmoud SS, Kumar A, Fang Q, Yu B. A Transformer Encoder and Convolutional Neural Network Combined Method for Classification of Error-related Potentials. In: 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE: Toronto, ON, Canada, Oct. 2023, 1–5.
- [9] Ren G, Kumar A, Mahmoud SS, Fang Q. A deep neural network and transfer learning combined method for cross-task classification of error-related potentials. *Frontiers in Human Neuroscience*. 2024;18:1394107.
- [10] Jalilpour S, Müller-Putz G. A framework for Interpretable deep learning in cross-subject detection of event-related potentials. *Engineering Applications of Artificial Intelligence*. 2025;139:109642.
- [11] Chan HTJ, Wimmer M, Šimić I, Müller-Putz GR, Veas E. Informing EEG-Based Error Decoding With Explainable AI. In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE: Vienna, Austria, Oct. 5, 2025, 4550–4557.
- [12] Wei W, Qiu S, Zhang Y, Mao J, He H. ERP prototypical matching net: A meta-learning method for zero-calibration RSVP-based image retrieval. *Journal of Neural Engineering*. 2022;19(2):026028.
- [13] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*. 2018;15(5):056013.
- [14] Chavarriaga R, Millán JDR. Learning From EEG Error-Related Potentials in Noninvasive Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2010;18(4):381–388.
- [15] Ehrlich SK, Cheng G. A Feasibility Study for Validating Robot Actions Using EEG-Based Error-Related Potentials. *International Journal of Social Robotics*. 2019;11(2):271–283.
- [16] Ferrez PW, Millán JDR. Error-Related EEG Potentials Generated During Simulated Brain-Computer Interaction. *IEEE Transactions on Biomedical Engineering*. 2008;55(3):923–929.
- [17] Wyman CG. *ERP-XTTN: Interpretable cross-attention ERP classifier*. Version v1.0.0. 2026. [Online]. Available: <https://doi.org/10.5281/zenodo.20087551>.