An EEG Conformer Model for Error Feedback During Human-Robot Interaction

Jinpei Han, Yinxuan Li, Xiao Gu, A. Aldo Faisal

***Abstract*— Identifying a brain signal that enables the detection of incorrect execution in human-robot interaction (HRI) is considered a holy grail for real-time systems. A major challenge in achieving this is the inherent imbalance caused by the sparsity of error-related potential (ErrP) events in streaming electroencephalogram (EEG) data, which often leads models to learn irrelevant features and perform poorly. Thus, while deep learning-based ErrP detection has seen considerable advancements, the variability in individual user reaction times introduces labeling errors, complicating model adaptation to new subjects. Moreover, most deep learning methods are developed and validated on discrete, offline experiments using pre-defined windows, which fail to translate effectively to continuous, real-time HRI. Addressing these challenges is crucial to improving the robustness and adaptability of real-time ErrP detection in practical HRI applications. Here, we develop a causal EEG conformer framework, combining a convolutional neural network (CNN) encoder and a transformer with causal attention for real-time prediction of ErrP signals during HRI. We evaluated our ErrP model in a pseudo-online environment in both inter-session and inter-subject cross-validation settings for exoskeleton assistive robotics. Our model demonstrated superior performance in decoding accuracy and efficiency, showcasing better generalization for real-world dynamic HRI applications.**

I. Introduction

Human-robot interaction (HRI) plays an important role in rehabilitation and assistive robotics for patients with spinal cord injuries, neurodegenerative diseases, or strokes. By enabling patients to regain lost motor functions or adapt to new ways of performing daily activities, effective HRI can significantly enhance their quality of life. A critical aspect of successful HRI is accurately recognizing the intention of patients, which is essential for the control of robotic assistive devices [1]. To facilitate this, as well as to promote more effective interaction between humans and robots, human-in-the-loop learning becomes highly important [2]. As illustrated in Figure 1, this approach uses the continuous feedback loop between the human user and the robotic system, allowing the robot to adapt and improve its responses based on the user's intentions and reactions.

Jinpei Han is with the Brain & Behaviour Lab, Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom. (email:j.han20@imperial.ac.uk). Yinxuan Li is with the Brain & Behaviour Lab, Department of Bioengineering, Imperial College London, London SW7 2AZ, United Kingdom. (email:yinxuan.li21@imperial.ac.uk). Xiao Gu is with the Computational Health Informatics Lab, University of Oxford, Oxford OX3 7DQ, United Kingdom. (e-mail: xiao.gu@eng.ox.ac.uk). A. Aldo Faisal is with the Brain & Behaviour Lab, Department of Computing and Bioengineering, Imperial College London, London SW7 2AZ, United Kingdom. Chair in Digital Health & Data Science, University of Bayreuth, Bayreuth, Germany. (email: aldo.faisal@imperial.ac.uk). We acknowledge funding from the UKRI Turing AI Fellowship to AAF (EP/V025449/1).

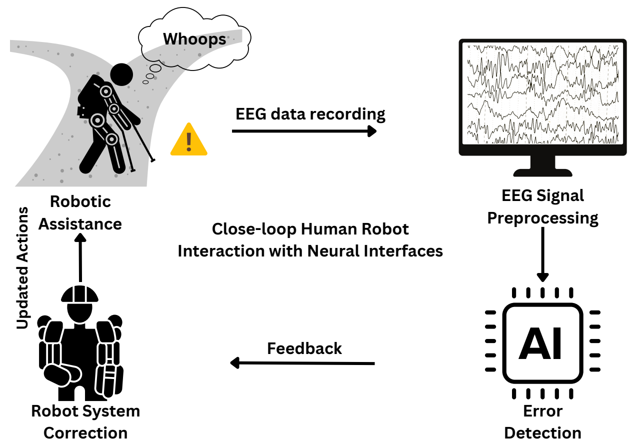

Fig. 1: Illustration of Close-loop Assistive Robotic System with Auto Error Detection.

Within HRI in assistive robotics, real-time error recognition is essential since errors made by the robot can result in undesired actions that the user neither intends nor wishes to perform. Such errors can lead to frustration, decreased trust in the robotic system, and, in severe cases, potential harm to the patient [3]. By detecting these errors, the robot can immediately correct its course, thereby improving safety, effectiveness, and user satisfaction.

Error recognition in conventional HRI systems involves monitoring discrepancies between expected and actual outcomes based on the robot's actions and user feedback. These systems often rely on external sensors, such as cameras [4], eye tracking [5], or voice recognition [6] to detect errors. They are particularly effective in controlled environments where the robot can accurately interpret a limited set of user actions or gestures. However, the reliance on explicit commands or pre-defined gestures can limit the naturalness of the interaction [7]. As a result, these methods can struggle with real-time error detection, particularly in dynamic or unstructured real-world situations where user intentions and feedback are not straightforwardly interpretable [8].

Brain-computer interfaces (BCI) offer promising methods for detecting such errors through the identification of error-related potentials (ErrPs) in the electroencephalogram (EEG) signals of the user. BCI systems with EEG enable direct access to the patient's brain activities in real time without pre-defined responses. ErrPs are specific neural responses that occur when a person recognizes an error, whether it is their own or that of an external system, such as the assistive robotic system. Detecting ErrPs from EEG signals offers a natural and direct way of identifying errors during HRI without conscious effort from the user, making it a seamless

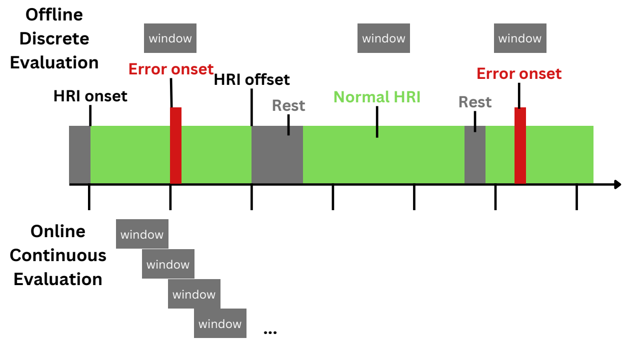

Fig. 2: Illustration of Offline Discrete Evaluation vs. Online Continuous Evaluation.

and less intrusive method of communication [9].

Earlier works employed domain features of the error-related negativity (ERN), such as the power of theta, beta, and gamma frequency bands of the EEG signal to train machine learning (ML) models for ErrPs detection [10]. However, the non-stationary, limited spatial resolution and low signal-to-noise ratio (SNR) of the EEG signal make it challenging to generalize to new subjects or environments [11]. More recently, advances in deep learning have enabled automatic feature extraction and prediction from EEG signals without specific domain knowledge. Deep learning approaches based on convolutional neural network (CNN) [12], [13], long short term memory (LSTM) [14] and graph neural network (GNN) [15] have outperformed conventional classifiers [16], on multiple tasks such as motor imagery classification and emotion recognition.

Despite these advancements, deep learning-based real-time ErrP detection presents inherent and significant challenges that have been overlooked in the existing work. First of all, given the nature of real-time detection from streaming EEG recordings, there is a substantial sparsity of error events, leading to severe imbalance. This would lead conventional deep learning models to learn a lot of task-irrelevant features, which hamper their performance [17]. In addition, variability in reaction times across different individuals can introduce labeling errors within the dataset, making it difficult for the models to adapt to new subjects. Furthermore, previous deep learning approaches were developed and validated on isolated events within pre-defined time windows, where each event is treated independently and the ErrP is assumed to occur at the center of the windows, as shown in Figure 2. While this discrete evaluation setting is effective in controlled, offline experiments, it struggles to generalize to online HRI settings where continuous, real-time detection is required [18]. For instance, despite the high offline accuracy achieved by the Ensemble model proposed by [19], a high false positive (FP) rate is observed in the pseudo-online setting due to the variable position of the ErrP within the window and the lack of long-range temporal information. These challenges make it difficult for existing deep learning models to train and generalize in real-world HRI applications effectively.

In this work, we dedicate our efforts to addressing the above challenges to develop effective real-time ErrP detection during HRI in the real world. Our contributions are follows:

A Causal EEG Conformer Model for ErrP Detection. We introduced an EEG Conformer model that leverages CNNs to learn frequency features from small EEG windows and transformer layers to capture long-term dependencies between these windows. The causal masking in the transformer layer ensures that attention calculations are unidirectional, preventing the use of future windows and making the model suitable for real-time online applications.

Two-stage Discrete Pre-training and Sequential Learning Framework. We proposed a two-stage training framework where the CNN module is first pre-trained on a set of balanced, discrete windows to learn meaningful features, followed by fine-tuning on continuous sequences to capture long-term dependencies.

Imbalance Strategies for Sparse ErrP Events. Throughout the two stages, we introduced a series of strategies, including resampling, label imbalance loss, and label smoothing strategy, to tackle the inherent label imbalance issue. These methods have proven to be critical and effective solutions for training real-time ErrP detection models.

Based on the integration of the above, we conducted extensive experiments in pseudo-online settings to validate our model's performance and compared it against existing methods in the literature. Extensive experiments demonstrated that these strategies significantly improved model robustness, reducing FPs and achieving higher balanced accuracy across inter-subject and inter-session tasks.

II. Related Work

A. Error-Related Potentials in EEG Signals

ErrPs are specific patterns of neural activity that occur when an individual detects an error in their own actions or the actions of an external system, such as a robot. According to cognitive neuroscience research [20], [21], ErrP consists of error-related negativity (ERN), which typically arises within 100 ms after an error is perceived, and positive deflection known as error positivity (Pe), occurs around 200 to 500 ms after the error.

ErrPs are primarily characterized by analyzing the theta (4–8 Hz), beta (18–30 Hz), and gamma (60–90 Hz) frequency bands [10]. Earlier works [22] used traditional signal processing methods to extract the changes in EEG signal frequency bands and train a classifier. However, challenges such as inter-subject variability, noise in EEG data, and the need for real-time processing continue to pose significant hurdles in this field [17], [23].

B. Deep Learning for EEG Decoding

Advances in deep learning have enabled automatic feature extraction and prediction from EEG signals without specific domain knowledge. Varied deep learning architectures, such as CNN [12], [13] and GNN [15], have outperformed conventional classifiers [16] on multiple classification tasks, including motor imagery classification and emotion recognition. Sequential models like LSTM [14] and transformer [24] have also proven effective in capturing long-term dependencies in EEG monitoring tasks such as sleep staging and clinical diagnosis.

These models have also been applied specifically for discrete ErrP classification. For instance, CNN [25], [26] and GNN [27] have shown improved accuracy over traditional approaches for ErrP classification. However, they primarily focus on discrete ErrP classification for isolated EEG trials rather than continuous prediction over a longer duration. Furthermore, in real-world settings, since robots constantly move and interact with the robotic system, it presents additional challenges for robust ErrP detection to handle such dynamics. Additionally, most validation datasets for these models are collected in controlled environments where there are a balanced amount of error and non-error trials, which is not the case for in-the-wild real-time continuous recordings.

C. Error Feedback during HRI

In assistive robotics, timely and accurate error feedback is essential for ensuring safety, efficiency, and user satisfaction. Effective error detection allows robots to adjust their behavior in real time, preventing potential harm to users. Cameras have been utilized to detect errors by analyzing head and shoulder movements [4] or facial expressions [28]. These vision-based methods can be improved by integrating other modalities, such as voice recognition [5], [6], which enhances the robot's ability to interpret user reactions and detect errors more accurately.

However, relying on explicit feedback from users can be limiting, as it may not always be consistently expressed or easily observable. Researchers [9], [29] have explored using ErrPs from EEG signals as implicit feedback, leveraging the brain's cognitive responses during interactions. While promising, these studies collected EEG signals from participants when they were observing the robot rather than during physical interactions, where muscle movement and task complexity introduce artifacts, complicating ErrP detection.

III. METHODS

A. Dataset Description

We used the IJCAI'23 CC6 Competition HRI dataset provided by [30]. The complete dataset consists of EEG and EMG recordings from eight healthy subjects who participated in an experiment involving the detection of errors introduced by an active orthosis device. However, we only used the EEG data in our experiment for our brain-machine interface system.

The EEG data were recorded using a 64-channel 10-20 EEG system (LiveAmp64 from Brain Products GmbH) at 500 Hz. The Participants were fitted with an active orthosis on their right arm and were instructed to detect and respond to errors in the orthosis movement by squeezing an air-filled ball in their left hand. The experiments consist of 10 runs of orthosis movement, each with 30 trials, including six randomly introduced errors of opposite directions per run. The annotations include the time point of the start, end, and the error onset (if any) of each interaction.

We first re-referenced the signal to the global average and down-sampled the data to 100Hz. A band-pass filter between 0.1 and 30 with Firwin filters was then applied to the signal to filter out as many motion artefacts as possible, followed by z-score normalization.

B. Problem Definition

Considering the ErrP EEG classification task, we denote the input as $\boldsymbol{x} \in \mathbb{R}^{C \times T}$, where C represents the number of EEG channels, and T represents the time length of the EEG segment. The label is denoted as $y_t \in \{0, 1\}$, where $y_t = 0$ represents a normal EEG segment at time t, and $y_t = 1$ indicates an error-related EEG segment.

Prior experimental settings in this context have several limitations, hindering their applicability in the real world.

1. Conventional Data Collection and Usage. Most previous studies utilize data collected in controlled lab environments, where an equal amount of positive (error) and negative (normal) samples are available. In these settings, participants observe the robot making mistakes without physically interacting with it, resulting in cleaner EEG data with fewer noise artifacts. This setup simplifies the classification task but does not fully capture the complexity of real-world HRI scenarios.

2. Offline Validation. Previous studies often rely on offline validation methods, classifying discrete windows of EEG data without considering the temporal dependencies in the user's brain signals. These approaches may inadvertently utilize clues after the ErrPs signals for prediction, such as muscle movements that occur when the participant squeezes the ball for feedback. These methods, while effective for controlled experiments, may fail in real-time HRI applications where continuous data is processed.

To improve real-world applicability, we followed the Pseudo-Online validation approach suggested by [18]. The Pseudo-Online validation approach ensures that the model cannot rely on the participant's future response clues (such as muscle movements) for error prediction.

C. Overview of Causal EEG Conformer Model

Our Causal EEG Conformer model, illustrated in Figure 3, consists of several key components designed to handle sequential EEG data for ErrP classification. The model begins with a CNN encoder block E, which uses convolutional layers to extract features using a sliding window approach along the entire EEG sequence. The input sequence $\boldsymbol{x} \in \mathbb{R}^{C \times T}$ is divided into smaller windows of size w with an overlap of v between consecutive windows. This results in $s = \left\lfloor \frac{T-v}{w-v} \right\rfloor$ windows, each of size w. The CNN encoder E is applied to each window $\boldsymbol{x}_i \in \mathbb{R}^{C \times w}$, where $i \in [1, s]$, to extract features from each segment of the sequence. This process yields a sequence of tokenized feature vectors $\boldsymbol{h}_i = E(\boldsymbol{x}_i) \in \mathbb{R}^d$, where d is the dimension of the extracted feature vector. These tokenized feature vectors are then concatenated into a sequence $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_s] \in \mathbb{R}^{s \times d}$.

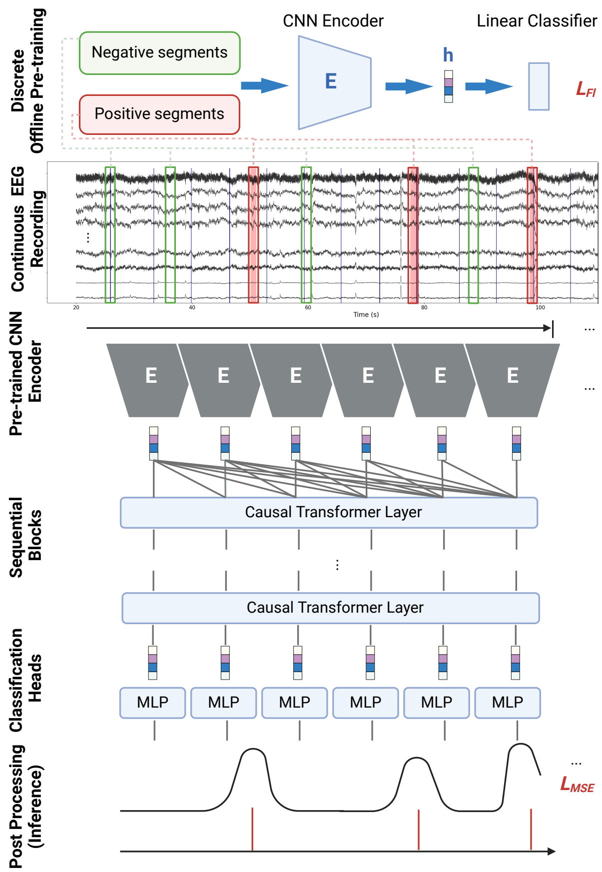

Fig. 3: **An illustration of our proposed Causal EEG Conformer Model.** The framework starts with discrete CNN encoder pre-training, where positive segments were sampled using the error onsets. During continuous decoding, the frozen CNN processes overlapping EEG windows to extract tokenized features, which are concatenated and encoded with positional embeddings. These are passed through transformer blocks with causal attention and a classification head.

The tokenized feature sequence $\boldsymbol{H}$ is then encoded with positional embeddings $\mathbf{p} \in \mathbb{R}^{s \times d}$ to ensure the temporal order of the sequence is preserved, allowing the model to understand the progression of events over time. The sequence is then passed into transformer blocks with causal attention mechanisms. Causal attention ensures that the features at time i only attend to the historical features from previous time steps $i' < i$ and not any future information, making the model suitable for real-time applications. Finally, the output sequence from the transformer blocks is fed into a linear classification head, which performs ErrP classification at each time point along the continuous EEG sequence. The resulting prediction sequence is then post-processed to pinpoint the error onset.

To effectively guide the model in performing real-time ErrP detection, we introduce two-stage training strategies to first conduct discrete pre-training of the CNN encoder and subsequently perform sequential downstream training for the causal transformer.

D. *Discrete Pre-training*

Due to the sparsity of the error signal within the data sequence, the feature tokenizer needs to be pre-trained to ensure it can encode meaningful information. Existing deep learning models, including transformers, often fail to handle sparse features effectively [31]. Sparse features, such as ErrP signals with infrequent occurrences, can impede the model's ability to learn meaningful representations.

Neuroscience research has shown that ErrP signals typically occur around 100-500 milliseconds after an error is perceived. Therefore, we used $w = 100$ which is a 1-second window following the error onset as positive samples. For negative samples, we randomly selected non-overlapping 1-second windows from the rest of the data, ensuring that these windows do not contain error-related potentials.

To address the imbalance between positive (error) and negative (normal) samples, we oversampled the positive samples and undersampled the negative ones to create balanced batches during training. The CNN encoder E was pre-trained to minimize focal loss, which is particularly effective for handling imbalanced datasets. The focal loss is defined as: $\mathcal{L}_{\text{fl}}(p) = -\alpha(1-p)^{\gamma}\log(p)$, where p is the predicted probability of the linear classification layer of h for the correct class, α is a weighting factor to balance the positive and negative samples, and γ is the focusing parameter, which adjusts the loss contribution for well-classified samples. By employing focal loss, we reduce the impact of easy-to-classify examples and ensure that the CNN encoder E learns to focus on difficult-to-detect error signals, improving its ability to encode meaningful features during pre-training. This discrete pre-training phase is critical for preparing the feature tokenizer to handle the sparsity of ErrP in the continuous data sequences in the next stage.

E. *Sequential Downstream Training*

Directly applying a discrete model with a moving window on sequential EEG data can result in poor performance due to the loss of historical brain signal information and varying reaction times across individuals. Each person may experience different amounts of delay after an error onset, which can introduce labeling errors into the dataset. This variation in reaction times leads to misaligned labels and reduces the accuracy of error detection. To address this issue, we applied label smoothing to account for the uncertainty and variability in error onset timing.

Given the label sequence as $\boldsymbol{y} \in \{0, 1\}^T$, a sequence of length T, the positive class (1) is only assigned to a sparse segment immediately following the error onset. However, this sharp and binary labeling is problematic, as the exact moment of error detection may vary. To smooth these labels, we repeatedly convolved the original label vector with a skewed Gaussian kernel, which is defined as:

$$\hat{\boldsymbol{y}}_k = \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(k-t)^2}{2\sigma^2}\right) \Phi\left(\alpha\frac{k-t}{\sigma}\right) y_t \quad (1)$$

where $\hat{\boldsymbol{y}}_k$ is the smoothed label at time k, σ is the standard deviation of the Gaussian kernel, α is the skewness factor of $\Phi(\cdot)$, and y_t is the original label at time t. The skewed Gaussian kernel has a longer tail on the right to allow a gradually decreasing probability for slower reaction times. This convolution ensures that the label vector transitions smoothly, replacing sharp binary values $\boldsymbol{y_t} \in \{0,1\}$ with continuous values $\hat{\boldsymbol{y}}_{\boldsymbol{t}} \in [0,1]$.

Then, both the input sequence $\boldsymbol{x} \in \mathbb{R}^{C\times T}$ and the smoothed label sequence $\hat{\boldsymbol{y}} \in [0,1]^T$ are divided into smaller windows of size w with an overlap of v between consecutive windows. This results in $s = \left\lfloor \frac{T-v}{w-v} \right\rfloor$ windows, each of size w. The pre-trained CNN encoder E processes each window $\boldsymbol{x}_I \in \mathbb{R}^{C\times w}$, where $I \in [1,s]$, to extract features from each segment of the sequence. This process yields a sequence of tokenized feature vectors $\boldsymbol{h}_i = E(\boldsymbol{x}_i) \in \mathbb{R}^d$, where d is the dimension of the extracted feature. These tokenized feature vectors are then concatenated into a sequence $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_s] \in \mathbb{R}^{s\times d}$.

During the Sequential Training phase, the CNN encoder E is frozen. The causal transformer block and the multi-layer perceptron (MLP) classification heads are trained to minimize the Mean Square Error (MSE) compared to the smoothed label sequence $\hat{\boldsymbol{Y}} \in [0,1]^{s\times w}$. This approach helps mitigate the effects of varying reaction times and labeling errors, allowing the model to learn more robust patterns in the EEG data. During inference, the prediction sequence is thresholded at 0.5 to convert into binary predictions of error during HRI.

F. Implementation and Validation Protocol

We followed the pseudo-online validation approach proposed by [18] and calculated the balanced accuracy (Acc), F1 score, and normalized Matthews Correlation Coefficient (nMCC) as follows:

$$\begin{aligned}
\text{MCC} &= \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},\\
\text{Balanced Accuracy} &= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right),\\
\text{F1 score} &= \frac{2\times TP}{2\times TP+FP+FN}
\end{aligned} \tag{2}$$

We implemented our model using PyTorch with Titan Xp. For training, we applied Adam as the optimizer, with a learning rate of 1e-3, betas as (0.9,0.999), and weight decay of 1e-4. The batch size was set as 32. We trained the downstream sequential model using a window size of $w = 100$ and an overlap of $v = 90$ on 6-second data segments, $T = 600$, after each interaction onset. The models were evaluated on unseen EEG input streams, with lengths ranging from 184.62s to 230.05s.

For baseline comparisons, we used the EEGNet [12], ResNet [23], MLP, logistic regression (LR), support vector machine (SVM), random forest (RF), and XGBoost bundled as an Ensemble model proposed by the winning team [19]. These discrete models were tested using the moving window approach on the test EEG sequence. We also used baseline models that combine EEGNet with a gated recurrent unit (GRU) [32] and LSTM [33] as backbone comparisons following our pre-training and sequential learning pipeline.

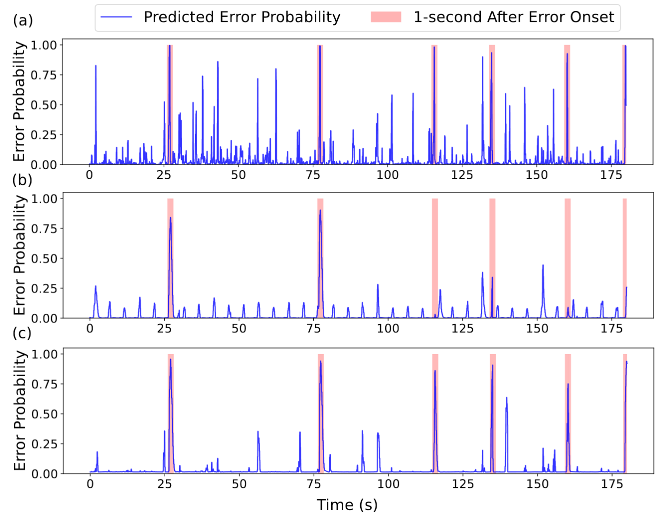

Fig. 4: **Examples of Pseudo-online Model Prediction Results.** (a) Ensemble Model (b) EEGNet-GRU Model (c) Our EEGNet-Transformer Model

To evaluate the generalization capability of different methods, we first tested both the inter-session 10-fold cross-validation by splitting the dataset into training and testing sets based on the runs. We also conducted Leave-One-Subject-Out (LOSO) cross-validations where all the runs of each unique subject are left out as a test set in each fold.

IV. Results and Discussion

A. Pseudo-Online ErrP Prediction Results

We present the inter-session 10-fold cross-validation and LOSO with the averaged balanced accuracies, F1 scores, and nMCC scores in Table I. Overall, the models with sequential layers achieved higher performance than the discrete models for pseudo-online prediction using moving windows. This shows that the historical brain response is essential for the online ErrP prediction within new incoming data. In particular, our proposed model, which combines a pre-trained CNN encoder with a transformer block, achieved the highest average performance in both inter-session and inter-subject validation tasks. This result demonstrated the transformer block's strong ability to capture the sparse key feature dependencies along the input context sequence. The Ensemble model with majority voting achieved the lowest balanced accuracy on the LOSO task due to the inter-subject variations in EEG patterns and reaction times. The covariance matrix features used to train classical machine learning methods could not be generalized to unseen users or unseen segments of the HRI.

We also present a set of examples for the predicted error probability in pseudo-online decoding tasks in Figure. 4. The red vertical bars represent the 1-second window after each error onset, which we treated as the ground-truth duration for ErrPs. The blue lines represent the predicted probability of each model in real time. In this example, although the Ensemble model accurately predicted all six errors, it has a high FP rate when thresholded at 0.5 as shown in Figure 4(a). This finding is also consistent with the results reported by

TABLE I: The Inter-Session and Leave-One-Subject-Out Cross-Validation Results on Pseudo-Online EEG Stream

Method	**Mode**	**Inter-Session**			**Inter-Subject**		
		Acc (%)	F1 (%)	nMCC (%)	Acc (%)	F1 (%)	nMCC (%)
EEGNet	Discrete	59.88 ± 1.46	23.84 ± 2.97	59.76 ± 1.73	57.92 ± 3.83	21.39 ± 7.24	59.33 ± 3.86
ResNet	Discrete	60.16 ± 1.02	29.02 ± 3.13	64.51 ± 2.22	58.16 ± 3.58	22.98 ± 8.58	60.78 ± 4.79
Ensemble*	Discrete	55.28 ± 0.68	17.79 ± 1.87	60.49 ± 1.29	54.28 ± 1.64	14.46 ± 4.51	58.60 ± 2.82
EEGNet-GRU	Continuous	61.12 ± 1.91	31.37 ± 4.63	65.86 ± 2.74	62.64 ± 6.44	33.20 ± 14.48	65.98 ± 7.69
EEGNet-LSTM	Continuous	60.68 ± 3.07	30.70 ± 3.07	65.49 ± 3.07	61.30 ± 4.37	31.20 ± 11.54	65.38 ± 6.94
ResNet-GRU	Continuous	60.76 ± 4.09	30.92 ± 9.58	66.36 ± 4.89	62.41 ± 4.96	34.49 ± 12.09	67.44 ± 6.93
ResNet-LSTM	Continuous	56.52 ± 2.24	20.58 ± 5.92	61.55 ± 3.16	58.72 ± 2.94	26.39 ± 7.63	64.20 ± 4.40
EEGNet-Transformer	Continuous	**72.79 ± 1.95**	**54.38 ± 5.11**	**76.87 ± 2.98**	<u>70.92 ± 2.50</u>	<u>48.47 ± 7.49</u>	<u>73.44 ± 4.28</u>
ResNet-Transformer	Continuous	<u>68.32 ± 3.72</u>	<u>46.86 ± 8.25</u>	<u>73.78 ± 4.21</u>	**74.14 ± 3.83**	**56.35 ± 8.78**	**77.85 ± 4.66**

*The Ensemble Model consists of a customized ResNet, MLP, LR, SVM, RF, XGBoost, and grid-search-optimized XGBoost [19].

Note: The best results are in **bold**, and the second-best results are <u>underlined</u>. All with Balanced Loader & Weighted Loss.

the winning team [19]. However, if the number of errors is unknown, as in a real-world scenario, the model may perform poorly. On the other hand, the EEGNet-GRU model has lower sensitivity, failing to identify four out of six errors as shown in Figure 4(b). Finally, Figure 4(c) demonstrates that our proposed EEGNet-Transformer model achieved the best pseudo-online prediction performance, correctly identifying all six errors with only one FP. The model effectively suppressed FPs during normal data segments, highlighting its superior ability to capture long-term dependencies in the sequence.

B. Ablation study

We further performed ablation studies to validate the effectiveness of each training design of our proposed method, with the results shown in Table II. The results showed that all components contributed to improving performance. In particular, the Skewed Gaussian Smoothing method, which converted the training objective from classification to regression, showed the highest performance increase, demonstrating its effectiveness in addressing data imbalance and labeling error issues.

TABLE II: Results of Ablation Study

	Ablation $\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$	**Acc (%)**	**F1 (%)**	**nMCC (%)**
1	✓					57.40 ± 1.09	23.43 ± 3.10	63.23 ± 1.70
2	✓	✓				59.88 ± 1.46	23.84 ± 2.97	59.76 ± 1.73
3	✓	✓	✓			66.47 ± 3.87	31.69 ± 9.03	64.31 ± 4.99
4	✓	✓	✓	✓		72.79 ± 1.95	54.38 ± 5.11	76.87 ± 2.98
5	✓	✓	✓	✓	✓	73.48 ± 1.76	55.62 ± 3.96	77.52 ± 2.26

$\mathcal{A}$: + Discrete CNN Encoder.
$\mathcal{B}$: + Balanced Loader & Weighted Loss.
$\mathcal{C}$: + Transformer blocks.
$\mathcal{D}$: + Skewed Gaussian Smoothing [$\alpha = 1$, $\sigma = 3$].
$\mathcal{E}$: + Fine-tunning of CNN Encoder.

C. Online Decoding Efficiency

We also compared the inference time of different models with different step sizes for 60s input sequences. As shown in Table III, our model achieved a faster decoding time compared to the Ensemble method and the ResNet-GRU model.

TABLE III: Decoding Efficiency for 60s Input Sequences with Different Step Sizes

Method	Decoding time for 60s input sequences (ms) step size 10	step size 20	step size 50
EEGNet	1.17 ± 0.03	1.19 ± 0.16	0.99 ± 0.25
ResNet	1.61 ± 0.23	1.69 ± 0.25	1.63 ± 0.57
Ensemble	1461.63 ± 56.03	561.11 ± 12.45	316.18 ± 13.90
EEGNet-GRU	1.74 ± 0.47	1.52 ± 0.20	1.49 ± 0.17
ResNet-GRU	54.52 ± 1.70	27.26 ± 0.47	12.19 ± 0.60
EEGNet-Transformer	8.64 ± 0.57	6.85 ± 0.75	7.07 ± 2.18
ResNet-Transformer	12.42 ± 0.43	8.77 ± 0.84	6.77 ± 0.75

D. Future Work

This study was conducted in a pseudo-online setting, and its effectiveness in real-world applications remains to be validated. To address this limitation, We are currently conducting experiments to deploy our proposed model into an assistive robotic system developed by our lab, which has demonstrated success in assisting users with reaching and grasping tasks [34]. Previously, when the robot made an incorrect action, the user had to wait for the robotic system to finish the action before any intervention. Our error detection model will enable the robot to detect and correct errors in real time during motor assistance. This will allow the robot to dynamically adjust its movements based on detected errors, ultimately improving safety and efficiency.

V. CONCLUSION

Data imbalance, feature sparsity, and variable response times pose significant challenges for ErrP detection in dynamic HRI scenarios. To address these issues, we proposed an EEG conformer framework that integrates a pre-trained CNN encoder with a transformer block employing causal attention to capture long-term dependencies in EEG.

Experimental results demonstrated that our model outperformed both classical discrete models and sequential deep learning approaches in inter-session and inter-subject cross-validation tasks, particularly in reducing false positive rates. The model's ability to maintain low false positive rates while accurately detecting error events underscores its robustness and adaptability. Our work advances ErrP detection for practical applications in assistive robotics and beyond. By enabling accurate and efficient error prediction, our approach contributes to safer and more responsive HRI in complex, real-world environments.

TABLE IV: Inter-Session Cross-Validation Results on Pseudo-Online EEG Stream

Method	Mode	Acc (%)	F1 (%)	nMCC (%)
EEGNet	Discrete	59.88 ± 1.46	23.84 ± 2.97	59.76 ± 1.73
ResNet	Discrete	60.16 ± 1.02	29.02 ± 3.13	64.51 ± 2.22
Ensemble*	Discrete	55.28 ± 0.68	17.79 ± 1.87	60.49 ± 1.29
EEGNet-GRU	Continuous	61.12 ± 1.91	31.37 ± 4.63	65.86 ± 2.74
EEGNet-LSTM	Continuous	60.68 ± 3.07	30.70 ± 3.07	65.49 ± 3.07
ResNet-GRU	Continuous	60.76 ± 4.09	30.92 ± 9.58	66.36 ± 4.89
ResNet-LSTM	Continuous	56.52 ± 2.24	20.58 ± 5.92	61.55 ± 3.16
EEGNet-Transformer	Continuous	**72.79 ± 1.95**	**54.38 ± 5.11**	**76.87 ± 2.98**
ResNet-Transformer	Continuous	<u>68.32 ± 3.72</u>	<u>46.86 ± 8.25</u>	<u>73.78 ± 4.21</u>

TABLE V: Leave-One-Subject-Out Cross-Validation Results on Pseudo-Online EEG Stream

Method	Mode	Acc (%)	F1 (%)	nMCC (%)
EEGNet	Discrete	57.92 ± 3.83	21.39 ± 7.24	59.33 ± 3.86
ResNet	Discrete	58.16 ± 3.58	22.98 ± 8.58	60.78 ± 4.79
Ensemble*	Discrete	54.28 ± 1.64	14.46 ± 4.51	58.60 ± 2.82
EEGNet-GRU	Continuous	62.64 ± 6.44	33.20 ± 14.48	65.98 ± 7.69
EEGNet-LSTM	Continuous	61.30 ± 4.37	31.20 ± 11.54	65.38 ± 6.94
ResNet-GRU	Continuous	62.41 ± 4.96	34.49 ± 12.09	67.44 ± 6.93
ResNet-LSTM	Continuous	58.72 ± 2.94	26.39 ± 7.63	64.20 ± 4.40
EEGNet-Transformer	Continuous	<u>70.92 ± 2.50</u>	<u>48.47 ± 7.49</u>	<u>73.44 ± 4.28</u>
ResNet-Transformer	Continuous	**74.14 ± 3.83**	**56.35 ± 8.78**	**77.85 ± 4.66**

References

[1] G.-Z. Yang, R. Riener, and P. Dario, "To integrate and to empower: Robots for rehabilitation and assistance," *Sci. Robot.*, 2017.

[2] Y. Guo, X. Gu, and G.-Z. Yang, "Human–robot interaction for rehabilitation robotics," *Digitalization in Healthcare: Implementing Innovation and Artificial Intelligence*, pp. 269–295, 2021.

[3] N. Lingg and Y. Demiris, "Building trust in assistive robotics: Insights from a real-world mobile navigation experiment," in *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, 2023, pp. 1–7.

[4] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi, "Head and shoulders: automatic error detection in human-robot interaction," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 181–188.

[5] E. Rosen, D. Whitney, M. Fishman, D. Ullman, and S. Tellex, "Mixed reality as a bidirectional communication interface for human-robot interaction," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 431–11 438.

[6] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2016, pp. 493–500.

[7] M. Awais and D. Henrich, "Human-robot collaboration by intention recognition using probabilistic state machines," *19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010)*, pp. 75–80, 2010.

[8] P. Ferrez and J. Millán, "Error-related eeg potentials generated during simulated brain–computer interaction," *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 923–929, 2008.

[9] S. K. Kim, E. A. Kirchner, A. Stefes, and F. Kirchner, "Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction," *Scientific reports*, vol. 7, no. 1, pp. 1–16, 2017.

[10] F. Iwane, A. Sobolewski, R. Chavarriaga, and J. d. R. Millán, "Eeg error-related potentials encode magnitude of errors and individual perceptual thresholds," *Iscience*, vol. 26, no. 9, 2023.

[11] P. K. Parashiva and A. P. Vinod, "Single-trial detection of eeg error-related potentials using modified power-law transformation," *Biomed. Signal Process. Control.*, vol. 67, p. 102563, 2021.

[12] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *JNE*, vol. 15, no. 5, 2018.

[13] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Hum Brain Mapp*, vol. 38, no. 11, pp. 5391–5420, 2017.

[14] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "Lstm-based eeg classification in motor imagery tasks," *IEEE TNSRE*, 2018.

[15] J. Han, X. Wei, and A. A. Faisal, "Eeg decoding for datasets with heterogenous electrode configurations using transfer learning graph neural networks," *JNE*, 2023.

[16] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE IJCNN*. IEEE, 2008, pp. 2390–2397.

[17] X. Gu, F. Deligianni, J. Han, X. Liu, W. Chen, G.-Z. Yang, and B. Lo, "Beyond supervised learning for pervasive healthcare," *IEEE Reviews in Biomedical Engineering*, 2023.

[18] I. Carrara and T. Papadopoulo, "Pseudo-online framework for bci evaluation: a moabb perspective using various mi and ssvep datasets," *Journal of Neural Engineering*, vol. 21, no. 1, 2024. [Online]. Available: https://doi.org/10.1088/1741-2552/ad171a

[19] Z. Ren, X. Xia, Y. Tang, B. Zhao, C. P. Wong, and D. Xiao, "Asynchronous detection of erroneous behaviors in human-robot inter-

action with eeg: A comparative analysis of machine learning models," *bioRxiv*, pp. 2023–09, 2023.

[20] W. J. Gehring, B. Goss, M. G. Coles, D. E. Meyer, and E. Donchin, "A neural system for error detection and compensation," *Psychological science*, vol. 4, no. 6, pp. 385–390, 1993.

[21] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "Erp components on reaction errors and their functional significance: a tutorial," *Biological psychology*, vol. 51, no. 2-3, pp. 87–107, 2000.

[22] S. Bhattacharyya, A. Konar, D. N. Tibarewala, and M. Hayashibe, "A generic transferable eeg decoder for online detection of error potential in target selection," *Frontiers in neuroscience*, vol. 11, p. 226, 2017.

[23] J. Han, X. Gu, G.-Z. Yang, and B. Lo, "Noise-factorized disentangled representation learning for generalizable motor imagery eeg classification," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[24] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.

[25] A. Farabbi, L. Mainardi *et al.*, "The subspace regularization method improves errp detection by eegnet in bci experiments," in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023)*, 2023, pp. 260–264.

[26] A. Kumar, E. Pirogova, S. S. Mahmoud, and Q. Fang, "Classification of error-related potentials evoked during stroke rehabilitation training," *Journal of Neural Engineering*, vol. 18, no. 5, p. 056022, 2021.

[27] A. Demir, T. Koike-Akino, Y. Wang, M. Haruna, and D. Erdogmus, "Eeg-gnn: Graph neural networks for classification of electroencephalogram (eeg) signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1061–1067.

[28] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling human response to robot errors for timely error detection," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 676–683.

[29] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus, "Correcting robot mistakes in real time using eeg signals," in *2017 IEEE ICRA*. IEEE, 2017, pp. 6570–6577.

[30] N. Kueper, K. Chari, J. Bütefür, J. Habenicht, T. Rossol, S. K. Kim, M. Tabie, F. Kirchner, and E. A. Kirchner, "Eeg & emg dataset for the detection of errors introduced by an active orthosis device," *Frontiers in Human Neuroscience*, vol. 18, p. 1304311, 2024.

[31] K. Ding, A. J. Liang, B. Perozzi, T. Chen, R. Wang, L. Hong, E. H. Chi, H. Liu, and D. Z. Cheng, "Hyperformer: Learning expressive sparse feature representations via hypergraph transformer," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2062–2066.

[32] H. Zhao, Y. Yang, P. Karlsson, and A. McEwan, "Can recurrent neural network enhanced eegnet improve the accuracy of erp classification task? an exploration and a discussion," *Health and Technology*, vol. 10, no. 4, pp. 979–995, 2020.

[33] C. Lee and J. An, "Lstm-cnn model of drowsiness detection from multiple consciousness states acquired by eeg," *Expert Systems with Applications*, vol. 213, p. 119032, 2023.

[34] A. Shafti and A. A. Faisal, "Non-invasive cognitive-level human interfacing for the robotic restoration of reaching & grasping," in *2021 10th International IEEE NER*. IEEE, 2021, pp. 872–875.