Human–AI Deferral under Limited Expert Availability: A Study in Intraoperative Ischemia Detection

Nihal Murali NIHAL.MURALI@PITT.EDU
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA, USA

Karim Elzokm KELZOKM@BU.EDU
Electrical and Computer Engineering
Boston University
Boston, MA, USA

Parthasarathy Thirumala THIRUMALAPD@UPMC.EDU
Neurological Surgery
University of Pittsburgh
Pittsburgh, PA, USA

Kayhan Batmanghelich† BATMAN@BU.EDU
Electrical and Computer Engineering
Boston University
Boston, MA, USA

Shyam Visweswaran† SHV3@PITT.EDU
Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA, USA

Abstract

Intraoperative neuromonitoring (IONM) during carotid endarterectomy (CEA) is used to monitor cerebral ischemia, but reliable interpretation depends on scarce expert neurophysiologists. Because many medical tasks are safety-critical, relying solely on artificial intelligence (AI) is impractical. We study how human–AI deferral methods perform in this setting and also introduce Learning to Augment (L2A), a simple method that selectively incorporates human input into AI predictions. We evaluate these methods on intraoperative electroencephalographic data from 400 CEA cases in a large U.S. academic health system. Unlike prior deferral work that focuses on non-clinical tasks and relies on expert humans, we use novice monitors with no prior IONM experience and only brief training. We use these novices as a conservative lower-bound test of whether non-expert human input contains complementary information, not as a proposed replacement for clinically trained personnel. Across methods, human–AI deferral shows that substantial gains over the AI alone can be achieved with minimal human involvement, even when the human is a novice. With just 5–10% novice involvement, L2A improves precision and sensitivity by ∼40%, and the area under the precision-recall curve (AUPRC) by ∼20% compared to the AI model alone. Importantly, while some deferral methods degrade with less experienced human input, others such as L2A remain robust with better calibration. We further observe that these methods produce sparse and temporally clustered deferral decisions, enabling long

periods without human intervention. These properties make human–AI deferral approaches practical in settings where expert neurophysiologists are not available.

1. Introduction

Carotid endarterectomy (CEA) is a common surgery performed to prevent stroke, but reduced blood flow during the procedure can cause cerebral ischemia, making timely detection critical (Sundt Jr et al., 1981; Blume et al., 1986). Intraoperative neuromonitoring (IONM) with continuous electroencephalography (cEEG) is the standard way to monitor for ischemia during CEA. The monitoring and interpretation is performed by neurophysiologists who are experts in this and are also scarce. While AI models to detect intraoperative ischemia have been developed (Mina et al., 2024; Visser et al., 1999; Kamitaki et al., 2021), prior work shows that AI alone is not sufficiently reliable for safety-critical use (Geirhos et al., 2020; Zech et al., 2018; Oakden-Rayner et al., 2020; Wynants et al., 2020; Obermeyer et al., 2019). Human–AI collaboration offers an alternative, but existing approaches either require continuous human involvement (Wilder et al., 2020; Pradier et al., 2021) or rely on deferral methods that assume access to expert decision-makers, often evaluated using simulated experts rather than real human monitors (Madras et al., 2018; Mozannar and Sontag, 2020; Verma and Nalisnick, 2022). In addition, most have been evaluated on vision and language tasks, with limited studies in clinical settings. This motivates evaluating human–AI deferral methods for intraoperative ischemia detection under clinical constraints, including limited expert availability. We additionally introduce Learning to Augment (L2A), a simple deferral method that combines AI and human predictions using learned weights and selectively incorporates human input only when needed. An overview of the proposed method is shown in Figure 1. Unlike hard deferral, L2A avoids fully relying on either source and remains effective even when human input is of lower quality. On intraoperative cEEG data from 400 CEA procedures, L2A improves the area under precision-recall curve (AUPRC) by ∼20%, sensitivity and precision by ∼40% over the AI alone, with just 5–10% human involvement. It maintains stable performance even when the human monitor is a novice (and not a neurophysiologist) and improves calibration relative to prior methods.

In cEEG monitoring, changes in signal patterns can indicate emerging ischemia and enable timely intervention (Sundt Jr et al., 1981; Blume et al., 1986). Interpretation of cEEG is typically performed by neurophysiologists who monitor the signals and alert the surgical team when abnormalities arise. However, the number of available neurophysiologists is limited. In the United States, only a few thousand certified neurophysiologists are available to oversee hundreds of thousands of procedures each year (UConn, n.d.). This gap creates challenges in providing consistent monitoring coverage, especially in resource-limited settings.

AI has been explored to support cEEG interpretation, with prior studies showing that AI models can detect ischemic patterns using quantitative EEG features (Visser et al., 1999; Kamitaki et al., 2021). However, deploying AI alone in safety-critical settings remains challenging due to bias, poor generalization, and limited real-world validation (Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021; Wynants et al., 2020). At the same time, fully manual monitoring is difficult to scale due to the need for sustained attention and limited expert availability. These challenges motivate hybrid human–AI systems that use limited human

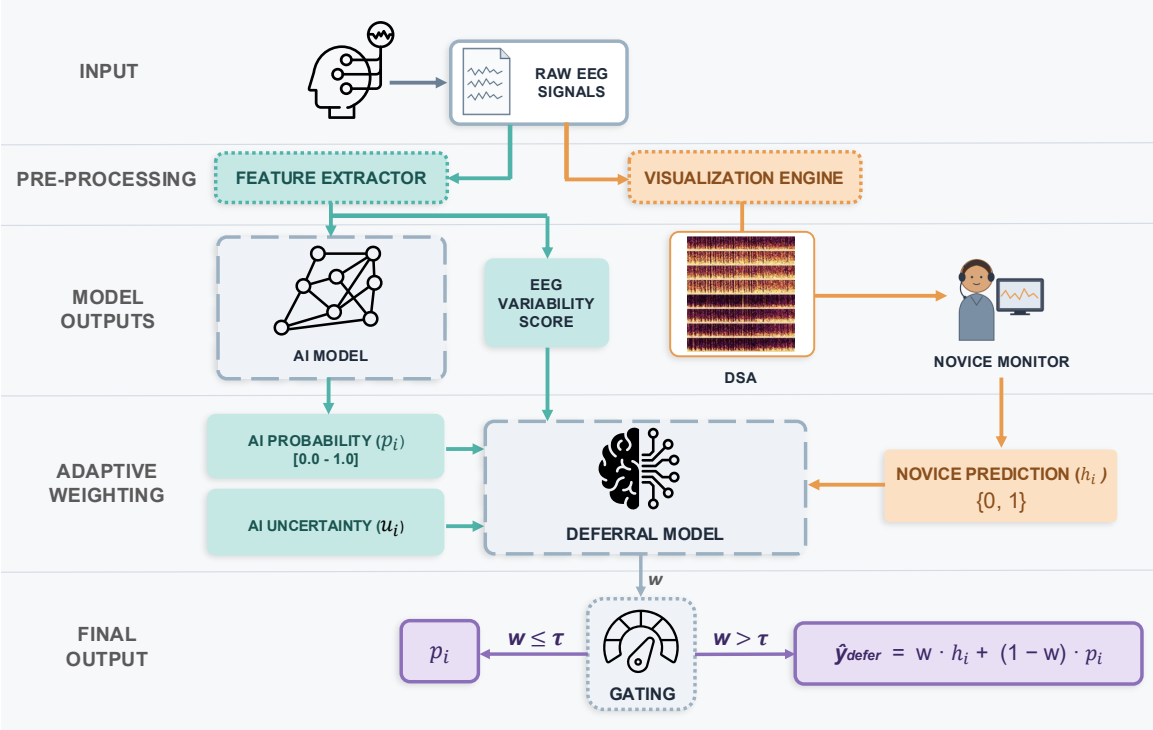

Figure 1: **Overview of the proposed method (L2A).** Raw cEEG signals are preprocessed to extract features for the AI model, which outputs a probability of ischemia and an uncertainty measure. A visualization engine presents the signals as a density spectral array (DSA) to the human monitor. A deferral model uses the AI outputs and EEG signal complexity to estimate a weight that determines how much to rely on human input. By thresholding this weight, L2A either combines AI and human predictions or relies on the AI alone. This allows human input to be used only when needed, improving performance while keeping involvement low.

input while retaining appropriate clinical oversight. One line of work studies human–AI teaming (Wilder et al., 2020; Pradier et al., 2021), where human input is always incorporated, but this requires continuous human effort. Another line of work studies human–AI deferral (Madras et al., 2018; Mozannar and Sontag, 2020; Verma and Nalisnick, 2022), where the AI predicts or defers entirely to a human. Prior work has shown that such deferral strategies can be effective when deferring to expert decision-makers, but they typically assume access to high-quality human input (Tailor et al., 2024) and have been evaluated primarily in non-clinical or controlled settings. However, in clinical practice, expert availability is limited and human input may come from less experienced monitors. It is therefore unclear how well these approaches perform when these assumptions are relaxed.

In this work, we study human–AI deferral for intraoperative ischemia detection under clinical constraints, including limited expert availability and variability in human expertise.

With minimal human involvement (5–10%), deferral methods improve significantly over both AI-only and human-only baselines by focusing human input on a small subset of intervals. We find that deferral methods remain effective despite substantial variability across novice monitors (mean Cohen's kappa 0.61), and that deferral decisions are sparse and tend to cluster within each case, resulting in long uninterrupted periods of hands-off monitoring. We additionally introduce Learning to Augment (L2A), a simple method that combines AI and human predictions using learned weights. L2A performs competitively with existing approaches and improves calibration, while remaining effective even when human input is unreliable. While current clinical guidelines emphasize that cEEG monitoring and interpretation require trained and credentialed personnel with substantial experience (López et al., 2023), our results suggest that even minimally trained monitors, far below this level, can provide useful complementary signal when combined with AI, potentially helping extend expert oversight.

Generalizable Insights about Machine Learning in Healthcare

Across many healthcare applications, AI models are often evaluated in isolation (Rajpurkar et al., 2017; Wu et al., 2022; Mina et al., 2024; Abdelhameed and Bayoumi, 2021). However, clinical data and decision-making are complex, and AI models remain imperfect, limiting their use in safety-critical settings. At the same time, relying entirely on human effort is not scalable, especially for tasks that require continuous monitoring or large-scale data review. Rather than treating AI and human expertise as competing alternatives, our results highlight the value of selectively combining the two. In particular, we find that strong performance can be achieved with minimal human involvement by focusing human input on a small subset of cases where it is most useful. This selective use of human input remains effective even when human expertise is limited or variable, suggesting that useful signal can be extracted from less experienced monitors when combined with AI. These findings point to an important design principle: the effectiveness of AI systems in healthcare depends not only on predictive accuracy, but also on how and when human input is incorporated. Methods that use human effort sparsely and strategically can improve performance while preserving practical workflows, including long uninterrupted periods of hands-off monitoring. Developing such human–AI deferral systems is a promising direction for translating AI models into real clinical practice.

2. Related Work

cEEG-Based Ischemia Detection. Prior work has explored quantitative EEG features for detecting cerebral ischemia during CEA. Visser et al. (1999) described characteristic spectral changes following artery clamping (a key step in CEA), and Kamitaki et al. (2021) showed that quantitative EEG metrics track post-clamp ischemic events. More recently, Mina et al. (2024) demonstrated that tree-based models using a rich set of quantitative EEG features can reliably identify ischemia; we build on this work. Beyond CEA, AI has been applied to cEEG for seizure detection, sleep staging, and neurological diagnosis, though applications to intraoperative ischemia remain limited. While these studies show that AI can detect ischemic patterns, they evaluate models in isolation and do not address how they should be used in practice, especially when neurophysiologists are not always available, and

errors may have dire consequences. This highlights the need for approaches that integrate AI with human input in safety-critical settings.

Limits of AI in Safety-Critical Settings. A consistent finding in clinical settings is that AI models deployed in isolation can fail in systematic and context-specific ways. Prior work has documented biases and disparities in real-world deployments (Obermeyer et al., 2019; Seyyed-Kalantari et al., 2021), as well as broader concerns about generalization and methodological rigor (Wynants et al., 2020). In addition, modern AI models are often miscalibrated and overconfident, especially under dataset shift (Guo et al., 2017; Ovadia et al., 2019), and may rely on spurious correlations that do not hold in practice (Geirhos et al., 2020). Taken together, these limit the use of standalone AI in high-stakes clinical workflows and motivate approaches that retain human judgment as a corrective signal. This naturally leads to hybrid human–AI approaches that combine model predictions with human judgment, rather than relying on AI alone in safety-critical settings.

Rejection Learning and Abstention. One early response to these limitations is to allow models to abstain on uncertain inputs, rather than forcing potentially incorrect predictions. This idea is formalized in rejection learning (Chow, 1970; Bartlett and Wegkamp, 2008; Cortes et al., 2016; Charoenphakdee et al., 2021) and selective classification (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; Gangrade et al., 2021; Acar et al., 2020), where a model may abstain from predicting on uncertain inputs under a coverage constraint. These approaches optimize the tradeoff between prediction accuracy and the fraction of cases answered. While abstention can improve safety by avoiding low-confidence predictions, it does not specify how decisions should be made after abstention and does not incorporate human input directly. This limitation motivates approaches that explicitly involve humans in the decision process.

Hard Deferral. We use the term human–AI deferral to refer broadly to methods that decide when and how to incorporate human input. Hard deferral is one such approach, where uncertain cases are explicitly routed to a human. In these methods, human–AI collaboration is framed as a routing problem: for each input, either the AI makes a prediction or the decision is handed off entirely to a human. Early work decoupled prediction and deferral (Raghu et al., 2019), while later approaches jointly optimize both using surrogate losses (Mozannar and Sontag, 2020; Verma and Nalisnick, 2022; Charusaie et al., 2022). Subsequent work has highlighted optimization challenges and theoretical limits of this formulation (Mozannar et al., 2023; De et al., 2021). In addition, these methods rely on binary routing, which does not allow the model to combine complementary information from AI and human predictions. As a result, performance depends on selecting a single source for each input, which may be suboptimal in settings with asymmetric error costs (Charusaie et al., 2024). This motivates approaches that allow joint use of human and AI predictions rather than strict routing.

Human–AI Teaming. A related line of work integrates human and AI outputs within a mixture-of-experts framework, combining both sources rather than selecting one. Wilder et al. (2020) jointly optimize team performance, while Pradier et al. (2021) propose a pref-

erential variant that biases decisions toward human rules. In these methods, human input is typically used across all or most cases, with the final prediction formed by combining both sources. This can improve performance, but requires continuous human involvement, which can be costly and infeasible in settings such as IONM where neurophysiologists are scarce. This highlights the need for human–AI deferral approaches that selectively incorporate human input only when it is most needed, reducing human effort while maintaining performance.

Soft Deferral. More recent work extends hard deferral by combining AI and human predictions instead of choosing one over the other. These methods incorporate human input selectively by combining AI and human predictions. Defer-and-Fusion (Charusaie et al., 2024) is one such approach, and has been shown to achieve lower error than hard deferral by leveraging both sources. However, prior work has mainly been evaluated on vision and language tasks and typically assumes access to human experts. In clinical settings, where expert availability is limited and human input may come from less experienced monitors, it remains unclear how well these approaches perform. In this work, we evaluate human–AI deferral methods under these conditions and introduce Learning to Augment (L2A), a simple method that combines AI and human predictions using learned weights.

3. Methods

We describe the proposed L2A approach for combining AI and human predictions in intraoperative cEEG monitoring. The method learns when human input is likely to improve the AI and selectively incorporates it into the final prediction.

3.1. Data and Problem Setup

We consider intraoperative cEEG data segmented into fixed non-overlapping 20-second intervals. Each interval is represented as a pair $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^{111}$ is a feature vector and $y_i \in \{0, 1\}$ is the ground truth ischemia label; details of the 111 features are provided in Appendix B. For each interval $\mathbf{x}_i$, the AI model outputs a probability $p_i \in [0, 1]$ indicating ischemia. We convert this into a binary prediction $\tilde{y}_i = \mathbf{1}\{p_i > 0.5\}$. A human (novice) provides a binary prediction $h_i \in \{0, 1\}$.

3.2. L2A Formulation

Our goal is to estimate whether the human prediction is more accurate than the AI prediction for a given interval. We define a binary target $d_i = \mathbf{1}\{|h_i - y_i| < |\tilde{y}_i - y_i|\}$, where $d_i = 1$ indicates that the human prediction is closer to the ground truth than the AI. For each interval, we construct an input vector $\mathbf{z}_i = (p_i, u_i, v_i)$, where p_i is the AI probability and $u_i = -\big(p_i \log p_i + (1 - p_i) \log(1 - p_i)\big)$ is the AI uncertainty (Shannon entropy). The EEG variability score is defined from the 111-dimensional feature vector for interval i. Let $s_i = \mathrm{std}(\mathbf{x}_i), \qquad r_i = \max(\mathbf{x}_i) - \min(\mathbf{x}_i)$, where both quantities are computed across the 111 features of that interval. These are then min–max normalized using statistics estimated on the training set. The final EEG complexity feature is $v_i = \frac{1}{2}\tilde{s}_i + \frac{1}{2}\tilde{r}_i$. Additional details are provided in Appendix B.

We train a Random Forest (RF) model $f(\cdot)$ to estimate $w_i = f(\mathbf{z}_i) \approx \mathbb{P}(d_i = 1 \mid \mathbf{z}_i)$, where $w_i \in [0, 1]$ represents the probability that the human is more accurate than the AI. We define a threshold $\tau \in [0, 1]$ to control the fraction of intervals where human input is used. The final prediction is given by:

$$\hat{y}_i = \begin{cases} w_i \cdot h_i + (1 - w_i) \cdot p_i, & \text{if } w_i > \tau, \\ p_i, & \text{otherwise.} \end{cases}$$

This formulation allows selective use of human input only when it is expected to improve performance.

3.3. Baselines

We benchmark L2A against a comprehensive set of human–AI deferral methods, including both hard and soft deferral approaches. The baselines considered are: Realizable Surrogate (RS; Mozannar et al. (2023)), One-vs-All Surrogate (OvA; Verma and Nalisnick (2022)), Compare Confidence (CC; Raghu et al. (2019)), Selective Prediction (SP; confidence thresholding), Mixture-of-Experts (MoE; Madras et al. (2018)), and two variants of Defer-and-Fusion (DaF-base and DaF-ext; Charusaie et al. (2024)). RS and OvA learn joint prediction and deferral policies using surrogate loss formulations, while MoE adopts a mixture-of-experts approach to combine human and model predictions. CC trains the classifier independently and defers based on a comparison between estimated human and model error. In contrast, SP is human-agnostic: it does not model or use human predictions, and instead defers solely based on model confidence, similar to rejection-based approaches. These methods correspond to hard deferral, where each interval is assigned fully to either the AI or the human. In contrast, DaF-base and DaF-ext represent soft deferral approaches that combine AI predictions with human input. DaF-base learns to choose between two options: the AI prediction alone or a combined AI and human prediction. In contrast, DaF-ext introduces a third option, allowing the model to choose among AI alone, human alone, and a combined AI and human prediction. Such joint use of AI and human predictions can achieve lower error than hard deferral, particularly under asymmetric error costs. These baselines allow us to evaluate human–AI deferral methods across varying levels of human expertise and to place L2A within this broader set of approaches.

4. Cohort

We describe the cEEG dataset, annotation procedure, and feature construction used in this study.

4.1. Cohort Selection

We used intraoperative cEEG data from patients undergoing CEA between 2009 and 2017. The dataset includes 400 patients, of whom 28 exhibited intraoperative ischemic events. Data were collected in a large U.S. academic health system, capturing real-world variability in clinical practice. cEEG signals were acquired at 500 Hz using an XLTEK Protektor32 IOM system (Natus Medical Inc.) from eight subdermal electrodes placed according to the 10–20 system, providing four bipolar channels per hemisphere (F3–P3, P3–O1, F3–T3,

T3–O1 on the left; F4–P4, P4–O2, F4–T4, T4–O2 on the right). Signals were filtered using a fourth-order Butterworth band-pass filter (0.5–45 Hz). For each patient, we extracted 10 minutes of cEEG data following carotid artery clamping, a period during which the risk of ischemia is highest. In this work, we define a case as the 10-minute cEEG segment from a single patient, and an interval as a non-overlapping 20-second segment within a case, which forms the unit of analysis. Across the 12,000 intervals in the dataset, 285 are labeled as ischemic (prevalence $\approx 2.38\%$), indicating a highly imbalanced setting with very few positive intervals. See Appendix A for full acquisition and preprocessing details.

4.2. Data Extraction and Labeling

Ground truth ischemia labels were derived from intraoperative notes recorded by neurophysiologists during IONM. These notes define time periods corresponding to ischemic events. Each 20-second interval was assigned a binary label by aligning its time window with the annotated ischemic periods. An interval was labeled ischemic if at least 10 seconds (50% of the interval) overlapped with an ischemic period; otherwise, it was labeled non-ischemic. In addition to expert-derived labels, each interval was independently labeled by four novices. Inter-rater agreement among novices is reported in Appendix E.

4.3. Novice Annotation

We recruited four engineering and data science trainees with no prior experience in cEEG or IONM. All novices completed a brief structured training program consisting of an introductory session by a neurophysiologist covering ischemic patterns and confounders, followed by a supervised practice session with feedback. Annotations were performed using a custom forward-only cEEG viewer designed to mimic real-time IONM. The interface advanced automatically in 20-second steps and allowed a limited look-back window (up to 120 seconds). This 120-second look-back window was selected in consultation with board-certified IONM professionals to reflect current clinical interpretation, in which ischemic changes may be confirmed using up to 2 minutes of evolving cEEG and density spectral array context; however, this retrospective study did not measure the resulting time-to-alert. For each CEA case, novices labeled both a pre-clamp baseline segment and a post-clamp segment. All cases were annotated by all four novices. Novices were blinded to expert annotations and intraoperative notes. In total, novices reviewed approximately 67 hours of cEEG data to annotate the full dataset. The novices served as experimental proxies for low-expertise human input and were not intended to represent personnel making independent clinical decisions.

4.4. Feature Construction

Each 20-second interval was represented by a 111-dimensional feature vector derived from quantitative cEEG measures capturing amplitude, spectral content, and interhemispheric asymmetry. Representative features include band power in standard frequency bands (delta, theta, alpha, beta, gamma), band power ratios (e.g., ADR, BDR, ABDTR), spectral edge frequency (SEF90), amplitude-integrated EEG (aEEG), and symmetry measures such as pdBSI. All features were computed as z-scores by normalizing to a 120-second artifact-free

pre-clamp baseline for each patient. Additional details of feature definitions are provided in Appendix B.

4.5. Experimental Setup

The dataset includes 400 CEA patients and 12,000 intervals of 20 seconds each, labeled for intraoperative ischemia. We perform 10-fold patient-level cross-validation, ensuring that all intervals from a given patient are assigned to the same fold. We compare L2A with the human–AI deferral baselines described in Section 3.3, spanning both hard and soft deferral approaches. Hard deferral methods (RS, OvA, CC, SP, MoE) assign each interval fully to either the AI or the human, while soft deferral methods (DaF-base and DaF-ext) combine both predictions. CC, SP, and L2A operate on precomputed AI probabilities and do not require differentiable training. We therefore use an RF model, which is known to perform well on tabular data and also gave the best performance on our cEEG features. The AI is trained using cross-validation, and out-of-fold predictions are used as fixed inputs for the deferral models. In contrast, RS, OvA, MoE, and DaF require differentiable, end-to-end training. For these methods, we use a two-layer multilayer perceptron (MLP) with 256 and 128 hidden units. This choice is appropriate given the tabular inputs, moderate dataset size, and strong class imbalance, where simpler models tend to be more stable and less prone to overfitting. The AI and deferral components are trained jointly on the training folds and evaluated on the held-out fold. All metrics are computed on held-out predictions. Because RF performs better than MLP on these data, absolute comparisons between these groups confound the deferral method with the underlying AI model. We therefore evaluate each method relative to its own 0%-deferral baseline. The cleanest method-matched comparisons for L2A are CC and SP, which use the same RF predictions. To control human involvement, we evaluate each method across deferral fractions from 0% to 100% in 10% steps, with an additional 5% point for finer resolution at low deferral levels. For a given fraction, intervals are ranked by a method-specific deferral score, and the top-ranked intervals receive human input. For L2A, this score is the learned novice weight, with higher weights indicating greater reliance on novice input. We additionally evaluate cross-novice transfer by training L2A on one novice and testing it on another across all 12 ordered novice pairs. We also estimate onboarding requirements by training L2A with progressively smaller patient-level subsets of novice labels. Full details are provided in Appendix F.

5. Results

We evaluate human–AI deferral methods along three dimensions: performance, calibration, and human workload. We vary the deferral level to control human involvement and, at each level, compute AUPRC, AUROC, and classification metrics (precision, sensitivity, and specificity). This assesses whether deferral methods retain the benefits of human–AI collaboration with minimal human involvement. Second, to study calibration, we evaluate the quality of predictions across different levels of human involvement. For each setting, we compute expected calibration error (ECE), Brier score, and logarithmic loss (log loss). This assesses how deferral methods affect calibration and whether predicted probabilities reflect true event likelihood. Third, to study human workload, we analyze how deferral decisions are distributed over time within each case. We measure interruptions and longest

uninterrupted free time per case to evaluate whether the system provides long, contiguous hands-off periods rather than frequent short interruptions.

5.1. Discriminative Performance under Deferral

Figure 2 shows results for a higher-performing novice whose standalone performance exceeds the AI model. At 0% deferral (red dashed line), all methods operate in the AI-only regime. As deferral increases, a larger fraction of intervals is assigned to the novice, with the novice-only performance shown as a green reference line. L2A, CC, and SP share the same starting point since they operate on fixed RF probabilities without retraining. In contrast, end-to-end methods train MLP models jointly with method-specific objective functions and architectural modifications, leading to different initial performance across methods. Accordingly, improvements from each method's own 0%-deferral baseline reflect the benefit of adding human input, whereas absolute differences between RF- and MLP-based methods should not be attributed solely to the deferral strategy.

Here, the novice substantially outperforms the AI models across most metrics. At an FPR of 1%, the novice, with ∼70 hours of annotation effort, achieves 83% sensitivity by identifying 237 out of 285 ischemic intervals, whereas the RF model with a higher FPR of 1.5% reaches only 53% sensitivity (151/285), and MLP-based methods can perform much worse at the same operating point (e.g., as low as 28/285, 10% sensitivity). Despite this gap, even small amounts of novice input (5–10%) are sufficient for most methods to recover a large fraction of the performance difference. For example, with 10% deferral (∼7 hours of novice effort), methods such as L2A, SP, and MoE achieve around 79% sensitivity (225/285) while operating at a lower FPR of 0.5%. At this level, precision also improves beyond the novice (0.8 vs 0.7), whereas the AI models alone remain closer to 0.5 precision. Overall, deferral methods substantially improve performance with minimal novice involvement. Among them, L2A performs strongly, particularly on AUPRC and AUROC, with improvements of 30–50% across key metrics.

Figure 3 shows results for a lower-performing novice who performs worse than the AI models on several metrics. However, this novice still has high sensitivity of 81% (231/285), compared to the RF model at 53% (151/285), but with much lower precision (0.34) and specificity (0.96), leading to a high FPR (4%). Despite this, deferral methods are still able to use this novice effectively. Using only 10% novice input, methods such as L2A and SP improve precision from 0.46 (RF) to 0.6, while reducing FPR from 1.5% to 1%, even though the novice alone has poor precision and high FPR. At the same time, sensitivity improves to 75% (214/285), closing much of the gap. Overall, even with a lower-performing novice, small amounts of selective human input are sufficient to improve both sensitivity and precision. As deferral increases further, a clear difference emerges between methods. Hard deferral methods degrade as more predictions are assigned to the novice, reflecting the novice's low precision and high FPR. In contrast, soft deferral methods such as L2A remain stable and do not show this drop, highlighting their advantage in settings where the human is unreliable. Overall, L2A remains stable as novice involvement increases and continues to improve performance even with a lower-performing novice, achieving gains of ∼35% in precision and sensitivity while reducing FPR by ∼30% with minimal human involvement. Additional results for the remaining novices are provided in Appendix D.

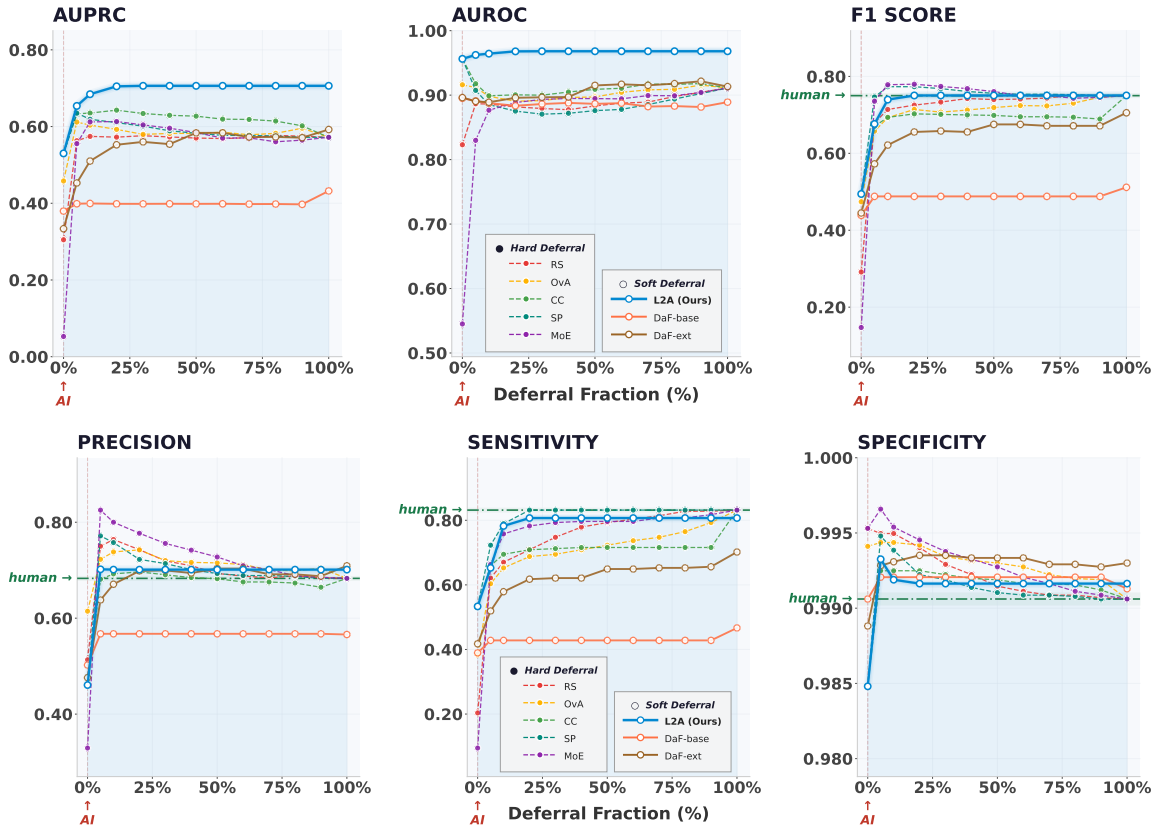

Figure 2: **Higher-performing novice: closing the gap between AI and human performance.** The novice (green line) substantially exceeds the AI-only baseline (red dotted line at 0% deferral) on key metrics such as precision and sensitivity. Despite this gap, small amounts of deferral (5–10%) allow most methods to recover much of the performance difference, with several methods approaching or exceeding novice-level performance at low deferral rates. L2A remains competitive with other methods across metrics.

5.2. Calibration under Deferral

We evaluate how calibration changes as the level of human involvement increases from 0% to 100%. Figure 4 shows calibration results for a lower-performing novice (top) and a higher-performing novice (bottom), evaluated using expected calibration error (ECE), Brier score, and log loss. Across both settings, L2A achieves better calibration than other methods, with lower ECE, Brier score, and log loss. We observe clear differences between hard and soft deferral. Soft deferral methods remain stable across deferral levels, whereas hard deferral methods are sensitive to human quality. For lower-performing novices, calibration degrades with increasing deferral as predictions increasingly reflect poorly calibrated human input. For higher-performing novices, calibration remains more stable and the gap between methods narrows. Overall, soft deferral methods are more robust when human performance

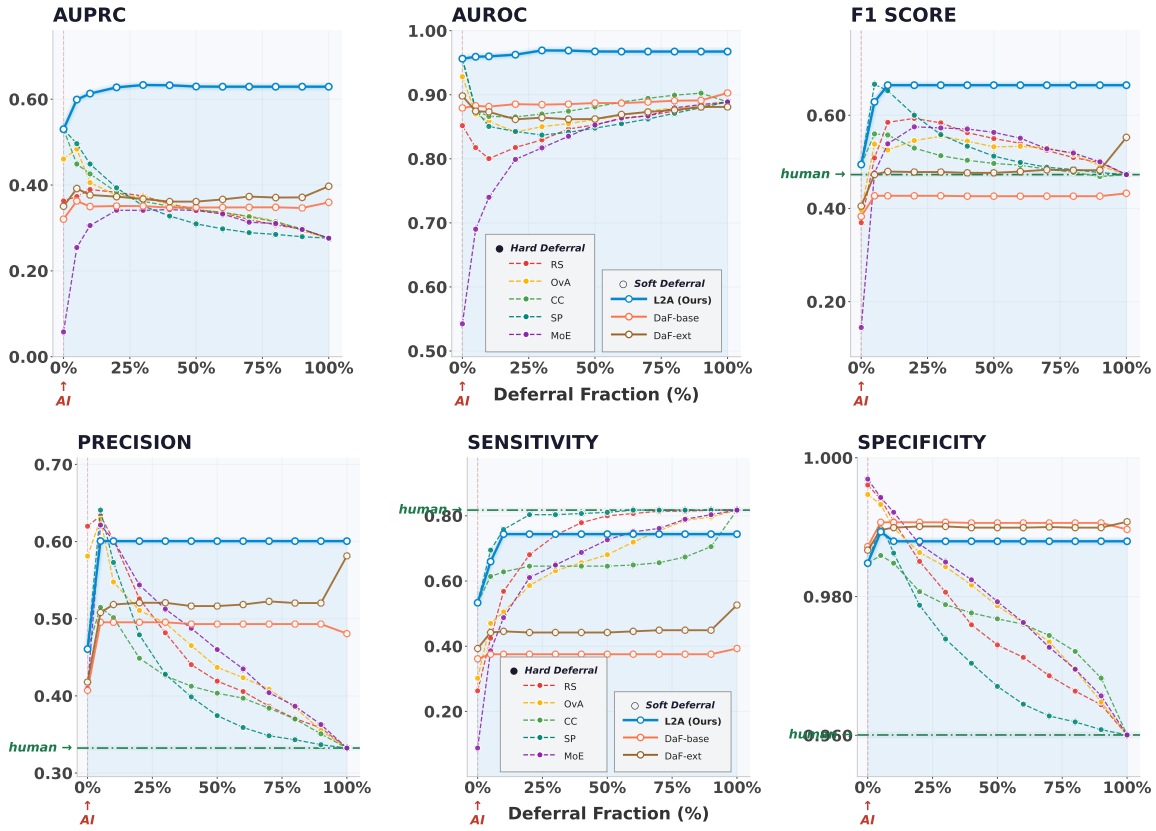

Figure 3: **Lower-performing novice: stable performance under unreliable human input.** The novice (green line) underperforms the AI-only baseline (red dotted line at 0% deferral) on several metrics. Despite this, small amounts of deferral still lead to clear performance gains for most methods. As deferral increases, hard deferral methods degrade and approach novice-level performance, whereas soft deferral methods like L2A remain stable even under suboptimal human input.

is poor, while differences reduce as human performance improves. L2A maintains strong calibration across both settings.

5.3. Human Workload and Contiguity

We evaluate human workload at a fixed deferral fraction of 10%. This operating point is chosen because all methods achieve near-peak performance at low deferral levels (5–10%), allowing comparison under minimal human involvement. We analyze how deferral decisions are distributed over time within each case. We measure (i) interruptions per case, defined as the number of switches from AI to novice within a case, and (ii) longest uninterrupted free time, defined as the longest continuous duration during which the novice is not required to intervene. Both metrics are averaged across cases and novices. These measures describe

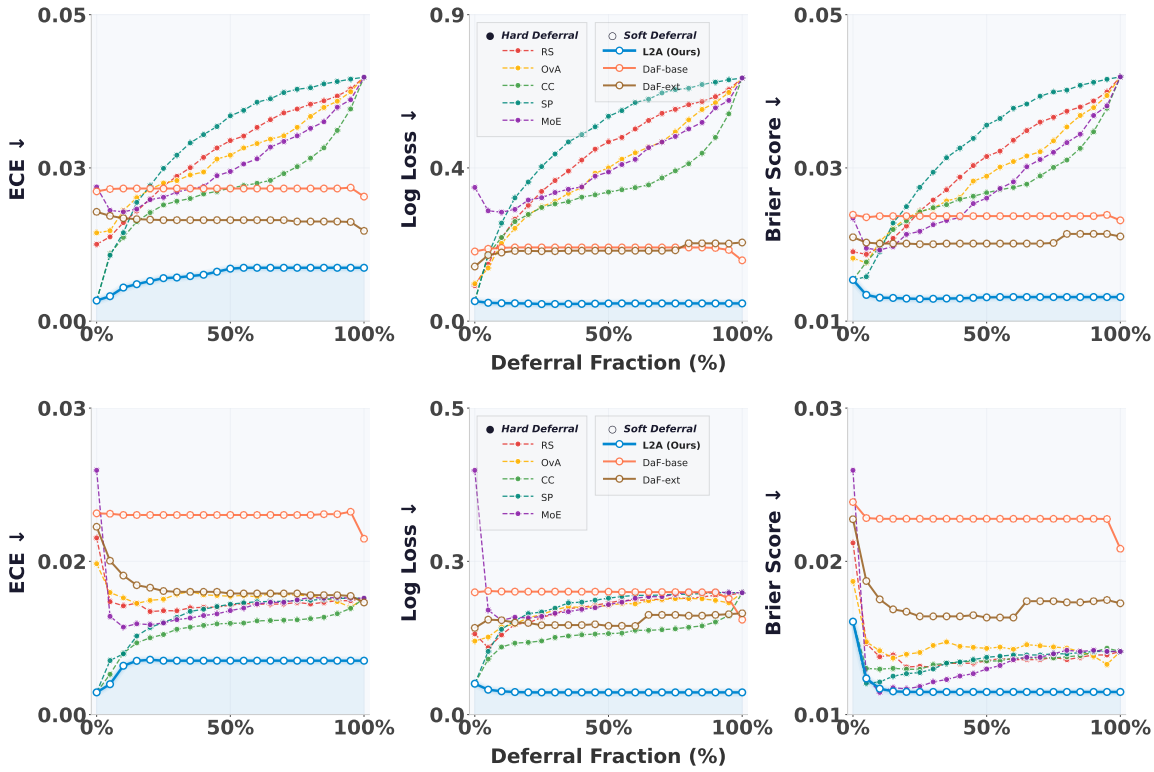

Figure 4: **Calibration across deferral levels (top: lower-performing novice; bottom: higher-performing novice).** Hard deferral methods progressively align with human-only calibration, which is poor when the novice has lower standalone performance. In contrast, soft deferral methods remain stable, as their continuous outputs are less affected by increased human involvement. L2A achieves the strongest calibration overall, with improvements in Brier score and log loss as deferral increases. While ECE shows a slight increase with higher human involvement, this change is modest compared to other methods.

simulated workload within our retrospective interface and should not be interpreted as validation of a clinical workflow.

Figure 5 shows that methods differ in how they structure human involvement. DaF-Ext yields the fewest interruptions and the longest uninterrupted free time, followed by SP. L2A results in approximately one interruption per case on average and maintains substantial uninterrupted free time. Hard deferral methods such as OvA and RS produce more frequent interruptions and shorter free time. Overall, simulated human involvement remains low across methods, with few interruptions per case and substantial uninterrupted free time.

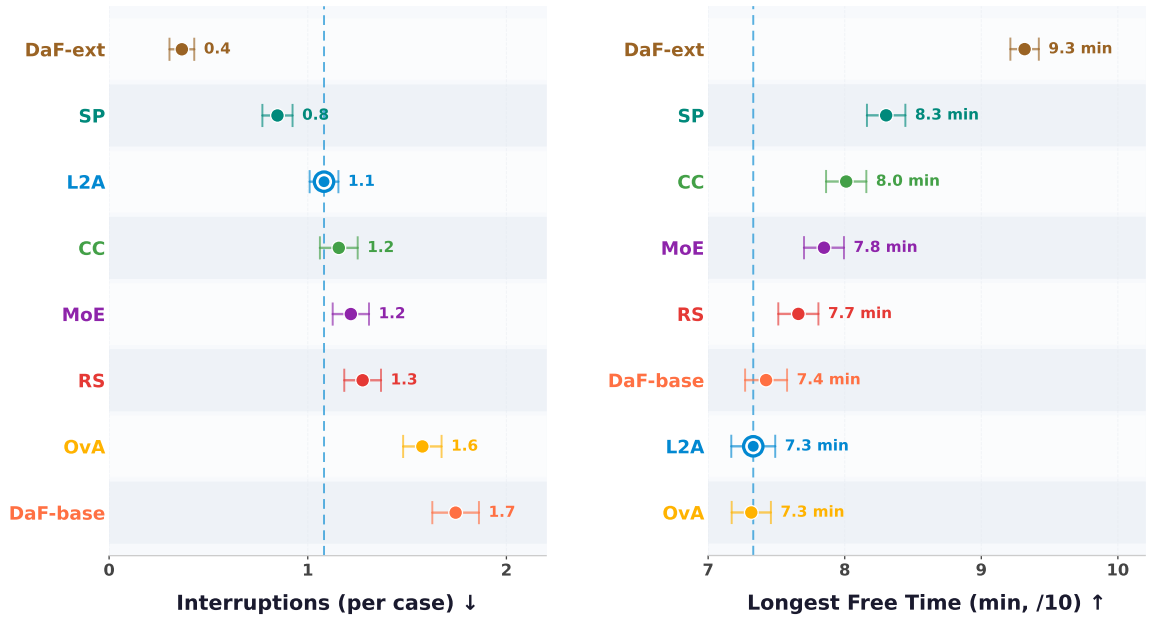

Figure 5: **Human workload and contiguity with 10% novice involvement.** Each case is 10 minutes long. Left: interruptions per case. Right: longest uninterrupted free time per case (minutes). Results are averaged across novices, with error bars showing 95% confidence intervals.

6. Discussion

This work addresses a key challenge in IONM during CEA: expert neurophysiologists are scarce, and AI models alone are not sufficiently reliable for safety-critical use. We study whether minimally trained novices can be combined with an AI model to improve performance while keeping human involvement low. Figures 2 and 3 show that deferral-based methods, including L2A, achieve substantial gains over AI-only performance with only 5–10% novice involvement. With minimal human involvement, these methods often exceed both AI-only and human-only baselines across most metrics. Notably, both hard and soft deferral methods identify a small subset of intervals where human input is most useful. As a result, most of the gains from human–AI collaboration are achieved with minimal human effort, enabling efficient use of limited clinical resources. Figure 2 shows a higher-performing novice whose performance surpasses the AI models on most metrics. Despite this gap, with only 5–10% novice involvement, deferral methods are able to recover most of the gains. This indicates that these methods are effective at identifying a small set of high-value intervals where human input is most useful, and leveraging them to improve performance with minimal effort. The steep deferral curves further highlight that most of the benefit comes from a small fraction of carefully selected intervals. When the human input is reliable, the difference between hard and soft deferral methods becomes smaller. In this setting, deferring to the human is less harmful, and hard deferral methods can perform comparably. This aligns with prior work on human–AI deferral, which often assumes access to expert

humans. Our results extend this understanding by showing that soft deferral methods are particularly useful when human performance is variable or suboptimal.

Figure 3 presents a more challenging setting with a lower-performing novice who is worse than the AI models on key metrics such as precision and specificity. Interestingly, even in this case, deferral methods, particularly L2A, are still able to improve performance. While the novice performs poorly overall, there exist specific intervals where the novice provides complementary information to the AI. Deferral methods are able to identify and exploit these intervals, allowing gains even from suboptimal human input. We also observe clear differences between hard and soft deferral strategies. Hard deferral methods produce binary outputs (AI or human) and degrade as deferral increases, eventually converging to human performance. In contrast, soft deferral methods combine AI and human predictions to produce probabilistic outputs, allowing them to remain stable even when the human is unreliable. This distinction also affects performance on ranking and calibration metrics such as AUPRC and AUROC, where probabilistic outputs are better suited. As a result, methods like L2A tend to achieve more stable performance. Overall, soft deferral provides a more robust way to incorporate suboptimal human input while still allowing flexible adjustment of operating points based on clinical requirements.

Another practical advantage of L2A, Compare Confidence, and Selective Prediction is that they operate on precomputed AI predictions and do not require retraining the AI model. This makes them easy to integrate with any pretrained system, which is important in clinical settings where models may already be deployed or access to training data is restricted. In such cases, retraining or replacing the AI model may be infeasible. In our setting, tree-based models such as RF performed best, which is consistent with their strong performance on tabular data under class imbalance ($\sim$2.38% positive intervals). In contrast, methods that require differentiable models are constrained to architectures such as MLPs, which performed slightly worse on this dataset. Despite this, all methods improved over AI-only performance, suggesting that the gains primarily come from effective use of human input rather than the choice of AI architecture. Across all novices, we observe consistent improvements over AI-only performance (see Appendix D for additional results). As shown in Appendix E, agreement among novices is only moderate (Cohen's $\kappa = 0.61 \pm 0.11$), with substantial disagreement on ischemic intervals (76.37% agreement on positives vs 97.85% on negatives). This indicates that human input is heterogeneous, with different novices exhibiting different behaviors and error patterns. Despite this variability, L2A transferred well across novices. At 10% deferral, within-novice and transferred models achieved similar mean performance (AUPRC 0.636 vs 0.641; AUROC 0.961 for both). Moreover, using only 36 labeled cases per novice monitor ($\sim$6 hours), compared with 400 cases ($\sim$67 hours), produced similar AUPRC (0.627 vs 0.631). These findings suggest that a new monitor could begin with a transferred policy and, if needed, personalize it using a modest labeled set; full results are provided in Appendix F.

The contiguity results in Figure 5 show that methods with similar discriminative performance can differ substantially in how they structure human involvement. At 10% novice involvement, all methods operate near peak performance, yet their impact on workflow varies. DaF-Ext yields the fewest interruptions and longest uninterrupted free time, while hard deferral methods (e.g., OvA, RS) exhibit more fragmented behavior with frequent switches and shorter free time. L2A occupies a middle ground, maintaining $\sim$1 interruption

per case while providing substantial uninterrupted free time. SP also performs well on contiguity, but does not incorporate human input and relies only on AI confidence, so its deferral pattern is independent of the human. These results highlight that contiguity is a distinct dimension of evaluation: strong contiguity does not necessarily imply effective human–AI collaboration. Low interruption rates and long uninterrupted periods may reduce monitoring burden, although this was not evaluated in a live clinical workflow. A more realistic future implementation could involve clinically embedded personnel, such as anesthesia staff after expert-led training, operating under expert oversight. Alternatively, limited deferral could help scarce experts oversee multiple operating rooms. Overall, methods achieve a favorable balance between performance and workflow, maintaining low workload with few interruptions and long hands-off periods. Notably, none of the methods explicitly optimize contiguity, yet they naturally exhibit desirable patterns at low deferral levels.

Finally, our results show that meaningful gains from human–AI collaboration can be achieved even without expert-level human input. Prior guidelines indicate that neurodiagnostic tasks span roles with varying training levels, from basic monitoring to expert interpretation (López et al., 2023). In our setting, novices with only a few hours of exposure, when combined with AI, achieved strong performance, suggesting that even minimal training can yield useful signal, despite real-world practice typically involving substantially longer preparation. We interpret this as a conservative feasibility test of whether low-expertise human input contains complementary information, not as evidence for independent monitoring by non-medical personnel. A clinically realistic deployment would retain expert oversight and could involve clinically embedded staff after appropriate training, with selective expert involvement potentially reducing workload and enabling multi-operating-room coverage. Overall, deferral-based approaches, particularly soft deferral methods such as L2A, offer a flexible framework for combining AI with human input in safety-critical clinical settings, but require prospective validation under appropriate clinical supervision.

Limitations This study has several limitations. First, annotations were collected retrospectively by non-medical novices without clinical context. Thus, the study does not establish the safety or feasibility of live IONM use. Novices served as a conservative lower-bound test, not as proposed independent clinical monitors. Second, only four novices participated. Although trends were consistent, larger, more diverse, and preferably multi-site studies are needed to assess variability across users and institutions. Third, each 20-second interval was analyzed independently, limiting assessment of temporal patterns and outcomes such as time to detection. Finally, our offline evaluation does not capture temporal changes in human or AI performance, including fatigue or distribution shift. Prospective deployment would require performance monitoring, periodic recalibration, and online or conformal updating.

References

Ahmed Abdelhameed and Magdy Bayoumi. A deep learning approach for automatic seizure detection in children with epilepsy. *Frontiers in Computational Neuroscience*, 15:650050, 2021.

Durmus Alp Emre Acar, Aditya Gangrade, and Venkatesh Saligrama. Budget learning via bracketing. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4109–4119. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/acar20a.html.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL http://jmlr.org/papers/v9/bartlett08a.html.

WARREN T Blume, GARY G Ferguson, and D Kent McNeill. Significance of eeg changes at carotid endarterectomy. *Stroke*, 17(5):891–897, 1986.

Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2972–3005. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/charusaie22a.html.

Mohammad-Amin Charusaie, Amirmehdi Jafari Fesharaki, and Samira Samadi. Defer-and-fusion: Optimal predictors that incorporate human decisions. In *5th Workshop on practical ML for limited/low resource settings*, 2024.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, 2016. URL https://api.semanticscholar.org/CorpusID:987180.

Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5905–5913, 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2179–2187. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/gangrade21a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *CoRR*, abs/1705.08500, 2017. URL http://arxiv.org/abs/1705.08500.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Brad K Kamitaki, Bin Tu, Stephen Wong, Anil Mendiratta, and Hyunmi Choi. Quantitative eeg changes correlate with post-clamp ischemia during carotid endarterectomy. *Journal of Clinical Neurophysiology*, 38(3):213–220, 2021.

Jaime R López, Judy Ahn-Ewing, Ron Emerson, Carrie Ford, Clare Gale, Jeffery H Gertsch, Lillian Hewitt, Aatif Husain, Linda Kelly, John Kincaid, et al. Guidelines for qualifications of neurodiagnostic personnel: a joint position statement of the american clinical neurophysiology society, the american association of neuromuscular & electrodiagnostic medicine, the american society of neurophysiological monitoring, and aset—the neurodiagnostic society. *Journal of Clinical Neurophysiology*, 40(4):271–285, 2023.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Amir I Mina. *Advancing Perioperative Stroke Prevention: Machine Learning-Based Electroencephalography Monitoring for Detecting Cerebral Ischemia During Carotid Endarterectomy*. PhD thesis, University of Pittsburgh, 2024.

Amir I Mina, Jessi U Espino, Allison M Bradley, Parthasarathy D Thirumala, Kayhan Batmanghelich, and Shyam Visweswaran. Detecting cerebral ischemia from electroencephalography during carotid endarterectomy using machine learning. *AMIA Summits on Translational Science Proceedings*, 2024:613, 2024.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Melanie F Pradier, Javier Zazo, Sonali Parbhoo, Roy H Perlis, Maurizio Zazzi, and Finale Doshi-Velez. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *AMIA Summits on Translational Science Proceedings*, 2021:525, 2021.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Thoralf M Sundt Jr, Frank W Sharbrouch, David C Piepcras, Thomas P Kearns, Joseph M Messick Jr, and William M O'Fallon. Correlation of cerebral blood flow and electroencephalographic changes during carotid endarterectomy: with results of surgery and hemodynamics of cerebral ischemia. In *Mayo Clinic Proceedings*, volume 56, pages 533–543. Elsevier, 1981.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *International Conference on Artificial Intelligence and Statistics*, pages 3475–3483. PMLR, 2024.

UConn. Intraoperative neuromonitoring and surgical neurophysiology, n.d. URL https://ionm.uconn.edu/. Accessed: 2026-02-28.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022.

GH Visser, GH Wieneke, and AC Van Huffelen. Carotid endarterectomy monitoring: patterns of spectral eeg changes due to carotid artery clamping. *Clinical neurophysiology*, 110(2):286–294, 1999.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Yinhao Wu, Bin Chen, An Zeng, Dan Pan, Ruixuan Wang, and Shen Zhao. Skin cancer classification with deep learning: a systematic review. *Frontiers in Oncology*, 12:893972, 2022.

Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Elena Albu, Banafsheh Arshi, Vanesa Bellou, Marc MJ Bonten, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

Appendix A. cEEG Data Collection and Processing

This section outlines how cEEG signals were recorded and processed. Continuous cEEG recordings were collected using XLTEK clinical monitoring systems (Natus Medical Inc.) at a sampling rate of 500 Hz. Eight subdermal needle electrodes were placed following the international 10–20 system. Signals were recorded using four bipolar channel pairs per hemisphere: F3–P3, P3–O1, F3–T3, and T3–O1 on the left, and F4–P4, P4–O2, F4–T4, and T4–O2 on the right. This bilateral anterior–posterior configuration provides balanced coverage of both hemispheres and is well suited for detecting unilateral changes associated with ischemia during CEA. All signals were acquired in a bipolar montage, representing voltage differences between electrode pairs, with no additional re-referencing.

During acquisition, signals were filtered using a hardware band-pass filter (0.5–70 Hz) along with a 60 Hz notch filter to reduce power-line interference. For downstream analysis, we applied a fourth-order Butterworth digital band-pass filter (0.5–45 Hz) to remove residual high-frequency noise and slow baseline drift. Signals were screened prior to analysis to identify issues such as saturation, drift, or inactive channels. Each recording was exported as a binary file along with a synchronized annotation log created during surgery by the monitoring neurophysiologist. The binary format was decoded to extract signal values, timing information, and metadata. The processed signals were then converted into CSV format and aligned with the annotation logs to assign labels. Artifact segments caused by electrocautery, motion, or surgical manipulation were identified using simple amplitude and variability checks and excluded.

For analysis, each recording was divided into 30 non-overlapping intervals of 20 seconds each, corresponding to approximately 10 minutes per patient. If one or more channels were unavailable in an interval, features were computed using the remaining valid channels. Less than 2% of intervals had missing or corrupted channels and were excluded. All results are reported at the interval level. Ischemia labels were derived from intraoperative annotations recorded by the expert neurophysiologist during surgery.

Appendix B. cEEG Feature Construction and Representation

This section describes how we derive 111 quantitative features from each 20-second interval, which serve as inputs to all AI and hybrid models (Mina et al., 2024; Mina, 2024).

Preprocessing

cEEG signals from eight bipolar channels (F3–P3, P3–O1, F3–T3, T3–O1, F4–P4, P4–O2, F4–T4, T4–O2) were sampled at 500 Hz and filtered between 0.5 and 45 Hz using a fourth-order Butterworth filter. Each 20-second interval contains approximately 10,000 samples per channel. Signals were checked for artifacts such as clipping, flat signals, or high-amplitude noise. intervals with clear artifacts were removed. No re-referencing, resampling, or interpolation was applied.

Spectral estimation

Table 1: **Key details of cEEG data collection and preprocessing.**

Component	Details
Recording platform	XLTEK intraoperative cEEG system (Natus Medical Inc.).
Sampling rate	500 Hz continuous acquisition.
Electrodes	Disposable subdermal needle electrodes.
Electrode layout	Eight electrodes placed using the 10–20 system. Left hemisphere: F3–P3, P3–O1, F3–T3, T3–O1. Right hemisphere: F4–P4, P4–O2, F4–T4, T4–O2.
Signal representation	Bipolar derivations (pairwise voltage differences), no re-referencing applied.
Hardware filtering	Band-pass filter (0.5–70 Hz) and 60 Hz notch filter during acquisition.
Preprocessing filter	Fourth-order Butterworth band-pass (0.5–45 Hz).
Data format	Raw binary recordings converted to CSV and aligned with time-stamped annotations.
Intervals	Signals split into 20-second non-overlapping intervals (30 per patient).
Use in analysis	Input for 111 extracted features capturing amplitude, frequency, and inter-hemispheric differences.

For each interval, the power spectral density (PSD) was estimated using Welch's method with 2-second windows, 50% overlap, and 0.5 Hz resolution. Power values were converted to decibel scale to stabilize variation and support ratio-based features. From the PSD, we derive a set of core spectral and amplitude-based features.

Frequency bands

We use standard intraoperative frequency bands within the 0.5–45 Hz range: δ (0.5–4 Hz), θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), and γ (30–45 Hz). Band power is computed by integrating the PSD over each band:

$$P_b = \int_{f_1}^{f_2} \mathrm{PSD}(f)\, df.$$

These band powers form the basis for downstream features.

Band-power ratios

To capture slowing of brain activity, we compute three ratios:

$$\mathrm{ADR} = \frac{P_\alpha}{P_\delta},$$
$$\mathrm{BDR} = \frac{P_\beta}{P_\delta},$$
$$\mathrm{ABTR} = \frac{P_\alpha + P_\beta}{P_\delta + P_\theta}.$$

Lower values indicate a shift toward slower frequencies, which is commonly seen during ischemia.

Spectral edge frequency

We compute SEF_{90} as the frequency below which 90% of the total spectral power lies:

$$\int_{0.5}^{f^*} \mathrm{PSD}(f)\, df = 0.9 \int_{0.5}^{45} \mathrm{PSD}(f)\, df.$$

A decrease in SEF_{90} reflects overall slowing of the signal.

Amplitude-based feature

For each interval, we compute an amplitude-integrated measure by smoothing the rectified signal and taking the difference between its upper and lower envelope. This captures changes in overall signal amplitude, which may decrease during ischemia.

Hemispheric asymmetry

We quantify left–right differences using a symmetry index computed across paired channels. For each pair and frequency, we measure the normalized difference in power and average across all pairs and frequencies:

$$\mathrm{BSI} = \frac{1}{N} \sum \frac{|P_L(f) - P_R(f)|}{P_L(f) + P_R(f)}.$$

Higher values indicate stronger asymmetry, which is expected in unilateral ischemia.

Feature aggregation

The above features are computed for each individual channel, giving 80 features. We then repeat the same computations on averaged signals from the left hemisphere, right hemisphere, and all channels combined (30 additional features). Finally, one symmetry feature is added, resulting in a total of 111 features per interval. This setup preserves both local and global signal information.

Table 2: **Summary of extracted feature types.**

Feature group	**Description**
Band power	Energy within standard frequency bands (delta to gamma).
Spectral ratios	Ratios of higher-frequency to lower-frequency power (e.g., ADR, BDR).
Spectral edge	Frequency below which most signal energy is concentrated (SEF_{90}).
Amplitude measure	Range of smoothed signal amplitude within an interval.
Symmetry index	Degree of difference between left and right hemispheres.

Table 3: **Breakdown of feature counts per interval.**

Category	**Count**
Per-channel features (8 channels × 10 features)	80
Left hemisphere averages	10
Right hemisphere averages	10
Global average (all channels)	10
Symmetry index	1
Total	**111**

Together, these features capture key aspects of cEEG signals, including overall power, frequency shifts, amplitude changes, and hemispheric differences. This provides a compact but informative representation for downstream modeling. In addition to these 111 features, we derive a simple summary measure, termed the EEG variability score, which captures dispersion within each interval and is used as an input to L2A.

B.1. Derived EEG Variability Score

In the L2A formulation, each interval is represented by three inputs: the AI probability p_i, the AI uncertainty u_i, and an EEG variability score v_i. The variability score is a simple feature derived from the 111-dimensional feature vector $\mathbf{x}_i$, designed to capture overall dispersion and range.

For each 20-second interval i, let $\mathbf{x}_i$ denote the 111 features extracted from the raw EEG signal. We compute two measures of variability:

$$\mathrm{std}_i = \mathrm{std}(\mathbf{x}_i), \quad \mathrm{rng}_i = \max(\mathbf{x}_i) - \min(\mathbf{x}_i).$$

We compute normalization constants from the training set:

$$s_{\min} = \min_i \mathrm{std}_i, \quad s_{\mathrm{range}} = \max_i \mathrm{std}_i - \min_i \mathrm{std}_i,$$

$$r_{\min} = \min_i \mathrm{rng}_i, \quad r_{\mathrm{range}} = \max_i \mathrm{rng}_i - \min_i \mathrm{rng}_i.$$

We then normalize each quantity:

$$\mathrm{std}_i^{\mathrm{norm}} = \mathrm{clip}\left(\frac{\mathrm{std}_i - s_{\min}}{s_{\mathrm{range}} + \epsilon}, 0, 1\right), \quad \mathrm{rng}_i^{\mathrm{norm}} = \mathrm{clip}\left(\frac{\mathrm{rng}_i - r_{\min}}{r_{\mathrm{range}} + \epsilon}, 0, 1\right),$$

where ϵ is a small constant for numerical stability.

The EEG variability score is defined as:

$$v_i = \tfrac{1}{2}\,\mathrm{std}_i^{\mathrm{norm}} + \tfrac{1}{2}\,\mathrm{rng}_i^{\mathrm{norm}}.$$

Appendix C. Patient and Dataset Characteristics

The dataset included 400 patients who underwent CEA for symptomatic carotid stenosis between January 2009 and December 2017. Each case contained 30 non-overlapping 20-second intervals, resulting in 12,000 intervals. Patient age, sex, and side of surgery are

summarized in Table 4. These non-PHI characteristics are reported to describe the study population and support assessment of generalizability.

Table 4: Patient and dataset characteristics of the study cohort.

Characteristic	**Study cohort**
Data type	Intraoperative cEEG
Number of patients (cases)	400
Number of intervals	12,000
Intervals per case	30
Age, years	Median (IQR): 70 (64–77) Range: 23–96
Sex	Male: 206 (51.5%) Female: 194 (48.5%)
Side of surgery	Right: 214 (53.5%) Left: 186 (46.5%)
Indication for surgery	Symptomatic carotid stenosis: 400 (100%)
Study period	January 2009–December 2017
Label type	Intraoperative labels and novice labels
Role in study	Primary performance analyses, including ROC and precision–recall curves, sensitivity, false-positive rate, precision, and calibration

IQR, interquartile range.

Appendix D. Additional Novice Results

To assess whether the main findings generalize beyond the two representative novices shown in the main text, we evaluate the remaining two novices in this appendix. Figures 6 and 7 show that the same overall patterns hold across both individuals. In particular, deferral methods consistently improve over AI-only performance at low levels of novice involvement, and L2A remains among the strongest and most stable methods across metrics. These results indicate that the gains from human–AI deferral are not specific to a particular novice, but persist across individuals with different performance profiles.

For this novice, performance is broadly comparable to the AI on some metrics, while offering higher sensitivity. As a result, most deferral methods achieve clear gains at low deferral fractions, particularly in sensitivity and F1 score. L2A shows the strongest overall discrimination, reaching the highest AUPRC and AUROC and remaining stable across

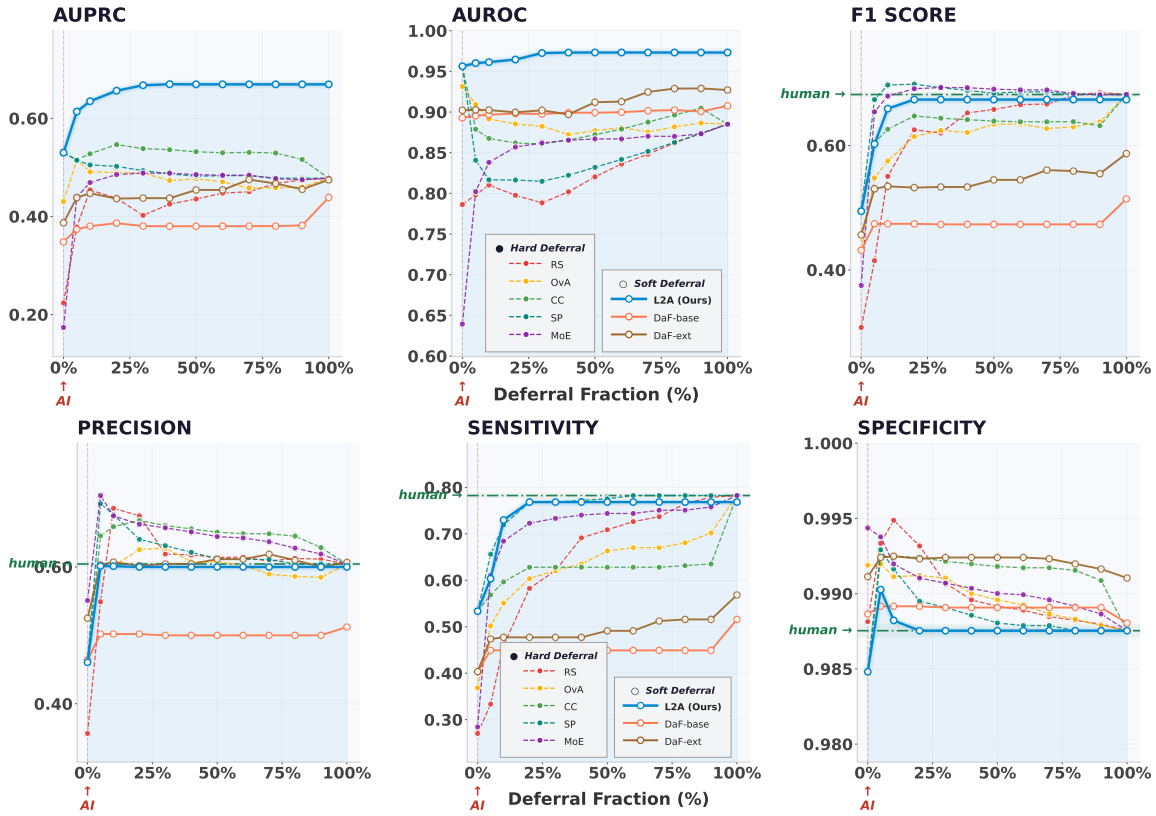

Figure 6: **Additional novice (Novice C): discriminative performance across deferral levels.** Deferral methods improve over AI-only performance at low deferral levels (5–10%), with L2A performing competitively and showing stable behavior across metrics.

deferral levels. Its gains occur primarily at low deferral, after which performance plateaus. In contrast, several hard deferral methods exhibit more variability as deferral increases. This suggests that useful complementary information is concentrated in a relatively small subset of intervals, and can be effectively leveraged with minimal novice involvement.

The second novice presents a different pattern, with mixed performance relative to the AI across metrics. Even in this setting, deferral methods are able to improve performance at low deferral levels. L2A again shows the strongest and most stable behavior, with consistent gains in AUPRC, AUROC, and F1 score. Hard deferral methods exhibit more variable trends as deferral increases, while soft deferral methods remain stable. This indicates that the benefits of soft deferral extend beyond settings with clearly strong or weak novices, and also apply when strengths and weaknesses vary across metrics.

Overall, these additional results reinforce three main observations. First, most of the gains from human–AI collaboration are achieved with small amounts of novice involvement (5–10%). Second, performance depends not only on overall novice quality, but also on whether the novice provides complementary information on a subset of intervals. Third,

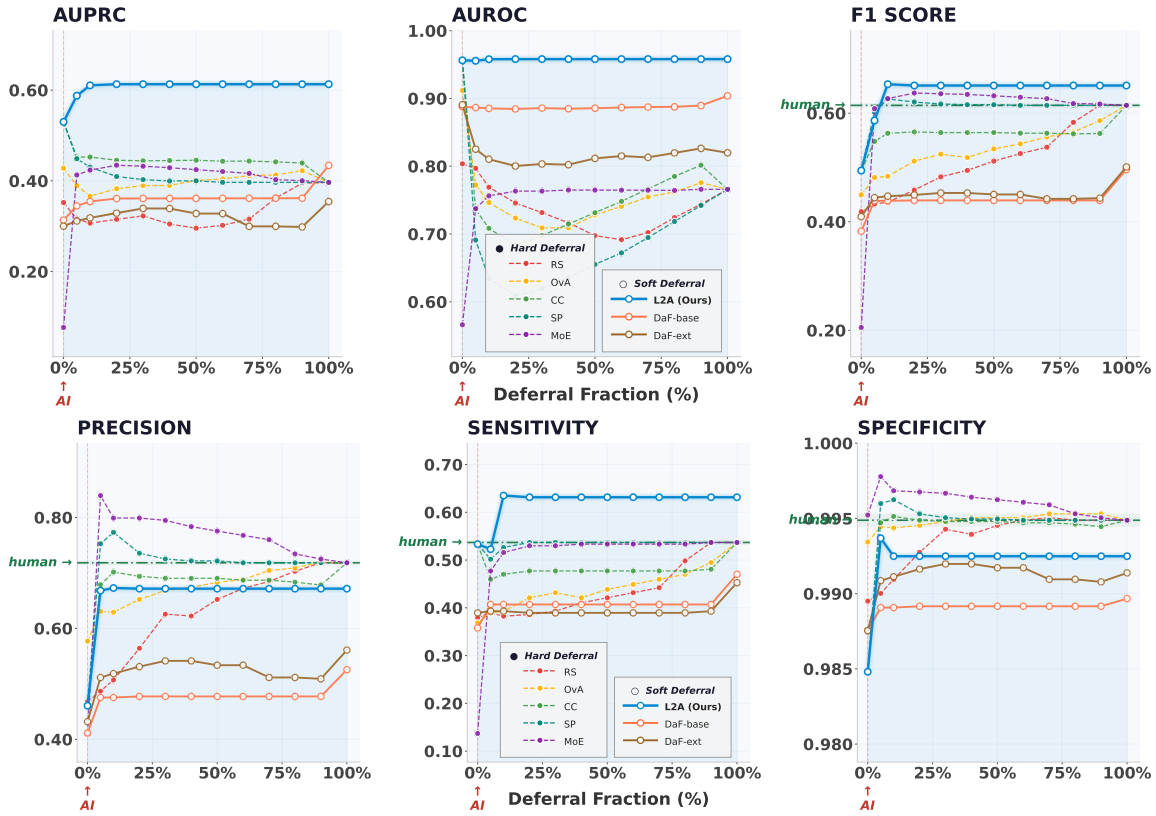

Figure 7: **Additional novice (Novice D): discriminative performance across deferral levels.** Despite mixed performance relative to the AI, deferral methods, particularly L2A, achieve consistent improvements at low deferral levels and remain stable as deferral increases.

soft deferral methods remain more stable than hard deferral methods as deferral increases, as they can downweight unreliable human input rather than fully replacing the AI. These findings further support the generality of the main results across different novice behaviors.

To complement the discriminative results for the two additional novices, Figures 8 and 9 show calibration across deferral levels using expected calibration error (ECE), log loss, and Brier score. The same broad pattern seen in the main text is observed here as well. Soft deferral methods remain more stable across deferral levels, whereas hard deferral methods show greater sensitivity to increasing novice involvement. In both additional novices, L2A maintains the lowest or among the lowest values across all three calibration metrics over most of the deferral range, indicating strong and stable calibration.

For Novice C, hard deferral methods show noticeable changes in calibration as deferral increases, particularly in log loss and Brier score. Some methods improve briefly at low deferral, but then become less stable as more novice input is incorporated. In contrast, soft deferral methods remain relatively steady across the full deferral range. Among them, L2A maintains particularly low ECE, log loss, and Brier score throughout, with only small

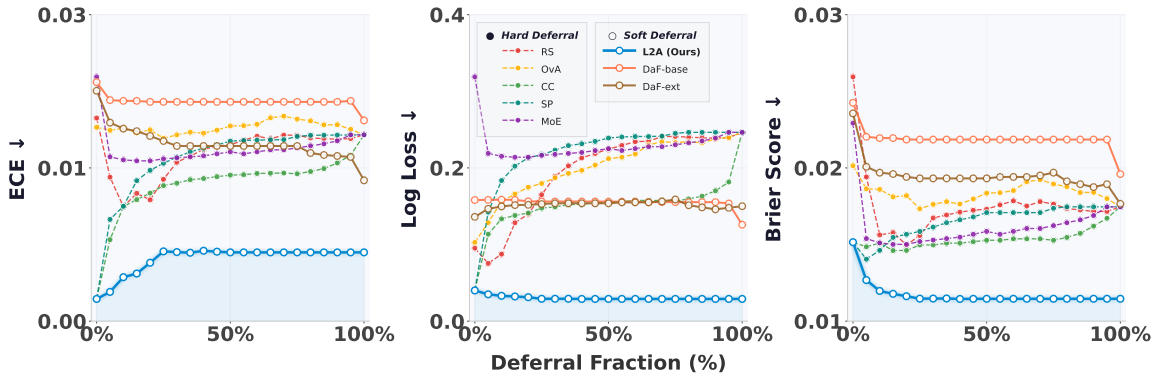

Figure 8: **Additional novice (Novice C): calibration across deferral levels.** Calibration is evaluated using ECE, log loss, and Brier score. Soft deferral methods show more stable calibration across deferral levels, while hard deferral methods vary more as novice involvement increases.

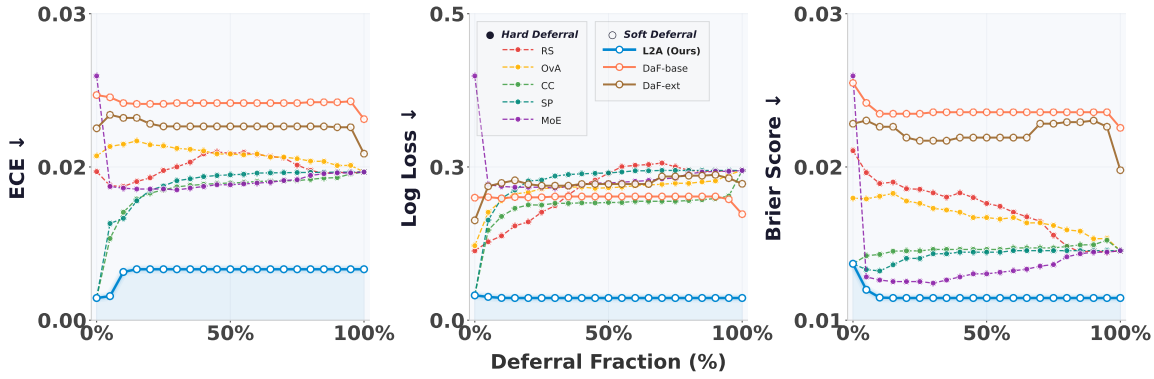

Figure 9: **Additional novice (Novice D): calibration across deferral levels.** Calibration is evaluated using ECE, log loss, and Brier score. Similar to Novice C, soft deferral methods remain more stable across deferral levels, whereas hard deferral methods show larger variation as novice involvement increases.

changes as deferral increases. This indicates that, for this novice, combining AI and novice input probabilistically preserves calibration more effectively than binary routing.

For Novice D, the same general behavior is observed, although the separation between methods is somewhat larger. Hard deferral methods show greater variation across deferral levels, especially in ECE and log loss, while soft deferral methods remain more stable. L2A again maintains very low values across all three calibration metrics and is largely unaffected by increasing novice involvement. These results further support the observation that soft deferral methods are more robust when novice predictions are incorporated progressively, and that L2A provides consistently strong calibration across different novice behaviors.

Overall, the calibration results for these two additional novices are consistent with the main paper. While discriminative performance varies across novices, the calibration advantage of soft deferral remains stable. This suggests that the improved calibration of L2A is not limited to the representative novices shown in the main text, but generalizes across a broader range of novice performance profiles.

D.1. Average Performance Across All Novices

To summarize performance across novice variability, Figure 10 reports the mean and sample standard deviation across all four novice monitors at each deferral fraction. For L2A, the 0% deferral point corresponds to the AI model alone, which achieved an AUPRC of 0.530, AUROC of 0.956, precision of 0.461, and sensitivity of 0.533. With only 5% novice involvement, the hybrid novice–AI system achieved an AUPRC of 0.614 ± 0.029, AUROC of 0.959 ± 0.003, precision of 0.643 ± 0.050, and sensitivity of 0.610 ± 0.063. These correspond to relative improvements over the AI model of 15.8% in AUPRC, 39.6% in precision, and 14.3% in sensitivity. At 10% novice involvement, L2A achieved an AUPRC of 0.636 ± 0.034, AUROC of 0.961 ± 0.003, precision of 0.644 ± 0.051, and sensitivity of 0.723 ± 0.063. Compared with the AI model alone, this represents absolute improvements of 0.106 in AUPRC, 0.005 in AUROC, 0.183 in precision, and 0.189 in sensitivity, corresponding to relative improvements of 19.9%, 0.5%, 39.8%, and 35.5%, respectively. These averaged results show that the benefits of the hybrid novice–AI system persist across all four novices and are not driven by a single high-performing individual.

Appendix E. Novice Label Variability and Agreement

We assessed consistency among novices using Cohen's κ, which accounts for agreement expected by chance, along with raw percent agreement. Because ischemic intervals are rare ($285/12000 = 2.38\%$), overall agreement can be misleadingly high. To address this, we also report agreement separately for true positives and true negatives, providing a clearer view of disagreement patterns.

Consistency across novices.

Across the 12,000-interval dataset, novices showed moderate to substantial agreement after correcting for chance (mean $\kappa = 0.61 \pm 0.11$ across six pairs), while raw agreement was very high (mean $97.34\% \pm 1.29\%$), this is driven mainly by agreement on non-ischemic intervals (Table 5). Agreement on true negatives was consistently high ($97.85\% \pm 1.41\%$), whereas agreement on true positives was notably lower ($76.37\% \pm 9.71\%$). This indicates that ischemic intervals are the main source of disagreement, while non-ischemic intervals are relatively easy to identify. Pairwise results further highlight this variability (Table 6). Cohen's κ ranges from approximately 0.44 to 0.76 across novice pairs, confirming that agreement is not uniform across individuals.

There was also clear variability across novice pairs, showing that novices do not behave identically. Some are more conservative, while others are more sensitive to potential ischemia.

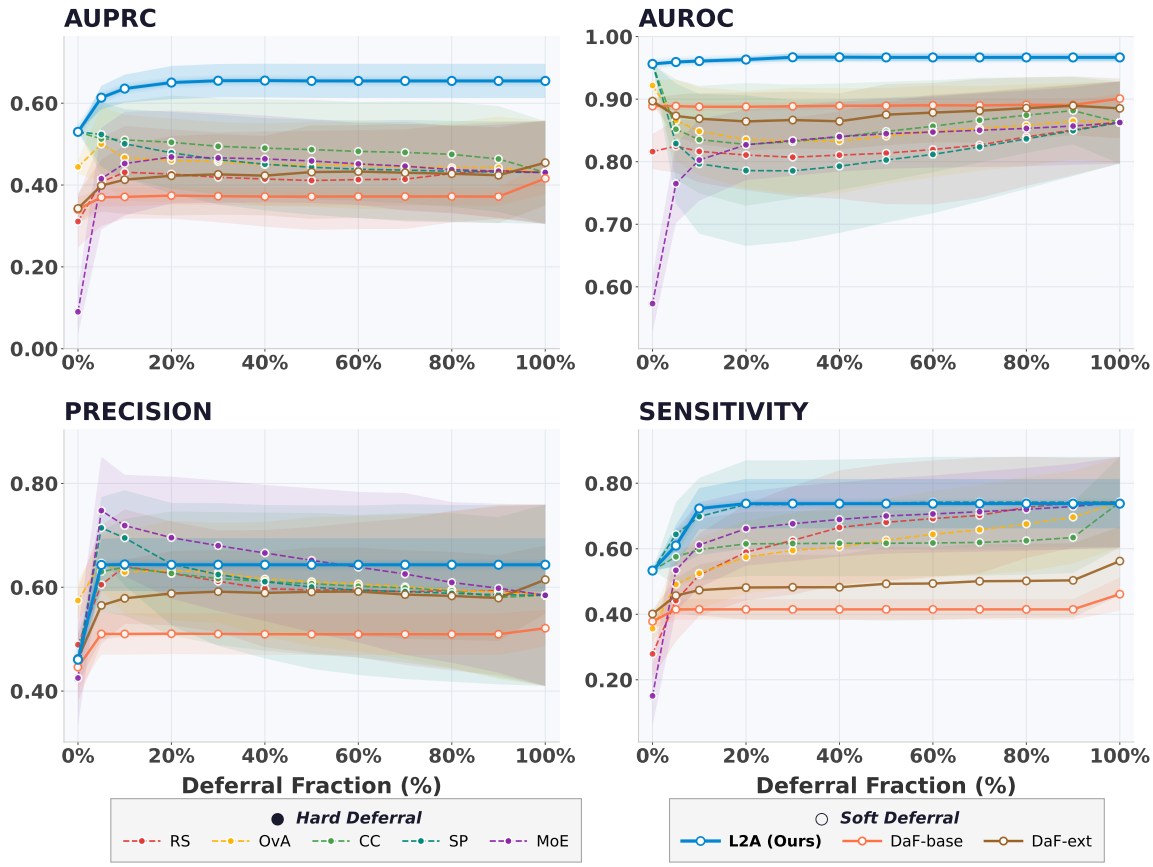

Figure 10: **Average discriminative performance across all four novice monitors.** Lines show the mean performance across novices at each deferral fraction, and shaded regions show ±1 sample standard deviation. At 0% deferral, L2A operates as the AI model alone. With only 5–10% novice involvement, the hybrid novice–AI system substantially improves mean AUPRC, precision, and sensitivity over the AI model while maintaining high AUROC.

This variability is important for our setting: it highlights that human input is heterogeneous, and any human–AI deferral system must be able to adapt to these differences rather than rely on a single fixed pattern of behavior.

Table 5: **Agreement statistics among novices (mean ± SD across 6 pairs).**

Measure	Value
Cohen's κ	0.61 ± 0.11
Agreement on positives	$76.37\% \pm 9.71\%$
Agreement on negatives	$97.85\% \pm 1.41\%$

Table 6: **Pairwise Cohen's κ among novices (dataset-1, $n = 12{,}000$ intervals).**

Novice pair	Cohen's κ
N1 vs N4	0.71
N1 vs N2	0.76
N1 vs N3	0.55
N4 vs N2	0.61
N4 vs N3	0.44
N2 vs N3	0.59

Appendix F. Cross-Novice Transfer and Monitor Onboarding

To assess whether L2A requires a separate deferral model for each novice, we trained the model using labels from one novice and evaluated it using labels from another. We considered all 12 ordered train–test pairs across the four novices and compared transferred models with models trained and evaluated on the same novice.

Cross-novice transfer showed little degradation. At 10% deferral, within-novice and transferred models achieved mean AUPRC values of 0.636 and 0.641, respectively; AUROC values were 0.961 for both, precision was 0.657 and 0.658, and sensitivity was 0.683 and 0.676. At 5% deferral, AUPRC was 0.614 within novice and 0.617 under transfer. Transfer effects were asymmetric: transfer from weaker to stronger novices sometimes improved performance, whereas transfer from stronger to weaker novices produced small decreases. Among pairs with decreased performance, the mean reductions were 0.026 in AUPRC and 0.0042 in AUROC. This transferability may arise because L2A is trained using novice-specific outcomes but relies on AI probability, uncertainty, and cEEG complexity as inputs, allowing it to identify regions where human input is broadly useful.

F.1. Monitor Onboarding and Data Efficiency

We quantified how much monitor-labeled data is needed to fit an L2A policy for a new novice monitor. Within each fixed 10-fold patient-level cross-validation split, approximately 360 of the 400 cases were available for training, while the remaining 40 cases were held out for testing. Therefore, 360 cases represents 100% of the available training data within each fold, not the full 400-case cohort. We subsampled complete cases from the training fold while preserving ischemic cases. The held-out test cases and AI predictions remained fixed; only the lightweight L2A policy was refitted. Each case required approximately 10 minutes of cEEG review.

At 10% deferral, L2A trained with 72 labeled cases per novice monitor ($\sim$12 hours) achieved an AUPRC of 0.627 ± 0.039, precision of 0.710 ± 0.075, and sensitivity of 0.590 ± 0.046. This was similar to using all 360 cases available for training within each fold, which achieved an AUPRC of 0.631 ± 0.032, precision of 0.720 ± 0.062, and sensitivity of 0.576 ± 0.054. Thus, 72 cases recovered approximately 96% of the full-training-fold AUPRC gain over the AI model. At 5% deferral, 72 cases achieved an AUPRC of 0.610 ± 0.036, compared with 0.612 ± 0.029 using all available training cases.

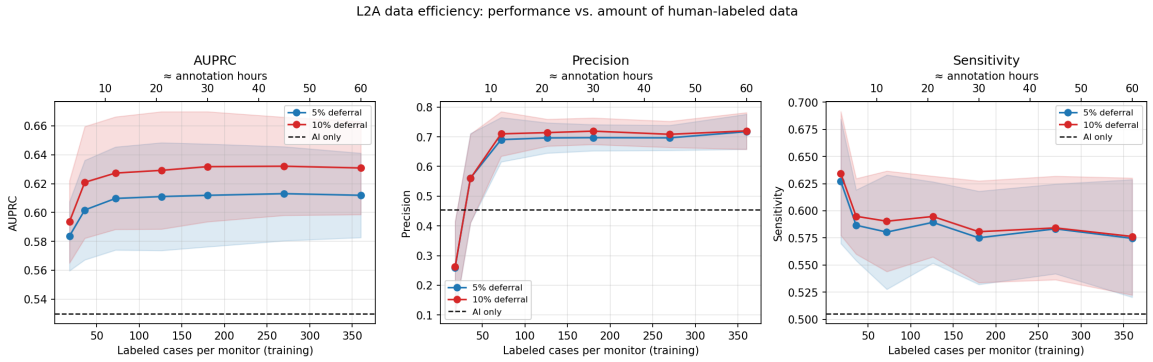

Figure 11: **L2A data efficiency for novice-monitor onboarding.** Performance at 5% and 10% deferral is shown as the amount of monitor-labeled training data increases. Points show means across four novices and five subsampling seeds, and shaded regions show ±1 SD. For the 360-case condition, all cases available for training within each cross-validation fold were used, while approximately 40 cases remained held out for testing. Because no subsampling was performed in this condition, it contains one run per novice. Dashed lines show the AI-only baseline.

Table 7: **Performance by amount of monitor-labeled training data.** Values are mean ± SD. Reduced-data conditions contain 20 runs across four novices and five subsampling seeds. The 360-case condition represents 100% of the available training data within each cross-validation fold and contains one run per novice.

		5% deferral			10% deferral		
Training cases	**Hours**	**AUPRC**	**Precision**	**Sensitivity**	**AUPRC**	**Precision**	**Sensitivity**
AI only	0	0.530	0.454	0.505	0.530	0.454	0.505
18	3	0.584 ± 0.024	0.260 ± 0.155	0.627 ± 0.057	0.594 ± 0.029	0.262 ± 0.154	0.634 ± 0.057
36	6	0.602 ± 0.034	0.561 ± 0.150	0.587 ± 0.033	0.621 ± 0.039	0.560 ± 0.149	0.595 ± 0.035
72	**12**	$\mathbf{0.610 \pm 0.036}$	$\mathbf{0.691 \pm 0.075}$	$\mathbf{0.580 \pm 0.053}$	$\mathbf{0.627 \pm 0.039}$	$\mathbf{0.710 \pm 0.075}$	$\mathbf{0.590 \pm 0.046}$
126	21	0.611 ± 0.037	0.696 ± 0.051	0.589 ± 0.037	0.629 ± 0.041	0.714 ± 0.045	0.595 ± 0.037
180	30	0.612 ± 0.036	0.697 ± 0.044	0.575 ± 0.043	0.632 ± 0.038	0.719 ± 0.045	0.581 ± 0.047
270	45	0.613 ± 0.033	0.696 ± 0.042	0.583 ± 0.041	0.632 ± 0.034	0.708 ± 0.044	0.584 ± 0.048
360	60	0.612 ± 0.029	0.717 ± 0.059	0.575 ± 0.054	0.631 ± 0.032	0.720 ± 0.062	0.576 ± 0.054

Performance largely plateaued after approximately 72 cases. Although 18–36 cases already improved AUPRC, precision was substantially more variable at these smaller sample sizes. These findings suggest a practical onboarding strategy in which a new monitor begins with a transferred L2A policy and optionally personalizes it using approximately 72 labeled cases. These labels are used only to fit the monitor-specific deferral policy; the AI model does not require retraining.

Appendix G. Baseline Implementation Details

We implemented RS, OvA, CC, SP, and MoE by adapting the public human–AI deferral benchmark.[1] DaF-base and DaF-ext were adapted from the authors' BeyondDefer repository,[2] where the `BeyondDefer` and `AdditionalBeyond` classes correspond to DaF-base and DaF-ext, respectively. Each method was trained separately for each novice monitor using identical patient-level folds.

For interval i, let $\mathbf{x}_i \in \mathbb{R}^{111}$ denote the cEEG features, $y_i \in \{0,1\}$ the ground-truth label, $h_i \in \{0,1\}$ the novice prediction, and $m_i = \mathbb{I}(h_i = y_i)$ an indicator that the novice is correct. For each test fold, the following fold was used for validation and the remaining eight folds were used for training. All 30 intervals from a case remained in the same fold.

Fixed AI model. CC and SP used the same precomputed patient-level out-of-fold AI probabilities as L2A. The AI model was a Random Forest classifier with 1000 trees, a maximum depth of 15, and random seed 42; the remaining classifier settings used their default values, with no class weighting. Let $p_i = P(y_i = 1 \mid \mathbf{x}_i)$ and $c_i^A = \max(p_i, 1 - p_i)$ denote its ischemia probability and confidence. No interval received a probability from a Random Forest model trained on its own case. Because CC and SP used these fixed probabilities, their AI components were not retrained within the deferral code.

Neural architecture and training. RS, OvA, MoE, DaF-base, and DaF-ext learned their own AI components from the 111 cEEG features. Each feed-forward component used two hidden layers with 256 and 128 units, respectively. Each hidden layer was followed by batch normalization, a ReLU activation, and dropout with probability 0.3. Linear weights used Kaiming initialization and biases were initialized to zero. For the DaF fusion component, a one-hot representation of the novice prediction was additionally provided as input.

Missing feature values were replaced using medians estimated from the training folds. Features were then standardized using training-fold means and standard deviations and clipped to $[-10, 10]$. Models were trained for 100 epochs using Adam with a learning rate of 5×10^{-4} and batch size 256. No learning-rate scheduler, weight decay, class weighting, or balanced sampling was used. Gradients were clipped elementwise to $[-1, 1]$, and all experiments used random seed 42. The validation fold was used for model selection where required; test labels were never used for checkpoint selection.

Realizable Surrogate. RS produced two class logits, g_{i0} and g_{i1}, and one deferral logit, $g_{i\perp}$. We used the realizable-surrogate loss with $\alpha = 1$:

$$\mathcal{L}_{\mathrm{RS}} = -\log_2 \left[\frac{\exp(g_{iy_i}) + m_i \exp(g_{i\perp})}{\exp(g_{i0}) + \exp(g_{i1}) + \exp(g_{i\perp})} \right].$$

No additional α search or rejection cost was used. Applying softmax over the three outputs gives π_{i0}, π_{i1}, and $\pi_{i,\perp}$. Intervals were ranked by $\pi_{i,\perp} - \max_{k\in\{0,1\}} \pi_{ik}$. The AI probability used for evaluation was obtained by applying softmax to the two class logits alone.

1. https://github.com/clinicalml/human_ai_deferral
2. https://github.com/AminChrs/BeyondDefer

Table 8: Scores used to rank intervals for novice involvement. Larger scores indicate greater preference for obtaining novice input.

Method	Ranking score	Output when selected
RS and OvA	$\pi_{i,\perp} - \max_{k\in\{0,1\}} \pi_{ik}$	Novice prediction h_i
CC	$q_i - c_i^A$	Novice prediction h_i
SP	$1 - c_i^A$	Novice prediction h_i
MoE	Learned gate probability g_i	Novice prediction h_i
DaF-base	$c_i^F - c_i^A$	Fusion probability $P_F(y_i = 1 \mid \mathbf{x}_i, h_i)$
DaF-ext	$\max(c_i^H, c_i^F) - c_i^A$	Novice prediction if $c_i^H \geq c_i^F$; fusion probability otherwise

One-vs-All Surrogate. OvA used the same three outputs but trained them using independent logistic losses. Defining $\ell(z,t) = \log_2(1 + \exp(-tz))$, the loss was

$$\mathcal{L}_{\text{OvA}} = \ell(g_{iy_i}, 1) + \sum_{k \neq y_i} \ell(g_{ik}, -1) + m_i \ell(g_{i\perp}, 1) + (1 - m_i)\ell(g_{i\perp}, -1).$$

The deferral score was $\pi_{i,\perp} - \max_{k\in\{0,1\}} \pi_{ik}$. OvA did not use an additional rejection-cost parameter.

Compare Confidence. CC retained the fixed Random Forest AI probability and trained a separate two-class network $P_H(h_i \mid \mathbf{x}_i)$ to predict the novice's label, using cross-entropy with the novice prediction h_i as its target. Let

$$q_i = \max_{k\in\{0,1\}} P_H(h_i = k \mid \mathbf{x}_i)$$

denote the confidence of this novice-simulation model. CC ranked intervals by $q_i - c_i^A$, deferring where the novice's prediction was modeled more confidently than the AI's own class prediction. When an interval was selected, the observed novice prediction h_i was used.

Selective Prediction. SP did not model novice behavior. It ranked intervals only by AI uncertainty, $1 - c_i^A = 1 - \max(p_i, 1 - p_i)$. Therefore, its deferral pattern depended only on the fixed AI probability and was identical across novices, although its final performance depended on the novice predictions used on selected intervals.

Mixture-of-Experts. MoE produced two class logits and one gate logit. Let a_{i,y_i} denote the softmax probability assigned to the correct class and $g_i \in [0,1]$ the sigmoid-transformed gate output. Its loss was

$$\mathcal{L}_{\text{MoE}} = (1 - g_i)\big[-\log_2 a_{i,y_i}\big] + g_i(1 - m_i).$$

The first term penalizes an incorrect AI prediction, while the second penalizes reliance on an incorrect novice. Although trained using a soft gate, MoE was evaluated as a hard-deferral method. Intervals were ranked by g_i, and selected intervals used the novice prediction.

Defer-and-Fusion. DaF used three learned components: an AI classifier $P_A(y \mid \mathbf{x})$, a novice-simulation model $P_H(h \mid \mathbf{x})$, and a fusion model $P_F(y \mid \mathbf{x}, h)$. The fusion model received the cEEG features and a one-hot representation of the novice prediction. The AI and expected fusion confidences were

$$c_i^A = \max_y P_A(y \mid \mathbf{x}_i)$$

and

$$c_i^F = \sum_{j \in \{0,1\}} P_H(h_i = j \mid \mathbf{x}_i) \max_y P_F(y \mid \mathbf{x}_i, h_i = j).$$

DaF-base ranked intervals by $c_i^F - c_i^A$. When an interval was selected, the observed novice prediction was supplied to the fusion model and the final output was $P_F(y_i = 1 \mid \mathbf{x}_i, h_i)$.

DaF-ext added a direct-novice option. An additional output estimated novice correctness, denoted by c_i^H. It ranked intervals by $\max(c_i^H, c_i^F) - c_i^A$. For a selected interval, the novice prediction was used directly when $c_i^H \geq c_i^F$; otherwise, the fusion probability was used.

Following the authors' implementation, the DaF components were trained jointly using summed one-vs-rest binary cross-entropy losses:

$$\mathcal{L}_{\text{DaF-base}} = \text{BCE}_{\text{ovr}}(A(\mathbf{x}_i), y_i) + \text{BCE}_{\text{ovr}}(H(\mathbf{x}_i), h_i) + \text{BCE}_{\text{ovr}}(F(\mathbf{x}_i, h_i), y_i).$$

For DaF-ext, the AI component contained an additional output with $m_i = \mathbb{I}(h_i = y_i)$ as its target. No additional fusion or consultation cost was used because novice involvement was controlled through the common fixed-fraction evaluation.

Fixed deferral fractions. We evaluated deferral fractions $\rho \in \{0, 0.05, 0.10, \ldots, 1.00\}$. Within each held-out fold containing n intervals, the round(ρn) intervals with the largest method-specific scores were selected. The score threshold was therefore the empirical fold-specific quantile required to attain the specified fraction; ground-truth labels were not used to choose selected intervals.

For hard-deferral methods, the final system probability was

$$\widehat{p}_i = \begin{cases} h_i, & \text{if interval } i \text{ was selected,} \\ a_i, & \text{otherwise,} \end{cases}$$

where a_i is the method's AI probability of ischemia. DaF-base and DaF-ext instead used the fusion probability when fusion was selected. Predictions and probabilities from the ten held-out folds were pooled before computing the reported metrics. Because the neural methods learned their own AI components, their values at 0% deferral correspond to their respective neural AI components, whereas CC and SP use the fixed Random Forest AI model.

Appendix H. Ethics Approval and Data Privacy

This retrospective study was approved by the Institutional Review Board under protocol STUDY18120055 ("Significance of Neurophysiological Changes"). The study received expedited approval and was determined to involve no greater than minimal risk. The analysis

used previously collected intraoperative cEEG recordings from patients undergoing CEA between 2009 and 2017. All data were de-identified before research analysis. The novice monitors accessed only de-identified cEEG data and did not have access to patient identifiers or protected health information.

Appendix I. Code Availability

The code supporting this work is publicly available at https://github.com/batmanlab/MLHC2026-Human-AI-Deferral/. The repository includes implementations for cEEG feature generation, model training, and the L2A deferral system. Patient-level data cannot be shared due to privacy and institutional restrictions. The code can be used to reproduce the analysis on independently approved cEEG datasets.