Check for updates

A Bayesian dynamic stopping method for evoked response brain-computer interfacing

Sara Ahmadi[1]*, Peter Desain[1,2] and Jordy Thielen[1]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands,
[2]MindAffect, Ede, Netherlands

Introduction: As brain-computer interfacing (BCI) systems transition fromassistive technology to more diverse applications, their speed, reliability, and user experience become increasingly important. Dynamic stopping methods enhance BCI system speed by deciding at any moment whether to output a result or wait for more information. Such approach leverages trial variance, allowing good trials to be detected earlier, thereby speeding up the process without significantly compromising accuracy. Existing dynamic stopping algorithms typically optimize measures such as symbols per minute (SPM) and information transfer rate (ITR). However, these metrics may not accurately reflect system performance for specific applications or user types. Moreover, many methods depend on arbitrary thresholds or parameters that require extensive training data.

Methods: We propose a model-based approach that takes advantage of the analytical knowledge that we have about the underlying classification model. By using a risk minimization approach, our model allows precise control over the types of errors and the balance between precision and speed. This adaptability makes it ideal for customizing BCI systems to meet the diverse needs of various applications.

Results and discussion: We validate our proposed method on a publicly available dataset, comparing it with established static and dynamic stopping methods. Our results demonstrate that our approach offers a broad range of accuracy-speed trade-offs and achieves higher precision than baseline stopping methods.

KEYWORDS

Bayes test, brain-computer interfacing (BCI), signal detection theory, dynamic stopping, early stopping, visual evoked potentials (VEP), c-VEP

1 Introduction

The field of brain-computer interfacing (BCI) has witnessed remarkable advancements in recent years, offering exciting opportunities to enable novel communication pathways between the human brain and external devices. Typically, BCIs rely on electroencephalography (EEG), because it is a neuroimaging technique that can non-invasively capture electrical activity generated by the brain and it is accessible for a broad range of applications. While being practical, the adoption of EEG in BCI also introduces significant challenges, primarily stemming from an inherently low signal-to-noise ratio (SNR) associated with EEG recordings. This makes it difficult to achieve a high classification accuracy and speed in EEG-based BCI systems, leading researchers to navigate intricate terrain in neuroscience, signal processing, and machine learning to harness the full potential of this technology.

The selection of the most relevant brain signal features in EEG, the so-called neural signatures, is a pivotal aspect of BCI design, dictating the efficacy and responsiveness of the system. Neural signatures in BCI can be broadly categorized into evoked and

oscillatory signals. Evoked signals, including the event-related potential (ERP), steady-state visually evoked potential (SSVEP), and code-modulated visual evoked potential (c-VEP), represent distinct neural responses that can be reliably triggered by external stimuli and have achieved very high performances in BCIs for communication and control (Gao et al., 2014).

An ERP is characterized by transient changes in electrical activity time-locked to specific external events. For BCI, typically the P300 ERP is used (Fazel-Rezai et al., 2012). On the other hand, the SSVEP is often used, which is hypothesized to capitalize on the brain's entrainment to visual stimuli flickering at specific frequencies (Tsoneva et al., 2015), so-called frequency-tagging (Vialatte et al., 2010). Finally, the c-VEP, a relatively new but promising neural signature, involves encoding information in the pseudo-random visual stimuli to evoke specific responses, so-called noise-tagging (Martínez-Cagigal et al., 2021). The exploration of these evoked signals holds the key to refining the precision and versatility of more effective and adaptive BCI.

Despite the promising potential of evoked signals in BCI, the persistent challenge of a low SNR in EEG measurements remains a significant hurdle. One strategy to mitigate this limitation involves extending the duration of the visual stimulation, allowing for a more robust extraction of evoked responses from the EEG signal. Indeed, it seems that BCIs that rely on visual evoked potential (VEP) paradigms do not suffer from BCI inefficiency (Volosyak et al., 2020). However, this remedy does introduce a critical trade-off between achieving a high classification accuracy and maintaining a desirable level of speed when using a BCI. The delicate balance between trial duration and classification accuracy necessitates careful consideration in BCI design, as trials with prolonged stimulation periods may impede the real-time responsiveness required for seamless and efficient user control.

The optimal balance between classification accuracy and speed in BCI systems is inherently dependent on the specific application at hand. As the spectrum of BCI applications continues to diversify, it becomes imperative for these interfaces to cater to a range of performance requirements tailored to distinct purposes. Consider, for instance, a BCI designed for a specific brain-controlled alarm to be raised by a patient in case of emergency. In this context, next to the overall high classification accuracy, minimizing the misses, i.e., the false negative rate, is of most importance. Conversely, in the context of a BCI developed for communication, a higher speed may take precedence, allowing for real-time interaction even at the expense of slightly lower accuracy, as they can be corrected for by the user or by post-processing the output for example using a language model (Gembler et al., 2019a,b). Thus in the scenario of a BCI speller, the inherent trade-off between accuracy and speed becomes a nuanced consideration, emphasizing the need for adaptive BCI systems that can be tailored to the specific demands of varied applications. Only then the utility across diverse application domains and users can be maximized.

While many conventional BCI systems adhere to fixed trial lengths, allowing users to issue commands only at predetermined intervals and leveraging a constant amount of EEG data, more sophisticated methods have emerged to optimize trial duration dynamically (Schreuder et al., 2013). Within this domain, two distinct categories, namely static stopping and dynamic stopping,

have garnered attention for their potential to enhance BCI performance. In static-stopping approaches, also referred to as fixed-stopping, the system learns an optimal stopping time based on training data, establishing a uniform termination point for all trials during online use. On the contrary, dynamic stopping, also known as early stopping or adaptive stimulation, introduces a (potentially calibrated) trial-by-trial decision-making process. In this paradigm, the system evaluates the level of confidence attained during each trial and determines whether the trial should be stopped and the classification be emitted. It is most useful if the set criterion and the performance reached are interpretable measures, like the expected probability of a mistake, accuracy, etc. If so, we call the method inherently calibrated, in contrast to methods that use an arbitrary score and a criterion that has to be gathered from training data. Stopping procedures introduce a valuable dimension to BCI design, offering avenues to minimize selection or stimulation duration while optimizing relevant performance metrics, thereby contributing to the continued evolution and refinement of BCI technology.

The literature on ERP BCIs has seen the proposal of various stopping methods, as extensively reviewed and compared by Schreuder et al. (2013). One straightforward approach mandates a minimum number of consecutive iterations with identical class predictions before finalizing the classification (Jin et al., 2011). Another method accumulates classification scores over iterations per class, accepting the classification if the largest sum surpasses a learned threshold (Liu et al., 2010). In contrast, Lenhardt et al. (2008) employ a dual-criteria system, requiring both the sum of classification scores to reach a learned threshold and the ratio of the best versus the second-best score to exceed a learned threshold, a multiplicative version of a margin rule. Schreuder et al. (2011) propose a method that tests whether the rank difference between the best and second-best median scores is greater than a learned threshold. Zhang et al. (2008) accumulate evidence for each class as the attended one, accepting the classification if the posterior probability of a class reaches a predetermined threshold. Höhne et al. (2010) utilize Welch's t-test on the scores of the best versus all other classes and emit the classification when significance is reached. Throckmorton et al. (2013) use a kernel density estimate to determine the probability density function for each class, accepting the classification if a predetermined threshold is reached. Finally, Bianchi et al. (2019) consider the distribution of scores with respect to the separating hyperplane.

Various stopping procedures have been proposed in the literature to optimize the termination of trials in c-VEP BCIs. One straightforward approach involves thresholding the maximum correlation ($\rho_1 \geq 0.4$) without the need for additional calibration (Spüler, 2017). Other non-calibrated methods fit at each potential stopping time a Beta (Thielen et al., 2017, 2021) or Normal (Martínez-Cagigal et al., 2023a,b) distribution to non-maximum correlations, and threshold ρ_1 based on the likelihood that it is an outlier in the distribution. Conversely, calibrated methods employ logistic regression to map the correlation vector to correct or incorrect classifications, with the resulting probability thresholded to emit a classification (Sato and Washizawa, 2015). Other calibrated procedures use the margin between the best and second-best class ($\rho_1 - \rho_2$), i.e., the difference between the

top two correlations (Thielen et al., 2015; Verbaarschot et al., 2021; Gembler and Volosyak, 2019; Gembler et al., 2020) or the distance to the decision hyperplane of a one-class support vector machine (Riechmann et al., 2015). Finally, two methods integrate classifications over time, one utilizing *p*-values of correlations with a learned threshold (Nagel and Spüler, 2019), and the other emitting a classification after a predefined number of consecutive identical classifications (Castillos et al., 2023). The wild variety of stopping procedures on the one hand contributes to the refinement of c-VEP BCIs, offering flexibility in adapting trial durations to optimize performance metrics. On the other hand, the differences make their evaluation hard and makes one desire the availability of a method that is best in all occasions.

While the aforementioned stopping methods in principle optimize BCI performance, they are severely limited. Most of them dependent on hyper-parameters that necessitate optimization with training data. Typically, this involves performing a cross-validated grid-search with candidate hyper-parameter values, and selecting those that maximize a performance metric like classification accuracy, information transfer rate (ITR), or symbols per minute (SPM). However, a potential drawback emerges from the user's perspective, as the chosen performance metric may not accurately reflect the system's efficacy for a specific application or user type. As mentioned earlier, in some applications the cost of different types of errors can be different. This means that the types of errors that the stopping method makes, regardless of its average performance, contribute to the effectiveness of the methods in terms of user-oriented efficiency. Therefore a more suitable approach is to optimize risk rather than average accuracy or average trial duration.

Another limitation of the existing stopping methods is that they only decide based on the confidence of some similarity score and ignore possible differences between the prior probability of each output. For example in a speller system, the trials selecting frequent letters can be decided based on a much lower confidence score compared to non-frequent letters.

To address these limitations, we propose a dynamic stopping method that focuses on minimizing the risk associated with each type of error. Our proposed method is novel as the problem of dynamic stopping has never been addressed with a model-based approach that can be intuitively tuned and control for different error types. The model also makes it possible to incorporate different prior probabilities for target outputs. Through an analytical model-based approach, we offer a more tailored and user-centric optimization that allows one to control the system's behavior using the algorithm's adjustable parameters. In turn, the BCI can be aligned with the nuanced requirements of different user preferences and applications.

In this paper, firstly, we show the full mathematical derivation of the proposed Bayesian dynamic stopping (BDS) method. Secondly, using a published open-access c-VEP dataset, we compare the proposed algorithm with other stopping methods and do so both in terms of conventional accuracy-time metrics as well as relevance-based performance criteria such as precision, recall and F-score. Despite that the model allows for non-equal priors, in this study, we focus on the risk optimization aspect and limit the formulation to equal prior scenarios.

2 Materials and methods

2.1 Dataset

We assessed the effectiveness of the proposed Bayesian dynamic stopping (BDS) framework using an openly accessible c-VEP dataset (Thielen et al., 2024). Detailed information about this dataset can be found in the original publication (Thielen et al., 2015). For our study, we exclusively used the part of the dataset labeled as "fixed-length testing trials", where 12 participants were involved in a copy-spelling task. EEG data were recorded from 64 Ag/AgCl active electrodes arranged according to the international 10-10 system, amplified by a Biosemi ActiveTwo amplifier, and sampled at a frequency of 2048 Hz.

During the experiment, participants interacted with a 6×6 matrix speller presented on a 24 in BENQ XL2420T LED monitor, with a refresh rate of 120 Hz and a resolution of 1920×1080 pix. The 36 cells within the matrix were displayed against a mean-luminance gray background. Each cell underwent luminance modulation using a pseudo-random binary sequence at full contrast, where a value of 1 represented a white background and 0 indicated a black background.

The stimulus sequences were carefully optimized subsets derived from a collection of Gold codes (Gold, 1967). Specifically, using training data, a participant-specific subset was chosen to minimize the maximum correlation between template responses within the subset (Thielen et al., 2015). Furthermore, these selected codes were assigned to cells in the matrix in a manner that ensured neighboring sequences maintained minimal correlation in their responses (Thielen et al., 2015). The chosen sequences were modulated to include flashes of only two durations: a short flash lasting 8.33 ms and a long flash lasting 16.67 ms. With a sequence length of $126 = 2 \times (2^6 - 1)$ bits and a presentation rate of 120 Hz, one cycle lasted 1.05 s.

Participants completed three identical runs, each consisting of 36 trials, one for each of the 36 cells, presented in a randomized order. Each trial started with a 1-second cue, highlighted in green, indicating the target cell. Following this, all cells began flashing with their respective stimulus sequences for a duration of 4.2 s, during which participants maintained fixation on the target cell. Subsequently, participants received feedback from an online classifier, with the predicted cell being highlighted in green for 1 s.

In summary, for each participant, the dataset contained 108 trials of a 4.2-second trial duration, including 3 repetitions of each of the 36 stimuli.

2.2 Bayesian dynamic stopping

In this section, we introduce the Bayesian dynamic stopping (BDS) framework. Let us consider $\mathbf{x} \in \mathbb{R}^T$ as the EEG data for a single trial and a single (e.g., spatially filtered, virtual) channel, comprising T samples. We conceptualize this data as the composite of the assumed underlying template EEG response $\mathbf{t}_y \in \mathbb{R}^T$, representing the source signal for the attended stimulus with a target label y, and a component of noise $\boldsymbol{\epsilon} \in \mathbb{R}^T$ drawn from a normal distribution $\mathcal{N}(0, \sigma)$, where $\sigma \in \mathbb{R}$ denotes the standard

deviation of the noise. Here, $\alpha \in \mathbb{R}$ represents a scaling factor.

$$\mathbf{x} = \alpha\mathbf{t}_y + \boldsymbol{\epsilon} \tag{1}$$

Let us define a similarity score f_i, calculated as the inner product between the EEG data $\mathbf{x}$, corresponding to the true class label y, and each of the templates $\mathbf{t}_i$, representing the (expected) source signal of the candidate stimuli $i \in \{1, \ldots, N\}$ from a set of N classes:

$$f_i = \mathbf{x}^\top \mathbf{t}_i = (\alpha\mathbf{t}_y + \boldsymbol{\epsilon})^\top \mathbf{t}_i = \alpha\mathbf{t}_y^\top \mathbf{t}_i + \boldsymbol{\epsilon}^\top \mathbf{t}_i \tag{2}$$

Since $\boldsymbol{\epsilon}$ is an uncorrelated Gaussian distributed noise source, f_i is also Gaussian distributed:

$$f_i \sim \mathcal{N}(\alpha\mathbf{t}_y^\top \mathbf{t}_i, \sigma \|\mathbf{t}_i\|) \tag{3}$$

Using the assumption that the observed EEG data $\mathbf{x}$ originated from class y, we can deduce that the similarity score f_i follows a distribution that is either a target or a non-target Gaussian:

$$f_i \sim \begin{cases} \mathcal{N}(\alpha b_1, \sigma_1) & \text{if } i = y \\ \mathcal{N}(\alpha b_0, \sigma_0) & \text{otherwise} \end{cases} \tag{4}$$

Here, the target distribution has a mean of $\alpha b_1 = \alpha\mathbf{t}_y^\top \mathbf{t}_y$ and a standard deviation of $\sigma_1 = \sigma \|\mathbf{t}_y\|$, and the non-target distribution has a mean of $\alpha b_0 = \alpha\mathbf{t}_y^\top \mathbf{t}_i$ and a standard deviation of $\sigma_0 = \sigma \|\mathbf{t}_i\|$.

At this point, we encounter two challenges. Firstly, we have to identify the most probable true target. Secondly, we need to decide whether we possess sufficient certainty to emit this true target, or to acquire more data, specifically, to increase the number of samples T.

Let us first consider the simpler two-class problem, where $N = 2$. Assuming that the target class is $y = 1$ with a prior probability of p_1, and the non-target class is $i = 0$ with a prior probability of p_0, we can define two hypotheses:

- $\mathcal{H}_1$: the observed score f_i originates from the target distribution $\mathcal{N}(\alpha b_1, \sigma_1)$.
- $\mathcal{H}_0$: the observed score f_i stems from the non-target distribution $\mathcal{N}(\alpha b_0, \sigma_0)$.

To determine the correct hypothesis, we use the Bayes criterion, which relies on two fundamental assumptions. Firstly, we possess knowledge of the prior probabilities associated with the source outputs, denoted as p_1 and p_0. Secondly, we know the costs assigned to each course of action, specified as follows:

- c_{00} is the cost of choosing $\mathcal{H}_0$ when $\mathcal{H}_0$ is true, i.e., no detection or true negative.
- c_{01} is the cost of choosing $\mathcal{H}_0$ when $\mathcal{H}_1$ is true, i.e., false reject or false negative.
- c_{10} is the cost of choosing $\mathcal{H}_1$ when $\mathcal{H}_0$ is true, i.e., false accept or false positive.
- c_{11} is the cost of choosing $\mathcal{H}_1$ when $\mathcal{H}_1$ is true, i.e., detection or true positive.

The Bayes test aims to determine the decision region in a manner that minimizes the risk $\mathcal{R}$, i.e., the Bayes criterion (van

Trees, 2004). Denoting the expected value of the cost as the risk $\mathcal{R}$, we have:

$$\begin{aligned} \mathcal{R} = &c_{00}p_0 P(\text{say } \mathcal{H}_0 | \mathcal{H}_0 \text{ is true}) + \\ &c_{01}p_1 P(\text{say } \mathcal{H}_0 | \mathcal{H}_1 \text{ is true}) + \\ &c_{10}p_0 P(\text{say } \mathcal{H}_1 | \mathcal{H}_0 \text{ is true}) + \\ &c_{11}p_1 P(\text{say } \mathcal{H}_1 | \mathcal{H}_1 \text{ is true}) \end{aligned} \tag{5}$$

The Bayes test results in a likelihood ratio test, as discussed in Section 2.2 of van Trees (2004):

$$\Lambda(f_i) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\underset{<}{>}}} \frac{p_0(c_{10} - c_{00})}{p_1(c_{01} - c_{11})} \tag{6}$$

where the likelihood ratio $\Lambda(f_i)$ is defined as follows:

$$\Lambda(f_i) = \frac{P(f_i|\mathbf{t}_1)}{P(f_i|\mathbf{t}_0)} = \frac{\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{f_i - \alpha b_1}{\sigma_1})^2}}{\frac{1}{\sigma_0\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{f_i - \alpha b_0}{\sigma_0})^2}} \tag{7}$$

Here, dealing with $N = 2$ classes and following Equation 4, we have $b_0 = \mathbf{t}_0^\top \mathbf{t}_1$, $b_1 = \mathbf{t}_1^\top \mathbf{t}_1$, $\sigma_0 = \sigma \|\mathbf{t}_0\|$, and $\sigma_1 = \sigma \|\mathbf{t}_1\|$.

In the context of our BCI problem setting, the risk $\mathcal{R}$ comprises only two components: the cost of a false positive (i.e., a misclassification) and the cost of a false negative (i.e., missing the target). Therefore, by setting $c_{00} = c_{11} = 0$, indicating no cost for correct detection, and introducing the cost ratio $\zeta = \frac{c_{10}}{c_{01}}$, the Bayes test can be formulated as follows:

$$\ln \Lambda(f_i) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\underset{<}{>}}} \ln\left(\frac{p_0}{p_1}\zeta\right) \tag{8}$$

Let us now transition to the more common scenario involving more than two classes. Interestingly, we can approach a multi-class problem with $N > 2$ as a two-class problem by treating it as a "one versus rest" problem. Here, we assume that any target class conforms to the target Gaussian distribution, while all non-target classes adhere to the non-target Gaussian distribution. Consequently, we are still confronted with the same two distributions, target and non-target, albeit with their parameters estimated as the averages of their respective distributions. Specifically, assuming an equal prior for all the N classes, the priors can be merged in two non-equal classes having a ratio of N, with the probability of a target being $p_1 = \frac{1}{N}$ and the probability of a non-target being $p_0 = \frac{N-1}{N}$. In this multi-class scenario, the mean and variances of the target and non-target distributions from Equation 4 become (see Figure 1):

$$\alpha b_1 = \alpha \frac{1}{N} \sum_i^N \|\mathbf{t}_i\|^2 \tag{9}$$

$$\alpha b_0 = \alpha \frac{1}{N^2 - N} \sum_i^N \sum_{j \neq i}^N \mathbf{t}_i^\top \mathbf{t}_j \tag{10}$$

$$\sigma_1^2 = \sigma^2 b_1 + \frac{1}{N} \sum_i^N (\alpha\mathbf{t}_i^\top \mathbf{t}_i - \alpha b_1)^2 \tag{11}$$

$$\sigma_0^2 = \sigma^2 b_1 + \frac{1}{N^2 - N} \sum_i^N \sum_{j \neq i}^N (\alpha\mathbf{t}_i^\top \mathbf{t}_j - \alpha b_0)^2 \tag{12}$$
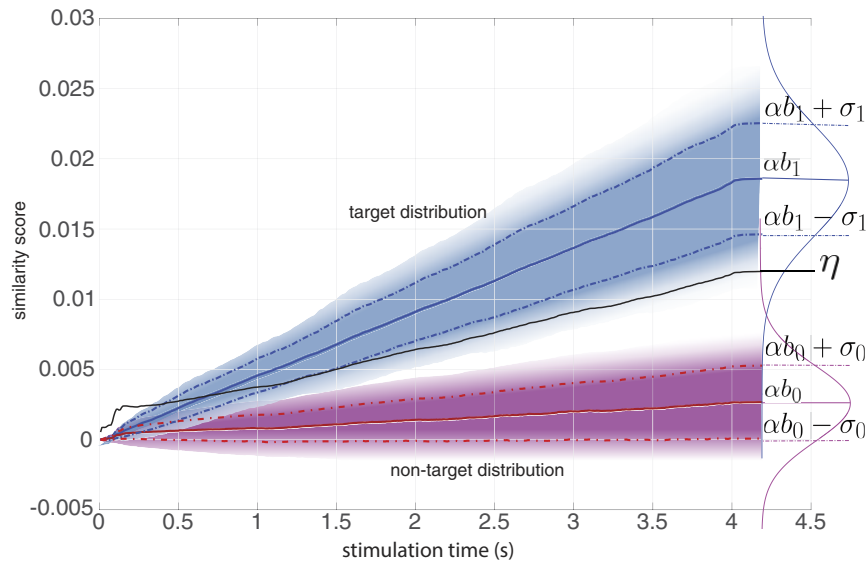
FIGURE 1
An example of the score distribution of target (blue) and non-target (pink) classes and how they change over stimulation time. Solid lines indicate the mean (αb_1 and αb_0) of the Gaussian distributions and dashed lines indicate the standard deviation (σ_1 and σ_0). The black solid line is the decision boundary η resulting from Equation 18 with $\zeta = 1$.

Given Equation 7, we now find:

$$\ln \Lambda(f_i) = \ln \frac{\sigma_0}{\sigma_1} + \frac{1}{2\sigma_0^2\sigma_1^2} \big((\sigma_0^2 - \sigma_1^2)f_i^2$$
$$-2\alpha(\sigma_1^2 b_0 - \sigma_0^2 b_1)f_i - \alpha^2(\sigma_0^2 b_1^2 - \sigma_1^2 b_0^2)\big) \qquad (13)$$

Now, the likelihood ratio test from Equation 8 can be reformulated to the following using a, b, and c as the coefficients of the quadratic formula:

$$af_i^2 + bf_i + c \mathop{\substack{\mathcal{H}_1 \\ > \\ < \\ \mathcal{H}_0}} 0 \qquad (14)$$

with

$$a = \sigma_1^2 - \sigma_0^2 \qquad (15)$$
$$b = -2\alpha(\sigma_1^2 b_0 - \sigma_0^2 b_1) \qquad (16)$$
$$c = -\alpha^2(\sigma_0^2 b_1^2 + \sigma_1^2 b_0^2) + 2\sigma_0^2\sigma_1^2 \ln \frac{\sigma_0}{\sigma_1(N-1)\zeta} \qquad (17)$$

Now, the decision boundary η is the root of the above quadratic polynomial:

$$\eta = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad (18)$$

such that the likelihood ratio test will result in the following test:

$$f_i \mathop{\substack{\mathcal{H}_1 \\ > \\ < \\ \mathcal{H}_0}} \eta \qquad (19)$$

2.2.1 Calibrating BDS

The framework that we have presented requires calibration to estimate the distribution parameters, given a set of labeled training data. Firstly, the template responses $\mathbf{t}_i$ are calculated for each of the N classes. For instance, for c-VEP data, one could use the "reference pipeline" (Martínez-Cagigal et al., 2021) or "reconvolution" (Thielen et al., 2015, 2021). Secondly, the template similarity score $\mathbf{t}_i^\top \mathbf{t}_j$ is calculated for each pair of stimuli at each stimulation time point, that is, for each of S decision window lengths. Thirdly, the scaling parameter α is obtained from a least square error minimization between $\mathbf{x}$ and $\mathbf{t}$. If the calibration data contain multiple trials, these are concatenated as if they were one long trial. Fourthly, a Gaussian distribution is fit on the residuals of the least squares fit, and the standard deviation of the noise σ is estimated as the standard deviation of this residual distribution. Fifthly, the distribution parameters $\mathbf{b}_0 \in \mathbb{R}^S$, $\mathbf{b}_1 \in \mathbb{R}^S$, $\boldsymbol{\sigma}_0 \in \mathbb{R}^S$ and $\boldsymbol{\sigma}_1 \in \mathbb{R}^S$ are calculated according to Equations 9–12, and the decision boundaries $\boldsymbol{\eta} \in \mathbb{R}^S$ are computed following Equations 15–18. For an illustration of these parameters, see Figure 1.

The decision boundary always defines the minimum risk decision area. The behavior of BDS is controlled by the hyper-parameter cost ratio ζ. With equal prior for target and non-target distributions and a cost ratio of $\zeta = 1$, the decision boundary would always be at the intersection of the target and non-target distributions. Instead, as can be concluded from Equations 17, 18, increasing the number of competing classes and as a result increasing the prior probability of the non-target class, results in shifting the decision boundary toward the target distribution.

According to Equation 8, the cost ratio ζ controls the cost of the two types of error. Therefore, $\zeta = 1$ means that both types of error have equal costs, hence the decision boundary will be the intersection of the two distributions. However, depending on the BCI application, the cost of false positive might be different than

```
 1: emit ← false
 2: t ← 1
 3: while emit = false do
 4:     Calculate the scores fᵢ(t)
 5:     if ∃i:fᵢ(t) > η(t)  or t = t* then
 6:         ŷ ← argmax(P(tᵢ|f_y))
                  i
 7:         emit ← true
 8:     else
 9:         t ← t+1
10:     end if
11: end while
```

Algorithm 1. The BDS procedure.

the cost of false negative. Imposing more cost on a false positive will result in a more reliable but slower system, while making a false negative more costly makes the system faster but more error-prone. The cost ratio parameter ζ in our formulation gives the user of the system the possibility of controlling this balance. Therefore, a cost ratio of $\zeta > 1$ means the cost of false positive is higher than false negative and therefore the decision boundary moves to the right and vice versa.

2.2.2 Applying BDS

Once all the model parameters are calculated, BDS can be used for dynamic stopping. At each stimulation time point, all f_i for $i \in \{1, \ldots, N\}$ will be compared with the decision boundary η. The index of the score that passes the test shows the winning class. If more than one score is above the threshold, the class with the highest likelihood will be chosen. If no score passes the test, the system waits for more data and repeats the test at the next stimulation time point.

BDS makes a decision as soon as the minimum risk criterion is met. Specifically, the algorithm terminates at the earliest stimulation time point t where at least one class's score f_i surpasses the minimum risk decision boundary η. Furthermore, it is possible to establish a maximum trial length t^*, after which the most probable class label is emitted (or potentially the trial is ignored and redone). The BDS procedure is outlined in Algorithm 1.

In summary, at each stimulation time point, the dynamic stopping algorithm makes a decision. If the score is below the threshold, it is a negative decision and in case the score is above the threshold, it makes a positive decision. Hence, there are four possible outcomes corresponding to the four courses of action listed in Section 2.2:

- No winning class is detected while the score of the true class is not winning (true negative).
- No winning class is detected while the score of the true class is winning (false negative or miss).
- The winning class is detected while the score of the true class is not winning (false positive).
- The winning class is detected while the score of the true class is winning (true positive or hit).

2.3 Baseline early stopping methods

We compared BDS with several other early stopping methods, including static and dynamic stopping techniques. Firstly, we contrasted BDS with the approach of conducting no early stopping, but instead employing a fixed trial length. We evaluated this by estimating a decoding curve for all trials and participants. The trial length could be regarded as a hyper-parameter for this method, and we assessed the accuracy achievable with each length. The trial length of 4.2 s, which uses the full trial duration, exploits all available data in the dataset, theoretically offering an upper bound on classification accuracy.

Secondly, we contrasted BDS with three static stopping methods. Each of these methods estimated a decoding curve on the training data and optimized a specific criterion to determine an optimized stopping time for all trials within a given participant. Using a 5-fold cross-validation approach on the training data, a classifier was calibrated on the training split and then tested on all validation trials ranging from 100 ms to 4.2 s in 100 ms increments.

The first static stopping method selected the stopping time at which the averaged decoding curve across folds achieved the highest accuracy for the first time. The second method chose the first instance where a predefined targeted accuracy was obtained. This targeted accuracy served as a hyper-parameter of the method. Lastly, the third method identified the stopping time at which the decoding curve achieved the highest information transfer rate (ITR).

Thirdly, we contrasted BDS with two dynamic stopping methods from previous studies: one relying on margins of classification scores (Thielen et al., 2015) and the other on a Beta distribution of the classification scores (Thielen et al., 2021).

Firstly, the margin method was a calibrated approach that considered the margin between the maximum and runner-up classification scores. During calibration, it learned a margin threshold for stimulation time points ranging from 100 ms to 4.2 s in 100 ms increments. These thresholds were learned to achieve a predefined targeted accuracy. During testing, the symbol associated with the maximum classification score was released as soon as the margin threshold was reached (Thielen et al., 2015).

Secondly, the Beta method did not require calibration. During testing, it estimated a Beta distribution based on all classification scores except the maximum one. Subsequently, it evaluated the probability that the maximum classification score did not originate from the estimated Beta distribution. If this likelihood surpassed a predefined targeted accuracy threshold, the classification was released (Thielen et al., 2021).

2.4 Analysis

2.4.1 Classification

For classification, we employ the "reconvolution CCA" template matching classifier (Thielen et al., 2015, 2021). Consider we have multi-channel, single-trial EEG data $\mathbf{X} \in \mathbb{R}^{C \times T}$, with C channels and T samples. To predict the label $\hat{y}$ for this trial, we perform template matching as follows:

$$\hat{y} = \arg\max_i f(\mathbf{w}^\top \mathbf{X}, \mathbf{r}^\top \mathbf{M}_i) \tag{20}$$

Here, $\mathbf{w} \in \mathbb{R}^C$ is a spatial filter, $\mathbf{r} \in \mathbb{R}^M$ is a temporal response vector with M samples, and $\mathbf{M}_i \in \mathbb{R}^{M \times T}$ is a structure matrix representing the onset, duration, and overlap of each event (e.g., flashes) in the ith stimulus sequence, where $i \in \{1, \ldots, N\}$ and N is the number of classes (i.e., the number of symbols in a matrix speller).

The function f represents the scoring function that calculates the similarity between the spatially filtered EEG data $\mathbf{w}^\top \mathbf{X}$ and the predicted template response for the ith stimulus $\mathbf{r}^\top \mathbf{M}_i$. Typically, f is the Pearson's correlation coefficient (Thielen et al., 2015, 2021). However, in this work, we use the inner product, as BDS relies on the inner product as the similarity score.

The reconvolution CCA method requires the spatial filter $\mathbf{w}$ and temporal response $\mathbf{r}$ to be learned from labeled training data. This is achieved through canonical correlation analysis (CCA) as follows:

$$\underset{\mathbf{w}, \mathbf{r}}{\arg\max}\, \rho(\mathbf{w}^\top \mathbf{S}, \mathbf{r}^\top \mathbf{D}) \qquad (21)$$

In this formulation, $\mathbf{S} \in \mathbb{R}^{C \times KT}$ is a matrix consisting of K concatenated training trials, and $\mathbf{D} \in \mathbb{R}^{M \times KT}$ is a matrix of stacked structure matrices that correspond to the labels of the training trials. CCA then identifies the spatial filter $\mathbf{w}$ and temporal response $\mathbf{r}$ that maximize the Pearson's correlation coefficient ρ between the spatially filtered EEG data and the predicted template responses.

2.4.2 Evaluation

We evaluated the proposed BDS method using all 108 trials from each of the 12 participants in the dataset, using 5-fold cross-validation. For each fold, 4/5 of the data served as the training set, on which the classifier was calibrated, resulting in the spatial and temporal filters necessary for classification. The parameters of BDS, including the scaling factor α, the noise standard deviation σ, the mean b_1, b_0 and standard deviation σ_1, σ_0 of the target and non-target distributions, and the decision boundary η as a function of stimulation time, were estimated from the same training data as described in Section 2.2. The remaining 1/5 of the data was used as the test set, where trials were classified according to reconvolution CCA once the stopping decision was made by BDS.

The baseline methods were evaluated using the same 5-fold cross-validation approach. For the static methods, the optimal trial length was determined using the training set in each fold. To assess whether the choice of similarity measure influences classification performance, we evaluated both the Pearson's correlation coefficient and the inner product. Among the baseline methods, the dynamic Beta stopping method does not support the inner product as a similarity measure, because its distribution range is bounded, which is not applicable to the inner product.

For each method, we varied their respective hyper-parameters to evaluate the changes in performance. The hyper-parameter for BDS was the cost ratio ζ. For the static methods, as well as the dynamic margin and Beta stopping methods, the targeted accuracy was used as the hyper-parameter.

To evaluate the performance of the early stopping methods, we considered several performance metrics. Firstly, we examined the conventional average accuracy versus average stopping time. However, since BDS aims to minimize risk rather than maximize accuracy, we also assessed performance using precision, recall, and

F-score, which are relevance-based performance metrics. Precision (also known as positive predictive value) is the fraction of true positives among all the positive decisions:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \qquad (22)$$

and recall (also known as sensitivity) is the ratio between the correct detections and all the detectable instances:

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \qquad (23)$$

We also looked at specificity (i.e., true negative rate):

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \qquad (24)$$

To compute the true/false positives/negatives, we examined all decisions made by the stopping procedures at each stimulation time point within each trial. With knowledge of the true label for each trial, we determined whether the highest similarity score at each stimulation time point corresponded to the true class. This enabled us to determine whether a positive or negative decision made by the stopping algorithm was true or false. Having the total number of true/false positive/negative decisions, we calculated precision, recall, specificity, and thereby the F-score:

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \qquad (25)$$

When interpreting these relevance-based measures, it is crucial to consider that the number of negative decisions in a stopping task far exceeds the number of positive decisions. This discrepancy arises because, for each trial, only one positive decision is made at the stopping time point, while negative decisions are made at all stimulation time points before that.

3 Results

Firstly, we investigated the performance of the proposed BDS method across various performance metrics. Figure 2 shows how the performance of BDS changes for each of the 12 participants in the dataset as the cost ratio between false positive and false negative errors ζ changes from 1e-10 to 1e10. The performance is evaluated using the six measures explained in Section 2.4 and averaged over the cross-validation folds (108 trials) for each participant.

As the cost ratio ζ increases, BDS becomes more accurate (higher accuracy) but slower (longer stopping time). The rate of these changes varies among participants. For example, for participants 1 and 6, altering the cost ratio has minimal impact on the accuracy. However, while the stopping time for participant 1 increases with a higher cost ratio, participant 6 shows much less sensitivity in stopping time.

Investigating precision, recall, specificity and F-score reveals several trends as the cost ratio ζ increases. Precision rises more sharply than accuracy, indicating that precision is highly sensitive to the cost ratio. Recall, on the other hand, does not change monotonically. For $\zeta < 1$, recall decreases as ζ increases, whereas
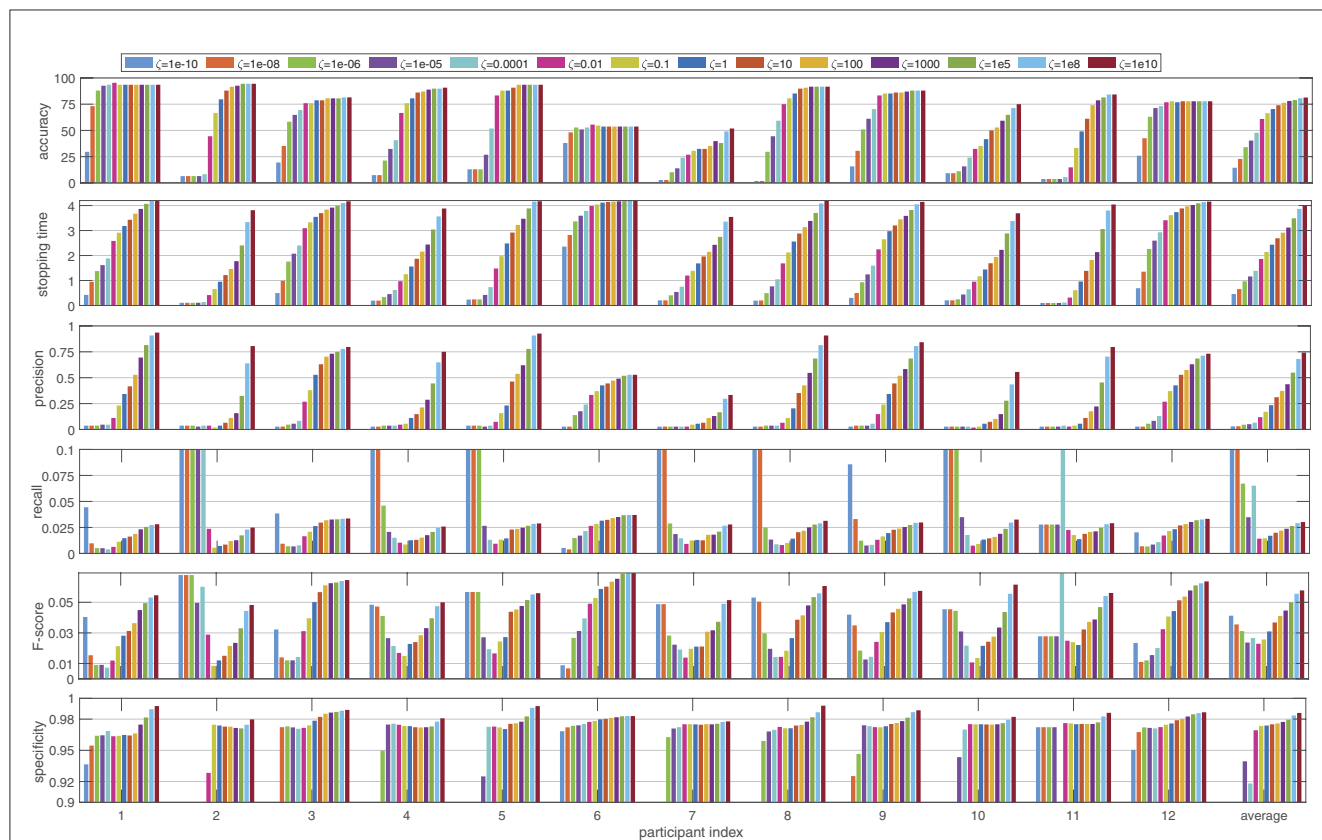
FIGURE 2
The performance of the Bayesian dynamic stopping averaged over the 108 trials for each participant, as the cost ratio ζ changes between 1e-10 and 1e10. The performance is measured in terms of accuracy and average stopping time as well as the relevance-based metrics precision, recall, F-score and specificity and is reported both per participant and averaged over the 12 participants.

for $\zeta > 1$ an increase in ζ leads to an increase in recall. A similar pattern is observed in the F-score, suggesting that due to the larger number of negative decisions, the F-score is more influenced by recall than by precision. Finally, specificity shows smaller variations compared to other metrics, but still shows an overall increase as ζ increases.

Secondly, we compared BDS with several baseline methods, including fixed, static and dynamic stopping methods, as described in Section 2.3. Figure 3 presents the results in terms of average accuracy versus average stopping time, while the hyper-parameters of each method vary.

For the fixed trial length, the hyper-parameter was the trial length, which varied between 0.5 s and 4.2 s. For the static stopping with accuracy optimization as well as dynamic stopping margin and beta, the hyper-parameter was the targeted accuracy, which ranged from 10% to 90%. The static methods that optimized accuracy or ITR did not have a hyper-parameter, hence they appear as single points in Figure 3 (+ for maximizing ITR and x for maximizing accuracy). All the baseline methods, except for dynamic beta stopping, are implemented both with correlation (dashed lines) and inner product (solid lines) as similarity scores.

As illustrated in Figure 3, using the correlation as the similarity score leads to higher performance for fixed length trials and static stopping methods when the trial duration exceeds 1 second. For the dynamic margin method, the performance appears largely

independent of the similarity score, except when the targeted accuracy exceeds 80%, where the inner product results in a slightly shorter average stopping time.

Among the dynamic stopping methods, the beta stopping method achieves the highest average accuracy in the shortest average stopping time. However, the obtained accuracy often deviates from the predefined targeted accuracy. For instance, when the targeted accuracy is set to 10%, the beta stopping method still achieves an accuracy of 61%, with the average stopping time not dropping below 1.3 s.

The dynamic margin method spans the shorter time area of the plot. With a targeted accuracy of 10% it achieves an average stopping time of 240 ms and an accuracy of 8.7% accuracy (which is still higher than the theoretical chance level 2.7%). As the targeted accuracy increases, the margin method's performance varies non-linearly. For instance, at a targeted accuracy of 98%, the method reaches an average stopping time of 2.4 s and an accuracy of 80.4%. However, when the targeted accuracy is reduced to to 90%, the resulting accuracy drops to 67%.

The static stopping method with targeted accuracy and inner product as similarity score results in a static stopping time of 0.5 s and an average accuracy of 30% when the targeted accuracy is set to 10%. For a targeted accuracy of 98%, the obtained accuracy is 79% with a static stopping time of 3 s.
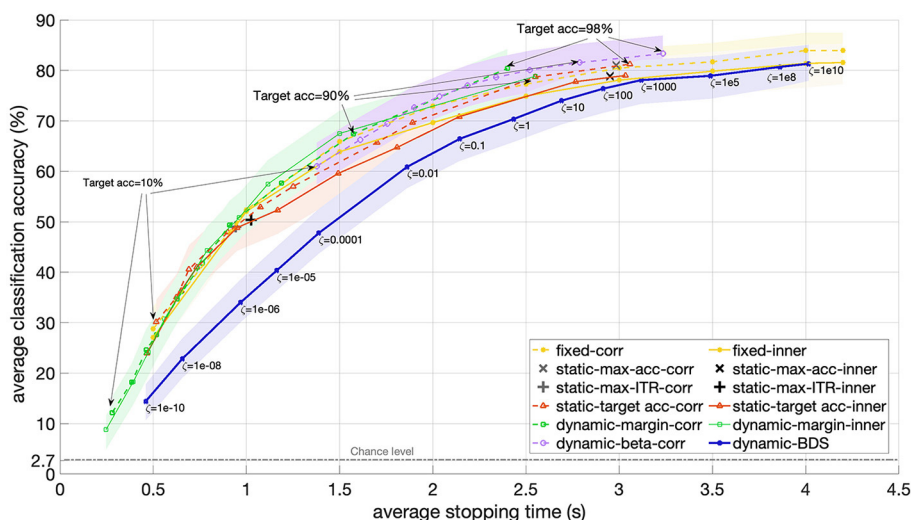
FIGURE 3
Accuracy versus stopping time averaged over 12 participants for BDS and static and dynamic stopping methods from the literature as the hyper-parameter of each method changes. For BDS (blue), the hyper-parameter is the cost ratio ζ ranging from 1e-10 to 1e10. For the static stopping methods with optimized accuracy (red), the margin (green), and beta (purple) dynamic stopping, the targeted accuracy varied between 10% and 98%. The static methods that maximize accuracy (gray ×) or ITR (gray +) do not have a hyper-parameter and therefore appear as one point in the figure. The performance of a fixed trial length from 0.5 s to 4.2 s is also included for reference (yellow). All but one of the baseline methods are implemented using both correlation (dashed) and inner product (solid) as the similarity score. The shaded areas indicate the 95% confidence interval.
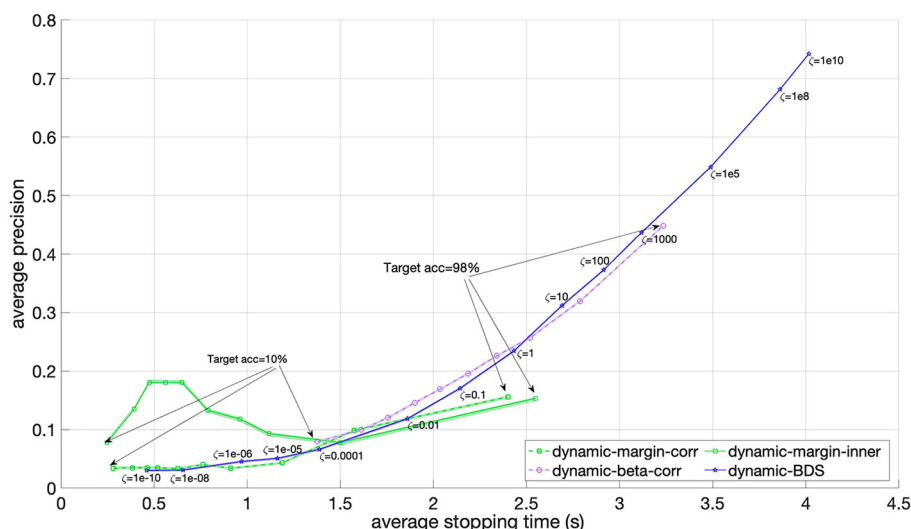
FIGURE 4
Precision versus stopping time averaged over 12 participants for the BDS, margin and beta dynamic stopping methods as the hyper-parameter of each method changes. For BDS (blue), the hyper-parameter is the cost ratio ζ which ranges from 1e-10 to 1e10. For the margin (green) and beta (purple) methods, the targeted accuracy ranges between 10% and 98%. The margin method is implemented using both correlation (dashed) and inner product (solid) as the similarity score. The 95% confidence interval is visualized using a shaded area, but they are in the range of 0.005 and therefore barely visible.

In summary, each of these baseline methods covers only a limited portion of the accuracy-time plane, restricting the range within which their hyper-parameters can control them. The average accuracy values obtained by these methods for similar average stopping times are not significantly different, as indicated by the overlapping 95% confidence intervals. Additionally, the targeted accuracy does not reliably predict the actual empirically obtained accuracy for many of the baseline methods.

The proposed BDS tends to result in a longer average stopping time for a similar accuracy level compared to the baseline methods. Or equally, for each average stopping time (reflecting speed), BDS yields a lower accuracy level. However, the difference in accuracy between BDS and static stopping for average stopping times longer than 1.5 s is not statistically significant, as indicated by the overlapping 95% confidence intervals (see Figure 3). Similarly, the difference in accuracy between BDS and beta stopping is not

significant for average stopping times above 2.5 s. Overall, the observed differences in accuracy were anticipated because BDS is designed to minimize risk, not error rate (hence, not to maximize accuracy). To assess the performance of each method in terms of risk, it is essential to examine precision for each method.

In Figure 4, the average precision across all 12 participants is plotted against the average stopping time for BDS, as well as for the two dynamic stopping baselines, i.e., the margin and beta methods. Notably, the margin method achieves a maximum precision slightly below 0.2, and the beta method obtains its highest precision of 0.44 at an accuracy of 83.3%. Instead, BDS shows comparable precision to both the margin and beta methods for average stopping times below 1.5 s, but beyond this, BDS surpasses the margin method in precision. Additionally, when the cost ratio $\zeta > 1$, implying a higher cost for false positive than false negatives, BDS's precision begins to outperform that of the beta method as well. These observed differences are statistically significant, indicated by the very small and non-overlapping 95% confidence intervals. This suggests that BDS excels in precision. Furthermore, BDS demonstrates the capability to increase its precision up to 0.75 by applying higher values of the cost ratio ζ.

4 Discussion

In this study, we proposed a model-based Bayesian dynamic stopping (BDS) method for evoked response BCIs and compared its performance with several existing static and dynamic stopping methods. The primary goal of this novel approach was to provide more control over the behavior of the stopping method, making it suitable for various BCI applications with different requirements. For example, applications sensitive to false positives, which result in miss-classifications, need a different dynamic stopping behavior compared to those where false negatives are more critical, which results in misses. We achieved a controllable precision-speed balance by assigning different cost values to each error type and minimizing the associated risk.

In general, dynamic stopping methods offer various advantages over fixed trial length and static stopping methods, making them a crucial area for exploration and improvement. With dynamic stopping, we can achieve more reliable decoding by allowing more data before making decisions, unlike static methods where decisions are made at a fixed time point regardless of the trial-specific confidence level. This adaptability is particularly beneficial given the non-stationary nature of EEG data. Dynamic stopping allows for faster classification of trials with high SNR, while granting more time for those with lower SNR. Additionally, dynamic stopping ideally facilitates the incorporation of a non-control state and asynchronous control by enabling decisions only when the certainty is sufficiently high, thereby enhancing the overall system responsiveness and user experience.

Examining the performance comparison between the baseline dynamic stopping methods and the static and fixed trial length methods (see Figure 3), it is evident that dynamic stopping methods achieve accuracy levels comparable to those of static, fixed and even optimized stopping methods for a given trial length.

While many dynamic stopping methods have been proposed for various types of evoked response BCIs, few offer inherently

calibrated hyper-parameters that intuitively and effectively control the accuracy-speed balance. Additionally, many existing methods focus on optimizing metrics such as accuracy, ITR or SPM, which do not necessarily reflect user-centric performance across different applications with varying sensitivities to different error types. Our proposed method addresses this issue by providing a hyper-parameter that allows for differential treatment of error types, enabling fine-tuned control over the accuracy-speed trade-off.

Many of the existing dynamic stopping methods have one or more hyperparameters that either need to be set manually or tuned by a search for optimal performance. BDS has only one hyperparameter, the cost ratio ζ, which can also be optimized by hyperparameter search to optimize a certain performance metric. However, importantly, because the cost ratio is a very intuitive parameter, it can be set by the designer given the requirements of the application. This contrasts with, for example, the threshold value in a threshold-based stopping method that does not have any intuitive connection to the application.

We evaluated our proposed method, BDS, using an open-access c-VEP dataset including 12 participants, each with 108 trials. Performance metrics were calculated for each participant individually and then averaged across all participants.

The per-participant results show that changing the hyper-parameter ζ, which is the cost ratio between false positive and false negative errors, allows control over the accuracy and speed of BDS. However, the effect of changing the cost ratio varies among participants. Such adjustment also affects relevance-based metrics: precision, recall and specificity. When $\zeta > 1$, implying a higher cost for false positive errors than for false negative errors, increasing ζ calls for a more precise classification, resulting in longer stopping times. Conversely, $\zeta < 1$ implies a higher cost for false negative than for false positive errors, so decreasing ζ leads to a faster system with shorter stopping times but reduced accuracy and precision (see Figure 2).

Recall, by definition, is the complement of the false negative error rate. The low recall values in our results indicate an imbalance between negative and positive decisions. This is because each trial only involves one positive decision, which stops the trial, whereas negative decisions are made at all stimulation time points before the stopping time. When $\zeta > 1$, increasing ζ leads to a higher number of true positives and more false negatives due to the longer waiting time. However, the increase in recall shows that the rise in true positives is larger than the increase in false negatives. Conversely, when $\zeta < 1$, decreasing ζ reduces the number of true positives and false negatives. The increase in recall as ζ decreases indicates that false negatives decrease faster than true positives. This behavior is observed in recall for most of the participants. Because of the imbalance between positive and negative decisions and the resulting disparity between the ranges of precision and recall, the F-score tends to follow the recall pattern. Finally, the high specificity values demonstrate that the number of true negatives exceeds false positives. Consequently, the effect of changing ζ is less pronounced in specificity due to this imbalance.

We investigated the performance of BDS and the baseline methods by examining how their hyper-parameters affect accuracy and speed. An intriguing observation is that for the methods with targeted accuracy as their hyper-parameter, adjusting the targeted accuracy does change the accuracy-speed balance but does not

guarantee achieving the targeted accuracy at either extreme time points. Moreover, each method spans a limited region in the accuracy-time plane, meaning that some accuracy or speed levels are unattainable for certain methods.

In contrast, the hyper-parameter of BDS allows for a wide range, from very fast and inaccurate, to very slow and accurate classifications, controllable by the cost ratio. However, for a given speed (average stopping time), the accuracy of BDS is typically lower than the other methods. Since BDS is designed to minimize risk rather than maximize accuracy, it is more appropriate to evaluate it in terms of precision.

Figure 4 compares the dynamic methods based on precision versus stopping time. The results show that BDS achieves comparable precision to the margin and beta methods within the speed levels they share. However, interestingly, BDS can achieve a much higher precision (up to 0.75) compared to the beta method's maximum achievable precision (0.32) by assigning a higher cost to false positive errors through a larger cost ratio ζ.

While the evaluation of our proposed stopping method on an offline dataset indicates the effectiveness of the BDS method in controlling the error types and its advantage in terms of relevance-based metrics, future research requires a more in-depth exploration of user experience and usability in an online setting to ensure the method's practicality in real-world applications.

It is essential to approach the results of our study on the novel proposed Bayesian dynamic stopping framework with caution and consider several important limitations. Firstly, in this study, we focused on carefully explaining the extensive math and description required for introducing our method and evaluated and compared BDS using a single c-VEP dataset, to provide empirical support for the effectiveness of the method. While BDS shows promise, its applicability to other evoked datasets such as ERP and SSVEP requires further investigation. Additionally, we believe BDS can easily be extended to other c-VEP datasets as well, generalizing beyond the specific stimulation sequences used here, because it holds no assumptions of the underlying structure in the data.

Secondly, it is worth noting that our comparison was limited to a handful of static and dynamic stopping methods. While numerous other methods exist in the literature, most of them either have a structure or make assumptions that limit them to a certain paradigm or setup. As an example of the limiting structure, the dynamic stopping method developed for SSVEP in Cecotti (2020), performs score thresholding as the basis for the stopping decision. The threshold however is defined as a function of the number of stimuli frequencies used, making the method only applicable to SSVEP paradigms. There are more threshold based stopping methods in which the threshold value is arbitrarily chosen based on eyeballing the data which makes it difficult to use them as a systematic comparison baseline (Nagel and Spüler, 2019). An example of a method with limiting assumptions, we can look at the dynamic stopping method proposed by Martínez-Cagigal et al. (2023b) for c-VEP. They assume a certain relationship between the templates that is only valid when circularly shifted codes are used. Therefore, the method is not directly applicable to all c-VEP datasets, for instance, those employing a set of Gold codes. Since adapting baseline methods to be more generally applicable is beyond the scope of

this paper, we limited our choice of baseline dynamic methods to those directly applicable to the c-VEP paradigm with Gold codes. Further research is necessary to comprehensively compare stopping methods, highlighting their underlying assumptions and application scenarios. This broader exploration can provide deeper insights into the strengths and limitations of different approaches, helping researchers and practitioners make informed decisions when selecting the most appropriate stopping method for their specific BCI applications.

Thirdly, BDS relies on template responses in the classifier, which are used to estimate the target and non-target distributions. While most standard decoding approaches involve some form of template (e.g., an event-related potential), others, such as deep learning frameworks, may not explicitly contain these templates.

BDS needs to be calibrated using training data as it requires knowledge of the parameters for the target and non-target distributions. However, if all non-target classes can be assigned to a single non-target distribution, the required training data can be limited to data from one class only. This is especially true when using the reconvolution CCA method, which can predict template responses to unseen stimulation sequences. Not only does this mean that BDS requires only limited training data, but it also means that BDS can handle any number of classes. Investigating the sensitivity of BDS performance to the amount of training data and its empirical performance with a small number of classes remains to be addressed in future work.

Moreover, if a zero-training approach for BDS is desired, its parameters could potentially be learned on the fly as data comes in, due to its minimal requirements on the shape of the training dataset. Nevertheless, an open empirical question remains regarding how much data is needed for BDS to learn a robust and accurate dynamic stopping rule.

While BDS allows for spanning a wide range of accuracy/precision-speed balances, the current implementation does not permit setting a specific targeted performance. However, the model does offer the potential to predict error rates using the available distributions, thereby influencing the decision of the stopping algorithm. This can help achieve desired performance levels in terms of expected error rates (total or specific error types) or precision. Formalizing and implementing such an extension to this method is an avenue for future work.

Another potential advantage of using a model-based dynamic stopping method like BDS is its ability to incorporate prior probabilities for different classes. Although in this study we focused on the case of equal priors, the model formulation allows for using non-equal priors as well. This is especially valuable in applications where some classes are much more frequent than others, enabling decisions on high-probability classes with less confidence compared to low-probability classes. An example of such an application is a speller used for typing. There are studies that use a language model to post-process the output of a BCI speller, thereby increasing the typing accuracy and speed (Gembler and Volosyak, 2019; Gembler et al., 2019b). Our proposed model offers the potential to directly integrate the language model within the dynamic stopping method, instead of using it for error correction. Integrating such priors into the stopping method is another avenue for future work.

In conclusion, BDS provides better control over false positive and false negative errors, enabling a broader range of accuracy-speed trade-offs. Additionally, BDS achieves a much higher level of precision compared to the baseline dynamic stopping methods. The proposed model opens promising directions for future work to harvest its potential more effectively.

Data availability statement

The dataset for this study, originally recorded by Thielen et al. (2015), can be found in the Radboud Data Repository, specifically Thielen et al. (2024). An implementation of BDS can be found in PyntBCI: https://github.com/thijor/pyntbci.

Ethics statement

The studies involving humans were approved by Ethical Committee of the Faculty of Social Sciences at the Radboud University Nijmegen. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study because a previously recorded dataset was used, involving human participants from which informed consent was taken.

Author contributions

SA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. JT: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Writing – original draft. PD: Conceptualization, Project administration, Supervision, Writing – review & editing.

Conflict of interest

PD is founder of MindAffect, a company that develops EEG-based diagnosis of perceptual functions.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bianchi, L., Liti, C., and Piccialli, V. (2019). A new early stopping method for P300 spellers. *IEEE Trans. Neural Syst. Rehab. Eng.* 27, 1635–1643. doi: 10.1109/TNSRE.2019.2924080

Castillos, K. C., Ladouce, S., Darmet, L., and Dehais, F. (2023). Burst c-VEP based BCI: Optimizing stimulus design for enhanced classification with minimal calibration data and improved user experience. *Neuroimage* 284:120446. doi: 10.1016/j.neuroimage.2023.120446

Cecotti, H. (2020). Adaptive time segment analysis for steady-state visual evoked potential based brain-computer interfaces. *IEEE Trans. Neural Syst. Rehab. Eng.* 28, 552–560. doi: 10.1109/TNSRE.2020.2968307

Fazel-Rezai, R., Allison, B. Z., Guger, C., Sellers, E. W., Kleih, S. C., and Kübler, A. (2012). P300 brain computer interface: current challenges and emerging trends. *Front, Neuroeng.* 5:14. doi: 10.3389/fneng.2012.00014

Gao, S., Wang, Y., Gao, X., and Hong, B. (2014). Visual and auditory brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 61, 1436–1447. doi: 10.1109/TBME.2014.2300164

Gembler, F., Benda, M., Saboor, A., and Volosyak, I. (2019a). "A multi-target c-VEP-based BCI speller utilizing n-gram word prediction and filter bank classification," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 2719–2724.

Gembler, F., Stawicki, P., Saboor, A., and Volosyak, I. (2019b). Dynamic time window mechanism for time synchronous VEP-based BCIs-performance evaluation with a dictionary-supported BCI speller employing SSVEP and c-VEP. *PLoS ONE* 14:e0218177. doi: 10.1371/journal.pone.0218177

Gembler, F., and Volosyak, I. (2019). A novel dictionary-driven mental spelling application based on code-modulated visual evoked potentials. *Computers* 8:33. doi: 10.3390/computers8020033

Gembler, F. W., Benda, M., Rezeika, A., Stawicki, P. R., and Volosyak, I. (2020). Asynchronous c-VEP communication tools-efficiency comparison of low-target, multi-target and dictionary-assisted BCI spellers. *Sci. Rep.* 10:17064. doi: 10.1038/s41598-020-74143-4

Gold, R. (1967). Optimal binary sequences for spread spectrum multiplexing. *IEEE Trans. Inform. Theory* 13, 619–621. doi: 10.1109/TIT.1967.1054048

Höhne, J., Schreuder, M., Blankertz, B., and Tangermann, M. (2010). "Two-dimensional auditory P300 speller with predictive text system," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (Buenos Aires: IEEE), 4185–4188.

Jin, J., Allison, B. Z., Sellers, E. W., Brunner, C., Horki, P., Wang, X., et al. (2011). An adaptive P300-based control system. *J. Neural Eng.* 8:036006. doi: 10.1088/1741-2560/8/3/036006

Lenhardt, A., Kaper, M., and Ritter, H. J. (2008). An adaptive P300-based online brain-computer interface. *IEEE Trans. Neural Syst. Rehab. Eng.* 16, 121–130. doi: 10.1109/TNSRE.2007.912816

Liu, T., Goldberg, L., Gao, S., and Hong, B. (2010). An online brain-computer interface using non-flashing visual evoked potentials. *J. Neural Eng.* 7:036003. doi: 10.1088/1741-2560/7/3/036003

Martínez-Cagigal, V., Santamaría-Vázquez, E., and Hornero, R. (2023a). "Toward early stopping detection for non-binary c-VEP-based BCIs: A pilot study," in *International Work-Conference on Artificial Neural Networks* (Cham: Springer), 580–590.

Martínez-Cagigal, V., Santamaría-Vázquez, E., Pérez-Velasco, S., Marcos-Martínez, D., Moreno-Calderón, S., and Hornero, R. (2023b). "Nonparametric early stopping detection for c-VEP-based brain-computer interfaces: A pilot study," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Sydney: IEEE), 1–4.

Martínez-Cagigal, V., Thielen, J., Santamaria-Vazquez, E., Pérez-Velasco, S., Desain, P., and Hornero, R. (2021). Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review. *J. Neural Eng.* 18:061002. doi: 10.1088/1741-2552/ac38cf

Nagel, S., and Spüler, M. (2019). Asynchronous non-invasive high-speed BCI speller with robust non-control state detection. *Sci. Rep.* 9:8269. doi: 10.1038/s41598-019-44645-x

Riechmann, H., Finke, A., and Ritter, H. (2015). Using a cVEP-based brain-computer interface to control a virtual agent. *IEEE Trans. Neural Syst. Rehab. Eng.* 24, 692–699. doi: 10.1109/TNSRE.2015.2490621

Sato, J., and Washizawa, Y. (2015). "Reliability-based automatic repeat request for short code modulation visual evoked potentials in brain computer interfaces," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE), 562–565.

Schreuder, M., Höhne, J., Blankertz, B., Haufe, S., Dickhaus, T., and Tangermann, M. (2013). Optimizing event-related potential based brain-computer interfaces: a systematic evaluation of dynamic stopping methods. *J. Neural Eng.* 10:036025. doi: 10.1088/1741-2560/10/3/036025

Schreuder, M., Rost, T., and Tangermann, M. (2011). Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Front. Neurosci.* 5:112. doi: 10.3389/fnins.2011.00112

Spüler, M. (2017). A high-speed brain-computer interface (BCI) using dry EEG electrodes. *PLoS ONE* 12:e0172400. doi: 10.1371/journal.pone.0172400

Thielen, J., Farquhar, J., and Desain, P. (2024). *Broad-Band Visually Evoked Potentials: Re(con)volution in Brain-Computer Interfacing.* Radboud University. doi: 10.34973/1ecz-1232

Thielen, J., Marsman, P., Farquhar, J., and Desain, P. (2017). Re(con)volution: accurate response prediction for broad-band evoked potentials-based brain computer interfaces. *Brain-Comp. Interf. Res.* 6, 35–42. doi: 10.1007/978-3-319-64373-1_4

Thielen, J., Marsman, P., Farquhar, J., and Desain, P. (2021). From full calibration to zero training for a code-modulated visual evoked potentials for brain-computer interface. *J. Neural Eng.* 18:056007. doi: 10.1088/1741-2552/abecef

Thielen, J., van den Broek, P., Farquhar, J., and Desain, P. (2015). Broad-band visually evoked potentials: Re(con)volution in brain-computer interfacing. *PLoS ONE* 10:e0133797. doi: 10.1371/journal.pone.0133797

Throckmorton, C. S., Colwell, K. A., Ryan, D. B., Sellers, E. W., and Collins, L. M. (2013). Bayesian approach to dynamically controlling data collection in P300 spellers. *IEEE Trans. Neural Syst. Rehab. Eng.* 21, 508–517. doi: 10.1109/TNSRE.2013.2253125

Tsoneva, T., Garcia-Molina, G., and Desain, P. (2015). Neural dynamics during repetitive visual stimulation. *J. Neural Eng.* 12:066017. doi: 10.1088/1741-2560/12/6/066017

van Trees, H. L. (2004). *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory.* Hoboken, NJ: John Wiley & Sons.

Verbaarschot, C., Tump, D., Lutu, A., Borhanazad, M., Thielen, J., van den Broek, P., et al. (2021). A visual brain-computer interface as communication aid for patients with amyotrophic lateral sclerosis. *Clini. Neurophysiol.* 132, 2404–2415. doi: 10.1016/j.clinph.2021.07.012

Vialatte, F.-B., Maurice, M., Dauwels, J., and Cichocki, A. (2010). Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Prog. Neurobiol.* 90, 418–438. doi: 10.1016/j.pneurobio.2009.11.005

Volosyak, I., Rezeika, A., Benda, M., Gembler, F., and Stawicki, P. (2020). Towards solving of the illiteracy phenomenon for VEP-based brain-computer interfaces. *Biomed. Phys. Eng. Express* 6:035034. doi: 10.1088/2057-1976/ab87e6

Zhang, H., Guan, C., and Wang, C. (2008). Asynchronous P300-based brain-computer interfaces: a computational approach with statistical models. *IEEE Trans. Biomed. Eng.* 55, 1754–1763. doi: 10.1109/TBME.2008.919128