Journal of Applied Data Sciences
Vol. 7, No. 3, September 2026, pp. 2232-2248
ISSN 2723-6471

Explainable EEG-Based Framework for Student Attention Classification with LLM Instructional Recommendations

Samaher Dawood[1,*], Wafaa Alsaggaf [2], Miada Almasre[3]

[1,2,3]Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia

[1]Computer and Information Technology Department, Digital Technical College for Girls in Jeddah, Technical and Vocational Training Corporation (TVTC), Saudi Arabia

(Received: February 5, 2026; Revised: April 10, 2026; Accepted: July 12, 2026; Available online: August 9, 2026)

Abstract

Student attention is a central factor in learning effectiveness, yet most existing electroencephalography-based attention monitoring systems focus exclusively on classification accuracy without addressing interpretability or practical instructional utility. This paper presents a unified framework that integrates attention classification, explainable artificial intelligence, and large language model-driven instructional recommendations to bridge the gap between physiological monitoring and actionable teaching decisions. The framework derives binary attention labels from electroencephalography spectral features using the Theta-Beta Ratio combined with age-adjusted Monastra thresholds, applied here as population-level heuristic guidelines rather than individual diagnostic criteria. Five machine learning classifiers were trained and evaluated under stratified five-fold cross-validation on a publicly available dataset of 983 samples with precomputed electroencephalography spectral power features across five frequency bands. Gradient Boosting achieved the strongest performance, reaching 98.3% accuracy, 97.8% balanced accuracy, and a macro F1-score of 97.6%, with particularly reliable detection of the low-attention class. These figures represent an upper bound, given the shared spectral origin of the labeling mechanism and model inputs, and the absence of subject-level identifiers in the dataset. To ensure transparency, Shapley Additive Explanations analysis was applied at both global and local levels, identifying Beta and Theta band activity as the dominant predictors of attention state. A large language model component then maps each classified attention state to a concrete, knowledge-grounded instructional strategy drawn from peer-reviewed educational literature, providing teachers with an actionable pedagogical recommendation rather than a classification label alone. The proposed framework advances the field by combining classification accuracy, model interpretability, and practical instructional guidance within a single deployable system, providing a foundation for intelligent, transparent, and educationally responsive attention monitoring in real classroom settings.

Keywords: EEG, Attention Classification, Explainable AI, SHAP, Large Language Model, Machine Learning, Instructional Recommendations

1. Introduction

Student attention plays a central role in how effective learning takes place. When attention drops, students struggle to follow explanations, retain information, and engage with course material, and this tends to show up directly in academic performance [1], [2], [3]. The problem becomes harder to manage as classrooms grow larger and more learning moves online, where teachers have far less visibility into how engaged their students actually are [4]. There is therefore a real and growing need for tools that can assess attention in an objective, reliable way and help teachers respond appropriately [5], [6].

One of the most promising approaches for this purpose is electroencephalography (EEG). Unlike self-report measures or behavioral observation, EEG captures brain activity directly and continuously, making it well-suited for real-time cognitive monitoring. The signals it produces can be broken down into frequency bands, including alpha, beta, and theta, each of which reflects different aspects of mental state and cognitive engagement [7]. Among the derived measures, the Theta–Beta Ratio (TBR) has attracted particular interest. It quantifies how dominant theta activity is related to beta, and research has consistently shown that higher TBR values correspond to reduced attention, while lower values are associated with alertness and sustained focus [6], [8].

Machine learning offers a natural fit for classifying attention states from EEG data, and several studies have demonstrated good results. However, a practical problem remains that most of the models used are black boxes. They

*Corresponding author: Samaher Dawood (sdawood0008@stu.kau.edu.sa)
DOI: https://doi.org/10.47738/jads.v7i3.1379

can produce accurate predictions, but they give little insight into why a particular student was classified as inattentive, which features drove the decision, or how confident the model is [9], [10]. For a tool to be genuinely useful in educational settings, it needs to be not only accurate but also interpretable teachers and administrators need to understand and trust what the system is telling them.

The introduction of Explainable Artificial Intelligence (XAI) techniques enables a better understanding of how models reach their decisions. Local explanation methods aim to make individual model predictions interpretable by approximating the model's behavior in the vicinity of a specific instance, rather than attempting to explain the entire model globally [11]. Recent XAI surveys have established interpretability and transparency as essential requirements for machine learning systems to earn trust and achieve responsible deployment in real-world settings [9].

Interpretability in machine learning is not a single well-defined concept; researchers approach it from different angles depending on what they need to understand and why. Lipton [10] draws a useful distinction between understanding how a model works globally and understanding why it made a specific prediction for a specific input. Both matter, but in practice, the second is often more actionable. What a teacher needs to know is not the model's architecture, but why this particular student was flagged as inattentive right now [11].

Among the available explanation methods, SHAP has become one of the most widely used. It is rooted in cooperative game theory and works by calculating how much each input feature contributed to a given prediction. What makes it appealing in practice is that it produces explanations that are both mathematically sound and readable by non-experts a rare combination [12].

Alongside these developments in explainability, large language models have opened new possibilities for translating analytical outputs into human-readable guidance. Models such as GPT-3 and GPT-4 can take structured information like a predicted attention state and its SHAP explanation and produce contextually appropriate text in natural language [13]. In an educational setting, this creates an opportunity to bridge the gap between what a classifier outputs and what a teacher can act on [14].

Despite these advances, a clear and persistent gap remains: existing EEG-based attention classification systems focus almost exclusively on predictive accuracy and rarely provide interpretable explanations or actionable instructional guidance. Teachers cannot act on a black-box prediction alone; they need to understand why a student was flagged as inattentive and what to do about it. No existing system provides all three components: accurate classification, transparent explanation, and concrete pedagogical recommendation within a unified framework. This paper addresses that gap directly.

This study pursues three interconnected objectives. First, it develops an EEG-based framework for accurately classifying learners' attention levels using multiple machine learning models. This classification layer is then extended through the integration of XAI techniques, enabling both global and local explanations for model predictions. Finally, large language model components are incorporated to generate context-aware and pedagogically meaningful recommendations, bridging the gap between automated attention detection and actionable instructional guidance.

The rest of this paper is structured as follows. Section 2 reviews related work on EEG-based attention classification and explainable AI. Section 3 describes the proposed framework and methodology. Section 4 presents experimental results and discusses the findings. Section 5 concludes the paper and suggests directions for future research.

2. Literature Review

Over the past few years, there has been a growing interest in using EEG signals alongside machine learning to study attention in educational settings [7], [15]. EEG has a practical advantage over many other methods; it is noninvasive, relatively easy to set up, and provides continuous real-time data about brain activity, which makes it well-suited for tracking cognitive states like attention during learning tasks [7]. Table 1 summarizes a selection of studies in this area, highlighting their objectives, features, classifiers, and reported accuracy.

Table 1. Summary of selected studies in EEG-based attention classification.

Study	Objective	Features	Classifiers	Accuracy	Explainability	Recommend strategy
[15]	Classify attention states (focused, unfocused, drowsy)	DWT - statistical features	XGBoost, RF, kNN	99%	No	No
[5]	Adaptive learning based on attention	Wavelet - PLI connectivity	kNN	87%	No	No
[8]	Multi-level attention classification	Entropy - frequency ratios	XGBoost	95.36% (2-class)	No	No
[6]	Attention detection in online learning	PSD -TBR -EEG bands	RF, SVM, kNN	96%	No	No
Proposed Work	Interpretable and attention classification	EEG spectral bands	LR, SVM, RF, Gradient Boosting, kNN	98.3%	Yes	Yes

The authors of [5] developed a neurological system based on electroencephalography (EEG) to differentiate between high and low attention states using a k-nearest neighbor (kNN) classifier. The proposed system achieved 87% accuracy by using wavelet-based analysis for feature extraction and integrating it with graph theory-based connectivity measures. While the integration of connectivity measures is methodologically interesting, the system lacks interpretability and provides no mechanism for translating predictions into teaching decisions.

Both supervised and unsupervised approaches have been applied to EEG signal classification. Al-Nafjan and Al-Dayel [6] developed a BCI system for detecting students' attention in e-learning environments. Their method uses PSD features and derives indices such as TBR and other frequency-based measures. The RF achieved the highest accuracy of 96% of the models (kNN, SVM, and RF) evaluated. However, the system provides no interpretability mechanism and does not connect its outputs to instructional recommendations, limiting its practical utility in real classroom settings.

Similarly, a brain-computer interface system was proposed by Katon Suwida, Shintami Chusnul Hidayati, and Riyanarto Sarno [15] to classify attention into states of focus, lack of focus, and drowsiness. Although the system achieved 99% accuracy, it relied on DWT features without any explainability mechanism, and the multi-class framing was not connected to instructional action

In addition, an EEG-based method for classifying attention levels using dynamical complexity analysis and XGBoost was introduced by Wan, Cui, Gao, and Gu [8]. Attention was divided into four categories (high, medium, low, and resting) based on calibrated response time (C-RT). The model achieved an accuracy of 81.39% for four categories, 80.42% for three categories, and 95.36% for binary classification using five-fold cross-validation, based on beta/theta and beta/alpha ratios, as well as nonlinear features.

A critical examination of these studies reveals two consistent limitations. First, none of the reviewed works incorporates explainability mechanisms; all models function as black boxes, providing predictions without insight into which EEG features drove individual classifications. This is particularly problematic in educational contexts where trust and transparency are essential for adoption. Second, none of the reviewed systems connects classification outputs to actionable teaching decisions. Attention labels are treated as endpoints rather than inputs to pedagogical action. The proposed framework directly addresses both limitations by integrating SHAP-based explainability and LLM-driven instructional recommendations within a unified pipeline.

Along with teaching strategies to enhance students' attention and engagement during learning activities, table 2 categorizes these strategies by level of interest, distinguishing between strategies that restore attention in low-focus situations and those that maintain and deepen engagement in high-focus situations. For example, in the case of a low level of attention, strategies such as movement breaks and rest periods are necessary to regulate cognitive load and restore focus, thus improving concentration and performance of learning activities [16],[17].

Interactive strategies, including polls, instant feedback, and pop-up questions embedded in videos, are also recommended to help regain students' attention by encouraging active engagement and sustained engagement with the content [18], [19]. These strategies are especially effective when attention declines due to fatigue or negative learning conditions.

When students are already engaged and focused, the challenge shifts from restoring attention to maintaining and deepening it. Simply keeping them busy is not enough; the tasks need to match their cognitive capacity and push them further. Open-ended and thought-provoking questions work well here because they require students to think beyond surface-level recall and sustain mental effort over time [20]. Peer teaching serves a similar function: when students explain concepts to each other, they are forced to organize and articulate their understanding, which tends to strengthen it [21].

Socratic questioning and structured debate take this a step further by demanding that students not just recall or explain but actively defend and evaluate positions, a process that builds critical thinking alongside sustained attention [22]. Error analysis, particularly when paired with immediate corrective feedback, is another underused strategy in this category. Asking students to find and fix mistakes requires them to engage actively with the material rather than passively receive it, and this kind of exploratory processing tends to produce deeper learning outcomes [23]. Similarly, timed challenges introduce an element of urgency that can sharpen focus and encourage students to monitor their own thinking more carefully [20].

Table 2. Instructional strategies mapped to student attention levels

Attention Level	Instructional Strategy	Description	Reference
Low	Movement Breaks	Brief physical activities designed to restore attention and reduce cognitive fatigue	[16]
Low	Rest Breaks	Short cognitive pauses that facilitate recovery of attentional resources	[17]
Low	Pop-up Questions	Embedded prompts within learning materials to re-engage learners and promote active processing	[18]
Low	Interactive Videos	Multimedia environments that support engagement and self-regulated learning	[19]
Low	Polling	Real-time response systems that stimulate participation and re-focus attention	[24]
Low	Immediate Feedback	Instant responses that reinforce engagement and guide learner focus	[24]
High	Challenge-Based Tasks	Open-ended, cognitively demanding problems that sustain deep engagement	[20]
High	Debates	Structured argumentative activities that enhance critical thinking and evaluation skills	[20]
High	Error Analysis	Tasks that require analyzing and correcting errors to deepen understanding	[23]
High	Timed Challenges	Time-constrained activities that maintain focus and increase cognitive intensity	[20]
High	Peer Teaching	Collaborative learning, where students explain concepts, promotes active knowledge construction	[21]
High	Socratic Questioning	Guided questioning techniques that stimulate analytical reasoning and reflection	[22]

A pattern that runs through much of the recent literature on decision support systems is the trade-off between performance and transparency. As models have become more powerful, they have also become harder to interpret, and in many domains, including education, this matters a great deal. A system that flags a student as inattentive but cannot

ISSN 2723-6471

explain why it is difficult to trust and even harder to act on. The practical and ethical concerns here are real: if educators are expected to make decisions based on model outputs, they need some basis for understanding those outputs [9], [10].

This problem has driven significant interest in explainable AI (XAI) over the past decade. The core idea is straightforward; rather than treating the model as a finished product, XAI methods open it up and ask which inputs drove which outputs. Lundberg and Lee's SHAP framework is probably the most widely adopted approach in this space. It draws on cooperative game theory to assign each feature a contribution score for each prediction, producing explanations that are both rigorous and readable [12].

Yet despite this progress, most EEG-based attention studies have not caught up. The focus in that literature remains heavily on accuracy, which classifier performs best, and which feature set gives the highest F1. Interpretability is treated as secondary, if it is addressed at all. The result is a body of work filled with well-performing black-box models that would be difficult to deploy in any real educational environment, where teachers need to understand and trust the system they use.

There is a second gap that is equally important. Even in studies that do discuss attention levels, the findings rarely connect to what should happen next. Instructional strategies do exist in the literature, such as movement breaks, interactive prompts, peer teaching, and timed challenges, but they are presented as static recommendations rather than dynamic responses to a student's current cognitive state. A student who is deeply focused needs something different from a student who is mentally drifting, and current systems do not make that distinction in real time.

Taken together, these two gaps point to a clear need: a framework that classifies attention from EEG signals, explains those classifications in terms a teacher can understand, and links them to concrete instructional responses. That is what this paper sets out to provide.

3. Methodology

This section describes how we built and evaluated the proposed framework. The work proceeded through a series of connected stages. We started by selecting a dataset containing EEG spectral features and defining which of those features were relevant to attention monitoring. From there, we constructed binary attention labels using a physiologically grounded labeling approach. We then trained five machine learning classifiers and compared their performance to identify the best-performing model. Once selected, that model was passed through SHAP-based explainability analysis to understand what was driving its predictions. Finally, those explanations were fed into an LLM-based component that mapped the predicted attention state to a specific instructional strategy. The full pipeline is shown in figure 1. The overall research process of the proposed framework is summarized in algorithm 1.

Algorithm 1. EEG-based attention classification and recommendation framework.

Algorithm 1: EEG Attention Classification and Recommendation Framework

```
Input:  {Alpha, Beta, Theta, Delta, Gamma}
Output: attention_label, recommendation
BEGIN
  // Label Generation
  FOR each sample s in dataset DO
     TBR(s) ← Theta(s) / (Beta(s) + ε)
     IF TBR(s) > threshold_high THEN label(s) ← "Low"
     ELSE IF TBR(s) < threshold_low THEN label(s) ← "High"
     ELSE Remove(s)
  END FOR
  // Model Training & Selection
  FOR each C in {LR, SVM, RF, GB, kNN} DO
     performance(C) ← CrossValidate(C, k=5)
  END FOR
  best_model ← argmax(performance, "Macro F1")
  // Prediction & Explanation
```

ISSN 2723-6471


```
  attention_label ← best_model.Predict(sample)
  shap_values ← TreeExplainer(best_model, sample)
  // Recommendation
  candidates ← KB.Filter(attention_label)
  recommendation ← GPT4o_mini(attention_label, shap_values, candidates)
  Display(attention_label, shap_values, recommendation)
END
```

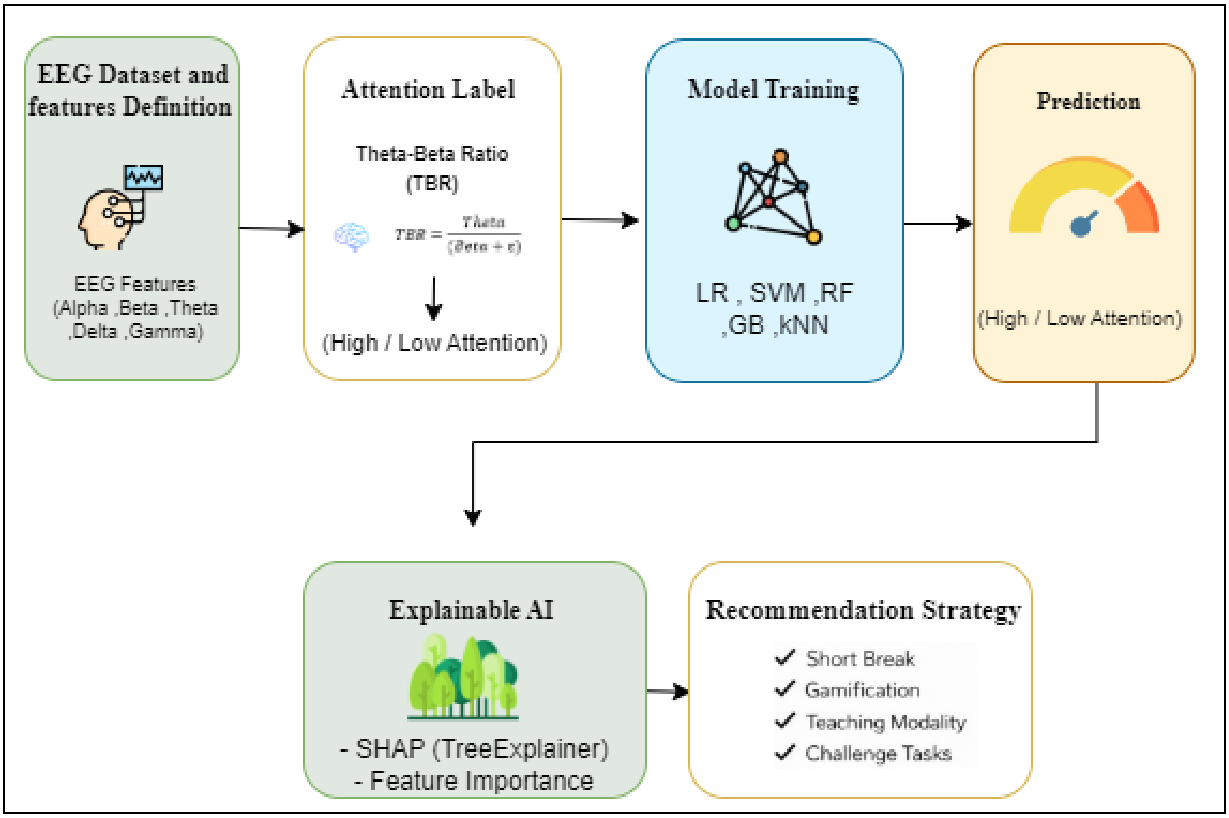


Figure 1. The overall proposed framework

3.1. Dataset Description and Features Definition

The study used a publicly available dataset from Kaggle (https://www.kaggle.com/datasets/ziya07/student-engagement-using-biosensor-technology) containing 1,000 samples with precomputed EEG spectral power features. The dataset includes five frequency bands: Alpha, Beta, Theta, Delta, and Gamma, along with participant age, which ranged from 16 to 30 years. These age groups align with the Monastra thresholds we used for attention labeling, which we describe in the next section. Because the dataset provides precomputed EEG spectral power features rather than raw signals, several methodological details that would ordinarily be reported are unavailable. These include the recording device, number of channels, sampling rate, preprocessing protocol (filtering, artifact removal, normalization), and the feature extraction method used to compute spectral band power values. These details are not documented in the dataset source and cannot be verified or reproduced from the available data. This represents an inherent limitation of working with third-party precomputed data and is acknowledged as a constraint on reproducibility and signal quality evaluation throughout this paper.

The five spectral bands each capture a different aspect of cognitive activity. Alpha (8–13 Hz) tends to reflect relaxed, internally focused states [7], [25]. Beta (13–30 Hz) is the band most consistently linked to active attention and alertness [7], [25]. Theta (4–8 Hz) relates to working memory and cognitive control, and elevated theta is often associated with reduced external focus [7], [25]. Delta (0.5–4 Hz) dominates during deep sleep but can appear under low-arousal conditions as well [25]. Gamma (>30 Hz) is associated with higher-order cognitive integration [25]. Using all five bands together gave the model a richer and more complete picture of each student's cognitive state than any single band could provide on its own.

ISSN 2723-6471

We excluded non-EEG variables such as demographic and behavioral attributes from the model inputs. Our reasoning was straightforward: we wanted the classification to depend entirely on physiological signals, not on characteristics like age or gender that could introduce confounds or reduce generalizability.

After applying the Monastra-based labeling procedure described in Section 3.2, 17 samples fell into an intermediate attention category and were excluded. The final dataset contained 983 usable samples — 757 high-attention (77.0%) and 226 low-attention (23.0%). This imbalance was expected given the nature of the thresholding approach, and we accounted for it by prioritizing balanced accuracy and macro F1-score as our main evaluation metrics rather than overall accuracy alone.

3.2. Attention Label Construction

To label each sample as high or low attention, we used the Theta–Beta Ratio (TBR), a well-established EEG index that captures the relative balance between theta and beta brain activity. The ratio is calculated as:

$$TBR = \frac{Theta}{(Beta + \varepsilon)} \tag{1}$$

ε is a small constant added to prevent division by zero. The reasoning behind this measure is straightforward: when theta activity is high relative to beta, the brain is typically in a less engaged, more internally focused state. When beta dominates, the person tends to be alert and externally focused. This relationship has been documented consistently across EEG research [6], [8]. We chose TBR-based labeling rather than using the engagement label already present in the dataset because that label had no documented derivation methodology. We had no way to verify what it measured or how it was constructed, so we built our own labels from scratch using a more transparent and scientifically referenced approach.

To convert continuous TBR values into discrete categories, we applied the age-adjusted thresholds from the Monastra framework. These thresholds were selected because they constitute the most widely cited normative EEG framework linking TBR values to attentional states across developmental age groups, and because the dataset participants fall within the 16–30 age range covered by these norms, making them the most scientifically grounded and transparent basis available for this labeling task. Although originally developed in a clinical context, we used them here as heuristic guidelines rather than diagnostic criteria. They provide a basis grounded in established neuroscience, but do not account for individual variability in EEG patterns arising from factors such as sleep, fatigue, or baseline cognitive differences, and should therefore be understood as a population-level approximation rather than subject-specific calibration.

One important design decision was to keep TBR out of the model's input features entirely. Since TBR was used to construct the labels, including it as a predictor would have created a form of information leakage; the model would essentially be predicting labels from the same signal used to create them. To avoid this, we trained all classifiers exclusively on the five raw EEG spectral band features. Samples that fell into an intermediate attention category, neither clearly high nor clearly low, were excluded. These 17 borderline cases would have introduced ambiguity into the classification task without adding useful signal. After removing them, we were left with 983 samples split into two classes: 757 high-attention and 226 low-attention.

It is worth noting that attention is a multidimensional construct. The decision to model it as a binary variable, high or low, is a deliberate simplification rather than a theoretical claim that attention is inherently dichotomous. We adopted this framing for three pragmatic reasons. First, the Monastra thresholding approach naturally produces a low/high partition with an excluded intermediate zone, making binary labels a direct consequence of our labeling strategy. Second, binary classification is more straightforward to validate reliably with the sample size available (983 instances after filtering). Third, from a practical classroom perspective, the most actionable distinction for a teacher is between students who need immediate intervention (low attention) and students who can be further challenged (high attention). We acknowledge that this simplification discards information about fine-grained variations in attentional states. Future work should explore multi-class or regression-based formulations that treat attention as a continuous or ordinal variable.

3.3. Model Training and Evaluation

The five models were selected to provide a diverse comparison across different learning paradigms: Logistic Regression as a linear baseline, SVM (RBF kernel) for its effectiveness in nonlinear feature spaces, kNN as a distance-based benchmark, and Random Forest and Gradient Boosting as ensemble methods known for strong performance on tabular data. All models were implemented using scikit-learn with default hyperparameters. This choice was deliberate:

given the relatively small dataset size (983 samples), extensive hyperparameter tuning risked overfitting and would complicate fair cross-model comparison.

Model performance was evaluated using stratified 5-fold cross-validation, which preserves class distribution across folds and provides a reliable estimate of generalization performance. Evaluation metrics included overall accuracy, balanced accuracy, macro F1-score, weighted F1-score, and per-class F1-scores. Balanced accuracy and macro F1 were treated as primary metrics given the class imbalance of 77% high-attention versus 23% low-attention samples. Gradient Boosting achieved the highest performance across all primary metrics and was therefore selected as the predictive model for the subsequent explainability and recommendation stages.

3.4. Prediction Stage

After training, the classifiers were used to predict students' attention levels (low and high). The best-performing Gradient Boosting model was selected for use in the next stage. This prediction step provides a basis for further analysis, where the model's behavior is examined more closely using interpretation techniques.

3.5. Explainable Artificial Intelligence

A model that classifies attention accurately but cannot explain its reasoning has limited value in a real classroom. Teachers and school administrators are not going to act on outputs they cannot understand or verify, and they should not have to. This is why we treated interpretability as a core requirement of the framework rather than an optional add-on. We used SHAP to open up the Gradient Boosting model and examine what was driving its predictions [12], specifically via the TreeExplainer method.

SHAP was selected over alternative explainability methods such as LIME for three reasons. First, it is rooted in cooperative game theory, providing theoretically guaranteed consistency and local accuracy. Second, the TreeExplainer method is specifically optimized for tree-based models, making it computationally efficient without approximation. Third, SHAP provides both global and local explanations within a unified framework, enabling the dual-level analysis conducted in this study. A quantitative comparison with other XAI methods is acknowledged as a limitation and is recommended as a direction for future work.

We ran the analysis at two levels. Globally, we aggregated SHAP values across all training samples to get a picture of which EEG features the model consistently relied on most essentially, what it learned to pay attention to. Locally, we computed SHAP values for individual predictions, which allowed us to see exactly how each feature contributed to a specific student being classified as high or low attention. This local view matters in practice. A global summary tells you that Beta activity is important in general; a local explanation tells you that this particular student was flagged as inattentive because their Beta power was unusually low at that moment.

3.6. Instructional Strategy Mapping

The proposed framework uses instructional strategy mapping to establish systematic relationships that connect various learner attention levels to specific teaching methods. The mapping is grounded in the existing literature, which demonstrates that different teaching strategies are differentially effective depending on students' cognitive states, as previously presented in table 2. The system classifies strategies based on their functional role, either supporting attention recovery or sustaining engagement.

Instructors use restorative strategies to regain focus and reduce cognitive exhaustion when students show reduced attention. The combination of movement breaks with short rest periods helps students achieve better concentration because it enables them to restore their attentional capacity [16], [17]. Interactive methods, which include polling and immediate feedback together with pop-up questions in instructional materials, help students interact with the learning material while participating actively in their studies [18], [24]. The strategies function effectively to prevent students from becoming disengaged, which happens when they participate in passive educational activities or spend extended periods working on mental tasks.

When students maintain high levels of concentration, instructors use advanced learning elements to maintain student interest while helping them achieve a deeper understanding. Students who participate in challenge-based tasks and open-ended problem-solving activities develop skills to think at higher levels while maintaining mental engagement [20]. Peer teaching enables students to create knowledge through explanation and collaborative work [21]. Students develop critical thinking skills through Socratic questioning and debate-based activities because these methods require them to analyze their answers before giving justification [22].

This mapping provides a structured foundation by clearly defining the relationship between attention states and instructional strategies. It also serves as the basis for the subsequent recommendation strategy, where appropriate interventions are dynamically selected in response to learners' cognitive conditions.

It is acknowledged that the mapping between attention states and instructional strategies remains heuristic in nature. While each strategy is grounded in peer-reviewed educational literature, as shown in table 2, the mapping criteria were not formally optimized or empirically validated within this study. The selection of strategies for each attention level was based on their documented functional role, either restoring attention or sustaining engagement as established in the existing literature. Formal empirical evaluation of strategy effectiveness and adaptive optimization of the mapping criteria are identified as important directions for future work.

3.7. Recommendation Strategy

Once the model has classified a student's attention state and SHAP has explained which EEG features drove that classification, the final step is to turn that information into something a teacher can use. We did this by connecting the classification output to a structured knowledge base (KB) of instructional strategies, organized by attention level as shown in table 2, and using an LLM to select the most appropriate strategy from that set. It should be noted that while the recommendations are grounded in peer-reviewed educational literature through the predefined knowledge base, they have not been formally evaluated for pedagogical effectiveness with actual educators. Empirical validation through user studies with teachers is acknowledged as a key limitation and is identified as the most important direction for future work.

The selection process works as follows. The system takes the predicted attention label high or low and filters the KB to the strategies mapped to that level. It then passes the filtered candidates, along with the SHAP-derived explanation and a cognitive interpretation of the student's EEG pattern, to GPT-4o-mini via the OpenAI API. GPT-4o-mini (OpenAI, 2024) was selected for this component due to its instruction-following capability and cost efficiency. It was used exclusively for strategy selection and recommendation generation within the predefined knowledge base and was not employed in any other part of this study. The model is instructed to choose exactly one strategy from the provided list and to justify its choice using evidence from the KB definition. It cannot invent new strategies or draw on knowledge outside the predefined set. This constraint was deliberate, and it was important to us that every recommendation remain grounded and verifiable.

The LLM was provided with a structured prompt containing three components: (1) the predicted attention label — high or low, (2) the SHAP-derived feature importance values and their plain-language cognitive interpretation, and (3) the filtered list of instructional strategies from the knowledge base corresponding to the predicted attention level. The model was explicitly instructed to: select exactly one strategy from the provided list, justify its selection using evidence from the KB definition, and provide step-by-step implementation guidance. It was strictly constrained from generating recommendations outside the predefined strategy set. This constraint ensures consistency and auditability of all generated outputs. Formal evaluation of prompt consistency, bias, and reliability of generated suggestions remains a direction for future work.

The EEG feature pattern influences which strategy gets selected within the filtered set. For instance, a student showing elevated delta activity alongside low beta, a pattern suggestive of fatigue or low arousal, would likely receive a rest break recommendation. A student with high theta and moderate beta, which tends to reflect wandering attention rather than fatigue, might instead be directed toward an interactive prompt or a pop-up question embedded in the learning material. Along with the selected strategy, the system generates a brief explanation of why that strategy fits the student's current state and how it can be applied in practice.

It should be noted that the prototype has not been formally evaluated through a usability study involving actual educators. It represents a proof-of-concept demonstration of the end-to-end framework rather than a production-ready deployment tool.

3.8. System Implementation

To show that the framework works in practice and not just on paper, we built an interactive prototype using Python and Streamlit. The interface is designed with the teacher in mind; it is not a research dashboard full of technical outputs, but a tool that an instructor could realistically sit in front of during or after a class. The teacher begins by selecting a sample student from a slider. The system immediately displays that student's EEG spectral band values across all five frequency bands, their TBR-derived attention label, and the model's prediction. From there, the interface shows the

SHAP waterfall plot for that specific student, which visualizes exactly how each EEG feature contributed to the classification. This local explanation is presented alongside a plain-language cognitive interpretation, a summary that translates the SHAP output into something meaningful without requiring any knowledge of machine learning.

The LLM component sits at the end of this pipeline. Once the attention state and SHAP explanation are available, GPT-4o-mini selects the most appropriate instructional strategy from the knowledge base and generates a structured recommendation that includes a justification grounded in the KB and step-by-step guidance on how to apply it. The model is always attached to the predefined strategy list; it cannot generate recommendations outside that set, which keeps the outputs consistent and auditable.

On the technical side, the backend runs on scikit-learn for model inference and the SHAP library for explanation generation. The full system from EEG input to instructional recommendation runs as a single Streamlit application, which makes it straightforward to deploy and extend. The user interface is shown in figure 3.

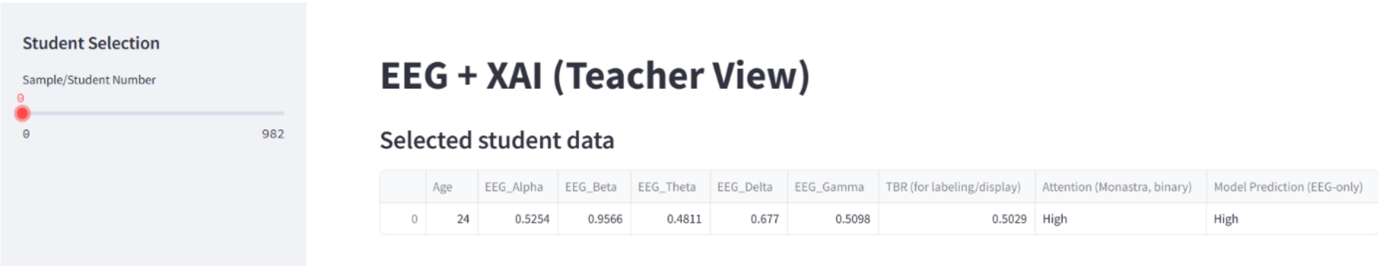


Figure 3. User interface

4. Results and Discussion

After applying the Monastra-based labeling procedure and removing the 17 intermediate-attention samples, we were left with 983 usable samples 757 high-attention (77.0%) and 226 low-attention (23.0%). The class imbalance was notable but not unexpected given how the TBR thresholds partition the data, and we accounted for it by treating balanced accuracy and macro F1 as our primary metrics rather than overall accuracy. We trained and compared five classifiers under stratified 5-fold cross-validation. The results are summarized in table 3. All five models performed reasonably well, but there were meaningful differences between them, particularly in the low-attention class, which is the harder and more important one to get right.

Gradient Boosting came out on top across all metrics. It reached an overall accuracy of 0.983 ± 0.006 and a weighted F1 of 0.983 ± 0.006, but what stood out more was its performance on the low-attention class, specifically an F1 of 0.963, which was the highest of any model. This is likely because Gradient Boosting builds sequentially, with each tree correcting the mistakes of the previous one, which tends to work well when the decision boundary is nonlinear, and one class is harder to capture than the other. Random Forest and Logistic Regression both performed well overall. Random Forest reached 0.975 accuracy, and Logistic Regression, somewhat surprisingly, achieved the highest balanced accuracy of the group at 0.982 ± 0.010, suggesting it handled the class imbalance particularly well at the aggregate level. That said, as we discuss in the confusion matrix analysis below, Logistic Regression's behavior on the low-attention class deserves a closer look. SVM and kNN lagged the other three, especially on low-attention samples. kNN produced the weakest low-attention F1 of 0.885 ± 0.032, which reflects a well-known limitation of distance-based methods when class sizes are unequal, the majority class dominates the neighborhood, and minority samples get pulled toward the wrong prediction.

Table 3. Model Performance Comparison

Model	Accuracy	Balanced Acc	Macro F1	Weighted F1
Logistic Regression	0.973 ± 0.016	0.982 ± 0.010	0.963 ± 0.020	0.973 ± 0.015
Gradient Boosting	0.983 ± 0.006	0.978 ± 0.007	0.976 ± 0.008	0.983 ± 0.006

Model	Accuracy	Balanced Acc	Macro F1	Weighted F1
SVM (RBF)	0.958 ± 0.015	0.967 ± 0.012	0.944 ± 0.019	0.959 ± 0.014
Random Forest	0.975 ± 0.009	0.965 ± 0.008	0.964 ± 0.013	0.975 ± 0.009
kNN	0.950 ± 0.013	0.912 ± 0.027	0.927 ± 0.020	0.949 ± 0.014

These results should be interpreted as an upper bound on true generalization performance for two reasons. First, the absence of participant identifiers means we could not enforce subject-independent splits; the same individual may appear in both training and test folds, which would inflate estimates relative to performance on completely unseen subjects. Second, the high classification performance is partly attributable to the shared spectral origin of the labeling mechanism and model inputs. Although TBR itself was strictly excluded from model inputs to prevent direct leakage, the models learned from raw band power values that share a spectral origin with the labeling signal, representing a legitimate but inherently related classification task. This may lead to optimistic performance estimates compared to deployment on completely unseen individuals in real-world settings.

4.1. Class-wise Performance

Table 4 presents detailed class-wise performance, including F1 scores for both high- and low-attention classes.

Table 4. Class-wise F1-scores for high- and low-attention classes.

Model	F1 High	F1 Low
LR	0.982 ± 0.010	0.944 ± 0.030
SVM (RBF)	0.972 ± 0.010	0.916 ± 0.028
kNN	0.968 ± 0.008	0.885 ± 0.032
Gradient Boosting	0.989 ± 0.004	0.963 ± 0.012
RF	0.983 ± 0.006	0.945 ± 0.019

All models show improved performance for detecting the high-attention class because this class appears more frequently in the dataset. The evaluated models show Gradient Boosting and RF as their best performers because these models achieved F1-scores above 0.94. The kNN model demonstrates poor performance in low-attention detection because it produced the lowest F1-score among all testing models. The SVM system shows moderate performance because it cannot identify subtle differences between different attention states. The results support the selection of the Gradient Boosting model because it functions best when the system needs to detect low-attention states.

4.2. Confusion Matrix Analysis

The confusion matrices show how well the system works to classify different attention levels. The results for the evaluated models are illustrated in figure 4. The evaluated models show that Gradient Boost delivers its best performance with only slight errors in both class categories. LR shows perfect sensitivity in detecting low-attention samples, although it slightly overestimates this class. RF achieves high performance results while SVM shows consistent performance with slightly increased error rates. kNN exhibits lower effectiveness when it comes to identifying low-attention states.

Notably, Logistic Regression achieved perfect sensitivity on the low-attention class with zero false negatives, which may reflect overfitting to the decision boundary rather than genuine generalization, further supporting the selection of Gradient Boosting, whose error distribution is more balanced across both classes.

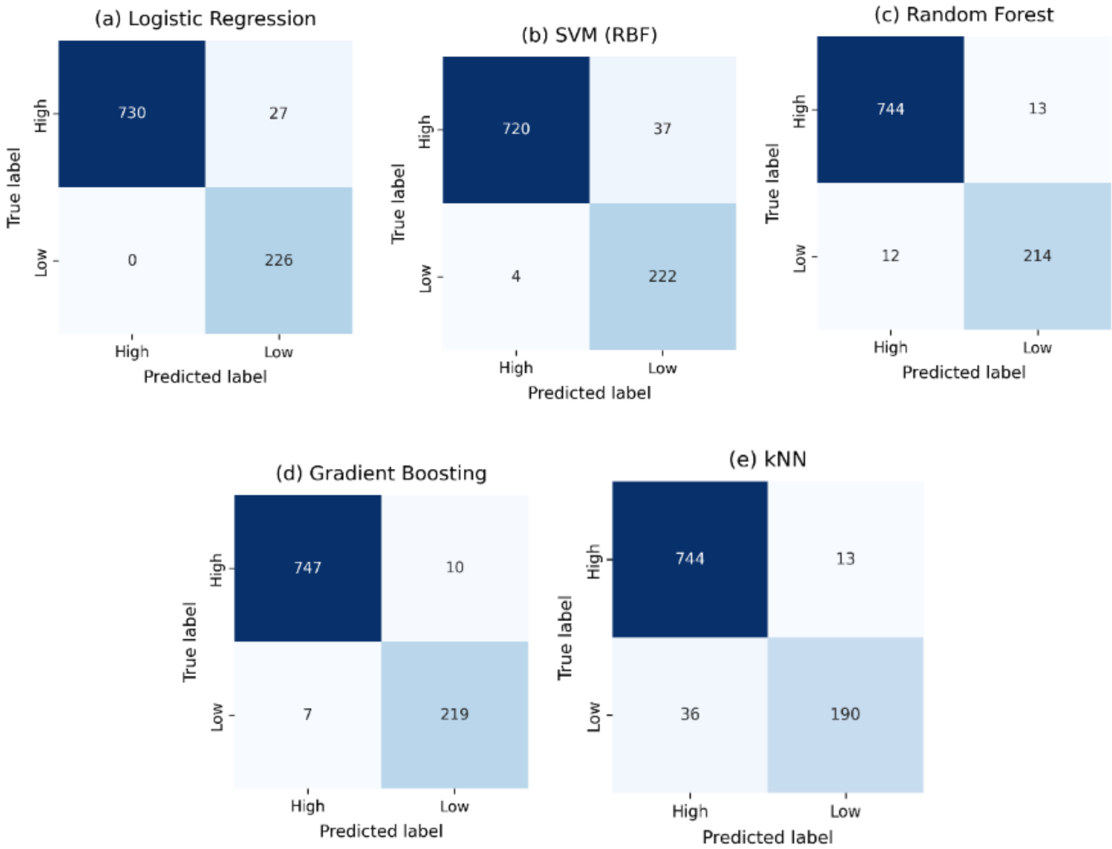
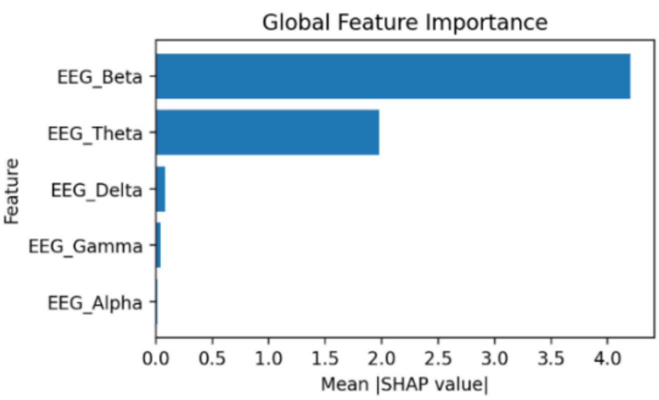


Figure 4. Confusion matrices of the evaluated classification models.

4.3. XAI Results

Global model examination was conducted using SHAP via the TreeExplainer method. The mean absolute SHAP values from all samples were used to calculate the global feature importance of the model. The results shown in figure 5 demonstrate that Beta band activity is by far the most influential feature, with a mean absolute SHAP value of 4.2013, more than twice the combined value of all remaining features. Theta ranked second at 1.975, while Delta (0.0824), Gamma (0.0404), and Alpha (0.0132) showed negligible contributions, confirming that Beta and Theta are the primary drivers of attention classification in this model.



	Feature	MeanAbsSHAP
1	EEG_Beta	4.2013
2	EEG_Theta	1.9755
3	EEG_Delta	0.0824
4	EEG_Gamma	0.0404
0	EEG_Alpha	0.0132

Figure 5. Global SHAP feature importance

To further examine the directional influence of each EEG feature, a SHAP beeswarm plot was generated as shown in figure 6. Each point represents one sample, colored by feature value (red = high, blue = low), with horizontal position indicating the direction and magnitude of impact on the model output. The plot reveals that high Beta activity (red dots) consistently produces large positive SHAP values, driving predictions toward the high attention class, while low Beta values (blue dots) push toward the low attention class. Theta activity exhibits the opposite directional pattern: high Theta values (red dots) produce negative SHAP values associated with low attention, while low Theta values support high attention predictions. This bidirectional relationship between Beta and Theta is fully consistent with the Theta–Beta Ratio framework underlying the label construction and aligns with established neuroscience literature on attentional regulation. Delta, Gamma, and Alpha remain tightly clustered around zero in both directions, confirming their negligible contribution to attention classification.

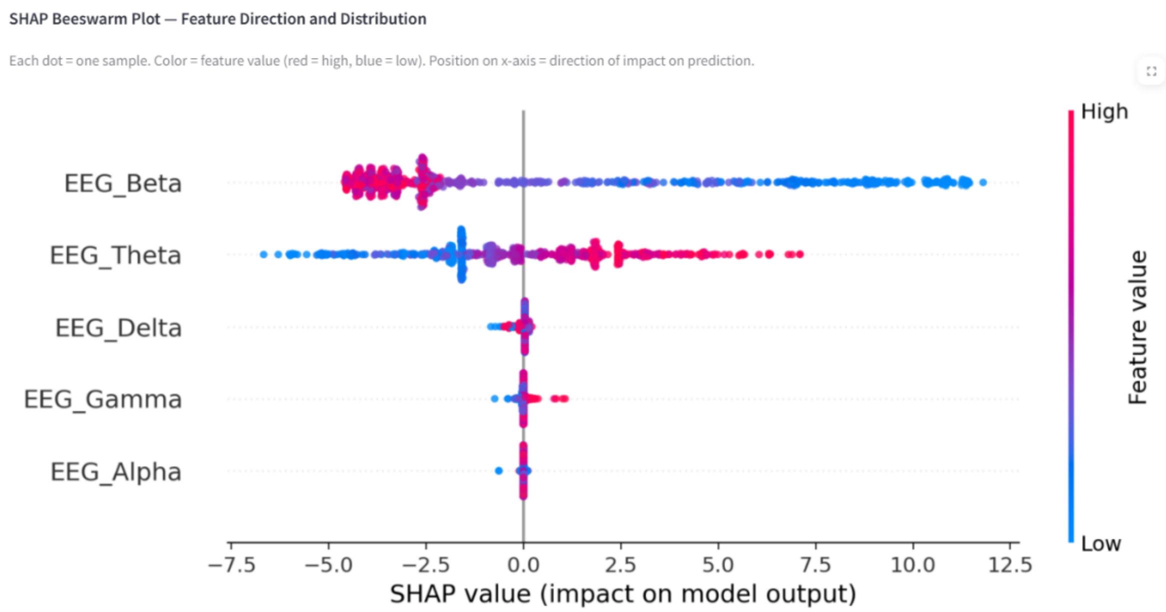


Figure 6. beeswarm plot showing the direction and distribution of each feature's impact across all samples.

To further interpret the model at the individual level, local SHAP analysis was conducted for selected samples. The results are illustrated in figure 7 using a waterfall plot, which explains how each EEG feature contributes to the final prediction. At the local level, Beta activity exerted the strongest influence on individual predictions, consistently driving classifications toward the high-attention class. Theta activity contributed to a secondary predictive influence, with a notably smaller effect magnitude than Beta, Delta, Alpha, and Gamma bands, which produce only slight effects, making their impact on prediction results very small. The consistency between global and local SHAP analyses confirms that Beta and Theta bands are the main factors that determine attention classification.

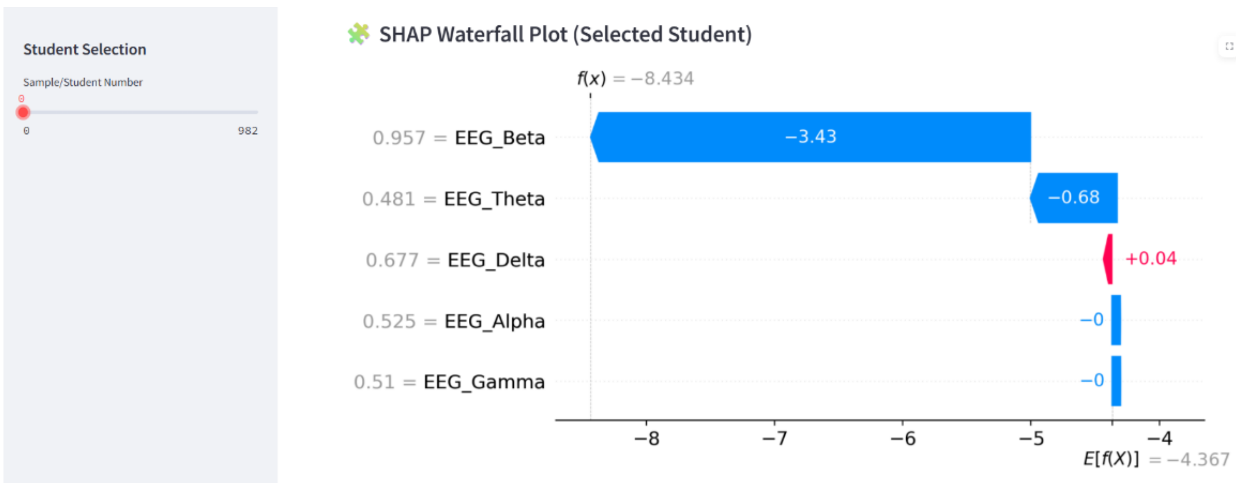


Figure 7. Local SHAP explanation using a waterfall

The cognitive interpretation from SHAP analysis shown in figure 8 provides contextual information about how learners focus their attention. The LLM uses these insights to determine which teaching methods to use.

ISSN 2723-6471

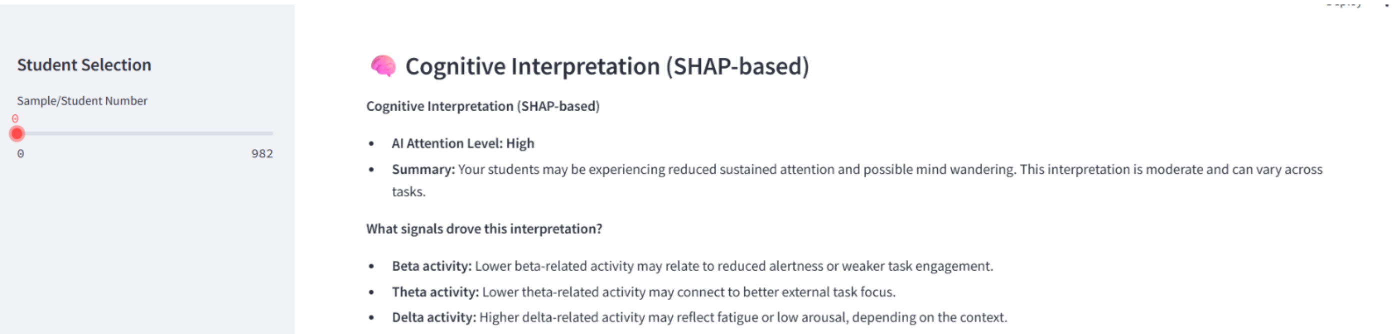


Figure 8. Example of LLM-generated cognitive interpretation and instructional strategy selection based on SHAP explanations and detected attention state.

4.4. Recommendation Strategy Results (LLM-based)

To enhance the adaptability of the proposed framework, LLM was integrated to support the selection of instructional strategies based on SHAP-derived explanations and cognitive interpretations. As illustrated in figure 9, the system recommends strategies that include (1) justification for the chosen strategy and (2) step-by-step instructions for implementation in the learning context. Each recommendation is supported by a reference from a pre-defined knowledge base, ensuring that the output remains pedagogically grounded. For example, the "timed challenge" strategy was chosen for the student's high attention level. The LLM justified this choice by linking it to the student's cognitive state, stating that short, time-restricted tasks can help maintain focus, re-engage attention, and improve responding.

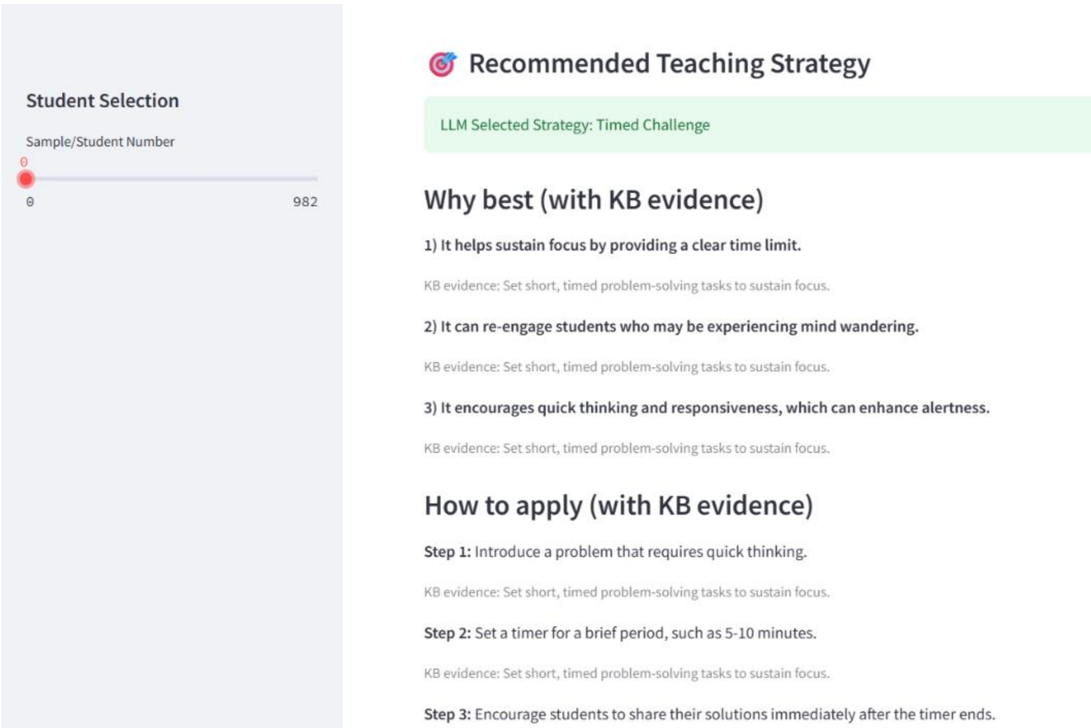


Figure 9. LLM-generated strategy recommendation based on attention classification and SHAP analysis

Beyond selecting the strategy, the system generates concrete implementation steps for instance, introducing a time-limited problem, setting a fixed duration, and prompting students to share their answers immediately afterward. This moves the output from an abstract recommendation to something a teacher can act on in the next five minutes.

Taken together, the results paint a consistent picture. Gradient Boosting handled the classification task well, and the SHAP analysis gave us a clear and interpretable account of why Beta and Theta activity dominated the predictions at both the global and local levels, which lines up with what we know about these bands from neuroscience. The LLM component then translated those classified and explained outputs into grounded instructional recommendations, completing the pipeline from raw EEG signal to practical classroom action.

ISSN 2723-6471

What we found most encouraging was not the accuracy figures themselves, but the consistency between layers. The features that mattered globally were the same ones that explained individual predictions. The recommendations that came out of the LLM were grounded in the same KB entries across different students and sessions. This kind of internal coherence is important if a system like this is ever going to be trusted in a real educational environment.

4.5. Limitations

Several limitations should be acknowledged. First, regarding dataset and signal-level constraints, the dataset does not include acquisition details such as device type, number of channels, sampling rate, or preprocessing protocol, as it provides precomputed spectral features rather than raw EEG signals. This limits reproducibility at the signal level and prevents evaluation of preprocessing quality, including artifact removal and filtering.

Second, regarding generalization and performance bounds, the absence of participant identifiers means subject-independent splits could not be enforced; the same individual may appear in both training and test folds, meaning reported figures represent an upper bound rather than true generalization performance. While cross-fold variance was low (accuracy: 0.983 ± 0.006), this consistency should be interpreted cautiously rather than as confirmation that overfitting is absent, given that the absence of subject-level identifiers limits the diagnostic value of cross-validation in this setting. The framework has also been evaluated on a single dataset collected under controlled conditions with participants aged 16 to 30 years, and performance may vary across different age groups, educational levels, learning tasks, and EEG hardware setups.

Third, regarding labelling assumptions, the Monastra thresholds were developed for clinical populations and applied here as heuristic guidelines rather than diagnostic criteria. As noted in Section 3.2, these thresholds do not account for individual variability in EEG patterns and should be understood as a population-level approximation rather than subject-specific calibration. Combined with the binary framing of attention, which simplifies what is inherently a continuous and multi-dimensional construct, these choices introduce approximations that future work should refine through adaptive or continuous labelling approaches.

Fourth, regarding system evaluation, the proposed framework is a proof-of-concept prototype intended to demonstrate the feasibility of integrating EEG, XAI, and LLM components within a unified pipeline. The Streamlit prototype has not been subjected to a formal usability evaluation, and no user study involving educators was conducted. The prompting strategy and output consistency of the LLM component have not been formally evaluated, and end-to-end latency under real classroom conditions remains untested. These represent the most important directions for future work.

5. Conclusion and Future Work

This paper addresses a gap that has persisted in the EEG-based attention literature: most existing work stops at classification and never connects its outputs to what a teacher should actually do. We present a framework that connects the EEG signal to the attention label, the attention label to the SHAP explanation, and the SHAP explanation to a concrete, knowledge-grounded instructional recommendation delivered through an LLM component.

The classification results were strong. Gradient Boosting outperformed the other four models across all primary metrics, reaching 98.3% accuracy and a weighted F1 of 98.3%, with particularly reliable detection of low-attention samples, the class that matters most for timely intervention. These figures should be interpreted as an upper bound on true generalization performance, as the absence of subject-level identifiers prevented subject-independent validation, and the shared spectral origin of the labeling mechanism and model inputs further contributes to optimistic estimates.

The SHAP analysis told a clear and interpretable story. Beta and Theta band activity were by far the most influential features, both globally and in individual predictions. The beeswarm plot made directional logic explicit: high Beta values consistently pushed predictions toward high attention, while high Theta values did the opposite. This pattern maps directly onto the physiological rationale behind the Theta–Beta Ratio and provides evidence that the model's decisions reflect genuine neurophysiological signal rather than statistical artifacts.

There is still meaningful work ahead. The most important next step is a structured usability study involving actual educators to evaluate whether the LLM-generated strategies are perceived as useful, whether the SHAP explanations are understandable, and whether the system changes how teachers respond to attention problems in practice. Beyond that, the framework needs to be tested on larger and more diverse datasets, ideally with live EEG data collected in real classroom conditions rather than controlled settings. Subject-independent validation using leave-one-subject-out evaluation on datasets where participant identity is tracked should also be a priority. Future work should additionally explore multi-class or regression-based attention modelling approaches that move beyond the binary framing adopted here, enabling more nuanced and personalized instructional responses. We see this work as a foundation, not a finished product.

6. Declarations

6.1. Author Contributions

Conceptualization: S.D.; Methodology: S.D.; Software: S.D.; Validation: S.D., W.A.; Formal Analysis: S.D.; Investigation: S.D.; Resources: W.A.; Data Curation: S.D.; Writing – Original Draft Preparation: S.D.; Writing – Review & Editing: S.D., W.A., M.A.; Visualization: S.D. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

"The dataset used in this study is publicly available on Kaggle at: https://www.kaggle.com/datasets/ziya07/student-engagement-using-biosensor-technology".

6.3. Funding

The research was funded by KAU Endowment (WAQF) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Zhang, “Cognitive Functions of the Brain: Perception, Attention and Memory,” arXiv, vo. 2019, no. May, pp. 1-12, May 30, 2019, *arXiv:* arXiv:1907.02863. doi: 10.48550/arXiv.1907.02863.

[2] J. Sweller, “Cognitive Load During Problem Solving: Effects on Learning,” *Cognitive Science*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[3] A. Diamond, “Want to Optimize Executive Functions and Academic Outcomes? Simple, Just Nourish the Human Spirit,” in *Minnesota Symposia on Child Psychology, 1st ed., P. D. Zelazo and M. D. Sera, Eds., Wiley*, 2013, pp. 203–230. doi: 10.1002/9781118732373.ch7.

[4] N. Bergdahl, "Engagement and disengagement in online learning," Computers & Education, vol. 188, Art. no. 104476, pp. 1-21, 2022. doi: 10.1016/j.compedu.2022.104476

[5] M. K. Syed, H. Wang, A. A. Siddiqi, S. Qureshi, and M. A. Gouda, "EEG-Based Attention Classification for Enhanced Learning Experience," *Applied Sciences*, vol. 15, no. 15, pp. 8668-8679, 2025, doi: 10.3390/app15158668.

[6] A. Al-Nafjan and M. Aldayel, “Predict Students’ Attention in Online Learning Using EEG Data,” *Sustainability*, vol. 14, no. 11, pp. 6553-6565, 2022, doi: 10.3390/su14116553.

[7] R. H. C. e Souza and E. L. M. Naves, “Attention Detection in Virtual Environments Using EEG Signals: A Scoping Review,” *Frontiers in Physiology*, vol. 12, Art. no. 727840, pp. 1–20, 2021, doi: 10.3389/fphys.2021.727840.

[8] W. Wan, X. Cui, Z. Gao, and Z. Gu, “Frontal EEG-Based Multi-Level Attention States Recognition Using Dynamical Complexity and Extreme Gradient Boosting,” *Front. Hum. Neurosci*., vol. 15, no. Jun., pp. 1-15, Jun. 2021.

[9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” *ACM Comput. Surv*., vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[10] Z. C. Lipton, "The mythos of model interpretability," *ACM Queue*, vol. 16, no. 3, pp. 31–57, 2018. doi: 10.1145/3236386.3241340

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM*, vol.. 2016, no. Aug., pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc*., vol. 2017, no. 1, pp. 4765–4774, 2017, doi: 10.48550/arXiv.1705.07874.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, no.1, pp. 1877–1901, 2020, doi: 10.48550/arXiv.2005.14165.

[14] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Art. no. 102274, pp.1-12, 2023.

[15] K. Suwida, S. C. Hidayati, and R. Sarno, "Application of Machine Learning Algorithm for Mental State Attention Classification Based on Electroencephalogram Signals," in *Proceedings of the 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia*, vol. 2024, no. Feb., pp. 354–358, 2023, doi: 10.1109/ICCoSITE57641.2023.10127825.

[16] C. L. Peiris, G. O'Donoghue, L. Rippon, D. Meyers, A. Hahne, M. De Noronha, J. Lynch, and L. C. Hanson, "Classroom Movement Breaks Reduce Sedentary Behavior and Increase Concentration, Alertness and Enjoyment During University Classes: A Mixed-Methods Feasibility Study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, Art. no. 5589, pp. 1–13, May 2021, doi: 10.3390/ijerph18115589.

[17] P. Ginns, K. Muscat, and R. Naylor, "Rest Breaks Aid Directed Attention and Learning," *Educational and Developmental Psychologist*, vol. 40, no. 2, pp. 141–150, Jul. 2023, doi: 10.1080/20590776.2023.2225700.

[18] M. E. Haagsman, K. Scager, J. Boonstra, and M. C. Koster, "Pop-up Questions Within Educational Videos: Effects on Students' Learning," *J Sci Educ Technol*, vol. 29, no. 6, pp. 713–724, Dec. 2020, doi: 10.1007/s10956-020-09847-3.

[19] M. S. Doğru, F. Yüzbaşıoğlu, and I. Arpaci, "The effect of interactive videos enhanced with pop-up questions on teacher candidates' learning performance in science," *Research in Science & Technological Education*, vol. 43, no. 1, pp. 312–327, Jan. 2025, doi: 10.1080/02635143.2023.2272820.

[20] M. Prince, "Does active learning work? A review of the research," *Journal of Engineering Education*, vol. 93, no. 3, pp. 223–231, 2004, doi: 10.1002/j.2168-9830.2004.tb00809.x.

[21] M. Stigmar, "Peer-to-peer Teaching in Higher Education: A Critical Literature Review," *Mentoring & Tutoring: Partnership in Learning*, vol. 24, no. 2, pp. 124–136, Mar. 2016, doi: 10.1080/13611267.2016.1178963.

[22] X. Chen, L. Wang, X. Zhai, and Y. Li, "Exploring the effects of argument map-supported online group debate activities on college students' critical thinking," *Front. Psychol*., vol. 13, Art. no. 856462, no. 1, pp. 1-12, 2022, doi: 10.3389/fpsyg.2022.856462.

[23] J. Metcalfe, "Learning from Errors," *Annu. Rev. Psychol*., vol. 68, no. 1, pp. 465–489, Jan. 2017, doi: 10.1146/annurev-psych-010416-044022.

[24] J. C.-Y. Sun, "Influence of polling technologies on student engagement: An analysis of student motivation, academic performance, and brainwave data," *Computers & Education*, vol. 72, no. Mar., pp. 80–89, Mar. 2014, doi: 10.1016/j.compedu.2013.10.010.

[25] J.-W. Wang, D.-W. Zhang, and S. J. Johnstone, "Portable EEG for assessing attention in educational settings: A scoping review," *Acta Psychologica*, vol. 255, no. May, pp. 1-13, May 2025, doi: 10.1016/j.actpsy.2025.104933.