

ORIGINAL RESEARCH

Open access

Mapping Understanding through Biometric Signals: AI Approaches to Cognitive and Affective Phases

Omar Zayed^{1*}, Leila Ibrahim¹, Ahmed Mansour²

Abstract

The connection between the cognitive and emotional components of comprehension has remained largely unexamined because of a shortage of dependable techniques for assessing feelings and emotions during educational activities. Concentrating on five distinct stages of comprehension—nascent understanding, misunderstanding, confusion, emergent understanding, and deep understanding—this investigation presents an AI-based method to assess subjective comprehension by tracking the physiological activity visible in facial expressions. To examine these stages, 103 individuals worked on 15 riddles in a remote setting while their facial expressions were video recorded. Action units (AUs) for each stage occurrence were quantified utilizing AFFDEX software. Recurring AU configurations linked to each stage were subsequently pinpointed by implementing six supervised machine learning algorithms. Unique AU configurations were identified for all five stages, with gradient boosting machine and random forest models delivering the greatest predictive precision. These outcomes indicate that physiological activity can be utilized to dependably assess comprehension. Furthermore, they propose an innovative methodology for assessing and encouraging comprehension in instructional environments, as well as designing adaptive educational technologies and tailored learning interventions. Future research should investigate how the physiological indicators of comprehension stages simultaneously mirror and impact their related cognitive workflows, alongside the generalizability of these outcomes across varied demographics and educational environments (A suite of AI tools was employed in the development of this paper: (1) ChatGPT4o (for writing clarity and reference checking), (2) Grammarly (for grammar and editorial corrections), and (3) ResearchRabbit (reference management)).

Keywords Machine learning, Subjective understanding, Physiological activity, Facial expressions, Action units (AUs)

*Correspondence:

Omar Zayed
omar.zayed@cu.edu.eg

¹ Department of Cognitive Science and Biometric AI, Faculty of Medicine, Cairo University, Cairo, Egypt

² Department of Affective Computing and Signal Processing, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Introduction

Comprehension is arguably the foundation of education. Its importance is acknowledged, both instinctively and explicitly, by learners pursuing conceptual clarity, instructors cultivating critical thinking rather than rote learning, and policymakers introducing programs such as the “Transforming American Education: Learning Powered by Technology” initiative from the U.S. Department of Education [1]. This recognition is fueled by the constructive results that comprehension generates, including insights into what is meaningful and appropriate, fresh paths for action, and advancement in individual growth. Consequently, it is vital to resolve deficiencies in what is known about comprehension, particularly within educational psychology and other fields that guide instructional reform. Only then can learners achieve their maximum academic and individual potential, instructors cultivate comprehensive comprehension, and policymakers implement successful guidelines and benchmarks for performance.

Although difficult to define, comprehension is frequently characterized in two ways: First, as a cognitive workflow that entails grasping the meaning of information, enabling people to utilize their knowledge successfully [2, 3]; second, as a subjective sensation—that something “makes sense”—which allows us to proceed with assurance [4, 5]. However, both viewpoints lack the integration of the alternative perspective, a blend long supported by educational, developmental, and social psychologists [6–12]. The difficulty of incorporating sensations and emotions into cognitive workflows has been a subject of psychological debate since the scholarship of William James [13], who emphasized their subjective and inward-looking nature. Historically, self-report measures were chiefly utilized to explore the emotional component of comprehension. However, when these techniques were found to be untrustworthy, they—and the analysis of sensations and emotions linked to comprehension—were largely set aside.

More recently, the scientific perspective on sensations and emotions has progressed [14–17]. Damasio contends that while emotions are subconscious, involuntary responses to indicators, sensations are the aware reflections of these emotions, arising from the internal perception of physiological variation. Through this framework, positively and negatively valenced sensations merge bodily responses into subjective encounters, mirroring shifts in mental states and driving modifications in behavior based on evaluations of circumstances relative to individual objectives or requirements [18–20]. For Damasio, sensations function within an embodied cognition structure, linking sympathetic and parasympathetic nervous system responses to cognitive workflows like perception and decision-making. Within this structure, the brain dynamically controls and translates the body's physiological reactions to its surroundings in an ongoing feedback loop [7, 14, 21]. This brain–body–environment connection creates an intricate, adaptive, self-organizing framework where cognitive workflows, sensations, and emotions surface as mutually dependent traits [22, 23].

Technological breakthroughs in quantifying physiological reactions, paired with machine learning, now offer the instruments vital for analyzing the emotional component of comprehension. Facial expression interpretation software such as AFFDEX [24] leverages deep learning to track AUs—movements of facial musculature—in real time according to the Facial Action Coding System [15]. AUs indicate parasympathetic and sympathetic nervous system engagement, and unique blood circulation patterns in the facial skin correspond to specific emotions [25]. Consequently, sensations are displayed in facial expressions. If this holds true, the sensation of comprehension can be dependably assessed.

While striving toward comprehension, students may pass through five separate stages that denote alterations and advancement in their comprehension: nascent understanding, misunderstanding, confusion, emergent understanding, and deep understanding [20]. These stages are fluid and do not necessarily occur in a rigid order; students may cycle between them as they meet fresh information or confront obstacles. This workflow positions comprehension along a spectrum, distinguished by changing degrees of depth and complexity [26]. For instance, a student beginning to analyze general relativity might commence with a nascent understanding marked by inadequate prior knowledge, a restricted capacity to build meaningful links, and trouble determining pertinent information [27]. Afterward, misunderstandings might surface from partial or erroneous details, potentially causing an inflated belief in their own comprehension [28] or from the sensation of being stuck despite fulfilling an external benchmark of accuracy [4]. Upon investigating further, they might experience confusion when clashes or paradoxes emerge between their baseline knowledge and fresh details [29]. Nevertheless, an emergent understanding takes shape by synthesizing this information through cognitive processing and sudden insights [30, 31]. As this comprehension progresses, it remains partial and changeable. Still, with persistent study and contemplation, deep understanding can be reached, marked by comprehensive and enduring knowledge of the topic, successful application of knowledge, and the capacity to explain ideas articulately [32].

This investigation is positioned within our wider research goal of pinpointing physiological indicators of sensations and emotions and examining how these affective configurations mirror and impact cognitive workflows in education. Here, we analyze comprehension as a blend of cognitive workflows and emotional encounters, suggesting that the sensation of comprehension arises from conscious internal perception of physiological variation. Thus, we concentrate on subjective comprehension, which is rooted in individual experience, rather than objective comprehension, which relies on validated knowledge and structures. To guide this exploration, we delineate the workflow of comprehension using five simplified stages: nascent understanding, misunderstanding, confusion, emergent understanding, and deep understanding. We

intend to pinpoint the sensations matching each stage and suggest that each stage is linked to specific AU configurations. To examine this suggestion, we monitored these stages as individuals solved word puzzles, using facial expression interpretation and machine learning algorithms to pinpoint their linked AU configurations.

The core contributions of this investigation are: (1) creating a dependable structure for assessing the sensation of comprehension, rooted in the concepts of embodied cognition; (2) advancing the blending of cognitive workflows and emotional encounters linked to subjective comprehension; and (3) adding to the creation of an AI instructor capable of monitoring stages of comprehension without relying on direct communication, thereby improving educational results.

Materials and Methods

Procedure

Participants finished this research online using Zoom during the COVID-19 pandemic. To guarantee premium video data, individuals were positioned squarely in front of their webcams with proper facial illumination, alongside further requirements (e.g., absence of facial obstructions, a firm surface for their laptop, no mobile phones or tablets, no green-screen tools or background movement). Individuals were provided a Qualtrics link where they completed a consent document and demographic survey before trying to solve 15 word puzzles. While participants solved the puzzles, they broadcasts their displays via Zoom, enabling real-time observation of their advancement and the delivery of input during the session. To maintain uniformity and a impartial emotional environment, the lead author ran every session, following a uniform procedural text and puzzle-specific input. Video capture commenced when individuals began the first puzzle, with the administrator's video and microphone shut off to diminish any prospective impact on individual performance.

Individuals were allotted three minutes to resolve each puzzle, granting ample time for reading, analyzing, and answering the puzzle while remaining cautious of cognitive burnout. After three minutes, individuals were prompted to deliver an answer and rationale. Once an answer and rationale were shared, spoken input was provided, pointing out whether the solution was correct or incorrect. If an individual could not deliver a correct answer, the accurate solution and explanation were then stated aloud. Before proceeding to the subsequent puzzle, individuals were asked if they were already familiar with the correct solution to the ongoing puzzle to establish whether it ought to be left out. After finishing the puzzles, individuals were debriefed and paid CAD 10 via electronic transfer. The Zoom video files were subsequently processed using AFFDEX and AUs were quantified (Figure 1).

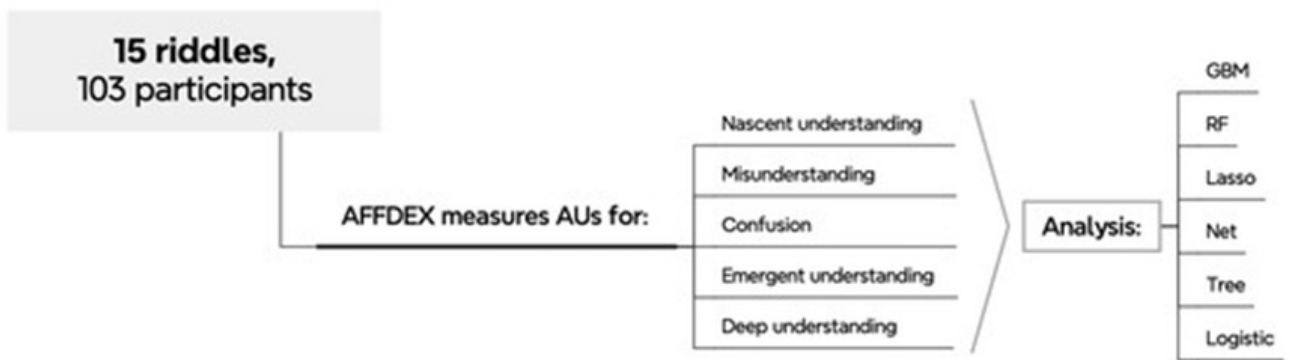


Figure 1. Simplified flowchart illustrating the methodology.

Participants

A total cohort of 103 individuals was enrolled in this investigation, of which 78 (75.7%) were female, presenting a mean age of 24 years ($SD = 4.3$). Digital channels (e.g., Facebook), student-affiliated networks (e.g., University of Toronto Psychology Student's Association), and word-of-mouth recommendations were utilized for participant recruitment. The majority of the cohort consisted of students actively enrolled in undergraduate courses at the University of Toronto.

Phases of understanding

To evaluate the phases of understanding, participants solved 15 systematically selected riddles, with each task structured to possess a single, unambiguous solution and an accompanying rationale. Riddles serve as an effective task paradigm due to their capacity to prompt occurrences of each phase, present cognitive challenges, and elicit emotional responses, while concurrently promoting brief responses, justifications, and active participant engagement. The riddles, collected from academic articles [33, 34] and digital repositories, varied across both difficulty levels and problem types. For each riddle, individuals were required to submit a written response, a justification, and their degree of confidence using a four-point scale (1 = Not certain, 2 = Somewhat certain, 3 = Certain, 4 = Very certain). A four-point scale was originally implemented to analyze whether nuanced response gradations could assist in differentiating between specific phases. Ultimately, no analytical distinction was established between "Certain" and "Very certain", indicating that a three-point scale would have been adequate.

Three separate techniques were utilized to operationalize the phases. First, the solution, justification, confidence level, and elapsed time prior to submitting a response were recorded from the written data. Although this method may have oversimplified the underlying phases, it maintained an analytical focus on the core dimensions of comprehension—objective accuracy and the subjective sensation [4]. Furthermore, it enabled a distinction between phases such as misunderstanding and deep understanding by measuring the confidence levels of the participants. Second, observational coding was applied to examine incorrect riddle solutions. Finally, AFFDEX software was implemented to automatically detect one specific phase.

Nascent understanding was operationalized as the provision of an incorrect solution or an invalid justification combined with a lack of confidence (1 = Not certain).

Misunderstanding was operationalized as the provision of an accurate solution and justification paired with a low level of confidence (2 = Somewhat certain). Submissions containing a correct solution and justification but a lack of confidence (1 = Not certain), along with submissions containing an incorrect solution and justification but high confidence (3 = Certain or 4 = Very certain), were excluded from analysis due to an insufficient volume of observations.

Confusion was identified utilizing AFFDEX, as the software automatically identifies this affective state based on its associated AU configurations. The three highest-intensity episodes of confusion per participant were selected for analysis to guarantee the inclusion of unambiguous observations while optimizing data collection efficiency.

Emergent understanding was operationalized as the provision of a correct solution and justification accompanied by a high level of confidence (3 = Certain or 4 = Very certain) with a response time of at least two minutes. This specific timeframe indicated active cognitive processing and persistent problem-solving effort. Emergent understanding was also determined through observational coding. Three independent evaluators assessed whether participants comprehended the solution and rationale to a riddle when it was explained, reviewing all occurrences across 71 participants. From these assessments, a codebook delineating the behavioral markers linked to episodes of unanimous coder agreement was structured. This codebook was subsequently utilized to code the data for the remaining 32 participants. Records that did not secure unanimous agreement among the evaluators were omitted from subsequent analysis. Additionally, the codebook was applied to validate the presence of emergent understanding within the written responses.

Deep understanding was operationalized as the provision of an accurate solution and justification alongside a high level of confidence (**3** = Certain or **4** = Very certain) within approximately 20 seconds of initial exposure to the riddle. This brief duration indicated cognitive ease and domain mastery in resolving the task.

In light of the exploratory framework of this research, only unambiguous instances of each phase were isolated for final analysis. For instance, observations were omitted if participants spoke while answering a riddle, as verbalization can distort facial expressions; if they recorded a riddle solution but continued brainstorming, which indicated ongoing cognitive processing; or if there was any indication that a phase was not being authentically experienced, such as laughter regarding an issue seemingly disconnected from the task, potentially indicating a confounding mental state.

Action units

AUs were quantified utilizing the AFFDEX module integrated within the facial expression analysis suite created by iMotions. AFFDEX implements FACS to code facial movements from pre-recorded media or in real time. FACS is highly compatible with conventional machine learning techniques that depend on feature extraction—a critical factor given the constrained number of phase observations available in this research. Conversely, deep learning models, which process raw images directly, necessitate significantly larger datasets, rendering them impractical for this specific context.

The AFFDEX system quantifies 20 discrete AUs, a neutral state, confusion, and primary emotional states. The software records the total frame count during which each AU, the neutral state, and the emotional metric exceed a fixed likelihood parameter. Structured on a continuum from 0 to 100, a value of '25' indicates a minor facial reaction, '50' a moderate response intensity, and '75' a pronounced reaction [35]. The likelihood parameter was fixed at 50 to maximize statistical reliability in tracking AUs while remaining sensitive enough to capture subtle facial adjustments. AFFDEX has been formally validated regarding its measurement accuracy and dependability in assessing AUs [24].

For written responses, AUs were tracked over a five-second duration starting from the moment participants began writing their ultimate answer to a riddle. AUs were evaluated prior to the initiation of writing only if unambiguous markers of a phase were apparent immediately beforehand. Analytical windows were lengthened if distinct indicators of a phase extended beyond the baseline five seconds. For observational coding, AUs were captured from the appearance of clear signs of emergent understanding until the expressions began to diminish. AUs were captured across the complete duration of every episode of confusion as logged by the AFFDEX system.

Data analysis

Interrater reliability

Interrater reliability was computed utilizing Gwet's AC1 [36] for the episodes of emergent understanding isolated via observational coding. Gwet's AC1 was selected because of its stability in environments characterized by high consensus among evaluators and its resilience against the Kappa paradox [37].

Machine learning

Supervised machine learning methodologies were applied to extract the AU configurations linked to each specific stage of comprehension. This approach prioritized high predictive accuracy, robustness when generalized to unseen data, and the prevention of overfitting. Six computational algorithms were executed: gradient boosting machine (GBM), random forest (RF), decision tree, lasso regression, elastic net regression, and logistic regression. GBM and RF represent ensemble strategies recognized for their capacity to model intricate, non-linear dependencies, thereby offering superior predictive yield. Decision trees provide highly interpretable architectures capable of isolating non-linear structures without requiring exhaustive data preprocessing. Lasso and elastic net regressions employ regularization frameworks for feature selection by applying penalties to less critical variables, which optimizes model interpretability and limits overfitting risks. Logistic

regression serves as a transparent baseline architecture for classification problems, enabling direct comparisons against more sophisticated models. Utilizing this broad array of algorithms ensured an analytical equilibrium among predictive accuracy, model interpretability, and computational overhead, rendering the workflow highly compatible with the anticipated complexity of the dataset.

The computational analysis was executed within RStudio version 2023.12.1-402. The GBM and RF frameworks were implemented using the 'caret' package (v6.0.94), the decision tree architecture was built via 'rpart' (v4.1.23), the lasso and elastic net formulations were calculated using 'glmnet' (v4.1.8), and the logistic regression classification was completed using 'nnet' (v7.3.19). The analytical pipeline is structured through the following operational phases:

1. **Splitting data:** Stratified 80–20 train–test partitioning was executed to accommodate the moderate class imbalance observed across the different phases.
2. **Feature scaling:** Feature normalization was implemented for the logistic regression, lasso, and elastic net models, given that these specific algorithms are highly sensitive to the scale of input variables. AUs were systematically standardized across both the training and evaluation subsets.
3. **Cross-validation setup:** A 10-fold stratified cross-validation methodology was deployed to secure resilient model validation. Cross-validation was deliberately bypassed for the logistic regression model, allowing it to function as a baseline benchmark reflecting conventional inferential statistical modeling.
4. **Selecting evaluation metrics:** Model validation was quantified using precision, recall, F1 score, and AUC. In explicit terms, precision computes the ratio of true positives relative to all positive predictions, reflecting the fidelity of positive classifications. Recall evaluates the ratio of true positives correctly isolated out of the entire pool of actual positives. The F1 score represents the harmonic mean of precision and recall, establishing a balance between classification accuracy and completeness. The AUC (Area Under the Curve) evaluates overall model discrimination across all operational classification thresholds, indicating its capacity to separate distinct classes. Throughout the cross-validation phase, performance metrics for each individual stage were averaged over the 10 folds to produce a consolidated performance evaluation. Weighted averages were computed for every metric based on the volume of observations available per phase. All classification metrics were derived utilizing the One-vs-Rest (OvR) methodology.
5. **Training baseline models:** Untuned baseline models for GBM, RF, decision tree, lasso, and elastic net were initially generated using default package configurations. Algorithms that displayed operational promise were selected for subsequent hyperparameter optimization.
6. **Hyperparameter tuning:** Model configurations were optimized through systematic hyperparameter refinement. This tuning targeted the hyperparameters and grid search fields most critical to boosting classification performance, while simultaneously accounting for processing efficiency and available computational resources.
7. **Test-set evaluation:** Lastly, the optimized model variants and the baseline logistic regression model were formally evaluated on the reserved testing subsets.

Results and Discussion

Inter-rater reliability

Three independent evaluators reviewed 241 recorded episodes of emergent understanding. Among these, unanimous consensus was achieved for 225 instances, yielding a 93% rate of agreement. Gwet's AC1 was calculated at 0.95, demonstrating an excellent level of interrater consensus.

Descriptive statistics

An aggregate pool of 1245 observations was compiled across the various phases of comprehension: 249 instances were labeled as nascent understanding, 171 as misunderstanding, 200 as confusion, 287 as emergent understanding, and 338 as deep understanding (Table 1).

Table 1. Number of observations for each phase of understanding, with the total number of instances across all phases.

Phase of Understanding	Observations
Nascent Understanding	249
Misunderstanding	171
Confusion	200
Emergent Understanding	287
Deep Understanding	338
Total Observations	1245

The heatmap presented in Figure 2 displays the mean distribution of active frames for the 20 monitored AUs and the neutral baseline state across each individual phase of comprehension. The color spectrum reflects the mean activation level, where warmer color profiles (red) denote elevated mean values and cooler color profiles (blue) signify reduced mean values. The observed AU configurations across the phases remain distinct, yet they display regions of structural overlap. Individual AUs, such as Brow Furrow during episodes of confusion and Smile during moments of emergent understanding, underscore the distinct expressive markers for each phase, while the neutral facial state exhibits high prevalence across all evaluated phases. In general, confusion and emergent understanding are characterized by elevated AU muscle activity, whereas nascent understanding, misunderstanding, and deep understanding exhibit minimized AU activation.

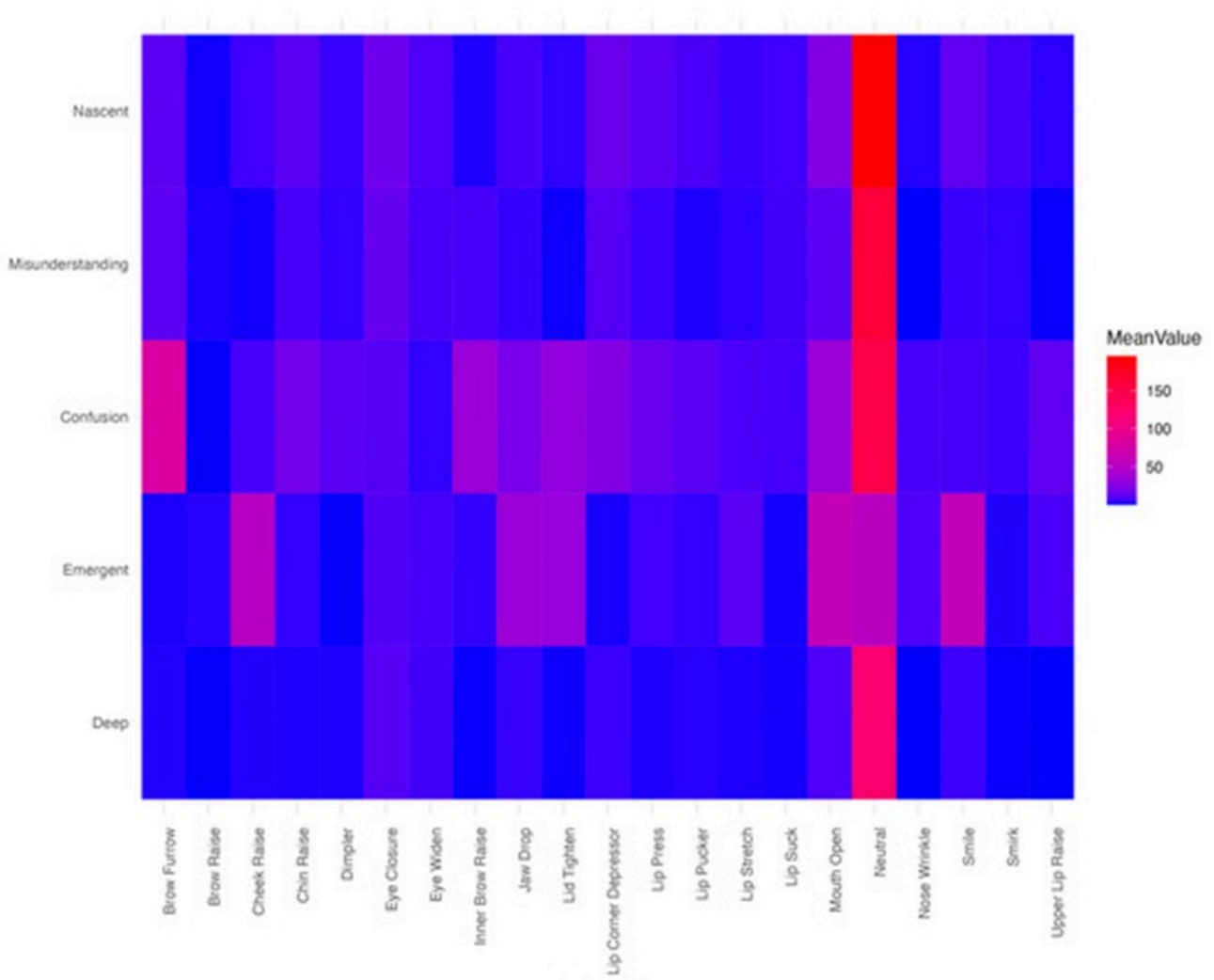


Figure 2. Average distribution of AU frames for each phase of understanding.

Photographic records of the facial configurations of participants for a representative instance of each distinct phase are provided (Figure 3), accompanied by illustrative records of participant solutions, think-aloud transcripts, justifications, and self-reported certainty parameters gathered during the riddle tasks (Table 2).

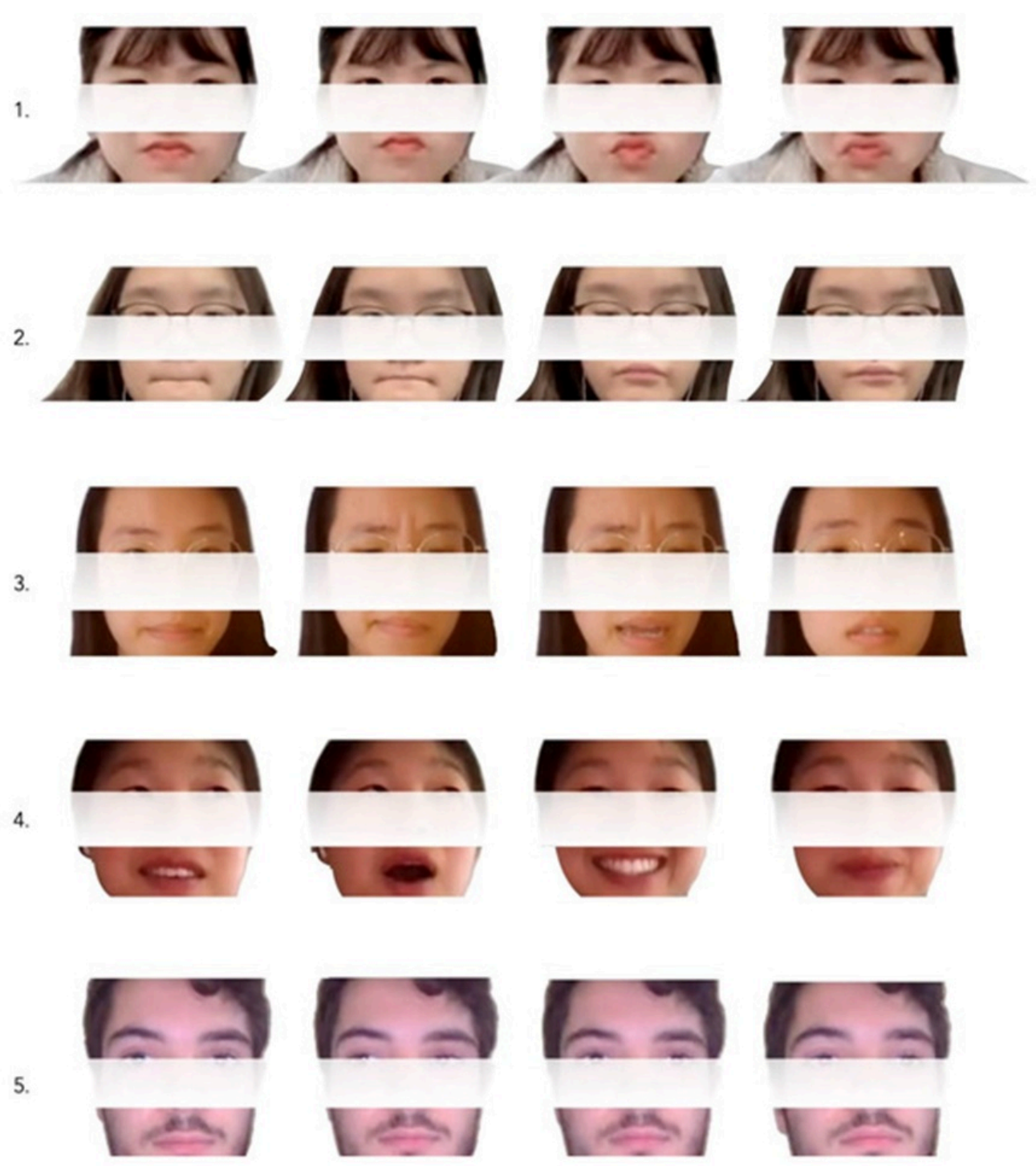


Figure 3. Facial expressions associated with (1) nascent understanding, (2) misunderstanding, (3) confusion, (4) emergent understanding, and (5) deep understanding.

Table 2. Participants' answers, think aloud, justifications, and levels of certainty in response to riddles across phases of understanding.

Phase	Riddle	Answer	Think Aloud	Justification	Level of Certainty

Nascent	Which month has 28 days?	July	Don't know	Not certain	1 (Not certain)
Misunderstanding	I have two coins that total thirty cents and one of them is not a nickel. What are the two coins?	Quarter and nickel	Tried combinations ($X + Y = 30$ cents); considered that "not a nickel" might refer only to one coin, explored possibilities like two dimes (20 cents), then reconsidered interpretation of the condition.	Interpreted "one is not a nickel" as possibly meaning only one coin is excluded, leading to the idea of a quarter paired with a nickel.	2 (Somewhat certain)
Confusion	You are in a dark room with a candle, a wood stove, and a gas lamp. You only have one match—what do you light first?	—	Unsure of what action is required or why anything must be lit first.	—	—
Emergent	I have two coins that total thirty cents and one is not a nickel. What are the two coins?	Quarter, nickel	Evaluated combinations (25 + 5, 25 + 10, etc.), initially confused but then recognized the logical trick in the statement.	Concluded 25 + 5 works; interpreted condition correctly as not restricting both coins.	4 (Very certain)
Deep	You are a cyclist in a cross-country race. Just before the finish line, you overtake the person in second place. In what place do you finish?	Second	Recognized race position logic immediately.	Passing the second-place cyclist means you take their position.	4 (Very certain)

Machine learning analysis

The training datasets comprised 998 individual observations, representing 80% of the aggregate data pool: 200 instances for nascent understanding, 137 for misunderstanding, 160 for confusion, 230 for emergent understanding, and 271 for deep understanding. The testing datasets consisted of 247 observations, representing 20% of the aggregate data pool: 49 instances for nascent understanding, 34 for misunderstanding, 40 for confusion, 57 for emergent understanding, and 67 for deep understanding (Table 3). Combined classification metrics for the unoptimized baseline models are documented in Table 4.

Table 3. Stratified 80–20 test–train splits for each phase of understanding.

Stage of Comprehension	Training Set Samples	Test Set Samples
Initial (Nascent) Understanding	200	49

Incorrect Understanding (Misunderstanding)	137	34
Confused State	160	40
Developing (Emergent) Understanding	230	57
Advanced (Deep) Understanding	271	67
Overall Total Observations	998	247

Table 4. Performance metrics—precision, recall, F1 score, and AUC—for five baseline models: GBM, RF, lasso, elastic net, and decision tree.

Metric	Nascent (GBM/RF/Lasso/Net/Tree)	Misunderstanding (GBM/RF/Lasso/Net/Tree)	Confusion (GBM/RF/Lasso/Net/Tree)	Emergent (GBM/RF/Lasso/Net/Tree)
Precision	0.84/0.85/0.83/0.83/0.35	0.87/0.88/0.86/0.86/0.19	0.93/0.94/0.9/0.9/0.6	0.97/0.97/0.95/0.95/0.6
Recall	0.93/0.88/0.93/0.93/0.28	0.98/0.94/0.99/0.99/0.06	0.95/0.94/0.97/0.97/0.44	0.95/0.94/0.92/0.92/0.6
F1 Score	0.88/0.86/0.88/0.88/0.3	0.92/0.9/0.92/0.92/0.2	0.94/0.94/0.93/0.93/0.5	0.96/0.95/0.94/0.94/0.6
AUC	0.75/0.72/0.73/0.73/0.66	0.7/0.66/0.68/0.68/0.65	0.92/0.91/0.83/0.83/0.74	0.97/0.97/0.94/0.94/0.6

The computational algorithms selected for hyperparameter tuning were GBM, RF, and lasso. Elastic net was bypassed for optimization because the ideal alpha value defaulted to 1 during training iterations where alpha ranges spanned from 0.1 to 1; this selection indicated an algorithmic preference for compressing coefficients fully to zero rather than to near-zero values. Furthermore, its functional performance was virtually identical to that of the lasso model.

The refinement of the GBM model was executed across two distinct phases. In the primary phase, an extensive grid search was performed over the following hyperparameter intervals: total tree count from 50 to 500 (using step steps of 50), interaction depth from 1 to 10 (using step steps of 1), learning rate coefficients of 0.01, 0.05, 0.1, and 0.2, and the minimum node size constraint for terminal nodes from 5 to 20 (using step steps of 5). In the secondary phase, a Bayesian optimization grid was applied over identical parameters across the following expanded ranges: tree count from 50 to 600, interaction depth from 1 to 12, learning rate values of 0.01, 0.05, 0.1, 0.2, and 0.25, and terminal node sizes from 5 to 20. The tuning framework for the RF model involved an exhaustive grid search regarding the number of features evaluated at each split configuration, spanning an operational range from 4 to 18 (using step steps of 1). Lasso optimization implemented a grid search for the regularization parameter lambda, testing 100 log-spaced values distributed between -3 and 3 .

For the finalized GBM configuration, the tree parameter ('n.trees') was fixed at 250, the maximum splitting depth ('interaction.depth') was set to 8, the learning rate factor ('shrinkage') was established at 0.01, and the minimum terminal node size requirement ('n.minobsinnode') was limited to 10. The optimized RF architecture utilized 5 random features at each split boundary ('mtry'). For the lasso regression, the regularization penalty coefficient ('lambda') was stabilized at 0.006.

Combined performance outputs for these optimized classification models are structured in [Table 5](#).

Table 5. Performance metrics—precision, recall, F1 score, and AUC—for three optimized models: GBM, RF, and lasso.

Metric	Nascent (GBM/RF/Lasso)	Misunderstanding (GBM/RF/Lasso)	Confusion (GBM/RF/Lasso)	Emergent (GBM/RF/Lasso)	Deep (GBM/RF/Lasso)
Precision	0.84/0.84/0.83	0.87/0.87/0.86	0.93/0.95/0.90	0.97/0.97/0.95	0.92/0.89/0.93
Recall	0.93/0.93/0.93	0.98/0.98/0.99	0.94/0.93/0.97	0.95/0.94/0.92	0.69/0.71/0.59
F1 Score	0.89/0.89/0.88	0.92/0.92/0.92	0.94/0.94/0.93	0.96/0.95/0.94	0.79/0.79/0.72
AUC	0.77/0.72/0.73	0.68/0.63/0.68	0.93/0.92/0.83	0.97/0.96/0.94	0.79/0.78/0.77

To evaluate performance on the test data partition, weighted averages are documented for the optimized GBM, RF, and lasso frameworks, alongside the baseline logistic regression model, ensuring a concise yet transparent disclosure of experimental findings. The GBM architecture attained a precision score of 0.91, a recall score of 0.87, an F1 index of 0.88, and an AUC value of 0.84. The RF configuration yielded a precision score of 0.91, a recall score of 0.88, an F1 index of 0.89, and an AUC value of 0.82. The lasso formulation established a precision score of 0.90, a recall score of 0.85, an F1 index of 0.86, and an AUC value of 0.80. The baseline logistic regression model recorded a precision score of 0.90, a recall score of 0.86, an F1 index of 0.87, and an AUC value of 0.79 (Table 6, Figure 4).

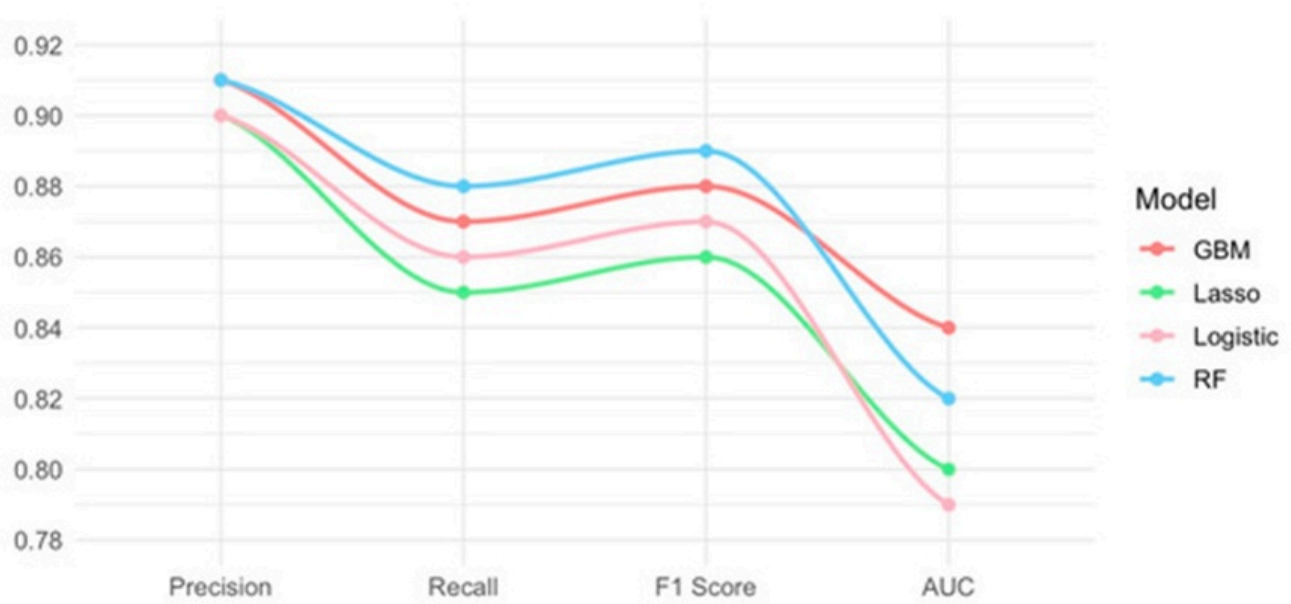


Figure 4. Line graph of test set weighted averages for precision, recall, F1 score, and AUC for optimized models and logistic regression.

Table 6. Test set weighted averages for precision, recall, F1 score, and AUC for optimized models and logistic regression.

Metric	Logistic	Lasso	RF	GBM
Precision	0.90	0.90	0.91	0.91
Recall	0.86	0.85	0.88	0.87
F1 Score	0.87	0.86	0.89	0.88
AUC	0.79	0.80	0.82	0.84

Figure 5 maps the confusion matrix illustrating the performance of the optimized GBM model when evaluated against the testing subset, selected here due to its status as the top-performing model and to optimize reporting clarity. This classification matrix exhibits robust predictive precision for instances of emergent understanding and confusion. However, comprehension phases with reduced representation in the testing partition, such as misunderstanding, display elevated error rates, which is likely a consequence of the constrained sample volume. The matrix further highlights classification difficulties when attempting to isolate deep understanding from nascent understanding, pointing to a prospective overlap among the facial features tied to these two distinct stages.

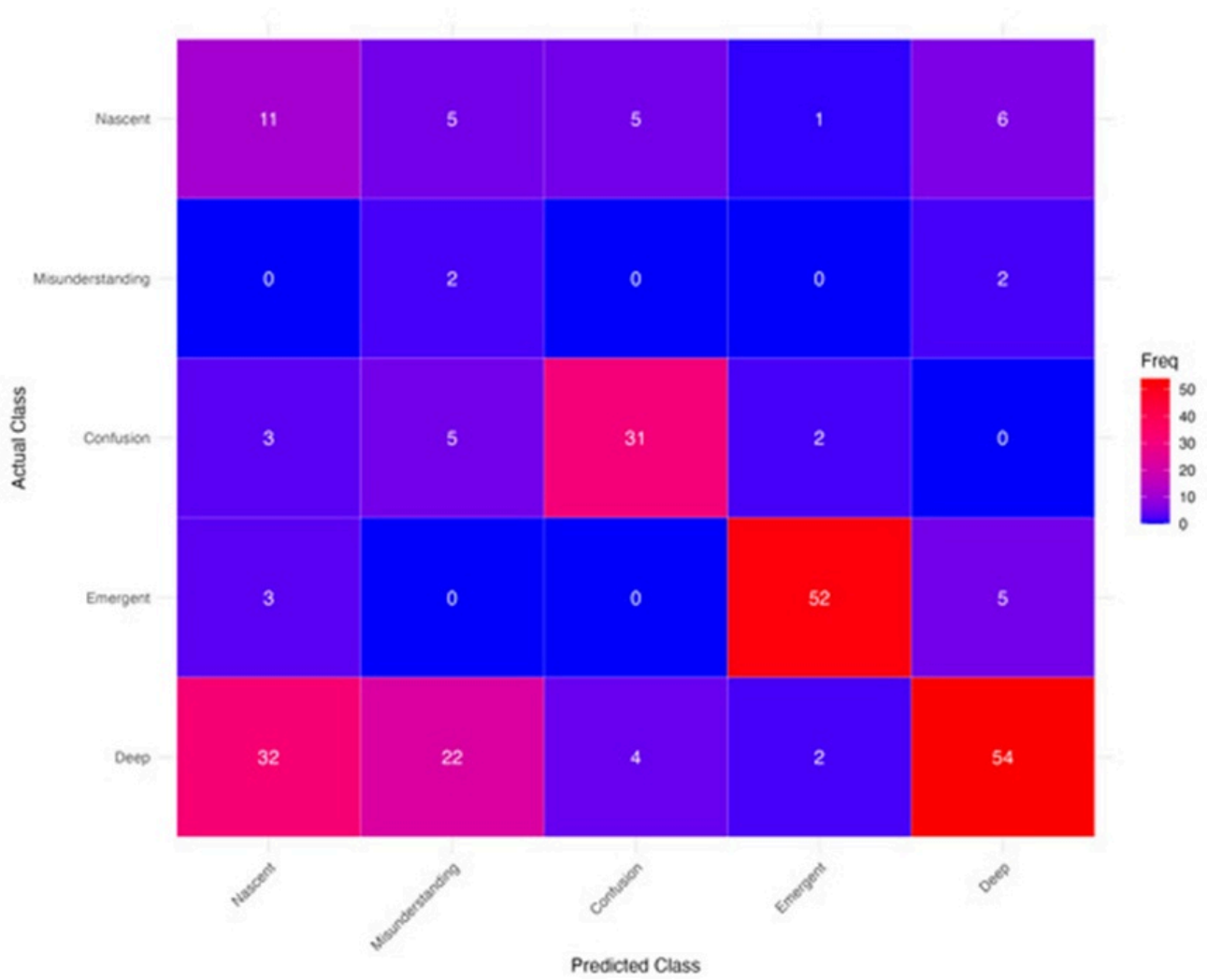


Figure 5. Confusion matrix for the optimized GBM model's performance on the test set.

A notable divergence exists between the weighted classification metrics computed for the evaluation dataset and the error distribution patterns revealed within the matrix. This discrepancy arises because weighted parameters overrepresent the performance of highly prevalent classes, thereby masking the classification difficulties the model faces with less frequent phases.

These outcomes suggest that subsequent studies could enhance accuracy by mitigating class imbalance through the integration of a more balanced dataset, alongside the refinement of feature selection and characterization techniques.

On the whole, the machine learning architectures exhibited robust predictive capabilities. Nevertheless, opportunities for improvement remain in terms of dependably identifying and separating specific stages of comprehension, particularly nascent understanding, misunderstanding, and deep understanding. While the optimized architectures yielded minor enhancements compared to their baseline counterparts, their overall performance metrics remained closely aligned. The predictive success of the linear frameworks implies that the classification boundaries between separate stages possess a degree of linear separability, occupy a continuous spectrum, and are primarily differentiated by the intensity of overlapping AU configurations. Concurrently, the efficacy of the non-linear frameworks demonstrates that these phases exhibit complex AU interactions that can be successfully exploited to a certain degree.

Although not explicitly detailed in the outputs, it is crucial to mention that experiments involving artificial training set expansion via SMOTE and random oversampling yielded heightened performance for both the GBM and RF architectures. On the contrary, the predictive accuracy of the lasso model fell drastically following SMOTE implementation and showed no change after random oversampling. This variation may stem from the fidelity of the synthetically generated data points; nonetheless, it implies that as AU configurations across distinct phases grow more complex, the predictive efficacy of linear models decreases. Notably, modifying the class weighting parameters produced no discernable impact on the holistic performance of the GBM, RF, and lasso architectures, indicating that these models were already utilizing the phase-specific patterns at their maximum potential capacity.

Feature importance metrics for each AU, scaled from 0 to 100, are provided for the optimized GBM framework (Table 7, Figure 6). Because these values closely mirror the importance distributions of the optimized RF architecture, the RF scores have been omitted. Lasso coefficients are also detailed for every monitored AU across each specific phase (Table 7, Figure 7).

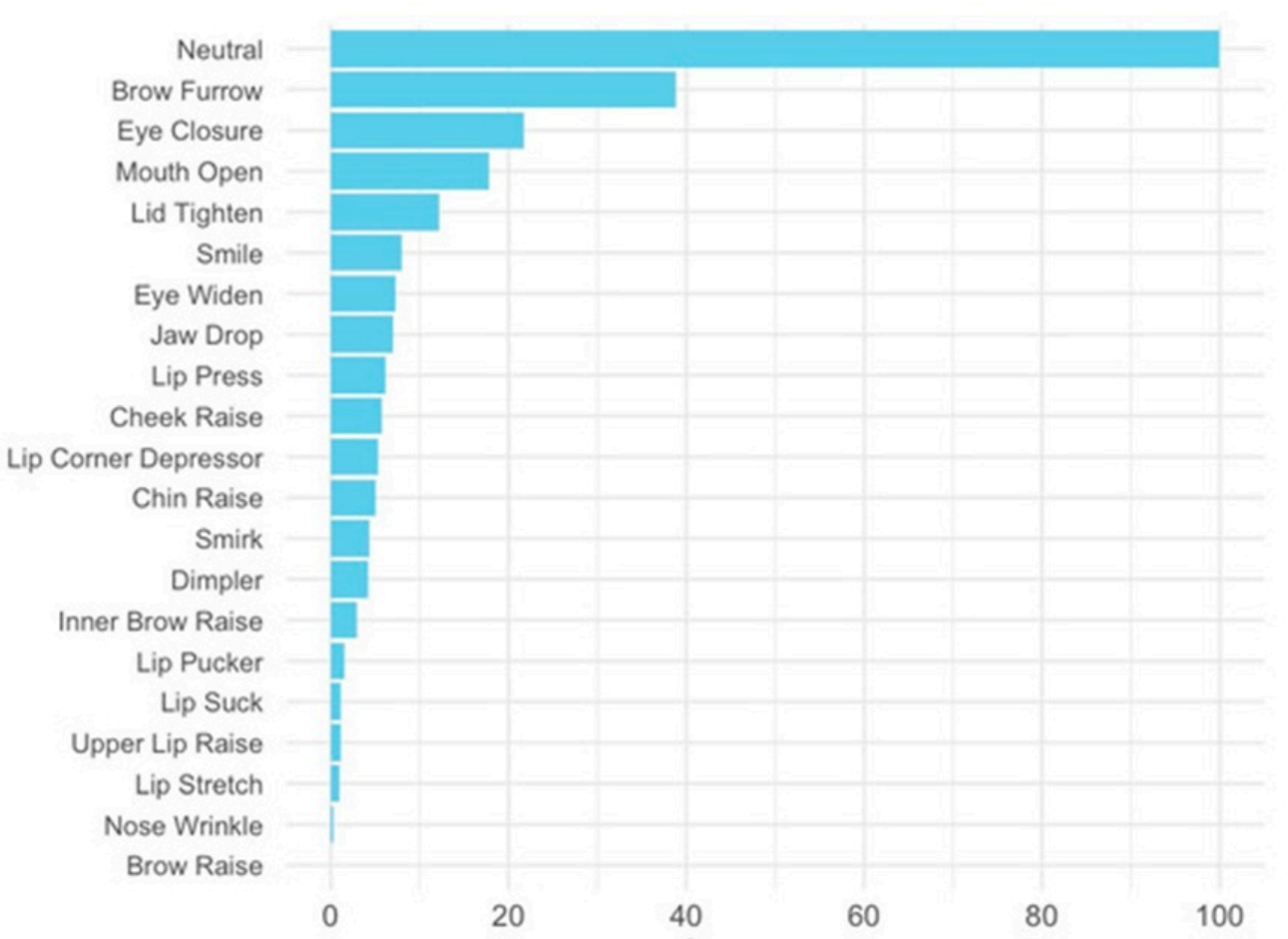


Figure 6. Bar graph of feature importance scores for AUs.

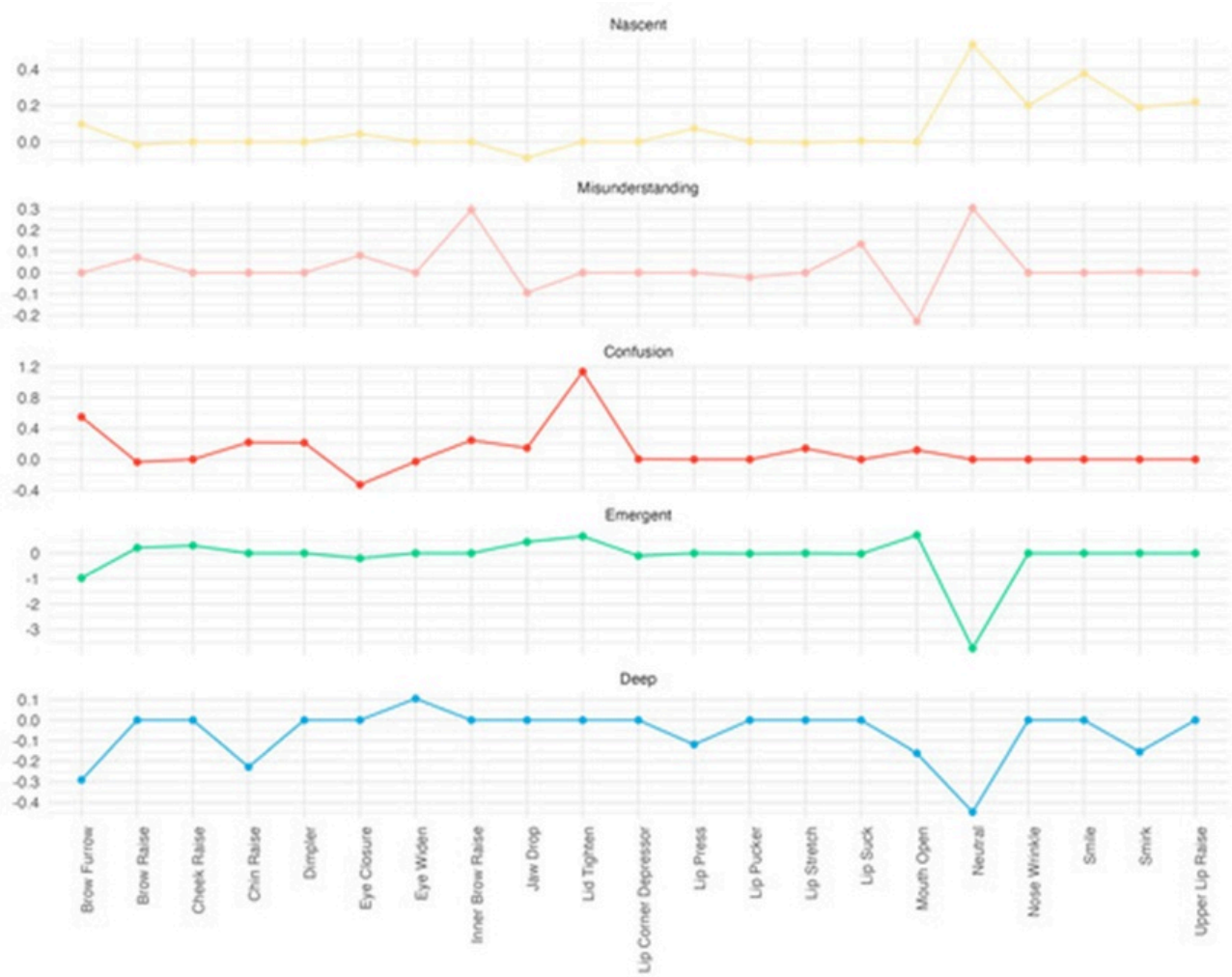


Figure 7. Line graphs of lasso coefficients for AUs per phase.

Table 7. Feature importance scores for AUs, and lasso coefficients for AUs per phase.

Action Unit	Deep	Emergent	Confusion	Misunderstanding	Nascent	Importance
Brow Furrow	−0.290	−0.976	0.550	0.000	0.098	38.80
Brow Raise	0.000	0.213	−0.035	0.072	−0.017	0.00
Cheek Raise	0.000	0.300	0.000	0.000	0.000	5.79
Chin Raise	−0.227	0.000	0.221	0.000	0.000	5.01
Dimpler	0.000	0.000	0.216	0.000	−0.001	4.25
Eye Closure	0.000	−0.201	−0.328	0.081	0.044	21.75
Eye Widen	0.105	0.000	−0.028	0.000	0.000	7.23
Inner Brow Raise	0.000	0.000	0.248	0.295	0.000	3.00

Jaw Drop	0.000	0.444	0.148	−0.094	−0.088	7.02
Lid Tighten	0.000	0.667	1.137	0.000	0.000	12.16
Lip Corner Depressor	0.000	−0.103	0.004	0.000	0.000	5.34
Lip Press	−0.119	0.000	0.000	0.000	0.074	6.10
Lip Pucker	0.000	−0.016	0.000	−0.022	0.003	1.56
Lip Stretch	0.000	0.000	0.142	0.000	−0.005	0.98
Lip Suck	0.000	−0.022	0.000	0.134	0.000	1.14
Mouth Open	−0.160	0.710	0.121	−0.228	0.000	17.82
Neutral	−0.447	−3.735	0.000	0.303	0.537	100.00
Nose Wrinkle	0.000	0.000	0.000	0.000	0.202	0.24
Smile	0.000	0.000	0.000	0.000	0.377	7.97
Smirk	−0.154	0.000	0.000	0.004	0.190	4.40
Upper Lip Raise	0.000	0.000	0.000	0.000	0.220	1.05

The feature importance rankings underscore the commanding role of the neutral facial state, which achieved a maximum score of 100 in separating the phases. This variable is likely the most critical indicator because the magnitude of physiological feedback fluctuates across different stages, and it serves as an aggregate index capturing the presence or absence of overall AU activity within any specific frame. The manual elimination of AUs deemed marginal based on low importance scores, near-zero values, or zero variance generated no impact on the holistic precision of the GBM and RF architectures.

The lasso coefficients illustrate the specific features isolated by the model following the regularization pipeline, thereby highlighting the most critical AUs for predicting each respective phase. The confusion phase exhibits substantial positive coefficients across multiple AUs, most notably Lid Tighten and Brow Furrow. For deep understanding, AU coefficients were compressed to zero for markers such as Jaw Drop, while demonstrating negative coefficients for markers like Brow Furrow. The emergent understanding phase displayed relatively uniform coefficients near zero, with a pronounced drop for the Neutral variable. The misunderstanding phase manifested low to moderate positive coefficients, including Lip Suck and Inner Brow Raise. Similarly, nascent understanding maintained relatively uniform coefficients near zero, with the exception of positive coefficients for markers such as Smile. Broadly, the lasso coefficients confirmed the presence of distinct AU configurations across the individual phases.

The experimental findings point to a credible link between unique physiological sensations, quantified via AUs, and specific stages of comprehension. This realization introduces three central themes for academic discussion.

First, a dependable methodology for assessing the sensation of comprehension is introduced, challenging historical pedagogical conventions. Educational and developmental theorists such as Bruner [6], Dewey [38], and Vygotsky [11, 39] have consistently highlighted the active engagement of the student within the educational process. Notwithstanding this long-standing advocacy, modern instructional environments regularly prioritize standardized curricula and objective assessment benchmarks while dedicating minimal attention to internal feelings, emotional states, and individual lived experiences—a practice reinforced by policymakers and academic researchers due to the historical complexities of quantifying subjective states [4]. Rooted in an embodied cognition paradigm, the data and methodologies of this current

study challenge that historical stance by providing an explicit and quantifiable baseline for evaluation via physiological indicators. Furthermore, embodied cognition serves as the foundation for our most compelling theories regarding consciousness, cognitive function, and emotional affect, underscoring the central position of the individual within these systems [22, 23]. Accordingly, while objective benchmarks of student performance remain vital for preserving institutional quality and educational accountability, they must be counterbalanced by an equivalent focus on the subjective internal experiences of learners, thereby addressing their distinct developmental needs and personal attributes.

Second, the structural integration of real-time cognitive workflows and emotional experiences linked to subjective comprehension is directly and reliably advanced. By merging the presented methodology—which records immediate transitions in mental states via physiological activity—with an analytical framework that accounts for the valence, rationality, and social consequences of cognitive states [40], alongside predictive processing models [41], the core intersection between the cognitive tracks, sensations, and emotions of subjective comprehension can be systematically evaluated. Specifically, the physiological signatures of varied emotional experiences linked to comprehension can be exploited to develop a more profound understanding of how cognitive processes, such as logical reasoning, manifest during the learning process.

Finally, this scholarship offers meaningful advancements for the architecture of an intelligent multimodal educational system that utilizes large language models to revolutionize tailored learning paradigms [42]. An automated tutor engineered to register sensations and identify exactly when they impede learning could guide students toward deep understanding in real time by identifying conceptual errors before they become deeply rooted, minimizing student frustration, driving active engagement, and equipping learners to independently track, regulate, and direct their own educational trajectories. Given the sensitive ethical boundaries involved, the engineering of such automated tutoring systems necessitates rigorous scrutiny to guarantee full compliance with user privacy rights, personal autonomy, and informed consent. The primary objective of these tools must center on fostering comprehensive personal development across all demographics, rather than executing data harvesting, reinforcing surveillance mechanisms, or prioritizing the commercial objectives of corporate entities.

The generalizability of these outcomes is constrained by the fact that the machine learning architectures were trained exclusively on optimized, clean datasets. Consequently, these models are anticipated to demonstrate diminished predictive performance when applied to noisy, real-world occurrences of each respective stage. Furthermore, the volume of observations within the evaluation subsets and test partitions was constrained—most notably for underrepresented stages such as misunderstanding—which additionally compromises the stability and generalizability of the classifiers. A further limitation stems from the demographic composition of the cohort, which consisted predominantly of female undergraduate students enrolled at a single academic institution. While the underlying stages are expected to align with uniform AU configurations irrespective of demographic variations, the homogeneity of this sample population may restrict the wider applicability of the findings.

The deliberate omission of data segments during which individuals engaged in verbal communication while solving the tasks may have inadvertently discarded significant behavioral indicators. Moreover, partitioning the process of comprehension into separate, sequential stages oversimplifies its inherently intricate nature. Although this methodology enabled the isolation of a distinct intersection between affective sensations and cognitive workflows, the classification system utilized for these stages of comprehension will likely require subsequent adjustment and optimization as academic scholarship advances. Finally, despite rigorous measures implemented to preserve procedural uniformity, the spoken input delivered after each task may have altered the subsequent reactions or facial configurations of the participants, thereby introducing a potential confounding factor into the analytical framework.

Subsequent investigations should prioritize the training of predictive models on more extensive, uncurated datasets to elevate performance thresholds across diverse environments, particularly within real-time frameworks. This direction corresponds with the intensifying academic focus on deploying ecologically valid data matrices that more accurately reflect the intricacies of authentic environments and human behaviors. This transition is especially apparent in domains

like affective computing, where repositories such as IEMOCAP and RAMAS are increasingly utilized to train analytical frameworks for more pragmatic deployment scenarios [43]. As more expansive datasets emerge, deep learning architectures—including Convolutional Neural Networks (CNNs) and Transformer-based configurations—should be engineered, given that they routinely surpass shallow models like RF and GBM and can process raw imagery directly [44]. This capability removes the requirement for feature extraction pipelines, such as tracking AUs via intermediate interfaces like AFFDEX. Furthermore, integrating multidimensional physiological streams through vocal signal analysis [45] and transdermal optical imaging (TOI, [25]) could boost classification accuracy while providing a more holistic characterization of these stages [46, 47]. Because this investigation operationalized misunderstanding as the subjective sensation of being stuck despite satisfying an external benchmark of objective accuracy, subsequent scholarship must also evaluate misunderstanding when characterized as an inflated degree of confidence paired with an incorrect solution. Lastly, future research must examine the explicit interplay between physiological indicators and cognitive workflows, such as logical deduction, across each separate stage, while replicating the findings of this study within an authentic academic environment.

Acknowledgements

None

Conflict of interest

None

Financial support

This research was funded by the Social Science and Humanities Research Council (SSHRC), Operating Grant #500427.

Ethics statement

This study was approved by the Social Sciences, Humanities & Education Research Ethics Board of the University of Toronto (protocol# 42280, date of approval: 04/03/2022) for studies involving humans.

Received: 23 Jan 2022

Revised: 18 Mar 2022

Accepted: 13 May 2022

Published online: 15 September 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Roumell EA, Walker J, Salajan FD. Lifelong learning and education policy in North America. In: Third International Handbook of Lifelong Learning. Cham: Springer; 2023. p.633-54.

https://doi.org/10.1007/978-3-031-19592-1_31.

Duke NK, Cartwright KB. The science of reading progresses: communicating advances beyond the simple view of reading. *Read Res Q.* 2021;56(S1):S25-44.

<https://doi.org/10.1002/rrq.411>.

Pearson PD, Palincsar AS, Biancarosa G, Berman AI, editors. Reaping the rewards of the Reading for Understanding Initiative. Washington (DC): National Academy of Education; 2020.

Olson DR. Making sense: what it means to understand. Cambridge: Cambridge University Press; 2022.

Olson DR. Ascribing understanding to ourselves and others. *Am Psychol.* 2024;79(7):920-7.

<https://doi.org/10.1037/amp0001244>.

Bruner JS. The process of education. Cambridge (MA): Harvard University Press; 1960.

Damasio AR. Descartes' error: emotion, reason, and the human brain. New York: Grosset/Putnam; 1994.

Gardner HE. Frames of mind: the theory of multiple intelligences. New York: Basic Books; 2011.

Goleman D. Emotional intelligence: why it can matter more than IQ. London: Bloomsbury Publishing; 2020.

Piaget J. The construction of reality in the child. London: Routledge; 2013.

Vygotsky LS. The collected works of L. S. Vygotsky: the fundamentals of defectology. Vol 2. New York: Springer; 1987.

Russell JA. A circumplex model of affect. *J Pers Soc Psychol.* 1980;39(6):1161-78.

<https://doi.org/10.1037/h0077714>.

James W. The principles of psychology. New York: Henry Holt; 1890.

Damasio A. Feeling and knowing: making minds conscious. New York: Pantheon Books; 2021.

Ekman P, Rosenberg EL, editors. What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). New York: Oxford University Press; 1997.

Barrett LF. How emotions are made: the secret life of the brain. Boston (MA): Houghton Mifflin Harcourt; 2017.

Lazarus RS. Emotion and adaptation. New York: Oxford University Press; 1991.

Frijda NH. The emotions. Cambridge: Cambridge University Press; 1986.

Frijda NH. The laws of emotion. Mahwah (NJ): Lawrence Erlbaum Associates; 2007.

Woodruff E. AI detection of human understanding in a Gen-AI tutor. *AI.* 2024;5(2):898-921.

<https://doi.org/10.3390/ai5020045>.

Damasio A. Self comes to mind: constructing the conscious brain. New York: Pantheon/Random House; 2010.

Kauffman SA. The origins of order: self-organization and selection in evolution. New York: Oxford University Press; 1993.

Kauffman SA. At home in the universe: the search for laws of self-organization and complexity. New York: Oxford University Press; 1995.

Stöckli S, Schulte-Mecklenbeck M, Borer S, Samson AC. Facial expression analysis with AFFDEX and FACET: a validation study. *Behav Res Methods*. 2018;50(4):1446-60.
<https://doi.org/10.3758/s13428-017-0996-1>.

Fu G, Zhou X, Wu SJ, Nikoo H, Panesar D, Zheng PP, et al. Discrete emotions discovered by contactless measurement of facial blood flows. *Cogn Emot*. 2022;36(7):1429-39.
<https://doi.org/10.1080/02699931.2022.2124960>.

Biggs JB, Collis KF. Evaluating the quality of learning: the SOLO taxonomy (structure of the observed learning outcome). New York: Academic Press; 1982.
<https://doi.org/10.1016/C2013-0-10375-3>.

Bartlett FC. Remembering: a study in experimental and social psychology. Cambridge: Cambridge University Press; 1932.

Johnson-Laird PN. Mental models: towards a cognitive science of language, inference, and consciousness. Cambridge (MA): Harvard University Press; 1983.

Piaget J. The origins of intelligence in children. Madison (CT): International Universities Press; 1952.

Vosniadou S, Brewer WF. Mental models of the earth: a study of conceptual change in childhood. *Cogn Psychol*. 1992;24(4):535-85.

Weinstein CE, Mayer RE. The teaching of learning strategies. In: Wittrock MC, editor. *Handbook of research on teaching*. 3rd ed. New York: Macmillan; 1986. p.315-27.

Barnett SM, Ceci SJ. When and where do we apply what we learn? A taxonomy for far transfer. *Psychol Bull*. 2002;128(4):612-37.
<https://doi.org/10.1037/0033-2909.128.4.612>.

Bar-Hillel M, Noah T, Frederick S. Learning psychology from riddles: the case of stumpers. *Judgm Decis Mak*. 2018;13(1):112-22.
<https://doi.org/10.1017/S193029750000886X>.

Toplak ME, West RF, Stanovich KE. Assessing miserly information processing: an expansion of the Cognitive Reflection Test. *Think Reason*. 2014;20(2):147-68.

Otamendi FJ, Sutil Martín DL. The emotional effectiveness of advertisement. *Front Psychol*. 2020;11:2088.
<https://doi.org/10.3389/fpsyg.2020.02088>.

Gwet KL. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. 4th ed. Gaithersburg (MD): Advanced Analytics, LLC; 2014.

Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:61.
<https://doi.org/10.1186/1471-2288-13-61>.

Dewey J. Experience and education. *Educ Forum*. 1986;50(3):241-52.
<https://doi.org/10.1080/00131728609335764>.

Vygotsky LS, Cole M. *Mind in society: development of higher psychological processes*. Cambridge (MA): Harvard University Press; 1978.

Tamir DI, Thornton MA. Modeling the predictive social mind. *Trends Cogn Sci*. 2018;22(3):201-12.
<https://doi.org/10.1016/j.tics.2017.12.005>.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci*. 2013;36(3):181-204.

<https://doi.org/10.1017/S0140525X12000477>.

Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274.

<https://doi.org/10.1016/j.lindif.2023.102274>.

Ryumina E, Dresvyanskiy D, Karpov A. In search of a robust facial expressions recognition model: a large-scale visual cross-corpus study. *Neurocomputing*. 2022;514:435-50.

<https://doi.org/10.1016/j.neucom.2022.10.013>.

Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Mark*. 2021;31:685-95.

<https://doi.org/10.1007/s12525-021-00475-2>.

Lu C, Zong Y, Zheng W, Li Y, Tang C, Schuller BW. Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Trans Audio Speech Lang Process*. 2022;30:2217-30.

Fairclough SH, Venables L. Prediction of subjective states from psychophysiology: a multivariate approach. *Biol Psychol*. 2006;71(1):100-10.

<https://doi.org/10.1016/j.biopsycho.2005.03.007>.

Zhao G, Zhang H, Liu Z, Zhang S, Zhang Y. Multiscale convolutional neural networks for affect recognition using EEG and peripheral physiological signals. *arXiv [Preprint]*. 2019. arXiv:1911.12918.