Machine Learning-Based Classification of Achievement Emotions: Integrating Control-Value Theory, Deep Learning, and Multi-Modal Physiological Fusion in Educational Environments

Himani Thakor
Ph.D. Scholar,
Department of Computer Science
Veer Narmad South Gujarat University, Surat, Gujarat, India
E-Mail: ht3991214@gmail.com

Abstract:

Theory (CVT) posits that achievement emotions arise from cognitive appraisals of control and value, directly modulating students' motivation, learning strategies, and academic performance. Traditionally, measuring these states relied on retrospective self-reports, which are susceptible to memory biases and response sets. This paper provides a comprehensive, unified synthesis of machine learning (ML) methodologies designed to classify and predict academic emotional states and student depression. We examine three main classification paradigms: short-text sentiment classification using Bidirectional Long Short-Term Memory (Bi-LSTM) networks with self-attention and external knowledge graphs; computer vision-based facial expression analysis using deep transfer learning with VGG-19; and multi physiological signal fusion (ECG, EMG, EDA, and respiratory signals) via Squeeze-and-Excitation Convolutional Neural Networks (SE-CNN). Our comparative benchmarking reveals that while text classifiers achieve 79.23% accuracy on classroom discussions and physiological fusion reaches 87.90% on the WESAD dataset, campus depression can be identified in Reddit posts with up to 98.98% accuracy using AdaBoost. Finally, we address critical ethical issues, algorithmic biases, and the fundamental "measurement problem" of treating dynamic human emotions as objective, quantifiable ground truths in automated decision systems.

1. Introduction and Theoretical Foundations

Emotions are not merely static traits or peripheral phenomena in the educational experience; they are dynamic, multidimensional variables that critically influence students' attention, cognitive resources, intrinsic and extrinsic motivation, information processing styles, and mental health. Following the "affective turn" in educational psychology, cognitive science, and human-computer interaction (HCI), researchers have increasingly focused on the quantitative modeling of academic emotions. At the core of this research is Pekrun's Control-Value Theory (CVT) of achievement emotions (Pekrun, 2024). CVT provides a comprehensive, integrative framework for the antecedents, outcomes, and regulation of emotions in academic contexts. According to CVT, achievement emotions are tied directly to achievement activities (e.g., studying, attending class) and achievement outcomes (e.g., success or failure on exams). These emotions are determined primarily by two proximal cognitive appraisals:

- **Control Appraisals:** The subjective perception of one's causal influence over achievement activities and outcomes. This includes self-efficacy expectations, action-control expectancies, and action-outcome expectancies.

- **Value Appraisals:** The perceived subjective importance of these activities and outcomes, which can be intrinsically or extrinsically valued, and positively valued (for success) or negatively valued (for failure).

$$E = [V, A, D]^T \quad (1)$$

To formalize these relations, the modern generalized CVT conceptualizes an academic emotional state as a three-dimensional vector containing valence (V), activation (A), and duration (D). In Equation (1), $V \in [-1, 1]$ represents emotional valence (positive vs. negative), $A \in [-1, 1]$ represents physiological activation or arousal (activating vs. deactivating), and $D > 0$ denotes the temporal duration of emotional maintenance. Furthermore, CVT proposes that the elicitation of a specific emotion is a function of the joint multiplicative interaction between perceived control ($C \in [0, 1]$) and perceived value ($Vapp \in [0, 1]$):

$$E = f(C, Vapp) = \gamma \cdot (C \times Vapp) \quad (2)$$

where γ is a scaling constant (Equation 2). For instance, when a student possesses high control and positive value over a task, they experience prospective enjoyment or excitement (positive activating). When control is low and negative value is high, test anxiety or prospective hopelessness (negative activating or deactivating) is aroused. Activity emotions such as boredom arise when the task lacks value ($Vapp \to 0$), which can manifest under conditions of underchallenge (high control relative to task demands) or overchallenge (low control relative to task demands). The downstream impact of these emotional states on academic achievement and cognitive-motivational performance (P) can be mathematically modeled as:

$$P = g(E, M, R) = WT\,[E, M, R]T - \delta interference(E, A) \quad (3)$$

where M represents learning motivation, R represents cognitive resources (e.g., working memory allocation), W is a vector of model weights, and δinterference is an interference function capturing the depletion of working memory capacity caused by the intrusive thoughts of high-activation negative emotions (e.g., test anxiety). Crucially, CVT highlights that achievement emotions exhibit relative universality—while their frequency, duration, and specific triggers vary across individual cultural, socioeconomic, and gender groups, the underlying functional mechanisms linking appraisals, emotions, and academic performance remain structurally equivalent across diverse populations.

2. Methodology I: Text-Based Academic Emotion Classification

Unstructured text data generated in modern online learning environments—such as course feedback, reflection logs, homework submissions, and discussion forum posts—contains a wealth of academic emotional information. To extract these latent states efficiently, researchers have moved away from traditional rule-based and manual feature-engineering techniques toward deep learning architectures. A prominent framework proposed by Yu Wang (2025) utilizes a Bidirectional Long Short-Term LSTM (Bi-LSTM) network combined with a Self-Attention Mechanism and external semantic knowledge graphs. The architecture addresses the contextual dependencies of language, capturing semantic shifts across long word sequences.

2.1 Bidirectional LSTM Layer

The input text is first tokenized, and each word is projected into a low-dimensional dense vector space using pre-trained text embeddings (e.g., Word2Vec or GloVe) combined with an external conceptual knowledge subgraph to produce a weighted "syntax-semantics" representation. The embedded sequence X = [x1, x2, . . . , xT] is processed simultaneously by a forward LSTM and a backward LSTM. The forward LSTM processes the text from left to right, computing the hidden states:

$$\overrightarrow{h}_t = LSTM_{fwd}(x_t, \overrightarrow{h}_{t-1}) \quad (4)$$

Concurrently, the backward LSTM processes the text from right to left, capturing future contextual dependencies to yield the hidden states:

$$\overleftarrow{h}_t = \mathrm{LSTM}_{bwd}(x_t, \overleftarrow{h}_{t+1}) \quad (5)$$

The forward and backward hidden states are concatenated at each time step t to form a comprehensive, context-aware hidden state:

$$h_t = [\, \overrightarrow{h}_t; \overleftarrow{h}_t \,] \quad (6)$$

2.2 Self-Attention Mechanism

To prevent the loss of critical emotional keywords (e.g., "anxious," "overwhelmed," "excited") in long sentences, a self-attention layer is integrated. The attention layer computes a scalar weight αt for each hidden state ht, reflecting its semantic importance relative to the student's emotional state. Let ut be a non-linear projection of the hidden state ht, and v be a learnable context vector that acts as a query to find important semantic patterns:

$$u_t = \tanh(W_a h_t + b_a) \quad (7)$$

The attention weight αt is computed using a SoftMax function to ensure a valid probability distribution over the sequence:

$$\alpha_t = \exp(u_t^T v) \,/\, \Sigma \exp(u_k^T v) \quad (8)$$

The final attention-weighted representation of the entire text sequence, denoted as the semantic vector s, is calculated as the weighted sum of all hidden states:

$$s = \Sigma\, \alpha_t h_t \quad (9)$$

The semantic vector s is then passed through a fully connected output layer, and a Softmax activation function computes the probability distribution over K distinct academic emotion categories:

$$P(y = c \mid X) = \exp(W_c^T s + b_c) \,/\, \Sigma \exp(W_j^T s + b_j) \quad (10)$$

where Wo and bo are the weights and biases of the output layer, respectively. Yu Wang's (2025) implementation validated this deep learning pipeline against traditional machine learning classifiers on a college student feedback dataset. The model achieved a high level of accuracy and robustness, outperforming unidirectional LSTMs by 7.83%, demonstrating the effectiveness of combining bidirectional sequential models with self-attention.

3. Methodology II: Computer Vision-Based Facial Expression Analysis

Facial expressions are among the most direct, observable, and natural channels through which human beings communicate emotions. In classroom settings, real-time monitoring of students' facial expressions provides educators with direct feedback on the collective emotional climate, indicating whether students are engaged, confused, or bored. Yuvaraj et al. (2023) developed a deep transfer learning framework based on a configured VGG-19 Convolutional Neural Network (CNN) to classify students' academic affective states from video data. VGG-19, pre-trained on the massive ImageNet dataset, acts as a highly generalized feature extractor, which is then fine-tuned on a custom classroom emotional dataset.

3.1 Dataset Construction and Preprocessing

A primary contribution of Yuvaraj et al.'s study was the assembly of a balanced hybrid dataset containing 8,674 static color images (128×128 pixels) across 9 academic emotional classes. The dataset was compiled from three publicly available databases and custom classroom recordings from the internet:

- **DAiSEE (Dataset for Affect States in E-Environment):** Contributed 3,935 images annotated for boredom, confusion, and engagement.
- **Raf-DB (Real-world Affective Faces Database):** Contributed 2,405 images of surprise, anger, and sadness.
- **EmotioNet:** Contributed 522 images.

Internet Classroom Videos: Contributed 1,812 images representing amusement and relief in natural learning environments.

Table 1: Class Breakdown and Data Sources in the Hybrid Academic Facial Dataset (Yuvaraj et al., 2023)

Emotion	DAiSEE	Raf-DB	EmotioNet	Internet Videos	Total
Anger	–	705	222	–	927
Amused	–	–	–	1000	1000
Boredom	1000	–	–	–	1000
Confusion	1000	–	–	–	1000
Engagement	1000	–	–	–	1000
Interest	935	–	–	–	935
Surprise	–	700	300	–	1000

Relief	–	–	–	812	812
Sadness	–	1000	–	–	1000
Total	**3935**	**2405**	**522**	**1812**	**8674**

The video frames underwent pre-processing to manage redundant frames. Utilizing frame-sampling, only 4 frames/second (an interval of 0.25 seconds) were extracted, as this sampling frequency preserves classification performance while reducing computational overhead. Faces were automatically detected, cropped, and resized to 128×128×3 dimensions.

3.2 VGG-19 Model Configuration

The VGG-19 architecture consists of 16 convolutional layers and 3 fully connected layers. Because training tens of millions of parameters from scratch require massive datasets, transfer learning was implemented. To adapt the pre-trained ImageNet model to classroom facial expression classification, the study employed a three-stage layer configuration strategy:

- **Layer Freezing:** The initial convolutional layers (Layers 1 through 12) were frozen. These front-end layers capture universal low-level visual features such as edges, corners, and textures, which are common to both ImageNet and human faces.
- **Layer Fine-Tuning:** The deep convolutional layers (Layers 13 through 19) were set to be trainable, allowing the network to adapt its higher-level semantic filters to the specific geometry of facial expressions.
- **Classification Head Reconstruction:** The original fully connected classification layers were replaced with a Multi-Layer Perceptron (MLP) containing dropout regularization (to prevent overfitting) and a final 9-neuron SoftMax layer corresponding to the nine target emotions.

The system was evaluated using 5-fold cross-validation. The configured VGG-19 model demonstrated outstanding stability and classification performance, achieving an average accuracy of 82.73% ± 2.26%, sensitivity of 82.55% ± 2.14%, and a high specificity of 97.67% ± 0.45%.

4. Methodology III: Multi-Physiological Signal Fusion

While textual and facial classification methods are powerful, they are susceptible to conscious manipulation (e.g., a student deliberately masking boredom with a fake smile, or a depressed student writing polite, positive feedback). To achieve an objective, unmaskable measurement of emotional states, researchers leverage physiological signals regulated by the autonomic nervous system. Wu et al. (2024) developed an objective sensing system for educational human-computer interaction (HCI) that captures and fuses four physiological signals using lightweight wearable sensors:

- **Electrocardiography (ECG):** Captures cardiac activity and autonomic arousal.
- **Electromyography (EMG):** Measures somatic muscle tension, particularly facial micro tensions.
- **Electrodermal Activity (EDA):** Reflects skin conductance and sympathetic nervous system activation.
- **Respiratory Signals (RESP):** Gauges respiratory rate and depth, which fluctuate under stress.

The system classifies four distinct human mental states: (1) Baseline State (relaxed, non-stimulus), (2) Stress State (high activation, anxiety), (3) Amusement State (positive activating), and (4) Meditation State (deep relaxation, regular respiratory patterns).

4.1 Feature Extraction and Hilbert-Huang Transform (HHT)

To extract features from non-stationary biological signals, the pipeline utilizes a combination of time-domain statistical analysis and time-frequency analysis. For time-domain analysis, statistical features including mean, mean square, first-order difference, and second-order difference are extracted over a 10-second sliding window. For non-stationary and non-linear signals like ECG and respiration, the Hilbert-Huang Transform (HHT) is employed. The first step of HHT is Empirical Mode Decomposition (EMD), which decomposes the complex, multi-component biological signal x(t) into a finite set of oscillatory components called Intrinsic Mode Functions (IMFs):

$$x(t) = \Sigma Ci(t) + rn(t) \tag{11}$$

where Ci(t) represents the i-th IMF, and rn(t) represents the final monotonic residual. An IMF must satisfy two conditions: the number of extrema and zero-crossings must differ by at most one, and the local mean of the envelope defined by local maxima and minima must be zero. The decomposition is performed iteratively via a sifting process. For each IMF Ci(t), a Hilbert transform is applied to compute the instantaneous frequency and amplitude, generating a local time-frequency spectrum:

$$H(\omega, t) = Re\ \Sigma\ ai(t)\ ej \int \omega\_i(t)\ dt \tag{12}$$

4.2 Feature Selection and SE-CNN Decision Model

To prevent feature redundancy and overfitting caused by high-dimensional inputs, Wu et al. implemented a rigorous, three-stage feature selection process:

- **Filter Method:** Initial feature space reduction by removing features with no significant statistical relationship to the target emotional states.
- **Wrapper Method (Genetic Algorithm):** Refined selection by utilizing a genetic algorithm (GA) to search for the optimal feature subset that maximizes classification accuracy.

- **Embedded Method:** Regularized regression (L1/L2 regularization) to penalize less informative parameters.

The selected features are fed into a Squeeze-and-Excitation Convolutional Neural Network (SE-CNN). The Squeeze-and-Excitation (SE) block is an architectural unit designed to improve representational capacity by performing channel-wise feature recalibration. The SE block operates in two steps:

- **Squeeze (Fsq):** Global average pooling is applied to squeeze the spatial dimensions H×W of the feature map uc into a channel descriptor z ∈ RC:

$$z_c = F_{sq}(u_c) = 1 / (H \times W) \Sigma \Sigma u_c(i, j) \quad (13)$$

- **Excitation (Fex):** A self-gating mechanism captures channel-wise dependencies using a simple bottleneck with two fully connected (FC) layers:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \quad (14)$$

where δ is the ReLU activation function, σ is the sigmoid gating function, W1 is a dimensionality-reduction layer and W2 is a dimensionality-increasing layer. The input feature maps are scaled by the computed channel weights sc to emphasize informative features and suppress irrelevant ones. Integrating an attention mechanism after the SE block provides high model interpretability, highlighting which biological indicators (e.g., ECG R-peaks, respiratory wave variations) drive the emotional classification.

5. Empirical Benchmarking and Comparative Analysis

To evaluate the absolute and relative efficacy of these different machine learning classification systems, we benchmark the performance metrics reported in the literature across different modalities. This includes text, facial, physiological, and adjacent campus depression-detection models. In the area of campus depression detection, Peter J. Yu (2023) developed a natural language processing (NLP) and machine learning pipeline to identify signs of depression in 7,680 social media posts from 13 college-related online subreddits. Using a keyword list containing base terms (e.g., depression, miserable, sad, unhappy, deject, despair, hopelessness), 552 posts were identified as depressive. Latent Dirichlet Allocation (LDA) topic modeling revealed three causal areas: Topic A (Institutions and programs), Topic B (Academic projects and assignments), and Topic C (College environment in general, representing 50% of the corpus). Yu evaluated four classifiers: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Adaptive Boosting (AdaBoost). Table 2 presents a unified empirical benchmark of all four primary methodologies evaluated in the literature.

Table 2: Comprehensive Performance Benchmarking of AI Emotion and Depression Classifiers

Model & Modality	Target States	Accuracy	Precision	Recall	F1 Score
Reddit NLP Classifiers (Yu, 2023)					
– AdaBoost	Binary (Depression)	98.98%	1.000	0.917	0.957
– SVM	Binary (Depression)	94.67%	1.000	0.995	0.990
– Random Forest	Binary (Depression)	93.53%	1.000	1.000	1.000
– Logistic Regression	Binary (Depression)	93.31%	0.956	0.844	0.755
Bi-LSTM + Attention (Wang, 2025)					
– Proposed Text Classifier	4-Class Academic Emotions	79.23%	–	–	75.66%
– Baseline LSTM	4-Class Academic Emotions	71.40%	–	–	64.92%
Facial VGG-19 (Yuvaraj et al., 2023)					
– Configured VGG-19 (Ours)	9-Class Academic Emotions	82.73%	–	82.55 %	–
– Pre-trained VGG-19	9-Class Academic Emotions	72.12%	–	71.85 %	–
SE-CNN Physiological (Wu et al., 2024)					
– WESAD Dataset	4-Class Mental States	87.90%	88.30%	87.71 %	87.71%
– MSSFT Dataset	4-Class Mental States	85.50%	86.20%	85.80 %	85.65%

5.1 Analysis of Findings

The empirical results reveal several critical trade-offs across modalities:

- **Text-Based Classifiers:** Demonstrate high accuracy in constrained, structured formats. The AdaBoost and Random Forest models of Yu (2023) achieve remarkable accuracy (98.98%) and F1 scores (1.000) on social media data because posts are written anonymously, encouraging candid, explicit disclosures of depressive feelings. In contrast, Wang's (2025) text classifier operating on student reflection logs and course feedback experiences lower performance (79.23%) due to the social desirability bias, where students underreport negative feelings in formal educational settings.

- **Visual Expression Classifiers:** Fine-tuning the configured VGG-19 model yields an 11% performance boost over the pre-trained model. However, real-world deployment remains constrained by computational latency. Processing a single frame containing one face takes 13.72 seconds total on standard hardware (including 8.78 seconds for face detection and 1.44 seconds for emotion classification), making high-frequency, real-time tracking of entire classrooms computationally expensive.
- **Physiological Fusion Classifiers:** Fusing multiple physiological signals (ECG, EMG, RESP, EDA) via SE-CNN results in high accuracy (87.90%) and F1 scores (87.71%) for 4-class emotional recognition, outperforming any single-signal classifier. However, the physical placement of wearable sensors (e.g., chest straps, electrodermal electrodes) is intrusive, potentially distracting students and introducing external noise or motion artifacts into the signal.

6. Ethics, Algorithmic Bias, and the Measurement Problem

Despite the promising accuracy of emotional AI classification systems, their deployment in high-stakes educational environments introduces serious ethical and operational risks. G.S. Smrithy et al. (2026) highlight three core challenges: the representativeness of data, algorithmic fairness, and the fundamental measurement problem.

6.1 The Fragile Assumption of Emotional Ground Truth

A fundamental limitation of Emotional AI systems is the Measurement Problem. These technologies rest on a fragile assumption: that human emotions can be treated as stable, objective, and universally measurable ground truth signals. In reality, emotional states are context-dependent, culturally mediated, and subject to immense individual variation. For instance, a facial analysis system might detect a lowered brow and tight lips, automatically inferring the emotional state as "anger" or "hostility." However, in an academic context, this exact expression frequently represents intense "concentration" or "deep thought." Treating observable, physical micro-expressions as direct, objective indicators of internal emotional states risks reifying highly uncertain algorithmic inferences into authoritative, irreversible decision inputs.

6.2 Algorithmic Bias Pathways and Sociodemographic Disparities

Upstream data representation issues directly translate into downstream algorithmic discrimination. Most facial expressions and sentiment datasets are heavily skewed toward Western demographic populations:

- **Western Expression Bias:** Predominantly Western-trained models fail to recognize culturally specific expressions where intense emotional suppression is valued.
- **Racial Disparities in Facial Analysis:** Foundations of commercial facial analysis software suffer from severe gender and racial accuracy gaps (Buolamwini & Gebru, 2018). Emotion recognition systems incorporating these biased facial detection pipelines propagate errors downstream. Studies demonstrate that systems systematically misclassify dark-skinned neutral faces, assigning significantly higher scores for "aggression," "fury," or "hostility" compared to light-skinned faces under identical neutral conditions.

6.3 Real-World Impact and Case Studies

The ethical dangers of deploying opaque emotion inference systems are illustrated by the case of HireVue, an AI-based recruitment platform. In 2019, HireVue faced severe public criticism and regulatory scrutiny for using automated video analysis—evaluating facial expressions, vocal tone, and word choice—to score job applicants' suitability. Because the algorithm was a "black box," neither the candidates nor HireVue's own developers could explain how specific emotional cues translated into hiring scores. Candidates were forced to provide biometrics without knowing how their data was stored or evaluated, raising serious privacy concerns. Crucially, critics argued that the system structurally penalized neurodivergent, disabled, or culturally diverse candidates whose emotional expressions did not align with the Western, neurotypical benchmarks embedded in the training data. Faced with legal scrutiny and public backlash, HireVue formally ceased using facial analysis in its assessments in 2021, shifting exclusively to verbal and written semantic evaluations. An adjacent case is Amazon's automated hiring system, which did not use emotion recognition but similarly automated resume screening. It was shut down in 2018 after it was discovered to actively discriminate against female applicants for technical roles because it was trained on historical resumes reflecting a male-dominated workforce. This case demonstrates that automated decision pipelines, if unchecked, amplify historical inequities.

6.4 Governance, EU AI Act, and Human-in-the-Loop Oversight

To mitigate these risks, developers must implement comprehensive algorithmic governance. The European Union's AI Act explicitly categorizes emotion recognition and inference systems as high-risk applications. The regulations establish strict legal requirements:

- **Mandatory Bias Audits:** Pre-deployment adversarial testing and continuous fairness audits to identify demographic accuracy disparities.

- **Transparency and Explainability:** Moving away from black-box models to Explainable AI (XAI) frameworks that provide logical, human-readable rationales for emotional assessments.
- **Human-in-the-Loop Oversight:** Mandating that automated emotion classifications never operate autonomously. In educational settings, algorithmic warnings of student depression or disengagement must act solely as supportive prompts for human educators and clinical professionals, preserving human dignity, privacy, and agency.

7. Conclusion and Future Directions

This paper synthesized three major paradigms in the machine learning-based classification of achievement emotions and student depression, grounded in Pekrun's Control-Value Theory. While text-based Bi-LSTMs, facial transfer CNNs, and multi-physiological SE-CNNs demonstrate impressive experimental accuracy, their real-world deployment faces significant technical and ethical hurdles. Future research must focus on:

- **Multi-Modal Fusion with Contextual Awareness:** Integrating textual, visual, and physiological signals into unified networks that account for the situational context of the classroom.
- **Privacy-Preserving Architectures:** Implementing federated learning and encrypted model training to process sensitive biological and biometric data without compromising student privacy.
- **De-biasing Algorithms:** Building highly diverse, representative datasets to eliminate sociodemographic accuracy disparities.

By combining robust technical design with strict ethical governance, emotional AI can be transformed from a speculative surveillance tool into a supportive, equitable asset for personalized learning and student mental health management.

References

1. Yu, P. J. (2023). A Machine Learning Method for Detecting Depression Among College Students. International Journal of Computer Applications, 185(24), 1–8.
2. Wang, Y. (2025). Analysis and application of college students' academic emotions based on deep learning and psychological status. Inderscience Online, 125–139.
3. Yuvaraj, R., Rajalakshmi, R., Dhanvanth, V., & Fogarty, J. (2023). Automated Classification of Student's Emotion Through Facial Expressions Using Transfer Learning. The European Conference on Education Official Conference Proceedings.
4. Wu, W., Zuo, E., Zhang, W., & Meng, X. (2024). Multi-physiological signal fusion for objective emotion recognition in educational human–computer interaction. Frontiers in Public Health, 12, 1492375.

5. Smrithy, G. S., Chandrasekaran, B., & J, O. (2026). Ethics and bias in emotional AI. Frontiers in Artificial Intelligence, 9, 1768696.
6. Pekrun, R. (2024). Control-Value Theory: From Achievement Emotion to a General Theory of Human Emotions. Educational Psychology Review, 36(83), 1–36.
7. Buolamwini, J., & Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81, 77–91.