How Teachers' Gestural Culture Influences Japanese Students' Emotions: A Machine Learning Approach

Yuka Nishi[1], Olivia Kennedy[2], Choi Dongeun[3], Noriaki Kuwahara[4]*
Graduate School of Science and Technology, Kyoto Institute of Technology, Kyoto, Japan[1, 3, 4]
Center for Social and Biomedical Engineering, Kyoto Institute of Technology, Kyoto, Japan[4]
Nagahama Institute of Bioscience and Technology, Nagahama, Japan[2]

***Abstract*—This study analyzes differences in teachers' gestural styles based on their culture and investigates how these differences are perceived to influence Japanese students' emotional responses by active observers. Classroom videos of Japanese- and English-native instructors were analyzed using MediaPipe for gesture tracking and DeepFace for facial emotion recognition. Ground-truth emotion labels were collected from four Japanese observers. Results show that Japanese and non-Japanese teachers' gesture dynamics differ in terms of range, rhythm, and symmetry. Japanese student observers perceived each group's gestures differently, with cultural familiarity playing a role in their shifts in emotion. Machine learning models trained on gesture features, facial emotion scores, and teacher background successfully predicted students' affective reactions. These findings highlight the importance of culturally sensitive nonverbal communication in education and demonstrate the potential of AI-based approaches for modeling student emotion in cross-cultural contexts. This study contributes a novel multimodal framework that integrates gesture dynamics, facial emotion recognition, and teacher cultural background to predict student affect, thereby highlighting the necessity of culturally adaptive affective computing in education.**

***Keywords*—Nonverbal behavior; affective computing; AI-based gesture recognition; cultural differences; multimodal analysis; cross-cultural education; emotion prediction**

I. Introduction

Understanding how teachers' nonverbal behavior influences student learning has become increasingly important, particularly in intercultural educational contexts. Among various nonverbal cues, gestures play a pivotal role in directing attention, emphasizing content, and conveying affective meaning. Prior research has shown that gestures not only supplement verbal explanations but also shape learners' cognitive and emotional engagement [1], [2].

However, most of the existing literature on classroom gestures has relied on qualitative observations, which lack precise or scalable measurement approaches [3] [4]. These limitations are especially notable when considering the effect of cultural differences in gesture usage. High- and low-context communication styles, as theorized by Hall [5], may lead to mismatches in how gestures are produced by teachers and interpreted by students. For example, Japanese students may not always correctly interpret gestures used by non-Japanese instructors [6].

Such variation has concrete implications for the classroom. East Asian instructors have been found to favor linking gestures for conceptual clarity, whereas Western educators tend to use more illustrative, expressive forms [7]. These differing gestural styles are not just cultural artifacts but pedagogical strategies that carry affective outcomes. Although the nonverbal behaviors that teachers use reduce psychological distance and have been associated with higher student motivation and engagement [8], such cues may be interpreted differently across cultures.

Beyond cultural and pedagogical dimensions, gestures also serve distinct semiotic functions. McNeill and Kendon classified gestures into types based on how they encode meaning, ranging from spatial reference (deictic) to rhythm and cadence (beat) [2] [4]. These types align with cognitive mechanisms such as mental imagery and discourse structuring [9] [10]. However, their effectiveness depends on how gestures and speech combine to convey meaning in real time [11], and on cultural display rules that modulate students' interpretations [5].

Coinciding with these cultural phenomena, technological advances in affective computing have enabled the quantitative analysis of emotional reactions in real-time classroom environments. Recent advances in AI-based technologies—such as those enabling real-time gesture tracking and facial affect recognition—offer scalable means to detect student emotions in classroom environments [12][13][14]. These methods, however, must account for cultural nuances in gesture production and emotional expression.

This study focuses on how teachers' culturally shaped gestural styles influence Japanese students' emotional perceptions during classroom interactions. Rather than analyzing gesture types per se, we examine how gesture dynamics vary between Japanese and non-Japanese instructors, and how those differences are perceived to affect changes in students' emotions by Japanese observers, simulating a classroom perspective. We also evaluate the extent to which these responses can be predicted using machine learning models trained on gestural and facial features. In other words, the central research question we address is how culturally conditioned differences in teachers' gestures influence Japanese students' affective responses, including the extent to which these responses can be systematically modeled using AI-based analysis.

II. Related Work

Understanding the role of nonverbal teacher behavior in shaping student engagement has been a long-standing point of interest to researchers. Gestures, facial expressions, and body

orientation contribute not only to the transmission of content but also to the affective climate of the classroom. Foundational studies have emphasized the importance of gesture in supporting conceptual understanding and memory formation [1][2]. However, most prior work in this area has relied on observational or qualitative methods [3][4].

Cultural context significantly shapes the form and interpretation of nonverbal behavior. Hall's theory on high- and low-context communication [5], laid the groundwork for understanding why gestures may be interpreted differently across cultures. Park et al. [6] further demonstrated that students from different cultural backgrounds may misinterpret hand gestures used by foreign teachers. Comparative educational studies also show that East Asian teachers tend to use more "linking" gestures to build logical connections, while Western teachers often gesture more illustratively [7].

From a pedagogical standpoint, higher levels of teacher immediacy—the nonverbal behavior used to reduce psychological distance— has been consistently associated with increased student motivation and engagement [8]. These behaviors include open gestures, sustained eye contact, and expressive facial expressions. However, what constitutes "immediacy" may vary by cultural norms, complicating how students interpret these behaviors [5].

Beyond their affective and cultural functions, teacher gestures also have distinct symbolic roles that help impart meaning in instructional discourse. McNeill [2] and Kendon [4] categorized gestures not just by their form, but by how they encode referential or abstract information. In this framework, deictic gestures refer to spatial or discourse objects, iconic gestures depict physical forms, metaphoric gestures give form to abstract ideas via spatial examples, and beat gestures offer rhythm and cadence. These categories align with cognitive processes such as mental imagery, thought activation, and discourse segmentation [9] [10]. In classroom contexts, these functions may influence students both in terms of comprehension and emotional response, especially when the gesture reinforces the accompanying verbal message [11].

In recent years, AI-based techniques have increasingly been applied to classroom research, enabling objective, fine-grained analysis of gesture and emotion. Computer vision systems like MediaPipe and DeepFace allow automatic tracking of hand motion and facial affect, making them useful as new tools for educational affective computing. Whitehill et al. [12] proposed a system to estimate a student's level of engagement via facial cues, finding that machine-measured engagement levels predicted learning outcomes better than prior academic performance. Dai et al. [13] extended this approach by combining multiple modes of tracking—including audio and transcript sentiment—to detect negative teacher emotion (e.g., frustration) across large video datasets. These developments point to the feasibility of real-time, AI-driven feedback systems that support teachers in monitoring classroom sentiment [14].

Although these systems still face challenges in recognizing culturally nuanced gestures or subtle emotions [13], their performance thus far marks a significant step toward quantifying affective dynamics in education. However, few studies explicitly incorporate the teacher's cultural background as a predictive feature when modeling student emotional responses. Combining both multimodal learning analytics and cross-cultural gesture, research enables a more holistic understanding of teacher–student interaction and could lead to change in both pedagogical design and teacher professional development.

While prior studies have examined cultural differences in nonverbal communication [6], few have systematically quantified how gestures differ by teacher cultural background and how those gestures are emotionally perceived by students. This study aims to fill this gap by employing a combination of gesture kinematics, facial emotion analysis, and teacher cultural identity in a unified predictive framework.

Recent advances in AI-based approaches have significantly enhanced the analysis of teachers' nonverbal behaviors and the evaluation of student affect. For instance, Yoon et al. proposed a MediaPipe-based system to quantify pointing gestures by monitoring asynchronous video lectures and demonstrated that in-service teachers adapted their gestures more effectively than pre-service teachers [15]. Similarly, Chen et al. developed a real-time gesture recognition framework using MediaPipe BlazePose to detect teachers' hand signals and deliver attentional cues to students with identified attention deficit problems; doing so achieved an F₁ score of approximately 0.88 [16]. Shou et al. introduced a multi-scale convolutional neural network (CNN) with fine-grained attention enhancements for recognizing student facial expressions in real classroom settings; reporting 93 % accuracy on in-school datasets [17]. Singh et al. gauged online learners' engagement with a multi-modal student attentiveness detection model (MMSAD) that fuses CNN-based facial expression analysis with speech activity detection [18]. Also using CNN-based emotion recognition, Salloum et al. demonstrated pipeline for real-time integration into online learning platforms; achieving 95% accuracy on the FER2013 benchmark [19]. These studies highlight the growing trend of leveraging MediaPipe, DeepFace, and deep learning models to monitor both teacher gestures and student emotions in educational environments. Such AI techniques provide a robust foundation for the work involved with in this study, which employs MediaPipe and DeepFace to perform a cross-cultural analysis of teacher gestures and their impact on Japanese students' affect.

Building on these developments, our study aims to bridge a crucial gap in cross-cultural gesture analysis literature by examining how culturally shaped gestural styles influence students' emotional perceptions in intercultural classrooms. While prior research has demonstrated the utility of AI-based systems for gesture tracking and emotion recognition, few studies have explicitly modeled the cultural background of teachers as a predictive factor in student affect; such cultural context is often treated as peripheral rather than integral to affective dynamics. We address this by integrating gesture kinematics, facial emotion features, and teacher cultural identity into a unified machine learning framework. This allows us to investigate not only behavioral differences across cultures, but also how these differences impact emotional interpretation by students. In doing so, we contribute to the development of affective computing systems that can adapt to culture, in order to better reflect the complexities of human communication in diverse educational settings.

III. Methodology: AI-Based Multimodal Analysis of Culturally Conditioned Gestural Impact

This study uses a multimodal analysis framework to investigate how cultural differences in teachers' gestural behavior influence Japanese students' emotional responses. Fig. 1 illustrates the overall workflow of the study, from video collection to modeling. As shown in Fig. 1, the procedure consists of five main steps:

- Video collection from Japanese and non-Japanese instructors
- Gesture clip extraction
- Observer-based affective labeling
- Feature extraction using MediaPipe and DeepFace
- Statistical analysis and machine learning modeling

Each step is described in detail in the following subsections.

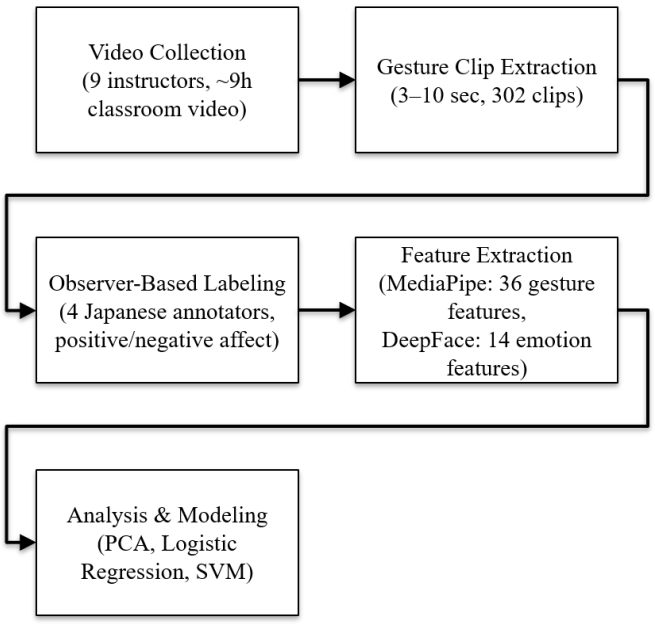

Fig. 1. Overall workflow of the study, from classroom video collection to predictive modeling of student affect.

We employed computer vision tools to extract gesture dynamics, analyzed student affect based on their facial expressions and observer labels, and built machine learning models incorporating both motion features and teacher cultural identity.

A. Dataset

Video data were collected from nine English language instructors (five native Japanese and four native English speakers) teaching Japanese university students in classroom settings. The total length of the recordings was approximately nine hours. The classroom videos were originally recorded as part of a study on spontaneous dynamic teacher emotion recognition by Kennedy et al. [20] and were reused in this study with appropriate ethical approval. Gesture clips of three to ten seconds were segmented from these videos based on visible hand movement, excluding clips where gestures were obstructed or out-of-frame. Only segments where the teacher was facing forward with both hands visible were included. In total, 302 clips were extracted. This dataset was selected because it captures authentic classroom interactions in a naturalistic setting, rather than staged or laboratory recordings, ensuring ecological validity. While the sample size is modest and limited to nine instructors, it provides a balanced comparison between Japanese and non-Japanese teachers in real educational contexts, which is essential for addressing our research questions.

B. Observer-Based Labeling

Four Japanese female annotators, aged 20 to 23, independently labeled each gesture clip by imagining how they, as a student, in the teacher's classroom, might emotionally respond to the teacher's gestures. Their ratings were based on their own perceptions, rather than on observed student reactions. Each annotator viewed the videos in randomized order and was not made aware of the teachers' cultural identities, language backgrounds, and machine-generated emotion scores (positive/negative). Because the focus of this study is gesture, the annotators made their judgments on soundless clips.

The initial consensus label for each clip was determined by majority—when three or more annotators labelled a clip the same way, their shared label was used. In cases of a tie (2 vs. 2), a decision tree model was trained on the majority-agreement cases to predict the most likely final label based on the pattern of individual annotator inputs. This model served to resolve ambiguous labeling outcomes in a reproducible and data-driven manner.

To ensure the validity of this modeling approach, we evaluated the decision tree model's ability to replicate majority-agreement labels through two complementary methods:

- Distribution-level comparison: We applied a chi-squared test and computed Kullback–Leibler divergence to compare the distribution of labels produced by the decision tree model to that of the majority-agreement labels.
- Sample-level comparison (two-two-split cases): We calculated prediction accuracy and Cohen's kappa between the decision tree model's predicted labels and the final consensus labels.

These evaluation methods were used to confirm that the decision tree model reliably approximated consensus labeling in tie situations. All final labels—whether decided by majority or through the decision tree—were then used as ground truth in subsequent supervised learning tasks, as described in Sections *E.2* and *E.3*.

C. Feature Extraction from Gestures

To quantify each teacher's gesture dynamics, we used MediaPipe Holistic (Zhang et al. [21]), which provides real-time tracking of body pose and hand landmarks. All features were computed relative to the teacher's nose landmark to normalize for their placement and the camera's framing. Below is a description of each feature group (totaling 36 dimensions):

1) Wrist Position Statistics (Left and Right): For each wrist, MediaPipe Holistic tracks the trajectory of its (x, y) coordinates

over time. From these trajectories, four descriptive statistics for the x-axis and four for the y-axis were computed—namely, the maximum, minimum, mean, and variance of the wrist position across all frames. These statistics capture the spatial extent (range), central tendency (average position), and variability (how much the hand moves) during the gesture.

2) Hand velocity statistics: Velocity was calculated by differentiating frame-by-frame wrist positions in both *x* and *y* directions. For each axis, we then derived the maximum, minimum, mean, and variance of these instantaneous velocities across the clip. These features reflect how quickly and consistently the teacher moves their hands. High maximum velocity, for example, indicates a fast gesture, while high variance suggests fluctuating speed.

3) Hand acceleration statistics: Acceleration was calculated by taking the temporal derivative of the velocity values along each axis. Like velocity, *x* and *y* accelerations were summarized using their maximum, minimum, mean, and variance across the clip. Acceleration metrics characterize the abruptness or smoothness of movements—large peaks in acceleration denote sudden starts or stops, whereas low variance implies a more uniform movement.

4) Positional correlations: To capture directional consistency and shape of motion, we computed the Pearson correlation coefficient between the *x* and *y* coordinates of each wrist throughout the gesture. A strong positive or negative correlation indicates that the wrists' motions are primarily diagonal, whereas a correlation near zero suggests more independent horizontal and vertical movement.

5) Velocity autocorrelations: Finally, we calculated autocorrelation values for the *x* and *y* components of velocity over short time lags. These correlations measure how similar the hand's speed is from one moment to the next, providing insight into rhythmic or repetitive motion patterns.

In total, these five groups produce a motion profile of 36 numerical features per gesture clip (8 wrist position statistics for each wrist=16, 8 hand velocity statistics=8, 8 hand acceleration statistics=8, 2 positional correlations=2, and 2 velocity autocorrelations=2). Each feature was normalized by clip duration and teacher height before summarization.

While other frameworks, such as OpenPose, have also been widely used for gesture tracking, MediaPipe Holistic was selected because it offers a lightweight and real-time implementation that is less computationally demanding while maintaining comparable accuracy in classroom environments.

This 36-dimensional motion profile, based on MediaPipe Holistic [21], serves as the basis for all subsequent analyses and predictive modeling.

D. Emotion Annotation and Predictive Modeling

This section describes how teacher gestures were analyzed to predict Japanese students' emotional responses, and how gesture characteristics differed systematically between Japanese and non-Japanese instructors. The analysis follows two complementary paths: 1) identifying cultural differences in gesture execution using machine learning, and 2) modeling students' affective responses.

1) Cultural style classification of teachers: To explore how gesture dynamics differ by cultural background, we trained a supervised classifier to distinguish between Japanese and non-Japanese teachers based on the gesture features extracted in Section *C*. This classification task served to identify which motion patterns (e.g., gesture range, rhythm, directional symmetry) are characteristic of each teacher group. The results of this analysis provide an empirical basis for comparing the gestural tendencies of different teacher populations.

2) Observer-based emotion annotation: Japanese students' emotional responses to each gesture were annotated using the four-person consensus protocol, described in Section *B*. Each clip was labeled as eliciting either a positive or a negative emotional reaction.

These consensus labels were then used as ground truth for evaluating whether machine learning models could accurately predict observer-inferred student affective responses based on gesture perception.

3) Prediction of student emotion from gesture features: We first trained a machine learning model to predict the binary emotion label (positive/negative) using gesture features alone. This allowed the model to assess whether the movement pattern of the teacher's gesture, independent of facial cues, was sufficient to explain how Japanese students emotionally responded.

Next, we trained a second model by incorporating both gesture features and facial emotion predictions from DeepFace [22]. DeepFace is a lightweight CNN-based emotion recognition framework capable of estimating seven basic emotion categories (happy, sad, fear, angry, disgust, surprise, neutral) from facial expressions in real-time. It was used to extract per-clip emotion probabilities, which were then averaged over time and used as input features alongside gesture motion. The goal in using this second model was to determine whether combining gesture dynamics with machine-estimated facial affect signals improves prediction performance.

4) Application to feedback and cultural interpretation: These predictive models can be used to provide actionable feedback to teachers by identifying gesture patterns that are more likely to be perceived negatively by students. Such insights can guide the creation of culturally responsive teaching strategies and help educators avoid nonverbal behaviors that may unintentionally cause discomfort or disengagement.

Likewise, the analysis of model coefficients and misclassified cases can help us understand how specific gestural differences between Japanese and non-Japanese teachers contribute to affective outcomes. By systematically comparing model behavior across teacher groups, we aim to identify which motion traits are positively or negatively received in a Japanese cultural context.

E. Implementation

All processing was performed in Python using MediaPipe, DeepFace, and scikit-learn. For gesture-based cultural classification (i.e., distinguishing between Japanese and non-Japanese teachers), we used both Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) to compare performance in modeling stylistic differences. For emotion prediction tasks, logistic regression was primarily used for its transparency, and in some cases, SVM was also applied to assess potential nonlinear relationships.

Although more complex deep learning models could potentially achieve higher predictive accuracy, we deliberately employed logistic regression and SVM because they are interpretable, less prone to overfitting with relatively small datasets, and allow clearer examination of how gesture and facial features contribute to cultural differences.

All of these features were standardized, and stratified cross-validation was applied throughout their application. No extensive hyperparameter tuning was conducted, and default settings and L2 regularization were used as appropriate.

IV. Results

This section presents the empirical findings in four stages. First, the reliability of emotion labels generated through observer annotation and machine-aided resolution are evaluated. Second, cultural differences in teachers' gestural dynamics based on motion features extracted from clips of classroom videos are analyzed. Third, how students' emotional reactions vary depending on the cultural background of the teacher are examined. Finally, machine learning models to predict students' affective responses from gestural and facial features are discussed and the models' performance across different input modalities and teacher groups compared.

A. Reliability of Emotion Labels

To ensure the validity of emotion labels used as ground truth for supervised learning, a multi-step evaluation of observer-labeled data was implemented, as outlined in Section III *B*.

While a majority-vote protocol was used to assign labels, in 30.8% of the samples (72 out of 234 clips), the labels were evenly split (2 vs. 2). To resolve these ambiguous cases, a logistic regression model was trained using the individual ratings from the 162 majority-agreement clips as input and their consensus labels as targets. The trained model was then applied to predict the most probable label in the tie cases.

To assess the validity of this model-based resolution strategy, we conducted two evaluations:

1) Distribution-level consistency: A chi-squared test comparing the emotion label distribution (positive vs. negative) in the predicted set versus the original majority-agreement set showed no significant difference ($p > 0.05$). Additionally, the Kullback–Leibler (KL) divergence between the two distributions was extremely low (0.012), indicating near-identical global distributions.

2) Sample-level agreement: To validate the logistic model's predictive power, we performed five-fold cross-validation on the majority-agreement cases. The model achieved an average accuracy of 95.4% and a Cohen's kappa score of 0.88, demonstrating high consistency with human consensus even in ambiguous cases.

These results suggest that the predicted labels for 2–2 split cases are statistically and behaviorally consistent with majority-vote annotations. All subsequent analyses, including classification and predictive modeling, used this unified set of final labels.

B. Cultural Differences in Gesture Dynamics in Gestures and Emotions

To investigate how nonverbal behavior and emotional expression differ between Japanese (J) and non-Japanese (N) teachers, we conducted a principal component analysis (PCA) on a set of 50 combined dimensions. This set consisted of 36 motion-related features derived from teachers' hand gestures and 14 facial emotion features extracted using DeepFace. All features were standardized before analysis. Component scores were then compared between cultural groups using Welch's t-test, applying Bonferroni correction to control for multiple comparisons.

1) PCA and explained variance: A total of 234 video clips were analyzed, each corresponding to either a Japanese teacher (J) or a non-Japanese teacher (N). For each clip, 50 features were extracted, consisting of:

a) 36 gesture features: These include the maximum, minimum, mean, and variance of left and right wrist positions and velocities in both X and Y directions, as well as wrist velocity correlations.

b) 14 affective features: These represent mean and variance of basic emotional expressions (anger, disgust, fear, happy, sad, surprise, neutral), extracted via facial analysis.

No missing values or filtering was applied. All features were standardized to zero mean and unit variance using StandardScaler from scikit-learn. Principal component analysis (PCA) was conducted to reduce dimensionality and explore patterns associated with cultural differences.

The PCA revealed that the first 14 components together accounted for 81.00% of the total variance in the feature space, with the first component alone capturing 23.2%. Table I shows the explained variance and cumulative variance for each component up to PC14.

2) Statistical comparison: Japanese vs. Native Teachers. We tested for cultural differences in each of the first 14 components using Welch's t-test. Table II presents the p-values (uncorrected and Bonferroni-adjusted) and effect sizes (Cohen's d) for each component. Three components—PC2, PC3, and PC9—showed statistically significant group differences after correction. Adjusted p-values were computed using Bonferroni correction. Values in bold indicate statistical significance (adjusted $p < 0.05$).

TABLE I. EXPLAINED VARIANCE OF PRINCIPAL COMPONENTS (UP TO 80%)

Principal Component	Explained Variance Ratio (%)	Cumulative Variance (%)
PC1	23.2	23.2
PC2	9.1	32.3
PC3	6.9	39.2
PC4	5.9	45.1
PC5	5.5	50.6
PC6	5.0	55.6
PC7	4.3	59.9
PC8	3.7	63.6
PC9	3.3	66.9
PC10	3.3	70.2
PC11	3.1	73.3
PC12	3.0	76.3
PC13	2.4	78.7
PC14	2.3	81

TABLE II. STATISTICAL COMPARISON OF PRINCIPAL COMPONENT SCORES BETWEEN JAPANESE AND NON-JAPANESE TEACHERS

Principal Component	p_value	adjusted_p	Cohen_d
PC1	0.206	>0.05	-0.16
PC2	**<0.001**	**0.001**	**0.52**
PC3	**<0.001**	**<0.001**	**0.82**
PC4	0.534	>0.05	-0.08
PC5	0.137	>0.05	-0.2
PC6	0.1	>0.05	-0.23
PC7	0.506	>0.05	0.09
PC8	0.057	>0.05	0.25
PC9	**<0.001**	**<0.001**	**0.6**
PC10	0.458	>0.05	-0.1
PC11	0.582	>0.05	-0.07
PC12	0.009	>0.05	-0.36
PC13	0.02	>0.05	0.31
PC14	0.732	>0.05	-0.05

1) Interpretation of significant components: We examined the five features with the highest absolute loadings for each significant component (PC2, PC3, PC9), as shown in Table III. Feature names correspond to variables such as RIGHT_WRIST_x_mean (i.e., a mean of the x-position of the right wrist) and velocity_x_correlation (i.e., the correlation coefficient between x-velocities of left and right wrists), etc. "Loading" denotes the PCA factor loading, representing how strongly each feature influences the component.

- PC2: Horizontal motion and facial fear: PC2 differentiates teachers along horizontal gestural dimensions and facial fear variability. Non-Japanese teachers tend to exhibit greater horizontal range and smoother motion in the x-axis (e.g., higher RIGHT_WRIST_x_mean and velocity_x_correlation), while Japanese teachers show higher variation in "fear" expression.
- PC3: Vertical Gesture Amplitude and Negative Affect: PC3 captures vertical gesture positioning and fluctuations in negative emotion. Non-Japanese teachers use broader vertical ranges and faster movements (higher RIGHT_WRIST_y_mean and Vy_mean), while Japanese teachers show greater variance in "sad" facial expression.
- PC9: Postural Stability and Positive Affect: PC9 reflects differences in vertical consistency and "happy" expression. Non-Japanese teachers show more stable hand elevation (higher RIGHT_WRIST_y_mean, lower position_y_correlation) and scored higher in both average and variability of "happy" expressions. Japanese teachers show more fluctuation in vertical posture.

TABLE III. TOP 5 LOADINGS OF SIGNIFICANT PCS

Principal Component	Feature	Loading
PC2	RIGHT_WRIST_x_mean	0.29
	velocity_x_correlation	0.28
	RIGHT_WRIST_x_max	0.28
	LEFT_WRIST_x_mean	-0.27
	fear_variance	-0.26
PC3	RIGHT_WRIST_y_mean	0.29
	RIGHT_WRIST_y_min	0.26
	sad_variance	-0.26
	angry_mean	0.25
	RIGHT_WRIST_Vy_mean	0.23
PC9	position_y_correlation	-0.31
	RIGHT_WRIST_y_mean	0.31
	LEFT_WRIST_y_variance	0.28
	happy_variance	0.28
	happy_mean	0.27

2) Interpretation and cultural implications: The PCA results reveal meaningful cultural contrasts in teachers' nonverbal and emotional behavior. Among the 14 principal components which, when combined, explained 81.0% of the variance, three components—PC2, PC3, and PC9—exhibited significant differences between Japanese and non-Japanese instructors.

- PC2 reflected differences in horizontal hand motion and fear-related facial variability; non-Japanese teachers showed greater lateral gesture range and motion consistency, whereas Japanese teachers exhibited more fluctuation in fear expression.
- PC3 involved vertical hand amplitude and negative affect; non-Japanese instructors used broader, faster

vertical gestures and expressed more "anger", while Japanese teachers showed greater variability in "sad" expression.

- PC9 was associated with postural stability and positive facial affect; non-Japanese teachers maintained higher and more consistent hand elevation with more pronounced "happy" expressions, whereas Japanese teachers exhibited greater variability in both gesture and affect.

These findings support the general view that low-context communicators (e.g., non-Japanese teachers) tend to favor explicit and clear nonverbal behavior [2] [5], whereas high-context communicators (e.g., Japanese teachers) rely more on subtle and context-sensitive expression [23]. The combination of gesture and facial features in this analysis allows us to better understand how teachers from different cultural backgrounds encode emotional signals during instruction.

A more detailed interpretation of these patterns—especially in relation to cultural display rules, teacher immediacy, and affective computing design—is provided in the Discussion (see Section V).

C. Student Emotion by Teacher Culture and Its Predictive Model

1) Group differences in student emotion: To investigate whether a teacher's culture influences their students' emotional responses, we compared the distribution of observer emotion labels (positive vs. negative) across teacher cultural backgrounds (Japanese or non-Japanese). A 2×2 contingency table was constructed (see Table IV), and both the chi-squared test and Fisher's exact test were performed.

TABLE IV. CROSS-TABULATION OF TEACHER CULTURE (JAPANESE VS. NON-JAPANESE) AND STUDENT EMOTION LABEL (POSITIVE VS. NEGATIVE)

Teacher Culture	Student Emotion: Positive	Student Emotion: Negative
Japanese (J)	58	45
Non-Japanese (N)	56	75

The results of the chi-squared test approached significance ($\chi^2 = 3.72$, $p = 0.054$, $df = 1$), and Fisher's exact test yielded a statistically significant result ($p = 0.048$, odds ratio = 1.73).

Although non-Japanese teachers are often seen as more expressive, here Japanese teachers more frequently elicited positive emotional responses (56.3% vs. 42.7%).

This suggests that cultural familiarity may take precedence over expressiveness in shaping students' emotional receptivity. To investigate whether Japanese students' emotional responses can be predicted from teachers' nonverbal behavior, we trained logistic regression model using 50-dimensional features derived from 36 gesture dynamics and 14 facial expressions (see Section IV *B*). Cultural labels were excluded from the models to focus solely on behavioral cues. Using five-fold stratified cross-validation, the model achieved a mean accuracy of 66.2% (σ = 7.5%), suggesting that nonverbal behavior alone provides moderate predictive power. As shown in Table V, however, accuracy differed substantially by teacher culture: 62% for Japanese teachers versus 76% for non-Japanese teachers.

TABLE V. ACCURACY OF LOGISTIC REGRESSION MODELS PREDICTING STUDENT EMOTION, BROKEN DOWN BY TEACHER CULTURE (JAPANESE VS. NON-JAPANESE)

Teacher Culture	Accuracy Mean	Accuracy Std
Japanese	0.62	0.061
Non-Japanese	0.76	0.090

To understand the source of this discrepancy, we compared model coefficients between two cultural groups. Table VI lists the five features showing the greatest differences. For non-Japanese teachers, positive student responses were more strongly associated with high hand velocity (e.g., RIGHT_WRIST_Vx_min, LEFT_WRIST_Vx_max) and consistent vertical motion (position_y_correlation). Meanwhile, for Japanese teachers, positional extremes such as RIGHT_WRIST_x_min played a more prominent role in positive student responses—possibly reflecting a culturally embedded preference for subtle, controlled gestures. These findings highlight the culturally situated nature of affective communication and underscore the value of adapting emotion-prediction models to cultural context.

TABLE VI. FIVE FEATURES WITH THE LARGEST COEFFICIENT DIFFERENCES BETWEEN LOGISTIC REGRESSION MODELS TRAINED ON JAPANESE-TEACHER DATA (J) AND NON-JAPANESE-TEACHER DATA (N)

Feature	J_Coefficient	N_Coefficient	Coefficient Absolute Difference
RIGHT_WRIST_Vx_min	0.82	-1.11	1.93
RIGHT_WRIST_x_min	-1.19	0.25	1.44
LEFT_WRIST_Vx_max	0.60	-0.79	1.39
position_y_correlation	0.37	-0.80	1.18
RIGHT_WRIST_x_mean	1.26	0.25	1.01

2) Summary and cultural reflection: The findings in this section (see also Table V) demonstrate that the cultural background of the teacher not only influences the distribution of students' emotional responses but also modulates the internal structure of the predictive model. Logistic regression classifiers, which are trained separately on Japanese and non-Japanese teachers' data, revealed notable differences in accuracy: student emotion was predicted with significantly higher precision when the teacher was a non-Japanese speaker. This suggests that the affective cues conveyed by native teachers were more consistently interpreted by Japanese students, leading to more stable input–output mappings in the model.

Interestingly, Japanese teachers—who are often described as emotionally reserved in high-context cultures [5] [23]—elicited less predictable emotional responses as perceived from a student's perspective. And yet, this does not imply that their gestures were affectively neutral or irrelevant. Rather, lower

gesture predictability may reflect greater contextual modulation or individual variation in how their nonverbal behaviors are perceived. In fact, this aligns with our findings in Section IV *B.2*, where Japanese teachers exhibited richer vertical gestural amplitude and a broader affective range than might be expected from cultural stereotypes.

This suggests that interactions with culturally dissimilar instructors—such as non-Japanese teachers—may enable otherwise restrained Japanese students to externalize their emotional responses more visibly. And so, our model not only predicts emotion, but also reveals how cultural contrast may enhance how well we perceive student affect in educational interactions. This is a particularly important consideration in intercultural classrooms, where emotional communication may otherwise remain latent.

Taken together, these findings indicate that predictive models must be both behaviorally precise and also culturally responsive. While non-Japanese teachers exhibited more overt and consistently interpreted nonverbal-affective behaviors, Japanese teachers demonstrated context-sensitive expressiveness that was no less impactful, though it may vary more. Overall, these results imply that culturally sensitive affective computing models must account for not only static behavioral norms but also dynamic pedagogical roles and expectations.

V. Discussion

This study sheds light on the complex interplay between cultural background, nonverbal behavior, and affective perception in classroom communication. The concept of nonverbal leakage—concealed emotions and intentions "leaking" through subtle facial and gestural cues [24]—helps explain why even small differences in teacher expressiveness can significantly affect student emotion. Building on prior studies in gesture typology [2][4], immediacy theory [8], and cultural context models [5][23], we quantitatively examined how gesture dynamics and facial affect from teachers of different cultural backgrounds influence Japanese students' emotional responses.

Although the emotion annotations in this study were not derived from actual student reactions, they were produced by native Japanese observers explicitly instructed to adopt the perspective of a student in the classroom. This imagined-student approach allowed us to assess affective responses to teacher gestures in a controlled and consistent manner. Prior work in affective computing and cross-cultural psychology supports the use of projected or observer-based affect labeling when direct feedback is impractical or ethically constrained.

Moreover, recent AI-based works such as quantifying gestures via MediaPipe [15], detecting real-time hand signals using BlazePose [16], and high accuracy recognition of student expressions [17][18]–demonstrates how automated, multimodal feedback can guide instructors. These advances support our use of MediaPipe and DeepFace, as well as point toward future systems that offer real-time, culturally sensitive classroom guidance.

A. Cultural Differences in Nonverbal Expressivity

Our findings confirm that cultural background shapes not only how teachers' gesture, but also how those gestures are emotionally interpreted. PCA results revealed that non-Japanese teachers tended to use gestures with broader horizontal and vertical range, smoother trajectories, and more expressive facial signals. On the other hand, Japanese teachers showed more localized motion patterns and higher variability in facial expressions such as sadness and fear. These patterns are consistent with Hall's theory that high-context cultures favor more subtle, less explicit communication styles [5].

Interestingly, although non-Japanese teachers demonstrated more clear and systematic nonverbal behavior, Japanese teachers were more likely to elicit positive affective responses (56.3% versus 42.7%). This suggests that cultural familiarity may play a stronger role in emotional receptivity than expressive intensity alone—a finding that challenges the assumptions derived from immediacy theory [8].

B. Predictive Modeling and Cultural Modulation

Logistic regression classifiers trained on gesture and affective features showed moderate accuracy in predicting student emotion (66.2%). However, when trained separately by teacher group, model accuracy was significantly higher for non-Japanese teachers (72.3%) than for Japanese teachers (64.8%). This suggests that non-Japanese teachers' affective cues were more consistently understood by students, whereas Japanese teachers' cues exhibited greater contextual or individual variation, possibly due to more nuanced display rules [23].

The top differentiating features between the two models—including RIGHT_WRIST_Vx_min, position_y_correlation, and RIGHT_WRIST_x_min—highlight how affective perception hinges on both intensity and spatial/rhythmic regularity. These quantitative findings resonate with earlier symbolic accounts of gesture [10] [11] that emphasize cognitive alignment and expressivity.

C. Cultural Contrast and Student Legibility

One particularly compelling implication is that interactions with culturally dissimilar instructors may enhance the visibility of student emotion. When students are exposed to low-context, overtly expressive behavior—typical of non-Japanese teachers—they may externalize their own emotional states more clearly. This could explain the more stable prediction accuracy for these cases, which would suggest that cultural dissimilarity may sometimes reduce ambiguity in affective signaling.

Conversely, Japanese teachers, often operating within a high-context communicative norm, may trigger more individualized and internalized affective interpretations, complicating the ability to predict affect. However, this should not be mistaken for lower emotional engagement. On the contrary, the data indicate that Japanese teachers express a broader emotional range, but with less behavioral regularity. This calls for more context-sensitive modeling approaches.

D. Implications for Culturally Adaptive Affective Computing

The implications of these findings extend beyond education to the broader field of affective computing. As previous literature has emphasized [12] [13], multimodal AI models must

capture both real-time behavior and contextual nuance. The results here suggest that affective computing systems should explicitly model cultural context—not only as a background variable but as a dynamic factor that modulates both expression and perception.

Future systems should adapt to the emotional norms, gestural practices, and display rules of the cultural environment in which they are used. This would allow such systems to move beyond accuracy benchmarks and toward human-aligned interpretability, which would then enhance trust and engagement in human–AI interaction across diverse populations.

VI. Limitations and Future Work

While this study offers novel insights into the relationship between teachers' nonverbal behavior and student emotion in cross-cultural classrooms, several limitations must be acknowledged.

First, the analysis was based on 234 video clips, selected from an initial pool of 302 after excluding segments that were too short, obstructed, or poorly framed. Although this filtering ensured data quality and consistency, the dataset was relatively small and drawn from a limited number of instructors—five Japanese and four non-Japanese teachers. This sample size restricts the generalizability of the findings, as it may not fully capture the diversity of instructional styles or cultural expression within each group. Future research should aim to expand both the number and variety of participating educators.

Second, emotional annotations were limited to a binary emotional scale (positive vs. negative), which simplifies the complexity of affective responses. Additionally, these annotations were not based on actual student reactions but were provided by four Japanese university students who imagined themselves as learners in the classroom. While this approach ensured cultural and procedural consistency, it introduces the possibility of projection bias and limits ecological validity. Future studies should consider using a more diverse annotation pool and explore the use of continuous or multidimensional emotion metrics.

Third, although facial expressions and gesture dynamics were combined in the predictive model, the analysis was conducted at the clip level, without fine-grained time sync between gestures and emotional shifts. Sub-clip or frame-level modeling might reveal more precise temporal cues associated with emotional perception, particularly in the case of micro-expressions or beat gestures.

Fourth, cultural differences were inferred solely from the teachers' identities (Japanese vs. non-Japanese), without explicitly modeling student cultural frameworks or individual differences. While this study assumed cultural homogeneity among students—all of whom were Japanese—the model did not incorporate any student-level variables, such as prior experience with foreign instructors, which may influence how nonverbal cues are perceived.

Fifth, logistic regression was chosen for its interpretability, but it may not capture more complex, nonlinear relationships between multimodal features and affective outcomes. Although suitable for initial exploration, future studies could employ deep learning or transformer-based architectures.

Finally, we acknowledge that, although their evaluations were consistent and culturally grounded, the homogeneity in gender and age of the emotion annotators in this study may introduce bias in emotion perception. Prior literature suggests that emotional sensitivity and interpretation can vary across genders. To enhance the robustness and generalizability of affective labeling, future studies should incorporate a more demographically diverse annotator pool.

Future research should explore more advanced AI architectures to overcome these limitations. For example, transformer-based multimodal models could capture temporal dependencies between teacher gestures and student facial expressions in finer detail. Additionally, evaluating AI-based affect estimation methods in online or hybrid learning environments would help generalize findings beyond face-to-face classrooms. Finally, incorporating student-level variables—such as prior exposure to foreign instructors—into predictive models could enable personalized, culturally adaptive feedback systems for real-time instructional support.

Taken together, these limitations point to several directions for future research. Expanding sample diversity—both among teachers and student annotators—will improve the generalizability of findings, while refining emotional representations through continuous or multidimensional metrics can capture subtler affective nuances. Enhancing temporal resolution to the frame level and incorporating student-level variables (e.g., gender, prior cross-cultural experience) will yield deeper insight into how culture shapes affective dynamics. Moreover, exploring teacher–student emotional co-regulation models in real-time classroom interactions could illuminate bidirectional affective processes. Such advances will not only support the development of adaptive, culturally responsive feedback systems but also enhance student engagement and emotional well-being in diverse educational settings.

VII. Conclusion

This study examined how differences in teachers' gestures in classroom settings based on their culture influence Japanese student observer's emotional responses. By combining gesture kinematics and facial emotion analysis with machine learning techniques, both the form and cultural context of nonverbal behavior were demonstrated to have a significant influence on affective perception and predictability.

While non-Japanese teachers were found to use more overt gestures that were easier for systems to interpret, Japanese teachers were more likely to elicit positive emotional responses, which highlights the role of cultural familiarity in affective engagement. Interestingly, the findings also suggest that cultural contrast may enhance the ability to read student affect. This raises important questions about how intercultural dynamics shape not only teacher expressivity but also student emotional visibility in classroom interactions.

Predictive modeling further revealed that non-Japanese instructors yielded higher emotion prediction accuracy, likely due to greater behavioral regularity, while Japanese teachers' gestures exhibited more contextual variability. This variability

was further reflected in the predictive models used, which achieved higher prediction accuracy for non-Japanese teachers. These differences underscore the need for culturally adaptive models that accommodate variation in signal consistency and contextual modulation.

These findings emphasize that differences in student perception are not merely a matter of gesture intensity or clarity, but are shaped by culturally embedded expectations, display rules, and interaction norms. By quantifying these dynamics and modeling affective responses, this study demonstrates that affective computing for education must be both multimodal and culturally adaptive.

By providing a scalable, quantitative framework for analyzing cross-cultural emotion perception in classrooms, this study contributes not only to research on gesture and education but also to the design of AI systems, that can be used to monitor human emotion. Future studies should build on these findings by incorporating student-level diversity, more discrete emotion labels, and real-time feedback mechanisms to support more inclusive and emotionally intelligent educational environments.

REFERENCES

[1] S. Goldin-Meadow, *Hearing Gesture: How Our Hands Help Us Think*, Cambridge, MA: Harvard University Press, 2003.

[2] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, Chicago, IL: University of Chicago Press, 1992.

[3] A. B. Hostetter and M. W. Alibali, "Visible embodiment: Gestures as simulated action," *Psychon. Bull. Rev.*, vol. 15, no. 3, pp. 495–514, 2008.

[4] A. Kendon, *Gesture: Visible Action as Utterance*, Cambridge, U.K.: Cambridge University Press, 2004.

[5] E. T. Hall, *Beyond Culture*, New York: Anchor Books, 1976.

[6] J. Park, M. Valstar, and M. Pantic, "Cultural differences in nonverbal communication: A study of manual gestures," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 134–147, 2012.

[7] G. S. Lee and A. K. Ng, "Linking gestures and cognition: A cross-cultural analysis of math classrooms," *Int. J. Educ. Res.*, vol. 98, pp. 1–14, 2019.

[8] A. L. Witt and R. Schrodt, "Teacher immediacy and student learning: A meta-analytic review," *Commun. Educ.*, vol. 55, no. 1, pp. 1–22, 2006.

[9] C. Müller, "Forms and uses of the palm up open hand: A case of a gesture family?," *Gesture*, vol. 1, no. 2, pp. 149–171, 2001.

[10] A. Cienki and C. Müller, Eds., *Metaphor and Gesture*, Amsterdam: John Benjamins, 2008.

[11] J. Mittelberg, "Gesture and speech from a semiotic perspective: A fresh look at the relationship between mimetic gestures and language," *Gesture*, vol. 9, no. 1, pp. 1–28, 2009.

[12] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," *IEEE Trans. Affective Comput.*, vol. 5, no. 1, pp. 86–98, 2014.

[13] W. Dai, C. Heffernan, and J. Davis, "Detecting Negative Teacher Affect in the Classroom Using Multimodal Signals," *Proc. Int. Conf. Learning Analytics & Knowledge (LAK)*, pp. 231–240, 2023.

[14] K. D'Mello and S. Graesser, "AutoTutor and affective learning," in *New Perspectives on Affect and Learning Technologies*, Springer, 2011, pp. 35–48.

[15] H. Y. Yoon, S. Kang, and S. Kim, "A non-verbal teaching behaviour analysis for improving pointing out gestures: The case of asynchronous video lecture analysis using deep learning," *J. Comput. Assist. Learn.*, vol. 40, no. 3, Aug. 2024.

[16] I. D. S. Chen, C. M. Yang, S. S. Wu, *et al.*, "Continuous recognition of teachers' hand signals for students with attention deficits," *Algorithms*, vol. 17, no. 7, Art. no. 300, Jul. 2024.

[17] Z. Shou, Y. Huang, D. Li, *et al*., "A student facial expression recognition model based on multi-scale and deep fine-grained feature attention enhancement," *Sensors*, vol. 24, no. 20, Art. no. 6748, Oct. 2024.

[18] R. Singh, E. Ramanujam, and M. N. Babu, "MMSAD – A multi-modal student attentiveness detection in smart education using facial features and landmarks," *J. Ambient Intell. Smart Environ.*, vol. 15, no. 1, Jan. 2025.

[19] S. A. Salloum, K. M. Alomari, A. M. Alfaisal, R. A. Aljanada, and A. Basiouni, "Emotion recognition for enhanced learning: Using AI to detect students' emotions and adjust teaching methods," *Smart Learn. Environ.*, vol. 12, Art. no. 21, Mar. 2025.

[20] O. Kennedy, N. Kuwahara, T. Noble, and C. Fukada, "Accuracy of Spontaneous Dynamic Teacher Emotion Recognition in Japanese College Students," under review, *Frontiers in Education*, 2025.

[21] F. Zhang, V. Bazarevsky, A. Vakunov, et al., "MediaPipe Hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[22] S. Serengil and A. Ozpinar, "LightFace: A hybrid deep face recognition framework," *Data Science and Applications*, vol. 1, no. 1, pp. 1–8, 2021.

[23] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivation and Emotion*, vol. 14, no. 3, pp. 195–214, 1990.

[24] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.