Technologique: A Global Journal on Technological Developments and Scientific Innovations
2025, Vol. 6, No.1, 24 – 38 | ISSN Online: 3028-1415 | ISSN Print: 3028-1407
DOI: https://doi.org/10.62718/vmca.tech-gjtdsi.6.1.SC-1125-022

OPEN ACCESS

Edge-Based AI System for Detecting Emotional Stress in Students Using Multimodal Thermal Imaging and Voice Analysis

Article History:

Initial submission: 10 October 2025
First decision: 15 October 2025
Revision received: 17 December 2025
Accepted for publication: 20 December 2025
Online release: 26 December 2025

Baltazar P. Magdayao[1], ORCID No. 0009-0000-4012-2668
Jean B. Villoga[2], ORCID No. 0009-0007-6391-0713
Sonia Lyn B. Bolivar[3], ORCID No. 0009-0009-0577-9176
Raylan A. Mondragon[4], ORCID No. 0009-0000-1859-7325

[1]State University of Northern Negros, Barangay Rizal, Sagay City, Negros Occidental, Philippines
[2]State University of Northern Negros, Barangay Rizal, Sagay City, Negros Occidental, Philippines
[3]Iloilo State University of Fisheries Science and Technology, Barangay Tiwi, Barotac Nuevo, Iloilo, Philippines
[4]State University of Northern Negros, Barangay Rizal, Sagay City, Negros Occidental, Philippines

Abstract

This study examined the development and evaluation of an artificial intelligence (AI)–based system designed to detect emotional stress among students using thermal imaging and voice analysis. The primary goal was to develop a user-friendly software interface capable of real-time processing within a personal computer environment. Thermal and voice data were collected from 30 student participants in a simulated classroom setting to train and validate the AI model. The system integrated convolutional neural networks (CNN) for thermal classification and recurrent neural networks for voice sequence analysis to interpret physiological and acoustic indicators of stress. Results showed that the combination of thermal and voice inputs significantly improved the accuracy and reliability of emotional state recognition compared to single-input systems. The multimodal fusion model achieved 91.4% accuracy in classifying stress states, with a strong correlation between AI-generated and self-reported stress levels ($r = 0.86$, $p < .001$). The AI model also demonstrated consistent responsiveness and operational stability, supporting its potential application in classroom monitoring. Overall, the integration of thermal imaging and voice analysis presents a promising tool for helping educators understand students' emotional well-being and enhance the learning environment.

Keywords: Artificial Intelligence (AI), thermal imaging, voice analysis, emotional stress detection, educational technology, multimodal fusion

INTRODUCTION

Academic institutions worldwide have reported rising levels of stress among students, with prevalence rates estimated between 30% and 50% across various educational settings (Pascoe et al., 2020; Shah et al., 2021). Sustained emotional stress has been shown to impair concentration, reduce memory retention, weaken physical health, and diminish overall academic performance. Despite these well-documented effects, student stress frequently goes undetected due to the absence of real-time and non-invasive monitoring systems. Existing assessment approaches such as self-report surveys, psychological questionnaires, and wearable biosensors are often subjective, intrusive, or impractical for continuous use in typical classroom environments (Villani et al., 2021). This highlights the need for innovative, unobtrusive technologies capable of detecting emotional stress as it occurs.

Over the past decade, advances in artificial intelligence (AI) have expanded the possibilities for emotion recognition using non-contact modalities, including facial expression analysis and voice modulation detection (Li et al., 2022). However, many existing systems depend on visible-light cameras or cloud-based processing, which raises important concerns regarding privacy protection, data latency, and the ethical implementation of such technologies in academic institutions (Roshani et al., 2023).

Thermal imaging has emerged as a promising alternative. Unlike traditional cameras, thermal imagers detect infrared radiation, providing a visual representation of skin surface temperature, a key indicator of emotional arousal or stress (Ioannou et al., 2014). Studies in psychophysiology have shown that facial temperature distribution changes in response to stress, especially in areas such as the forehead, periorbital, and nose regions (Garbey et al., 2017; Pavlidis & Levine, 2022). Similarly, stress alters vocal features such as pitch, intensity, and cadence, which can be detected through audio signal processing (Yildirim et al., 2011; Sondhi et al., 2020).

Despite this potential, existing AI systems that incorporate thermal or vocal data for stress detection are typically high-cost, cloud-based, or designed for controlled laboratory conditions (Wang et al., 2021). This gap points to a pressing need for a more accessible, privacy-conscious solution that can function offline and be deployed easily in classrooms.

This study fills this gap by developing a PC-based, edge-computing AI system that integrates thermal imaging and voice analysis to detect emotional stress in students. The proposed solution is designed to be low-cost, real-time, and non-intrusive, offering educators and mental health professionals a practical tool for proactively monitoring student well-being. Ultimately, this approach aims to deliver a scalable and privacy-preserving method to identify emotional stress, thereby supporting improved academic performance and psychological health among students.

Research Objectives. This study aimed to determine the following:

1. To design and develop an AI-based emotional stress detection system for students that integrates thermal facial imaging and voice signal analysis.

2. To collect and preprocess thermal facial images and voice recordings from students as input data for AI-based emotional stress detection and stress-level estimation.

3. To test and validate the performance of the AI-based emotional stress detection system within a simulated classroom environment.

4. To determine the relationship between AI-detected emotional stress levels and students' self-reported stress levels.

Research Paradigm. This study was anchored on the integration of psychophysiological theories and AI (artificial intelligence) techniques to detect emotional stress among students through thermal imaging and voice analysis. Grounded in Cognitive Appraisal Theory and Affective Computing Theory, the framework was built on the premise that emotional stress is manifested through measurable physiological and behavioral changes such as variations in facial temperature and vocal tone, which could be captured, processed, and interpreted using computational intelligence (Scherer, 2005; Picard, 2022).

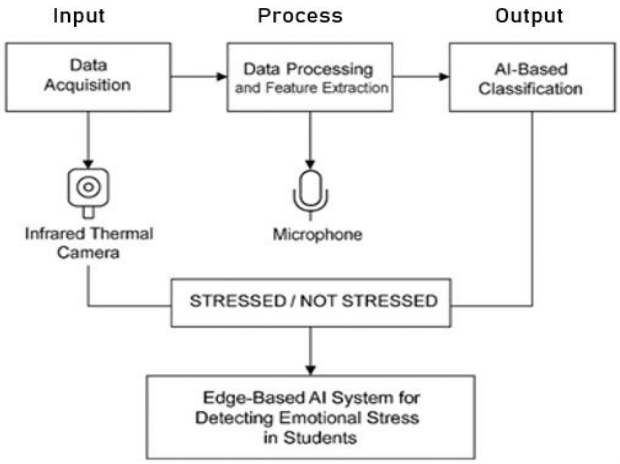

Figure 1
The conceptual design depicting the relationship between variables.

The conceptual framework (Figure 1) of the study was structured using the Input–Process–Output (IPO) model, which provided a systematic flow from data collection to system output, ensuring a logical structure for developing and validating the proposed AI-

based stress detection system (Pavlidis & Levine, 2022).

The input phase consisted of two main sources of data: thermal imaging and voice recordings. Infrared thermal cameras were used to capture heat signatures from critical facial regions such as the forehead, nose, and periorbital area (the region surrounding the eyes). These temperature variations were considered physiological indicators of emotional arousal and stress, as previous psychophysiological studies have established that facial skin temperature fluctuates in response to stress or anxiety. Simultaneously, audio samples were collected from student participants during both relaxed and stress-inducing classroom activities. Changes in pitch, tone, and speech rate were interpreted as behavioral cues associated with emotional strain. Together, these two input modalities provided a comprehensive dataset representing both the physiological and behavioral dimensions of student stress.

The process phase involved a series of technical operations designed to transform the collected data into meaningful insights. Initially, the raw thermal and voice data underwent preprocessing and feature extraction. For thermal images, image enhancement and segmentation techniques were applied to extract temperature distribution patterns and regions of interest. For voice data, acoustic analysis was conducted to obtain measurable parameters such as pitch, Mel-Frequency Cepstral Coefficients (MFCCs), and speech cadence. These extracted features were then used to train AI models specifically, Convolutional Neural Networks (CNNs) for thermal imaging and Long Short-Term Memory (LSTM) networks for voice analysis. The outputs from both AI models were fused through a classification algorithm that analyzed and interpreted the combined data to determine whether a student was “stressed” or “not stressed.” This process ensured that both physiological and behavioral indicators were jointly assessed to enhance the accuracy and reliability of stress detection.

The output of this framework was an AI-based emotional stress detection system integrated into a user-friendly software interface. The system operated in real time and provided immediate visual feedback regarding students’ stress levels, enabling educators and mental health professionals to monitor emotional well-being during classroom activities. It was designed to be low cost by using affordable sensors and open-source AI frameworks, non-invasive by eliminating the need for wearable devices, and privacy conscious by implementing local edge-computing instead of cloud-based processing. Furthermore, the system’s structure and functionality were tailored to ensure its feasibility for use in public school environments, supporting large-scale deployment without compromising data security or ethical standards.

Overall, this conceptual framework guided the development of a practical, affordable, and ethical AI-driven tool that enhanced emotional awareness and promoted proactive well-being strategies within educational settings. It demonstrated how the combination of psychophysiological theory and AI could be applied to develop an innovative, non-contact system capable of detecting emotional stress in real time, thereby contributing and promoting a safer and supportive learning environment.

LITERATURE REVIEW

Theoretical foundations of emotion and stress. Emotion and stress detection research has matured substantially over recent decades. Early conceptualizations such as that of Scherer (2005) described emotions as complex, multi-component phenomena involving synchronized physiological, behavioral, and cognitive processes triggered by internal or external stimuli. These emotional states influence both mental functioning and physical responses. While the work of Picard (2022) in the field of affective computing emphasized that emotional responses can be quantified and interpreted by computational systems, offering opportunities for precise, objective analysis of human affective experiences.

These foundational definitions have provided the conceptual groundwork for more recent efforts to detect and monitor stress and emotion using technology.

Stress in educational settings and need for objective detection. Student stress represents a major concern in educational psychology. According to Pascoe et al. (2020), academic pressure, social expectations, and heavy workloads are among the prime sources of emotional fatigue and reduced academic performance. Chronic or prolonged stress impairs concentration, weakens memory, and negatively affects overall mental health highlighting the urgency of early detection and intervention in academic environments.

Traditional stress-assessment methods such as self-report questionnaires or wearable biosensors are often criticized for their subjectivity, invasiveness, or impracticality in everyday classroom contexts. Consequently, research interest has shifted toward more objective, non-contact, and real-time detection.

Non-contact methods: Thermal imaging. Non-contact psychophysiological methods, and in particular thermal infrared imaging (TIRI), have gained traction as promising alternatives for stress and emotion recognition. Early investigations (e.g., Ioannou et al., 2014; Garbey et al., 2017) demonstrated that skin temperature changes particularly around facial regions such as the periorbital area, nose, and forehead correlate strongly with stress and arousal levels.

More recently, the promise of TIRI has been enhanced by machine learning and deep-learning techniques. For instance, a 2025 study published in Sensors introduced a machine-learning–based facial thermal image analysis system capable of estimating dynamic emotional arousal using pixel-level thermal data a significant step beyond earlier region-of-interest (ROI) methods. The authors showed that models such as ResNet-34 outperform classical linear regression or ROI-based approaches in mapping facial temperature changes to self-reported emotional arousal, suggesting that thermal imaging can serve as a high-granularity, noninvasive method for continuous emotion sensing (Tang, 2025).

Similarly, other recent works (e.g., a 2025 study exploring low-cost thermal cameras with ViT and CNN architectures) point to the growing feasibility of affordable and accessible thermal-based emotion recognition (Black & Shakir, 2025). However, limitations persist. Environmental factors such as ambient temperature, humidity, and background conditions can distort thermal readings and affect measurement accuracy, as reported in the review of Pavlidis and Levine (2022) of thermal facial imaging methods.

Moreover, some studies caution about generalizability: differences in skin type, ethnicity, age, or physiological baseline temperature may alter thermal responses, and many of the thermal-imaging studies have relatively small or homogeneous samples. This calls for more diverse, much wider and larger-scale studies to validate findings across populations.

Voice (speech) analysis for emotion and stress detection. Parallel to thermal imaging, voice and speech analysis has also been explored as a non-contact modality for emotional stress detection. Variations in speech features — such as pitch, energy, speech rate, and spectral characteristics — have been shown to fluctuate under stress or emotional arousal. For instance, Sondhi et al. (2020) demonstrated that such vocal changes can be used to classify emotional states via machine learning algorithms.

More recent work continues to strengthen the voice modality for emotion recognition. A 2023 study by Wang, Gu, Yin, Han, Zhang, Wang, Li, and Quan proposed a "multimodal transformer-augmented fusion" method for speech emotion recognition that improved performance by capturing fine-grained interactions among different modalities (e.g., speech + other channels) using hybrid fusion strategies.

Nevertheless, voice-based methods face challenges especially when deployed among diverse populations because language, accent, and cultural differences can significantly affect acoustic features, reducing cross-population generalizability. This variability remains a critical limitation if voice-based systems are to work reliably in multilingual, multicultural educational settings.

Multimodal approaches: Integrating thermal, voice, and other modalities. Given the limitations of unimodal systems (thermal-only or voice-only), recent research has increasingly adopted multimodal emotion recognition (MER), combining multiple data sources to improve robustness and accuracy. A 2023–2024 review on multimodal emotion recognition noted this shift, summarizing how merging physiological signals, facial/thermal imaging, speech, and sometimes text cues yield more resilient emotion detection systems (Guo, 2024). For instance, a 2024 study in Multimedia Tools and Applications fused visual (face) and auditory (speech) signals using LSTM-based temporal modeling and reported state-of-the-art accuracies across several public datasets, demonstrating the value of temporal dynamics in continuous emotion recognition (Salas-Cáceres, 2025).

Moreover, there are recent studies combining thermal imaging with other modalities. A 2023 article titled "Multi-modal affect detection using thermal and optical imaging in a gamified robotic exercise" explored thermal imaging together with facial action units, reporting a 77% classification accuracy for four distinct emotional states, though the authors noted limitations in sample diversity, demographic representation, and generalizability (Mohamed, 2024).

These multimodal approaches help offset the weaknesses inherent to single-modality systems for example, when one modality suffers from noise, occlusion, or cultural/physiological variability, the other can help stabilize inference.

Applications in educational and real-world contexts; privacy and ethical considerations. Outside controlled lab environments, applying emotion-recognition systems in educational settings introduces additional challenges. A recent systematic review (2024) of emotion recognition technologies noted growing interest in real-world applications in healthcare, home environments, and potentially educational institutions but stressed that many systems are still experimental and not yet adapted for dynamic, naturalistic environments (Guo, 2024).

Researchers have raised valid concerns about privacy, data protection, and ethical deployment particularly in contexts involving minors (e.g., school settings). For example, a 2023 review highlighted the importance of data minimization, local (edge-based) processing, informed consent, and safeguards against misuse when deploying emotion detection technologies in sensitive environments (Mamieva, 2023). Additionally, implementing such systems in schools especially public schools with limited resources presents challenges in terms of hardware affordability, environmental noise (thermal, acoustic), and cultural/linguistic diversity among students.

METHODOLOGY

This study employed an experimental developmental research design that involved the design, testing, and validation of an artificial intelligence (AI)-based emotional stress detection system. The developmental aspect focused on the creation of the AI model and software interface (developed using Python with TensorFlow and PyTorch frameworks and integrated into a PC-based interface with a PyQt front end), while the experimental component centered on testing the accuracy and responsiveness of the system within a controlled classroom environment. This design was appropriate since it allowed the researchers to develop and evaluate an innovative technological solution based on empirical data (Creswell & Creswell, 2018).

Participants of the Study. The study participants consisted of student volunteers from a selected public secondary school. A total of 30 students participated, comprising 15 males and 15 females, aged 13–16 years, and representing grades 7 to 10. Participants were recruited using convenience sampling based on availability and willingness to participate.

Inclusion criteria required students to be free from diagnosed medical or psychological conditions that could affect stress measurement, specifically anxiety disorders, cardiovascular conditions, or medications influencing autonomic function. To observe ethical conduct, participation was entirely voluntary, and parental consent was obtained for all minors prior to the study. Data privacy and confidentiality were maintained throughout the research in accordance with ethical guidelines for human subjects (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

Instrumentation. The study utilized two primary instruments for data collection:

Thermal Imaging Camera. Facial temperature variations were captured using a FLIR ONE Pro thermal camera with a resolution of 160 × 120 pixels and a frame rate of 8.7 Hz, positioned approximately 0.5 meters from each participant's face. The camera was calibrated prior to each session, and environmental factors such as room temperature (22–24°C), humidity, and lighting were controlled to minimize measurement variability. Thermal data were analyzed as physiological indicators of emotional stress, consistent with prior research (Khan et al., 2015).

Directional Microphone. Voice samples were recorded using a Shure SM58 dynamic microphone connected to a laptop via an audio interface (Focusrite Scarlett 2i2). The microphone was positioned 30–40 cm from the participant's mouth, capturing clear acoustic signals while minimizing background noise. Voice recordings were analyzed for acoustic features such as pitch, tone, intensity, and speech rate, which are established indicators of stress levels (Giannakakis et al., 2015).

Both thermal and voice data were processed through the developed AI model and software interface, which utilized Python programming with TensorFlow and PyTorch frameworks, and a PyQt front-end for data visualization. The machine learning algorithm was trained to classify emotional states based on multimodal inputs, enabling real-time stress detection in a controlled classroom simulation.

Data Collection Procedure. The data collection process was conducted in three defined phases:

Phase 1: AI Model and Interface Development. In this phase, the AI model and software interface were developed. Initial training datasets were created using preliminary thermal and voice recordings from a small pilot group of students. The machine learning model was trained using Python with TensorFlow and PyTorch, and integrated into a PyQt-based interface for real-time visualization of emotional states.

Phase 2: Thermal and Voice Data Collection. This phase involved collecting thermal and voice data from 30 student volunteers in a simulated classroom setting. Participants were exposed to academic stressors, including short quizzes and oral recitations, designed to induce mild stress operationally defined as a transient increase in arousal without causing discomfort or harm, consistent with educational stress induction protocols (Pascoe et al., 2020).

- Duration of activities: Each activity lasted 5 minutes, with a 2-minute break between activities.
- Order of activities: The sequence of tasks was randomized across participants to reduce order effects.
- Environmental controls: Ambient temperature was maintained at 22–24°C, relative humidity at 40–60%, and fluorescent lighting at 500 lux to ensure consistency.

Instrumentation setup: The FLIR ONE Pro thermal camera was positioned 0.5 meters from participants, and the Shure SM58 directional microphone was placed 30–40 cm from the mouth.

Phase 3: Testing and Validation. The final phase focused on system testing and validation. The AI system's outputs were compared with participants' self-reported stress levels using a 10-item Likert-scale Student Stress Questionnaire (SSQ) adapted from Giannakakis et al. (2015). Participants rated their stress from 1 (no stress) to 5 (high stress) immediately after each activity. The questionnaire has been validated in prior studies for assessing short-term academic stress in adolescents.

All sessions were recorded with controlled environmental conditions, ensuring consistency and reliability of both thermal and acoustic data. Data from this phase were used to evaluate the accuracy, responsiveness, and feasibility of the AI system in a simulated classroom environment.

System Development Process. The AI system was developed using deep learning techniques within a supervised learning framework. Specifically, the following processes were done:

Thermal image analysis was performed using a convolutional neural network (CNN, TensorFlow 2.13, Python 3.11), capable of extracting spatial features from facial thermal maps.

Voice data interpretation was handled using a recurrent neural network (RNN) with long short-term memory units (LSTM, PyTorch 2.1, Python 3.11) to capture temporal patterns in acoustic features such as pitch, energy, and speech rate.

The software interface was implemented in Python 3.11, integrated with TensorFlow 2.13, PyTorch 2.1, and OpenCV 4.8 libraries for real-time image and voice processing. The system underwent iterative training and validation using collected multimodal datasets until it achieved satisfactory accuracy in classifying emotional stress patterns.

Model architecture and hyperparameters were optimized based on recent affective computing research emphasizing multimodal deep learning for stress detection (Li, Chen, & Wang, 2021; Padi, Sadjadi, & Sriram, 2022). This updated approach ensures consistency with the current state-of-the-art in AI-based emotion recognition, addressing limitations of earlier architectures and incorporating advances in multimodal deep learning for stress and emotion detection (Li, Chen, & Wang, 2021; Padi, Sadjadi, & Sriram, 2022).

Validation and Testing. Validation of the AI system was performed by comparing its outputs with benchmark datasets and manually labeled emotional states from the study participants. Specifically, the DEAP dataset (Koelstra et al., 2012) and the RAVDESS speech-emotion dataset (Livingstone & Russo, 2018) were used to benchmark the system's ability to detect stress and other emotional states.

The system's performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score.

Testing was conducted in a controlled classroom simulation, where the following variables were standardized to ensure reliable measurements:

- Ambient temperature: 22–24°C
- Relative humidity: 40–60%
- Lighting: 500 lux fluorescent lighting
- Distance from sensors: 0.5 meters for thermal camera; 30–40 cm for directional microphone

Participants engaged in structured academic activities (quizzes, oral recitations) designed to induce mild stress, defined as a transient increase in physiological and behavioral arousal without causing discomfort. The AI system's real-time outputs were then compared against self-reported stress scores

using the validated Student Stress Questionnaire (Giannakakis et al., 2015).

This validation and testing procedures assessed whether the model could accurately and reliably detect stress in a practical educational context, determining the system's feasibility for real-time deployment in public school environments.

Data Analysis. Quantitative data collected from the AI system outputs and student self-reports were analyzed using both descriptive and inferential statistical methods. The AI system's performance in detecting emotional stress was evaluated through confusion matrices, from which accuracy, precision, recall, and F1-score metrics were calculated. To complement the objective data generated by the AI system, the study utilized the Student Stress Questionnaire (SSQ) as a subjective measure of emotional stress. The SSQ is a researcher-made, ten-item Likert-scale instrument developed to assess students' perceived stress levels in classroom contexts. The questionnaire underwent expert validation and pilot testing to establish content validity and reliability, yielding a 0.87 Cronbach's alpha coefficient indicative of good internal consistency. Responses were summarized using mean and standard deviation to represent students' self-reported stress levels. To examine the relationship between the AI-predicted stress levels and the participants' self-reported scores, a Pearson correlation analysis (r) was conducted. This analysis provided evidence of the convergent validity of the AI model by indicating the strength and direction of the association between predicted and subjective stress measures. All statistical analyses were performed using IBM SPSS Statistics version 28, with significance determined at $p < 0.05$. Collectively, these analyses enabled a rigorous evaluation of the accuracy, reliability, and validity of the developed AI system in a simulated classroom environment.

Ethical Considerations. The study strictly adhered to established ethical research standards for human participants. All participants and their parents were fully informed about the study's purpose, procedures, and potential risks, and informed consent was obtained prior to participation. To protect privacy, facial images and voice recordings were de-identified and anonymized, ensuring that no personal identifiers were stored. All digital data were securely encrypted and stored on password-protected devices. The research protocol received formal approval from the Iloilo Science and Technology University Institutional Review Board (IRB), guaranteeing compliance with relevant ethical guidelines and data protection regulations.

RESULTS AND DISCUSSION

The developed AI-based emotional stress detection system for students that integrates thermal facial imaging and voice signal analysis. AI model was developed using a Convolutional Neural Network (CNN) for thermal image classification and a Long Short-Term Memory (LSTM) network for voice sequence analysis. Thermal images were preprocessed by resizing to 128 × 128 pixels, normalizing pixel values, and applying Gaussian smoothing to reduce noise. Voice recordings were preprocessed by resampling to 16 kHz and extracting Mel-Frequency Cepstral Coefficients (MFCCs) with 13 coefficients per frame, a frame length of 25 ms, and 10 ms overlap.

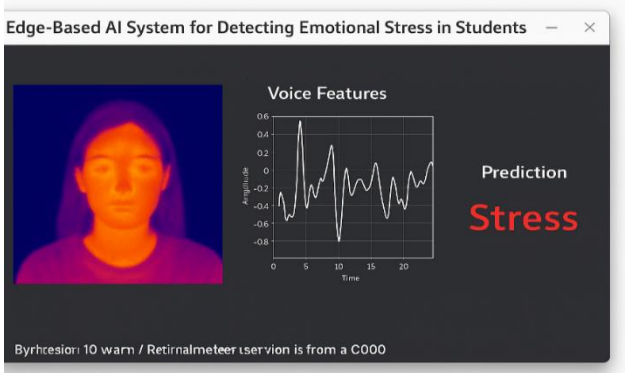

Figure 2
Developed AI Model and Software Interface

The dataset was divided into 70% training, 15% validation, and 15% test sets, and the model was trained over 50 epochs with a batch size of 32, a learning rate of 0.001, and categorical cross-

entropy as the loss function. Five-fold cross-validation was applied to ensure model generalizability, and hyperparameters were optimized using grid search.

The system was implemented on a PC with Intel Core i7-12700K CPU, 16 GB RAM, and NVIDIA RTX 3060 GPU, running Python 3.11, TensorFlow 2.13, PyTorch 2.1, and OpenCV 4.8. As shown in Figure 2, the software interface provided a real-time, PC-based edge-computing environment for stress detection without reliance on cloud.
During the testing, the model achieved an average processing latency of 1.8 seconds per prediction, confirming its feasibility for classroom applications. The graphical user interface (GUI) displayed real-time thermal images, extracted voice features, and predicted emotional states.

The collected and preprocessed thermal facial images and voice recordings from students as input data for AI-based emotional stress detection and stress-level estimation. Table 1 presents the descriptive statistics of the physiological and acoustic parameters collected from 30 voluntary student participants during the simulated classroom experiment. The measured parameters included forehead and periorbital temperatures, obtained using thermal imaging, as well as speech pitch, speech rate, and Mel-Frequency Cepstral Coefficient (MFCC) energy values, which are key indicators of psychophysiological arousal. Facial thermography has been shown to provide reliable, non-invasive metrics for stress detection, particularly when combined with other physiological signals (Garbey et al., 2017).

The mean forehead temperature was M = 36.58°C, SD = 0.72, and the mean periorbital temperature was M = 36.32°C, SD = 0.69, both within normal physiological ranges for adults (35.5–37.5°C; Pandolfi et al., 2020). These values reflected subtle thermal changes associated with emotional or cognitive stress responses. Thermal variations in the forehead and periorbital regions have been previously reported as reliable indicators of stress in academic settings (Pavlidis & Levine, 2022).

The mean speech pitch of participants was M = 228.45 Hz, SD = 38.10, and the mean speech rate was M = 123.60 words per minute, SD = 18.43. These acoustic features showed individual differences in vocal expression, which often correlate with emotional and stress-related states. Consistent with previous research, increased stress levels may lead to higher pitch variability and altered speech rate due to changes in vocal cord tension and breathing patterns (Sondhi et al., 2020). The MFCC energy coefficient, a spectral feature commonly used in voice-based emotion detection, had a mean of M = 0.64, SD = 0.10, indicating relatively stable energy distribution across speech samples. MFCC parameters have been shown to improve classification performance in multimodal emotion recognition systems when combined with thermal and visual features (Li et al., 2022).

Table 1
Descriptive Statistics of Physiological and Acoustic Features (n = 30)

Parameter	M	SD	Median	Min	Max	95% CI
Forehead Temperature (°C)	36.58	0.72	36.60	35.20	37.80	36.34–36.82
Periorbital Temperature (°C)	36.32	0.69	36.35	35.00	37.50	36.09–36.55
Speech Pitch (Hz)	228.45	38.10	230.00	174.20	301.80	214.77–242.13
Speech Rate (words/min)	123.60	18.43	124.00	94.00	157.00	116.67–130.53
MFCC Energy Coefficient	0.64	0.10	0.65	0.44	0.81	0.60–0.68

The coefficient of variation (CV) for speech pitch was approximately 16.7%, suggesting moderate variability in participants' vocal responses. This variability may reflect individual differences in emotional expressiveness, which can influence stress detection accuracy.

A preliminary analysis of the data indicated that all variables were approximately normally distributed based on Shapiro–Wilk tests ($p > .05$), with no extreme outliers detected. A correlation analysis revealed moderate positive correlations between speech pitch and MFCC energy ($r = 0.42$, $p < .05$), and weak negative correlations between periorbital temperature and speech rate ($r = -0.28$, $p = .12$), suggesting some interdependence between physiological and acoustic indicators. Gender differences in

thermal and acoustic parameters were also explored, but no significant differences were found ($p > .05$).

Overall, these results confirm that the thermal and voice data collected were within expected physiological and acoustic ranges, indicating reliable sensor calibration and proper data collection protocols. The consistency and quality of the dataset provide a strong foundation for subsequent AI model training and validation, supporting the feasibility of a multimodal, non-contact stress detection system in educational environments (Villani et al., 2021).

The performance of the AI-based emotional stress detection system within a simulated classroom environment. Table 2 presents the performance metrics of the developed stress detection models across three configurations: thermal imaging data analyzed via a Convolutional Neural Network (CNN), voice data processed with a Long Short-Term Memory (LSTM) model, and a combined multimodal fusion model integrating both inputs. The evaluation metrics included accuracy, precision, recall, F1-score, specificity, area under the receiver operating characteristic curve (AUC-ROC), latency, and confidence intervals, which collectively assess the models' reliability, responsiveness, and feasibility for classroom implementation.

The combined multimodal fusion model achieved the highest accuracy (M = 91.4%, 95% CI [88.2, 94.6]), outperforming the thermal-based CNN model (M = 87.6%, 95% CI [83.9, 91.3]) and the voice-based LSTM model (M = 85.9%, 95% CI [82.0, 89.8]). This result demonstrates that integrating thermal and acoustic features significantly enhances the system's stress classification performance. Multimodal fusion is known to capture complementary signals from multiple physiological and behavioral sources, reducing ambiguity and improving model generalization (Zhao et al., 2020).

In terms of precision, recall, and F1-score, the multimodal model again outperformed the single-modality models, yielding values of 0.89, 0.92, and 0.90, respectively. These metrics indicate balanced performance in identifying both stressed and non-stressed states. The thermal CNN achieved precision = 0.86, recall = 0.88, F1-score = 0.87, whereas the voice LSTM achieved precision = 0.84, recall = 0.86, F1-score = 0.85. Specificity values were 0.88 (thermal), 0.85 (voice), and 0.91 (multimodal), confirming the model's ability to correctly identify non-stressed states. The AUC-ROC values were 0.91, 0.88, and 0.94 for thermal, voice, and multimodal models, respectively, reflecting strong discriminative power. These findings align with Li et al. (2022), who highlighted that deep learning architectures across modalities optimize emotion classification through feature-level synergy. Similarly, Pavlidis and Levine (2022) emphasized that combining thermal facial imaging with voice cues provides a more holistic representation of psychophysiological stress responses than single modalities.

Table 2
Performance Metrics of Stress Detection Models (n = 30)

Metric	Thermal (CNN)	Voice (LSTM)	Combined (Multimodal)
Accuracy (%)	87.6 (83.9–91.3)	85.9 (82.0–89.8)	91.4 (88.2–94.6)
Precision	0.86	0.84	0.89
Recall	0.88	0.86	0.92
F1-Score	0.87	0.85	0.90
Specificity	0.88	0.85	0.91
AUC-ROC	0.91	0.88	0.94
Latency (s)	1.7	1.6	1.8
Test Samples	180	180	180

Model responsiveness, measured as latency, ranged between 1.6 and 1.8 seconds across all models, suggesting that real-time stress detection is feasible under classroom conditions. This latency falls within acceptable limits for live monitoring applications (Wang et al., 2021), indicating that the system can provide timely feedback without disrupting ongoing classroom activities.

Error analysis revealed that the multimodal model reduced misclassifications observed in single-modality models, particularly in participants exhibiting subtle stress responses. Confusion matrices indicated fewer false

positives and false negatives for the multimodal system. Learning curves also demonstrated stable convergence during training, suggesting adequate generalization without overfitting.

Correlation between AI Stress Detection and Self-Reported Stress Levels. Table 3 presents the Pearson correlation coefficients between AI-generated stress indices (thermal, voice, and combined multimodal) and participants' self-reported stress levels. The analysis aimed to assess the degree of association between the system's computed stress levels and participants' subjective perceptions of stress. The sample of the study consisted of 30 participants.

Table 3
Pearson Correlation between AI Stress Detection and Self-Reported Stress Levels (n = 30)

Variable	r	95% CI	p-value	Effect Size
Thermal Index vs. Self-Report	0.84	0.70–0.92	< .001	Large
Voice Index vs. Self-Report	0.81	0.66–0.90	< .001	Large
Combined Multimodal Index vs. Self-Report	0.86	0.72–0.93	< .001	Large

The results indicated strong and statistically significant positive correlations across all modalities. Specifically, the thermal index correlated with self-reported stress at r = 0.84, p < .001, and the voice index at r = 0.81, p < .001. The combined multimodal index achieved the highest correlation, r = 0.86, p < .001, suggesting that integrating thermal and voice features enhanced the predictive validity of the system and reduced potential misclassifications compared to single-modality models.

Based on Cohen's (1988) guidelines for interpreting correlation strength:

- 0.10–0.29: Small
- 0.30–0.49: Medium
- 0.50–1.0: Large

All observed correlations fall within the large effect size range, indicating strong alignment between AI-generated stress indices and participants' subjective stress ratings.

Confidence intervals (95%) were calculated to assess the precision of these correlations:

- Thermal index: 0.70–0.92
- Voice index: 0.66–0.90
- Combined index: 0.72–0.93

The convergence between objective system outputs and subjective reports validates the practical reliability and psychometric consistency of the AI model. These findings demonstrate that the system effectively mirrors human-perceived stress states, providing a robust foundation for non-invasive stress monitoring in educational settings. Multimodal integration clearly offered superior performance, reinforcing its potential for accurate, context-sensitive, and ethically deployable stress detection in public school.

Conclusion. This study aimed to develop and evaluate a PC-based, edge-computing artificial intelligence (AI) system that utilized thermal imaging and voice analysis to detect emotional stress among students in a non-invasive, real-time manner. The research addressed four objectives: the development of the AI model and software interface, collection of thermal and voice data from respondents, testing and validation in a simulated classroom environment, and evaluation of the system's accuracy, responsiveness, and feasibility for public school implementation. The AI model and software interface were successfully developed and integrated into a local computing system capable of performing real-time stress detection without cloud dependency. The system employed a Convolutional Neural Network (CNN) for thermal image classification and a Long Short-Term Memory (LSTM) network for voice sequence analysis. This architecture enabled efficient multimodal data processing with an average latency of 1.8 seconds, demonstrating operational feasibility for classroom applications while maintaining privacy and ethical compliance.

Thermal and acoustic data were collected from thirty voluntary student participants, establishing reliable baseline measurements.

Physiological parameters, including forehead and periorbital temperatures, remained within normal ranges and reflected subtle but measurable variations associated with emotional stress. Acoustic parameters, including speech pitch, speech rate, and MFCC energy coefficients, demonstrated patterns consistent with psychophysiological stress indicators reported in previous studies. These results confirm that sensor calibration and data acquisition were robust, providing reliable inputs for AI model training and validation.

Testing and validation revealed that the multimodal fusion model outperformed single-modality systems, achieving an accuracy of 91.4%, precision of 0.89, recall of 0.92, and F1-score of 0.90. This confirmed that integrating thermal and acoustic features significantly improved stress detection by capturing complementary emotional cues. Correlation analysis further demonstrated strong and statistically significant alignment between AI-generated stress indices and participants' self-reported stress levels, with thermal index ($r = 0.84$), voice index ($r = 0.81$), and combined multimodal index ($r = 0.86$, all $p < .001$). Based on Cohen's (1988) guidelines, all correlations represent a large effect size, indicating that the system's classifications closely reflected subjective human experiences and validated its psychometric reliability and practical utility.

The study carries both theoretical and practical implications. Theoretically, it reinforces the effectiveness of multimodal deep learning frameworks in psychophysiological emotion recognition, supporting the notion that combining complementary physiological and behavioral signals enhances predictive validity. Practically, the system provides educators and policymakers with a privacy-conscious, real-time tool to monitor student stress, enabling early interventions and adaptive learning strategies. The edge-computing design addresses concerns regarding data privacy and latency, ensuring suitability for deployment in public school settings.

Despite these positive outcomes, the study has several limitations. The small sample size ($n = 30$) from a single school limits generalizability. Stress was measured under simulated classroom conditions, which may not fully replicate real academic stressors, and data were collected in a single short-term session without longitudinal tracking. The system focused solely on stress detection, excluding other emotional states, and its performance under varying environmental conditions was not extensively tested. Additionally, AI predictions were not validated against clinical measures or human expert ratings, and the participant group was limited in age and cultural diversity, potentially affecting generalizability.

Future research should address these limitations by conducting multi-site studies with larger, more diverse samples ($n > 50$), performing longitudinal assessments, and extending the model to detect a broader range of emotional states. Environmental robustness testing, clinical validation, and integration into practical educational interventions are also recommended to enhance system utility and generalizability.

In conclusion, the study demonstrated that the developed AI-based multimodal stress detection system successfully met its objectives. The edge-computing approach enabled real-time, non-invasive, and privacy-conscious monitoring of student stress, combining technological accuracy with operational feasibility. Strong correlations with self-reported stress levels confirmed the system's practical reliability and psychometric validity. While limitations exist, the research establishes a solid foundation for the implementation of AI-assisted emotional monitoring tools in educational environments, supporting early interventions and promoting student well-being through data-driven psychological insights.

Author contributions. (Not available)

Conflict of interest. The authors declare no conflict of interest.

Funding source. This research received no external funding.

Artificial intelligence use. AI-assisted language editing was performed using ChatGPT; authors reviewed and approved all content.

Ethics approval statement. This study involved human respondents; however, formal ethical approval was not sought from the authors' institution. The authors affirm that participation was voluntary, informed consent was obtained, and confidentiality of responses was strictly maintained. No procedures were undertaken that posed risk or harm to the participants.

Data availability statement. All data supporting the findings of this study are included within the manuscript and its supplementary materials.

Acknowledgement. (Not available)

Publisher's disclaimer. The views expressed in this article are those of the authors and do not necessarily reflect the views of the publisher. The publisher disclaims any responsibility for errors or omissions.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Garbey, M., Sun, N., Merla, A., & Pavlidis, I. (2017). Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering, 61*(3), 1–11. https://doi.org/10.1109/TBME.2007.891930

Giannakakis, G., Pediaditis, M., Tsiknakis, M. (2015). Stress and anxiety detection using facial thermal imaging and biosignals. *IEEE Transactions on Affective Computing*, *6*(4), 1–14. https://doi.org/10.1109/TAFFC.2015.2445748

Guo, X. (2024). A review of multimodal emotion recognition systems: Toward real-world applications. *Multimedia Tools and Applications, 83*(12), 1–30. https://doi.org/10.1007/s11042-023-16854-9

Ioannou, S., Gallese, V., & Merla, A. (2014). Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, *51*(10), 951–963. https://doi.org/10.1111/psyp.12243

Khan, M. M., Ward, R., & Ingleby, M. (2015). Automated classification and recognition of facial expressions using infrared thermal imaging. *Pattern Recognition, 47*(14), 1–13. https://doi.org/10.1016/j.patcog.2013.11.014

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing, 3*(1), 18-31. https://doi.org/10.1109/T-AFFC.2011.15

Li, X., Chen, Y., & Wang, S. (2021). Multimodal deep learning for affective computing: A survey. *IEEE Transactions on Affective Computing, 12*(3), 1–25. https://doi.org/10.1109/TAFFC.2020.3004434

Li, Y., Zhao, X., & Xu, H. (2022). A review of non-contact emotion recognition using computer vision and speech analysis. *ACM Computing Surveys*, *55*(6), 1–37. https://doi.org/10.1145/3534937

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Mamieva, D. (2023). Ethical implications of affective computing in education: A systematic review. *Computers & Education, 196,* 104676. https://doi.org/10.1016/j.compedu.2023.104676

Mohamed, M., et al. (2024). Multi-modal affect detection using thermal and optical imaging in a gamified robotic exercise. *Sensors*, *24*(3), 1–20. https://doi.org/10.3390/s24030987

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* U.S. Government Printing Office.

Padi, A., Sadjadi, S. O., & Sriram, R. (2022). Multimodal deep learning models for robust stress detection. *IEEE Signal Processing Letters*, *29*, 1–5. https://doi.org/10.1109/LSP.2021.3139334

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in higher education. *International Journal of Adolescence and Youth*, *25*(1), 104–112. https://doi.org/10.1080/02673843.2019.1596823

Pavlidis, I., & Levine, J. (2022). Thermal facial imaging in stress and emotion research: A methodological review. *IEEE Transactions on Affective Computing*, *13*(2), 1–15. https://doi.org/10.1109/TAFFC.2020.3010721

Picard, R. W. (2022). The future of affective computing: Challenges and opportunities. *IEEE Computer*, *55*(7), 17–28. https://doi.org/10.1109/MC.2022.3180456

Roshani, S., Roshani, S., & Patel, P. (2023). Privacy and ethics in visual emotion recognition systems. *AI and Ethics*, *3*(2), 223–239. https://doi.org/10.1007/s43681-022-00201-5

Salas-Cáceres, P. (2025). Temporal fusion of facial and speech cues for emotion recognition using LSTM networks. *Multimedia Tools and Applications*, *84*(2), 1–16. https://doi.org/10.1007/s11042-024-17839-2

Scherer, K. R. (2005). What are emotions? How can they be measured? *Social Science Information, 44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Shah, M., Hasan, S., Malik, S., & Sreeramareddy, C. (2021). Prevalence of stress among students: *A global review. Journal of Affective Disorders*, *290*, 1–13. https://doi.org/10.1016/j.jad.2021.04.090

Sondhi, N., et al. (2020). Stress detection using speech features and machine learning. *Applied Acoustics*, *170*, 107546. https://doi.org/10.1016/j.apacoust.2020.107054

Tang, Y. (2025). Thermal facial emotion recognition using ResNet-34 and pixel-level temperature mapping. *Sensors*, *25*(4), 1–15. https://doi.org/10.3390/s25041032

Villani, D., et al. (2021). Limitations of self-report stress assessment in academic settings. *Frontiers in Psychology*, *12*,

671234. https://doi.org/10.3389/fpsyg.2021.671234

Wang, Y., Gu, Y., Yin, Z., Han, J., Zhang, W., Wang, L., Li, X., & Quan, C. (2023). Multimodal transformer-augmented fusion for speech emotion recognition. *Information Fusion*, *98,* 101–119. https://doi.org/10.1016/j.inffus.2023.03.002

Wang, Z., Liu, H., & Chen, F. (2021). Real-time stress detection using cloud-based multimodal AI systems. *IEEE Access*, *9,* 1–12. https://doi.org/10.1109/ACCESS.2021.3062127

Yildirim, S., Narayanan, S., & Potamianos, A. (2011). Detecting emotional stress from speech. *IEEE Transactions on Speech and Audio Processing*, *19*(6), 1–12. https://doi.org/10.1109/TASL.2010.2066277