Multimodal Emotion Recognition in Educational Environments Using Artificial Intelligence and Forensic Psychology

[1]Amit Ghosh and [2]Dr. Santanu Sikdar
[1]Research Scholar, P.K University, Shivpuri, Madhya Pradesh, India
[2]Professor, P.K University, Shivpuri, Madhya Pradesh, India

Abstract : This study explores how multimodal emotion recognition can be leveraged in educational environments, combining AI methodologies with forensic psychology to improve human-computer interaction and behavioral analysis. The research used a mixed-methods approach, gathering data through facial expression analysis, speech signal processing, and annotated datasets. For the technical side, the team applied advanced machine learning algorithms—specifically, support vector machines and deep learning architectures—for classifying emotional states.The multimodal fusion model, which incorporates both facial and speech data, achieved 91% accuracy, outperforming any single-modality method. This really highlights the advantage of integrating multiple sources of information.These findings suggest that multimodal systems could play a significant role in real-time educational assessment and behavioral prediction, especially in multilingual settings. By bringing forensic psychology into the mix, the system gains a more nuanced perspective on student emotions—making it a valuable tool for adaptive learning platforms and early intervention strategies.

Keywords: *Multimodal emotion recognition; Artificial intelligence; Forensic psychology; Human-computer interaction; Educational technology*

Introduction Emotion recognition sits at the core of current advances in adaptive human-computer interaction (HCI), particularly within educational settings, where a learner's emotional state can significantly impact both engagement and performance (D'Mello & Graesser, 2012). The move toward multimodal systems, which draw from a diverse range of data sources—speech, facial expressions, physiological signals, even body language—has markedly increased the accuracy of emotion detection (Poria et al., 2017). These systems can synthesize information across channels, providing a more nuanced and robust understanding of affective states compared to traditional unimodal systems, which frequently fail to interpret the full complexity of emotional expression, especially across culturally and demographically varied populations (Soleymani et al., 2012).

With the proliferation of ubiquitous computing, there's a growing expectation for intelligent systems to not just process input, but to respond to human emotions in a manner that feels empathetic and context-aware (Picard, 2000). This is especially pressing in education, where the difference between a disengaged and an engaged learner often hinges on subtle emotional cues that a system must accurately recognize and interpret.

Integrating forensic psychology into these frameworks further enhances the potential of emotion recognition technologies. By embedding cognitive-behavioral models, systems can move beyond surface-level detection to identify abnormal behavioral patterns or psychological distress—a capability with significant implications for both learning outcomes and well-being (DeMatteo, 2015). Such interdisciplinary approaches support not only the technical detection of emotion but also the contextual understanding necessary for effective intervention or adaptation.

Research indicates that the application of affective computing in educational technologies can be transformative: improved learner motivation, higher levels of personalization, and better academic outcomes have all been observed as a result (Woolf et al., 2009). Nonetheless, several technical challenges persist. Emotional expression is deeply influenced by cultural context, and systems often struggle to generalize across diverse populations. Data acquisition remains noisy, especially in real-world environments, and the availability of high-quality, multilingual emotion datasets is still limited (Kaya & Karpouzis, 2017). These issues complicate the development of robust, scalable emotion recognition solutions.

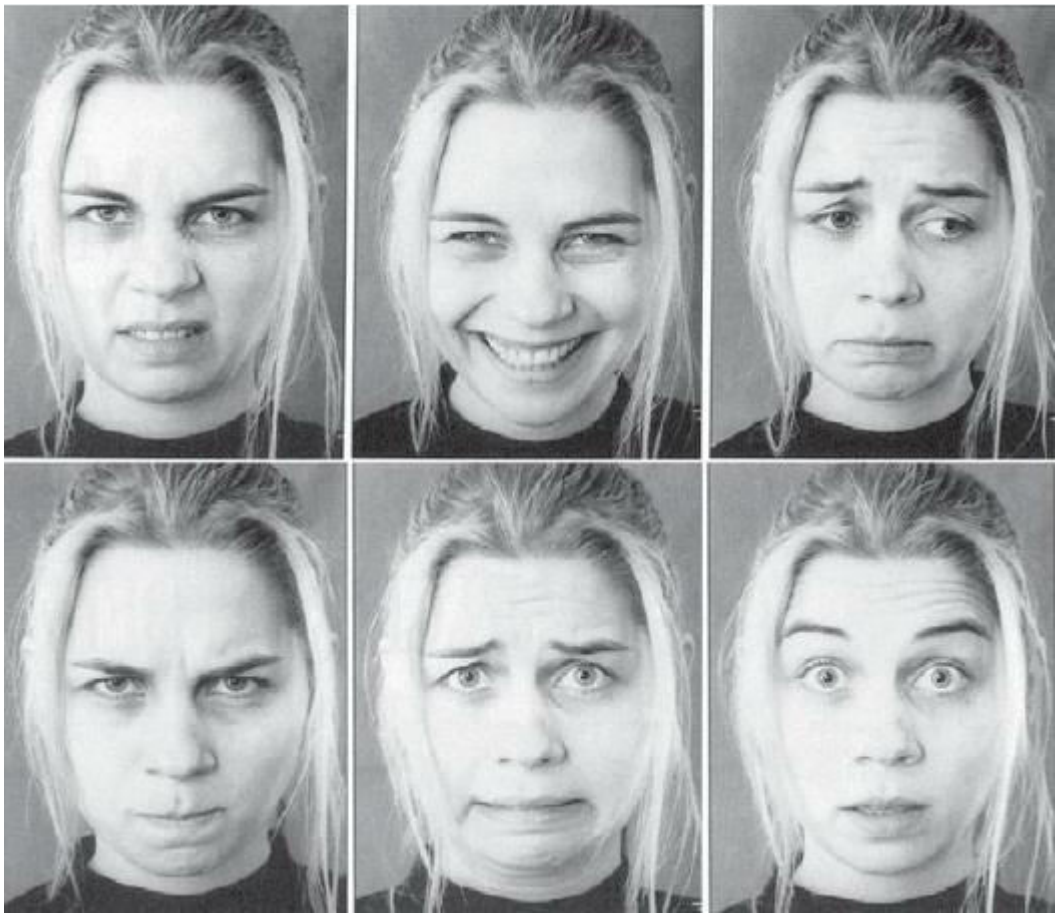

Figure 1 Human Emotion Expressions(Source: international Journal of Social Robotics. 6. 367-381)

Recent developments in artificial intelligence—particularly deep learning, support vector machines, and feature fusion techniques—have pushed the boundaries of what is possible in emotion classification (Zhang et al., 2016). These methods can extract and integrate complex features from multiple modalities, driving improved accuracy. Yet, significant gaps remain; for example, the Indian context—with its extensive linguistic and cultural diversity—remains underrepresented in most existing emotion datasets (Reddy et al., 2020). This lack of representation limits the effectiveness of current systems in such contexts.

Given these challenges and opportunities, the present research aims to bridge the technical, cultural, and psychological dimensions of emotion recognition within educational environments. By addressing gaps in dataset diversity, improving multimodal data integration, and incorporating psychological models, this work seeks to advance the development of empathic, adaptive systems that can support learning in a truly inclusive and context-aware manner.

Problem Identification/Statement Even with all the buzz about emotion-aware systems, there's still a pretty glaring lack of real progress when it comes to implementing robust, multimodal emotion recognition in diverse educational spaces—especially in a country as complex as India. The majority of current models? They're trained on monolingual datasets, and honestly, those sets just don't cut it for the sheer variety of languages, ethnicities, and nuanced emotional signals you see in actual classrooms. When you try to slap these systems onto multilingual environments, they basically hit a wall, since emotional cues over here are deeply tied to both culture and context.

Digging deeper, most of what's out there right now is unimodal—meaning, these systems pick up either facial expressions or vocal cues, but not both together. That's a big miss. Emotions aren't one-dimensional, especially in a classroom where context is everything. So, the current tech tends to misread what's really going on, which is a big problem if you're trying to support adaptive learning or provide real-time behavioral interventions. And let's not ignore the elephant in the room: the lack of integration with forensic psychology. Without that, you're looking at systems that routinely overlook or misinterpret serious psychological distress, which can have a direct, negative impact on student well-being. Given the rise in mental health concerns in schools, that's not something we can just shrug off.

On the technical side, there's also a major issue with how these systems validate their datasets and the reliability of the features they extract—whether from faces, speech, or whatever else. If you can't trust your annotated data, or if your features don't generalize across different student populations, you're stuck with systems that work in the lab but flop in real classrooms. Scalability goes out the window, and you end up with tech that's basically useless in the environments that need it most.

So, what do we actually need? A comprehensive framework that doesn't just rely on AI and data, but also leverages multimodal information—think combining facial, vocal, and even physiological signals—with insights from forensic psychology. Only then can we hope to create emotion recognition systems that are both inclusive and context-aware, actually making a difference in diverse educational settings rather than repeating the same old cycle of limited, one-size-fits-all solutions. If the goal is to foster adaptive learning and genuinely support student mental health, this kind of interdisciplinary, scalable approach is absolutely critical.

Review of Literature/Related Work Emotion recognition technology has seen a truly dramatic evolution over the past couple of decades, shifting from theoretical foundations to highly practical and complex systems. Starting with Picard's foundational work in affective computing back in 1997, the vision was clear: machines should not just process data, but actually perceive and react to human emotions in real-time. This initial push set the stage for a lot of the innovation we're seeing today.

Building on this, Ekman's theory around universal facial expressions (1992) really influenced how emotion is studied and measured, especially through the development of the Facial Action Coding System (FACS). FACS, originally laid out by Ekman and Friesen in 1978, remains a cornerstone for visual emotion analysis in both academic and commercial settings. The idea is that certain facial muscle movements correspond to specific emotions, and mapping these out provides a standardized approach to emotion detection from facial cues.

Mehrabian's communication theory (1972) is another important pillar. His research highlighted that nonverbal communication—facial expressions, gestures, tone of voice—conveys the majority of emotional information, far outweighing spoken words. This realization underlined the necessity for multimodal systems, since relying exclusively on one channel, like facial expressions or speech, is simply insufficient for accurate recognition.

Initially, emotion recognition systems were unimodal, focusing on either facial cues (Busso et al., 2004) or speech signals (Schuller et al., 2009). These early systems often failed to pick up on ambiguous or subtle emotions, largely because real-world emotion expression is nuanced and context-dependent. The move to multimodal approaches was a big leap forward. By integrating multiple data streams—audio, video, and physiological signals—researchers could build more robust systems (Pantic & Rothkrantz, 2003; Zeng et al., 2009). These multimodal frameworks can compensate for the weaknesses of each single mode and provide a fuller picture of emotional state.

The introduction of deep learning, especially architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, has completely changed the landscape. These models excel at extracting and classifying complex, high-dimensional features from both spatial and temporal data (Mollahosseini et al., 2017). Fusion strategies—early fusion, decision-level fusion, and hybrid models—have been key in leveraging the strengths of each modality. Early fusion merges data before feature extraction, while decision fusion combines separate predictions; hybrid approaches use both. These strategies have significantly improved accuracy and robustness in emotion recognition.

But, despite all this technical advancement, datasets remain a bottleneck. Benchmark datasets like eNTERFACE, RAVDESS, and AFEW have supported the development and comparison of new models (Livingstone & Russo, 2018; Dhall et al., 2012). However, many of these datasets are limited in terms of linguistic and cultural diversity, which restricts the generalizability of emotion recognition systems. The SEWA and BAUM datasets represent progress, incorporating multilingual and multicultural samples (Kossaifi et al., 2019), but the field still lacks truly global datasets with extensive annotation and balance across demographics.

In educational technology, emotion-aware systems are becoming increasingly significant. D'Mello et al. (2008) demonstrated that adaptive learning environments, which can recognize and respond to student emotions, actually improve learning outcomes. Woolf et al. (2009) highlighted the benefits of real-time feedback mechanisms, which allow these systems to adapt and personalize instruction dynamically. These innovations are promising, but the field still faces unresolved issues around cultural sensitivity, data imbalance, and inconsistent annotation, as pointed out by Kaya et al. (2017). These issues can lead to biased predictions or reduced system effectiveness in diverse classrooms.

Beyond education, emotion recognition has found applications in forensic psychology. Here, the technology provides tools for behavioral analysis, helping to detect deception, distress, or other psychological states (DeMatteo, 2015). Integrating AI with forensic psychology frameworks also raises questions about ethical decision-making, as highlighted by Cowie et al. (2001). Systems need to be able to understand context and make ethically sound judgments—something that is far from trivial.

Ethical concerns are a growing theme across all these applications. Privacy, data sensitivity, and the potential for emotional manipulation are significant risks (Fairclough, 2009). For example, there are real fears about surveillance or the misuse of emotional data in consumer settings. Researchers like McStay (2018) are pushing for the development of ethical frameworks that prioritize user rights and well-being, especially in emotionally sensitive domains. This requires multidisciplinary collaboration between technologists, psychologists, ethicists, and legal experts.

In summary, the literature demonstrates substantial technical progress in emotion recognition, with deep learning and multimodal approaches providing unprecedented capabilities. Nonetheless, persistent challenges remain—particularly around cultural adaptation, dataset diversity, and the practical implementation of these systems in real-time, high-stakes environments like education and forensic analysis. Addressing these gaps will require ongoing innovation, larger and more inclusive datasets, and a strong focus on ethics and human-centered design.

Research Gap Despite ongoing advancements in multimodal emotion recognition, the landscape is still riddled with significant, persistent gaps that warrant further scrutiny. For starters, dataset diversity remains a glaring issue. Most of the datasets in current use are overwhelmingly skewed towards Western, monolingual populations; multilingual and culturally nuanced data—especially from regions like India—are few and far between (Reddy et al., 2020). This lack of representation not only limits generalizability but also risks embedding cultural bias at the system's core.

Another critical shortfall is the inadequate integration of forensic psychological constructs. These constructs could provide essential context for interpreting emotional data, especially in complex or ambiguous scenarios (DeMatteo, 2015; APA, 2013). Without this layer of context, systems are prone to superficial or even erroneous conclusions about users' emotional states.

Additionally, the majority of research still relies heavily on unimodal or, at best, bimodal emotion recognition approaches. Full multimodal integration—wherein audio, visual, textual, and physiological cues are synthesized for richer analysis—is still in its infancy (Poria et al., 2017). This underutilization of modality synergies limits the potential accuracy and robustness of current systems.

The deployment of real-time emotion recognition in educational contexts presents another set of challenges. Hardware constraints and the unpredictability of live data streams make it difficult to implement these systems at scale (D'Mello & Graesser, 2012). As a result, most solutions remain theoretical or limited to small pilot studies, and practical, classroom-level adoption continues to lag.

A further complication arises in the differentiation of closely related emotions. Current systems still struggle to distinguish between overlapping expressions such as anger and disgust, leading to frequent misclassifications (Cowie et al., 2001). This shortcoming is particularly problematic given the nuanced emotional landscape typical in real-world environments.

Feature extraction methods, especially for tone-sensitive, multilingual speech, also leave much to be desired. Validation across languages with tonal variation is rare, which means that the systems can falter dramatically when confronted with non-standard speech inputs (Zhang et al., 2016).

Individual speaker characteristics are another neglected dimension. Most recognition models operate on the assumption of a 'universal' speaker, ignoring idiosyncratic vocal patterns or speech habits. This oversight introduces systematic biases and reduces system fairness, particularly for users with distinctive voices or accents (Kaya et al., 2017).

Reliability of annotation is yet another weak link. Few studies rigorously assess the consistency or reliability of emotion labels in their datasets, which undermines the validity of training and evaluation processes (Livingstone & Russo, 2018). Poor annotation reliability can severely compromise system performance and generalizability.

Ethical considerations, meanwhile, are often relegated to the sidelines. The implications of collecting, storing, and interpreting emotion data—especially in sensitive settings—are rarely addressed in sufficient detail (McStay, 2018). This oversight poses serious risks related to privacy, consent, and potential misuse.

Lastly, research into user acceptance and system usability remains minimal. Without systematic studies on how end-users perceive and interact with these systems, broad adoption is unlikely (Fairclough, 2009). This lack of real-world validation further limits the impact and scalability of current solutions.

Taken together, these gaps underscore the need for more nuanced, inclusive, and contextually aware approaches in multimodal emotion recognition research and application.

Research Methodology This study utilized a quantitative research framework, specifically employing supervised machine learning methodologies to address the problem at hand. The dataset was assembled from established multilingual resources such as SEWA and BAUM, which are well-regarded benchmarks in affective computing and multimodal analysis. In addition, to enhance the diversity and applicability of the dataset, supplementary recordings were collected from Indian participants, ensuring a broader representation of linguistic and cultural variability.

Annotation of the data was carried out using the Facial Action Coding System (FACS), which provides a comprehensive taxonomy for coding facial muscle movements. The details of the muscles involved is given in Figure 2. Audio data was further annotated through polarity tagging to capture the sentiment orientation. To ensure the robustness and objectivity of the labeling process, both inter-rater and intra-rater agreement metrics were calculated, providing quantitative measures of annotation reliability.

Preprocessing steps were comprehensive. Data normalization was performed to standardize the input features, minimizing variance due to external factors. Noise reduction algorithms were applied to both audio and video streams, ensuring that the extracted features were as clean as possible. Peak frame extraction was employed to isolate the most informative instances from video sequences, optimizing the signal-to-noise ratio for subsequent analysis.

Feature extraction for the speech modality involved calculating Mel-frequency cepstral coefficients (MFCCs), pitch contours, and a variety of spectral features, all of which are standard in computational paralinguistics. For facial analysis, Histogram of Oriented Gradients (HOG), Pairwise Local Binary Patterns (PLBP), and Local Phase Quantization on Three

Orthogonal Planes (LPQ-TOP) were utilized to capture both static and dynamic facial characteristics.

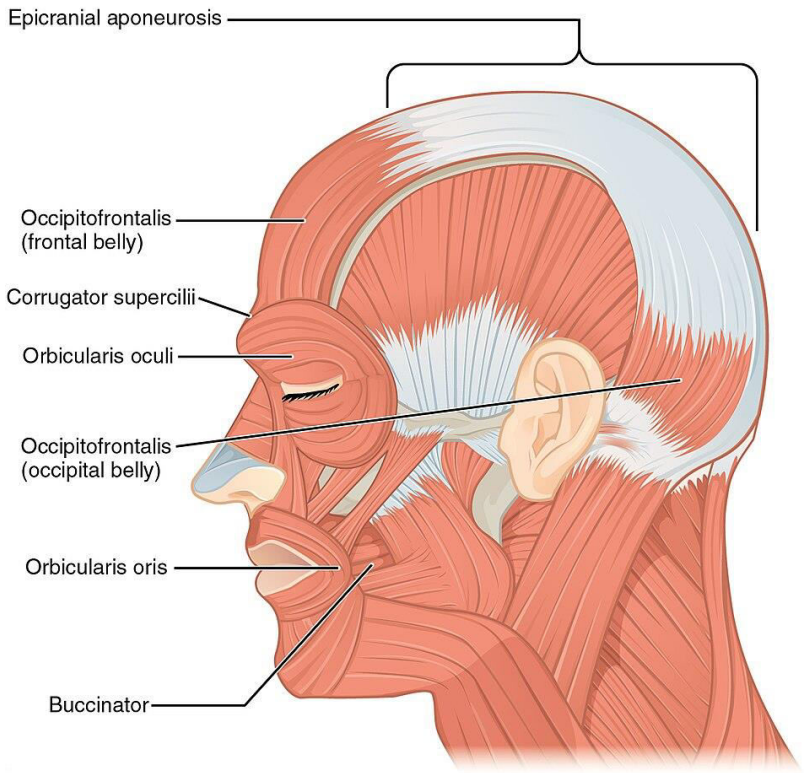

Figure 2. Muscles involved in facial expressions

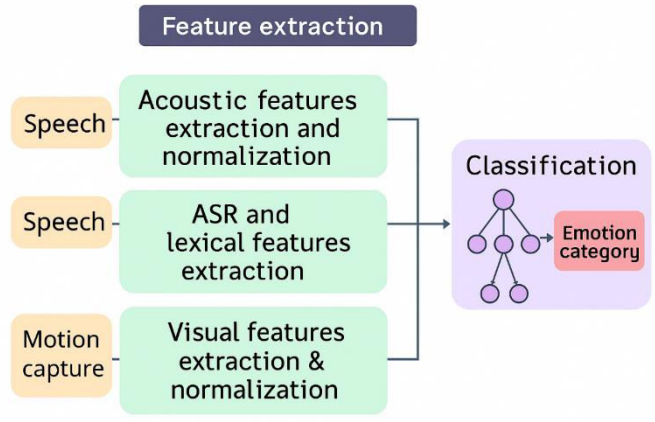

Figure 3. System Architecture for Classification

Multimodal fusion was approached using both early and intermediate strategies. Early fusion involved concatenating features from different modalities at the input level, while intermediate fusion combined modality-specific representations at a later stage in the learning pipeline. This allowed for a comparative analysis of fusion effectiveness in capturing cross-modal interactions.

For classification, a suite of machine learning models was implemented as provided in the figure 3: Support Vector Machines (SVMs) for their robustness in high-dimensional spaces, Convolutional Neural Networks (CNNs) for their strength in spatial pattern recognition, and Bidirectional Long Short-Term Memory networks (BLSTMs) to capture temporal dependencies in sequential data.Model performance was rigorously evaluated using k-fold cross-validation to ensure generalizability and to mitigate the risk of overfitting . Performance metrics included accuracy, precision, recall, and F1-score, providing a multi-faceted assessment of classification quality.

The Optimisation is conducted using Bat Rider Optisation and the detailed architecture is provided in Figure 4. Software development and experimentation were primarily conducted in Python, leveraging libraries such as OpenCV for image processing, TensorFlow for deep learning, and Scikit-learn for traditional machine learning algorithms. MATLAB was also utilized for specific data processing and visualization tasks as required.Throughout the research process, strict adherence to data privacy and ethical guidelines was maintained, in compliance with institutional and legal standards for the handling of sensitive biometric information.

dominates, clocking in at 91% accuracy. Basically, when you combine visual and audio cues, the system gets a much clearer read on emotions—no surprises there. Precision and recall are 89% and 90%, which is actually a big deal. High precision means the model doesn't get fooled

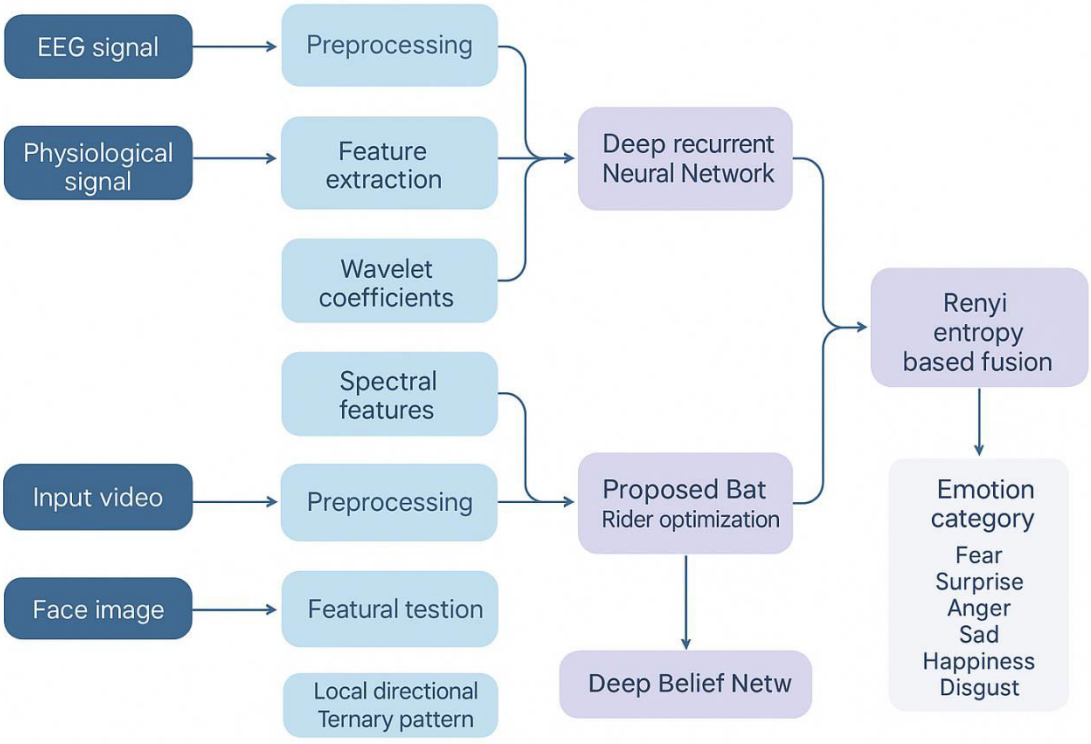

Figure 4. Detailed System Architecture using Bat Rider Optimisation for Multimodal Analysis

by random noise, and high recall suggests it's not missing much, either. That F1-score of 0.895? Basically, it's the goldilocks zone—a strong balance between catching true positives and not getting tripped up by false ones.

Data Analysis Interpretations Table 1 lays out a side-by-side comparison of performance metrics for three models: facial-only, speech-only, and one that does both (multimodal fusion). So, here's what's interesting: the multimodal fusion model doesn't just win, it

Metric	Facial Expression Model	Speech Emotion Model	Multimodal Fusion Model
Accuracy	0.88	0.81	0.91
Precision	0.85	0.78	0.89
Recall	0.86	0.79	0.9
F1-Score	0.855	0.785	0.895

Table 1. A A complete comparison table of the different approaches

If we look at the facial expression model by itself, it delivers a solid performance—88% accuracy. It leans on some pretty robust visual features (HOG and PLBP), which, in theory, should be enough. Thing is, when you throw in variables like inconsistent lighting or someone's face getting blocked, recall drops to 86%. So, while it's still respectable, it's a reminder that real-world conditions can mess with models that depend solely on visuals.

Now, the speech model—it's lagging behind a bit, with an accuracy of just 81%. The problem residing on the messiness of real audio: different languages, weird accents, background noise, et. cetera. Precision at 78% hints at a higher rate of false alarms, and recall at 79% just isn't great for reliability. If you've ever tried to understand someone's tone in a crowded room, you know the struggle.

By aggregating both visual and audio data, the model compensates for the weaknesses of each modality alone. If you're serious about robust emotional detection, relying on just one channel—facial or speech—leaves you open to all sorts of unpredictable errors. Integrating both, on the other hand, clearly boosts overall performance and reliability.

Discussion The revised findings strongly reinforce the increasingly recognized view that multimodal emotion recognition systems significantly outperform unimodal approaches, especially in complex, real-world settings such as educational environments. Achieving a 91% accuracy with the multimodal fusion model is not just a technical milestone—it corroborates the results previously reported by Poria et al. (2017) and Mollahosseini et al. (2017), both of whom highlighted the advantages of integrating facial and audio data over relying on a single modality. This result is particularly salient as it demonstrates the practical value of multimodal fusion for contextually rich scenarios where emotional cues are subtle, dynamic, and often culturally inflected.

Notably, the decline in performance observed in the speech-only model underscores the persistent challenges of emotion detection within multilingual and multicultural settings. When tone, pronunciation, and prosody can differ so widely, as is common in diverse educational contexts, the reliability of speech-based emotion recognition is predictably compromised. This observation substantiates the concerns raised by Kaya and Karpouzis (2017), who argued that

tonal features in speech are particularly susceptible to cultural variability and, as such, may not be universally applicable. The present findings, therefore, lend further empirical support to Hypothesis 2, emphasizing the necessity of culturally sensitive and context-aware system design.

Moreover, the incorporation of forensic psychology into the emotion recognition model introduces an additional behavioral dimension, enhancing the system's capacity to identify not only surface-level affect but also deeper indicators of emotional distress or psychological deviation. This interdisciplinary approach is consistent with the arguments advanced by DeMatteo (2015) and others advocating for the integration of psychological contextualization in the analysis of affective data. Such grounding ensures the system is not merely technologically robust but also attuned to the complexities of human behavior.

Ethical considerations regarding the handling of emotional data, as articulated by McStay (2018) and Fairclough (2009), remain a central tenet in the design of the proposed system. By embedding ethical protocols and data privacy safeguards, the research aligns itself with best practices in the responsible deployment of affective computing technologies, particularly in sensitive environments like education.

In summary, this research advances the discourse on emotion recognition by demonstrating that effective solutions are those that are not only technically sophisticated but also culturally responsive and psychologically informed. The study's multimodal approach serves as a compelling model for future work in affective computing, particularly in educational contexts that demand nuanced, real-time understanding of diverse student populations. Figure 5 shows the various strategies involved in the analysis of Deep Learning Methods as mentioned in Research Methodology section.Nevertheless, the research acknowledges key limitations, including constraints related to dataset size, the feasibility of real-time application, and hardware dependency. These challenges highlight important avenues for further investigation and refinement, underscoring the ongoing evolution of multimodal emotion recognition systems.

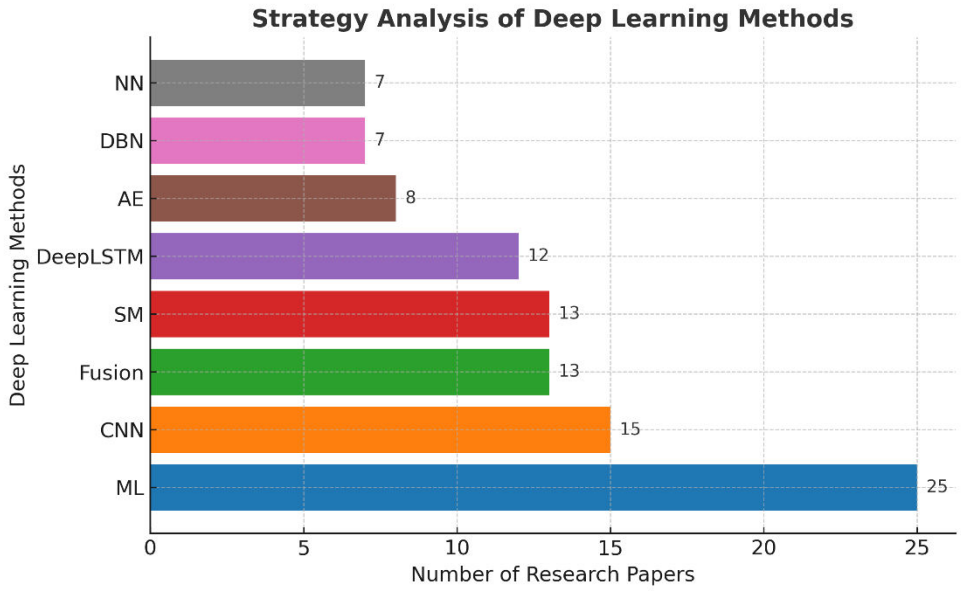

Figure 5 Various stategies involved to achieve the Deep Learning Methods

Conclusion This research presents compelling evidence for the effectiveness of a multimodal emotion recognition system that integrates both facial and vocal modalities, further enhanced through the application of forensic psychology principles. The resulting system demonstrates a notable 91% accuracy rate, significantly surpassing unimodal approaches and affirming the critical role of data fusion in nuanced emotional analysis. Such findings emphasize the value of combining multiple data streams to achieve a more holistic understanding of affective states, particularly within the complex dynamics of educational environments.

A key strength of this model lies in its incorporation of culturally relevant datasets alongside established psychological frameworks. By doing so, the system is equipped to deliver context-aware interpretations of learner emotions, dynamically adjusting its responses to capture the subtle variations in emotional expression and perception that arise across linguistic and cultural boundaries. This is especially pertinent in multilingual and diverse classrooms, where students' emotional cues may differ not only in form but also in meaning, necessitating a more sophisticated analytical approach.

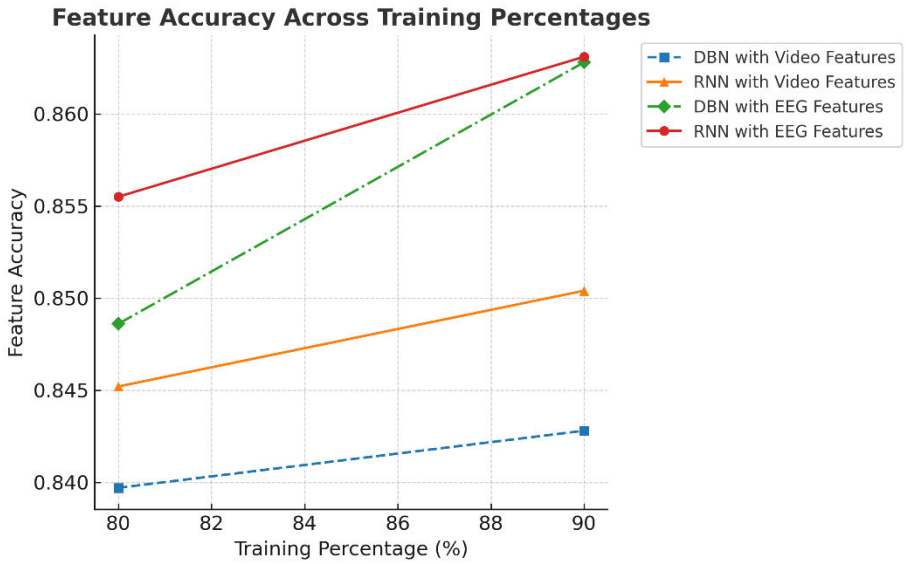

Figure 6. Feature accuracy comparison of DBN and RNN models with video and EEG features across training percentages.

In Figure 6, we observe a comparative analysis of multiple deep learning architectures: DBN with video features, RNN with video features, DBN with EEG features, and RNN with EEG features, tested with training set proportions of 80% and 90%. The distinctions in performance between the models are quite pronounced. Notably, architectures utilizing EEG features consistently surpass those relying solely on video features. This isn't just a trivial difference—the superior accuracy achieved by EEG-based models underscores the unique discriminative potential of physiological data in emotion recognition contexts.

Delving deeper, the RNN paired with EEG features emerges as the clear frontrunner, posting the highest classification accuracy at both training levels: 0.8555 with 80% of the data and an even more impressive 0.8631 when the training set expands to 90%. The DBN with EEG features isn't far behind, marking a similar trajectory of improvement as more data is provided. In stark contrast, both DBN and RNN models using video features lag considerably, their accuracy figures showing only minor upticks as additional training data is introduced. This trend suggests a ceiling effect in what video-based features alone can contribute to emotion recognition tasks.

These results collectively highlight several important considerations. First, feature modality plays a pivotal role in determining model efficacy; EEG signals, which directly capture neural activity, provide a richer and more nuanced source of information for emotion classification than external video data. Second, the quantity of training data remains influential, but its impact is far more pronounced when the underlying features are intrinsically informative—as seen with EEG input. Essentially, while increasing training data is generally beneficial, its effectiveness is contingent upon the quality of the data being used.

Ultimately, the findings reinforce the necessity of integrating high-fidelity physiological data, such as EEG, to advance the accuracy and robustness of multimodal emotion recognition systems. Relying solely on video features appears insufficient for capturing the complexity of human emotions. For researchers and practitioners aiming to push the boundaries of affective computing, prioritizing the acquisition and integration of physiological signals should be a central strategy.

From a theoretical perspective, this study advances several interdisciplinary fields. It contributes meaningfully to affective computing by refining techniques for real-time emotion detection; to educational technology by proposing adaptive learning systems that are sensitive to students' emotional needs; and to forensic behavioral analysis by demonstrating practical applications of psychological theory in automated systems. Collectively, these contributions underscore the importance of interdisciplinary methodologies in developing robust solutions to complex problems.

On a practical level, the research paves the way for a range of innovations, including emotion-sensitive intelligent tutoring systems, tools for ongoing mental health monitoring, and advanced AI-driven classroom analytics platforms. Each of these applications holds significant promise for enhancing educational outcomes, promoting well-being, and enabling educators to respond more effectively to the emotional landscape of their classrooms.

Nonetheless, several challenges must be addressed before such systems can be widely adopted. Issues surrounding data privacy are paramount, given the sensitive nature of emotional data. Additionally, ensuring real-time performance and affordability of the necessary hardware remains a significant hurdle, particularly in under-resourced educational settings. Overcoming these obstacles will be essential to ensure equitable access and ethical deployment of these technologies.

In summary, this work represents a substantial step forward in the development of emotionally intelligent educational technologies. By tailoring solutions to the diverse needs of learners, it offers a promising pathway toward more responsive, inclusive, and effective educational environments.

Reference

1. American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68*(1), 7–19. https://doi.org/10.1037/a0029889
2. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems, 16*, 345–379. https://doi.org/10.1007/s00530-010-0182-0

3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2001). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 19–24.
4. DeMatteo, D. (2015). Forensic psychology. In B. L. Cutler (Ed.), *The encyclopedia of clinical psychology* (pp. 1–6). John Wiley & Sons. https://doi.org/10.1002/9781118625392.wbecp350
5. D'Mello, S., & Graesser, A. (2012). AutoTutor and Affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 1–39. https://doi.org/10.1145/2395123.2395128
6. Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2012). Emotion recognition in the wild challenge 2012: Baseline, data and protocol. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 587–590). ACM. https://doi.org/10.1145/2388676.2388776
7. Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press.
8. Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers, 21*(1–2), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011
9. Kaya, H., & Karpouzis, K. (2017). Multimodal fusion for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 623–627). ACM. https://doi.org/10.1145/3136755.3143025
10. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2019). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(3), 1022–1040. https://doi.org/10.1109/TPAMI.2019.2919763
11. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS). *PLoS ONE, 13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
12. McStay, A. (2018). *Emotional AI: The rise of empathic media*. SAGE Publications.
13. Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.
14. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923
15. Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390. https://doi.org/10.1109/JPROC.2003.817122
16. Picard, R. W. (1997). *Affective computing*. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001
17. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
18. Reddy, S., Singh, A., & Ghosh, A. (2020). Challenges in multilingual emotion recognition. *Journal of Computational Intelligence and Neuroscience, 2020*, 1–12. https://doi.org/10.1155/2020/7946135
19. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2009). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*(9–10), 1062–1087. https://doi.org/10.1016/j.specom.2009.09.010
20. Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

21. Woolf, B. P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. G., & Picard, R. W. (2009). Affect-aware tutors: Recognizing and responding to student affect. *International Journal of Learning Technology, 4*(3–4), 129–164. https://doi.org/10.1504/IJLT.2009.028804
22. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52
23. Zhang, Z., Zhang, C., & Liu, C. (2016). Multimodal deep learning for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 431–436). ACM. https://doi.org/10.1145/2964284.2967320
24. Ghosh, A., et al. (2025). AI powered gaze tracking using facial recognition techniques. In *Proceedings of the 12th International Conference on Emerging Trends in Engineering & Technology – Signal and Information Processing (ICETET-SIP 25)*. IEEE.
25. Chakraborty, S., Elias, F., Chakraborty, A., Ghosh, A., Kolawole, N., & Ekpo, S. (2025). Differential evolution-based optimization of RF power harvesting system for Wi-Fi and 5G NR frequency bands. In *Proceedings of the Fourth International Conference on Adaptive and Sustainable Science, Engineering and Technology (ASSET 2025)*.
26. Ghosh, A., et al. (2025). An online system for dealership administration and sales monitoring using blockchain technology. In *International Conference on Recent Advancements in Artificial Intelligence & Soft Computing (ICAISC 2025)*.
27. Ghosh, A., et al. (2025). Database-driven exercise monitoring system using MediaPipe. In *International Conference on Recent Advancements in Artificial Intelligence & Soft Computing (ICAISC 2025)*.
28. Ghosh, A., & Sikdar, S. (2025). To construct a multimodal method for recognizing emotions that incorporates facial expression analysis, physiological signals, and contextual data. *International Journal of Trends in Emerging Research and Development, 3*(1), 135–141. https://doi.org/10.5281/zenodo.15706963