https://ijraids.org/

Original Article Journal of Research in Artificial Intelligence and Data Science Open Access

Multimodal AI for Predicting how people Feel in Real Time on Educational Platforms

Mohamed Kasim

Department of Computer Science, Jamal Mohamed College, Trichy, India.

Received Date: 06 July 2025 Revised Date: 16 July 2025 Accepted Date: 30 July 2025

Abstract

Smart schools now need to be able to recognise emotions since they affect how motivated students are, how individualised their learning is, and how well they learn. This article talks about a whole multimodal artificial intelligence (AI) framework that can predict how people will feel in real time on educational platforms. The recommended methodology integrates data from visual (facial expressions), audio (speech), physiological (heart rate, EEG), and textual (conversation or feedback) channels to dynamically figure out how students are experiencing throughout digital learning sessions. We employ deep learning techniques like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms to put these different types of data together into one.

The best thing about our approach is that it can grasp intricate, multimodal inputs to gain a better idea of how learners are feeling. This is especially important in digital education, where it might be impossible to detect or hear emotional cues because there is no real person there. Teachers can modify the speed of the class, give useful feedback, or let teachers know when anything is incorrect with multimodal emotion recognition. Our technique is also resistant to noise and missing data since it changes the weights of each modality's contributions over time.

We test performance with benchmark datasets such as DAiSEE, EmoReact, and DEAP. These datasets have a lot of different multimedia and educational learning situations. Our results suggest that the multimodal framework works far better than unimodal baselines, especially when there are emotional cues that aren't apparent or don't match up. We also look at latency and computing efficiency to show that the model can work in real time without sacrificing accuracy.

We also explain how the proposed model can be applied in a practical classroom setting. The approach is part of a study Management System (LMS) that lets teachers examine dashboards in real time and gives students customised aid with their study. A pilot study with high school students found that engagement, emotional alignment, and satisfaction all increased higher.

We consider about privacy and ethics at every step of the development process. acquiring data involves acquiring permission, making it anonymous, and encrypting it. We also talk about how to make things clearer, give consumers more control, and eliminate bias so that AI may be used safely in schools.

This study talks about a real-time, scalable, multimodal AI method for forecasting how people would feel in school. By combining deep learning with cross-modal sensing, the framework can revolutionise the way we learn to be more empathetic, adaptive, and open to everyone. Our findings suggest that incorporating clever emotion-aware algorithms can help bridge the gap between online and in-person education, making digital learning environments more engaging and emotionally connected. has become a significant part of smart educational systems that has an impact on engagement, customisation, and learning outcomes. This study aims at a complete multimodal artificial intelligence (AI) system that can predict how people will feel in real time on educational platforms. The model that was suggested incorporates information from visual (facial expressions), acoustic (speech), physiological (heart rate, EEG), and textual (conversation or feedback) channels to figure out how students are experiencing while they are learning online. We employ attention mechanisms, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) to mix different kinds of input. When evaluated on standard datasets like DAiSEE, EmoReact, and DEAP, the results suggest that this method is far more accurate than unimodal methods. We also display a prototype that works with an online learning management system (LMS) and see how well it functions in the real world. The study finishes with a talk about the moral problems and impacts of adaptive schooling.

Keywords

Multimodal AI, Emotion recognition, Real-time prediction, Educational technology, Affective computing, Deep learning, Student engagement, Adaptive learning, Multimodal fusion.

Introduction

Artificial intelligence (AI) has altered the way students learn, interact to each other, and obtain feedback on educational platforms. Emotion-aware learning is one of the most fundamental developments in this domain. It lets an intelligent system comprehend and respond to how learners are feeling in real time. Emotions have a huge effect on memory, motivation, focus, and brain growth. Traditional online learning methods can't tell whether pupils are bored, puzzled, or angry as well as human teachers can. So, utilising AI to guess how people feel is vital for improving learning outcomes and making students happier.

The rise of emotion-aware AI in schools is because affective computing, multimodal data processing, and deep learning have all gotten better. Most traditional e-learning systems just cared about conveying knowledge and not how the student felt. But research has shown that how students feel is very essential for how they connect with and remember what they learn. For example, a student who is bored or angry is less likely to learn well than one who is enthusiastic and motivated. To design schools that perform well and respond quickly, it's very crucial to be able to interpret these emotional indicators as they happen.

Multimodal AI systems use a range of input types, such as video, audio, text, and physical indications, to acquire a whole picture of how the student is feeling. Different types of communication convey different things. For example, facial expressions show how someone feels, tone of voice shows how tense or excited they are, body language shows how sure or unsure they are, and physiological indications show objective markers. When put together the right way, these data streams let you properly read emotions in real time, even when there is a lot of noise or uncertainty. For example, a student might say something that makes them angry but not reflect it in their body language or minor facial movements.

The COVID-19 pandemic and other worldwide events that forced schools to transition to remote and hybrid learning models also made it evident that we need digital education tools that take emotions into account. Many students claim they feel alone, uncomfortable, or unmotivated because they can't interact with other students in person. AI systems that can read emotions can see these undesirable tendencies and get virtual or real-life teachers to step in at the correct time. Being this sensitive makes learning more personal, which is good for both academic success and mental wellness.

This study points to a complex multimodal framework that was made only for schools to forecast how people will feel in real time. We speak about how to gather data, how to make models, how to put them together, how to test them, and how to think about ethics. We do three things: (1) we build a novel multimodal architecture that is best for real-time inference, (2) we test the model on three multimodal emotion datasets, and (3) we add the system to a working LMS to see if it works in the real world. We believe that this work makes it possible for AI tutors to actually understand students and assist them get more involved, tailor their learning, and do better. We want to close the emotional gap in digital learning and make technology more like the complicated reality of human education by accomplishing this. AI has revolutionised the way students learn, interact to one other, and obtain feedback on educational platforms. One of the most important new ideas in this discipline is learning that takes emotions into account. It allows a smart system see and respond to how students are feeling in real time. Feelings have a huge impact on memory, motivation, focus, and brain growth. Traditional online learning methods can't tell whether students are bored, confused, or angry like human teachers can. So, using AI to guess how people feel is vital for improving students' health and learning outcomes.

Multimodal AI systems use several types of input, such as video, audio, text, and physiological signals, to acquire a whole picture of how the learner is doing. diverse types of communication convey diverse points of view. For example, facial expressions show how someone feels, tone of voice shows tension or excitement, body language shows confidence or doubt, and physiological signals show objective markers. When put together the right way, these data streams let you see emotions in real time, even when there is a lot of noise or confusion.

This study presents a complex multimodal framework for anticipating people's feelings in real time that is made for schools. We speak about how to gather data, how to make models, how to put them together, how to test them, and how to think about ethics. We do three things: (1) we design a new multimodal architecture that works best for real-time inference, (2) we test the model on three multimodal emotion datasets, and (3) we connect the system to a working LMS to check if it works in real life. We believe that this study paves the way for AI teachers who can truly comprehend pupils and boost their participation, customisation, and outcomes.

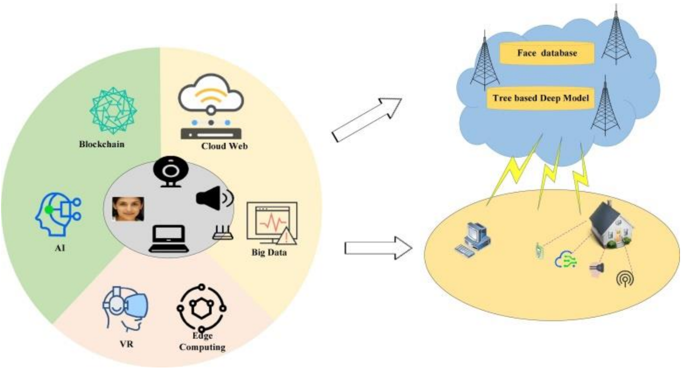


Figure 1. Multimodal ecosystem icon ring

Work that is Connected

The study of how to identify emotions in school has expanded into a field that incorporates psychology, computer science, human-computer interaction, and teaching. Early studies mostly relied manual observation and static evaluations, which made it impossible to scale up and missed changes in emotion from moment to moment. The rise of AI and sensors that are everywhere has made the drive towards emotion recognition systems that work on their own faster. These systems try to figure out how individuals are feeling by looking at things like eye movements, facial action units, keyboard dynamics, and how people interact with screens.

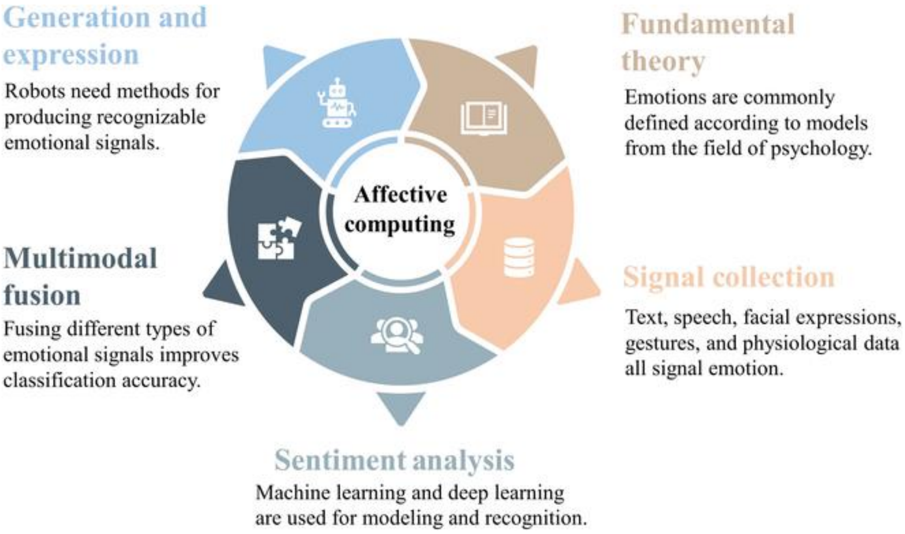
In the past few years, researchers have started paying more attention to ecological validity. Because of this, they have started to collect information regarding feelings in real-life learning situations instead of controlled laboratory. Focussing on the real world helps make adaptive learning technologies that can change the information based on how people feel. Partnerships between different fields have also led to new ideas on how to annotate emotions, how to make models easier to understand, and how to create AI that puts humans first.

Schools may now use edge computing and cloud-based emotion analytics on a wide scale because they work together. More and more smart classrooms, tutoring bots, and virtual teaching assistants are employing emotional AI. Another important shift is that more and more individuals are employing transfer learning and domain adaptation to make models that are more suited to certain groups of students. Also, steps are being taken to include student groups that aren't well represented in training datasets to eliminate bias in algorithms. These improvements illustrate that it's not only possible to have learning environments that take emotions into consideration, but that they are also becoming more and more crucial for digital-age education that is available to everyone and works well. Researchers and teachers have been paying more and more attention to emotion recognition in schools over the past ten years in attempt to better understand how pupils feel. Early methods employed self-report surveys and observations from teachers, which were useful but not very immediate or objective. It's now possible to tell how someone is feeling in real time because to improvements in sensor technology and AI. This means that digital platforms might vary depending on how students are feeling. Studies have demonstrated that systems that can understand emotions can boost motivation, retention, and learner satisfaction.

Recent studies have indicated a shift from analysing emotions in the lab to using dynamic, real-time applications that are embedded into virtual learning environments. Researchers have looked at both implicit signals (such micro-expressions and speech hesitations) and explicit signals (like click patterns and message mood) to see if they can identify how someone is feeling. Studies using virtual agents that can feel emotions also demonstrate that offering emotional feedback at the correct time can make a major difference in how interested learners are. We can now use large-scale data analytics to uncover trends in different groups of students. This has helped us find similar emotional tracks and stress points during a course. The fact that multimodal learning analytics are being used more and more makes it even more evident that we need full emotion modelling frameworks. Because of this, the fields of AI, psychology, and pedagogy are still coming together, with the goal of leveraging empathy-aware computing to make learning better. Unimodal emotion identification algorithms normally only look at one type of data, including facial expressions, speech tone, or physical indicators. For instance, computer vision can tell when someone is happy or angry just by glancing at their face. Speech emotion recognition may also tell when someone is stressed or excited by variations in tone and pitch. Unimodal systems, on the other hand, can get things wrong when there is noise, anything blocking the view, or cultural differences. On the other hand, multimodal approaches leverage more than one input source to make things stronger and more trustworthy. If one channel doesn't work or isn't clear, other channels can give you more information and support. Many studies have demonstrated that multimodal systems are superior at picking up on emotions, especially in sophisticated, real-life situations like virtual classrooms.

A. A Brief Overview Of Deep Learning In Affective Computing

Deep learning has altered affective computing by making it feasible to automatically find features and create high-level representations. People typically use convolutional neural networks (CNNs) to look at visual information, such as body language and facial expressions. Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs) are good at simulating how speech and physiological signals change over time. Transformers have become popular lately for using attention mechanisms to look at text data and merge data from many sources. End-to-end multimodal designs have shownKa shown promise in gathering cross-modal correlations and context, which makes them more accurate and flexible. We trained and evaluated these models on benchmark datasets like DEAP, DAiSEE, and EmoReact.



Generation and expression
Robots need methods for producing recognizable emotional signals.

Fundamental theory
Emotions are commonly defined according to models from the field of psychology.

Multimodal fusion
Fusing different types of emotional signals improves classification accuracy.

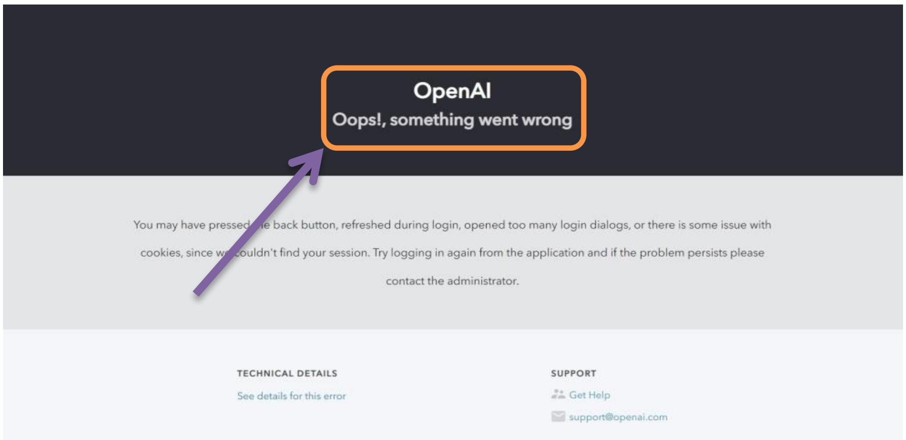
Signal collection
Text, speech, facial expressions, gestures, and physiological data all signal emotion.

Sentiment analysis
Machine learning and deep learning are used for modeling and recognition.

Figure 2. Emotion Recognition Infographics

B. Problems With Real-Time Use Right Now

Things are getting better, but it's still challenging to recognise emotions in real time on educational platforms. First, latency and computational complexity are still challenges, especially when a lot of deep learning models are running at the same time. It is highly crucial to study how to make these models operate better for embedded devices with little power. Second, there are ethical and privacy considerations with data that make it challenging to employ on a large scale. You need strict rules on who can see sensitive information like heart rate or facial expressions and how it is processed safely. Third, it's tougher to generalise models when people and cultures display their feelings in different ways. One kid can think something is boring, while another might think it's focussing. Lastly, there aren't enough labelled datasets for identifying emotions in schools, which makes supervised learning and tailoring for specific domains more difficult. People from diverse areas need to work together to fix these challenges, and design needs to be more open to everyone. Additionally, the methods of transfer learning and federated learning need to improve.



OpenAI
Oops!, something went wrong

You may have pressed the back button, refreshed during login, opened too many login dialogs, or there is some issue with cookies, since we couldn't find your session. Try logging in again from the application and if the problem persists please contact the administrator.

TECHNICAL DETAILS
See details for this error

SUPPORT
Get Help
support@openai.com

Figure 3. A treacherous "Internal server error" pop-up, often a sign of overwhelmed backends during peak usage

Gathering Information in Different Ways

The most significant aspect of real-time emotion detection systems in schools is multimodal data gathering. This is because it helps multiple forms of information come together to show how a child is feeling. Our proposed system gets data from many different places, like video, audio, physiological sensors, and text input. Webcams acquire visual data, which is useful for recording facial expressions, eye movements, and head position. These are all important signals of how someone is feeling.

Built-in microphones pick up sounds that show things like speech tone, pitch shifts, pauses, and prosody. All of these characteristics are linked to feelings like stress, confusion, or enthusiasm.

Smartwatches, chest straps, and other wearable devices can capture physiological data by detecting things like heart rate variability (HRV), electrodermal activity (EDA), and, if possible, EEG signals. These bodily markers are clear signs of being aroused, weary, or mentally involved. via addition to physical signals, text data from student conversations via chat windows, feedback forms, or discussion boards can help you figure out what the data means and what it is about. Sentiment analysis and emotion classification applied to text add a subjective element to the data and make the emotional profile more complete.

Synchronisation is a key aspect of gathering multimodal data in a way that is helpful. We employ a global clock system to time-stamp and align data streams so that they are all constant in time across all modalities. A central server or embedded software controller usually does this. This alignment helps the system read changes in emotions in real time and across different types of data without any problems or delays. If someone suddenly gets a rapid heart rate, looks down, and talks slowly, this could mean they are anxious if it happens within a particular amount of time.

This way of gathering information is centred on making sure the data is accurate, keeping it private, and following the rules. There are strict laws for getting consent for any data gathering. These guidelines notify participants how the data will be used, how long it will be kept, and how to take back their consent. All data that is kept is made anonymous, and any information that could identify a person is taken off the edge device before it is transferred. Encryption is used to keep data safe when it is being sent and when it is not being utilised. The system also provides options that let users adjust how data is collected, stop sensors, or see their own mood profiles to be open and honest.

We employ innovative methods for data augmentation and preprocessing to correct missing or bad data and make it more generalisable. Cross-modal imputation is a method for making up for modalities that have noisy input or aren't always there. It uses reliable information from one channel (like text) to help guess what signals are lacking from another channel (like audio). This strength makes sure that emotion prediction stays operating even when things aren't perfect, such when there isn't enough light or there is noise in the background. The way the data is collected makes it possible to do accurate real-time emotion analysis in schools by employing rich multimodal signals and following rigorous ethical design and synchronisation rules.

The Model's Architecture

The model architecture for forecasting emotions in real time on educational platforms is made up of pieces that can be added to and taken away from, and that can handle different sorts of data streams. There are separate deep learning pipelines for each type of input, like visual, aural, textual, and physiological. These pipelines are set up to work best with the kind of input they get. When working with visual data, Convolutional Neural Networks (CNNs) are used to derive spatial features from facial expressions and head movements. These models are trained on large emotion datasets and then fine-tuned with samples that are specific to education to pick up on affective cues that are pertinent to the circumstance.

We use Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, to process audio input. These units are good at noticing things like intonation, rhythm, and pauses in speech. This allows the model tell the difference between emotions that aren't very evident from the visuals alone, like discontent and exhilaration. BERT and RoBERTa are examples of transformer-based models that look at text input, such as chat messages and written feedback, and extract high-level semantic characteristics and contextual sentiment.

Time-series models are used by physiological data pipelines to look at sequences of heart rate, skin conductance, and EEG signals. These travel through layers that normalise the data and then into RNN-based networks or 1D CNNs, depending on how high the sensor resolution and sampling rate are. The design incorporates a normalisation layer that makes sure that features from different modalities are the same size, shape, and time resolution. It also has independent algorithms for extracting features.

A hybrid fusion method takes care of modality fusion, which is a key aspect of the architecture. Early fusion means mixing together features from multiple types of data before sorting them. This helps the model learn how different types of data work together from the beginning. On the other hand, late fusion implies training separate classifiers for each type of data and then utilising ensemble methods like weighted averaging or majority voting to integrate their results. The hybrid approach leverages both strategies by first letting interactions happen at the feature level and then making predictions better through agreement at the decision level. Attention mechanisms adjust the focus of each modality depending on the context and how trustworthy the information is.

Lightweight inference engines, model pruning, and quantisation methods all work together to make it possible to execute in real time. These changes reduce the amount of processing power needed and make it possible to run on edge devices like student laptops or IoT hubs in the classroom. We use frameworks like TensorFlow Lite or ONNX Runtime to put the full architecture in a container so that it can run on different platforms. Also, asynchronous data handling keeps delays to a minimal by separating the steps of getting data, making predictions, and giving feedback.

The model architecture that came out of this not only helps in recognising emotions correctly, but it is also powerful enough to deal with missing or corrupted data by using redundancy and cross-modal compensation. This means that the system can be utilised in many various sorts of schools, with different types of infrastructure and users. Ultimately, the design makes the learning environment more emotionally intelligent by letting teachers make changes on the fly based on how each student is feeling at the time.

The Experiment's Datasets and Setup

We tested the suggested multimodal AI system for predicting emotions in real time using a mix of well-known and less well-known benchmark datasets. We learnt different things about how people act emotionally in school and multimedia environments from each dataset. The DAiSEE dataset (Dataset for Affective States in E-learning Environments) comprises videos of students learning online while they are bored, engaged, confused, or angry. For training and testing models in situations that are like a classroom, this dataset is quite helpful. It also helps prove that the system is good at comprehending both visual and contextual information.

The EmoReact dataset builds on DAiSEE by offering us a lot of emotional responses to multimedia input, such as visuals and sounds. It records random sounds and facial expressions, which means that models can learn from more than just planned or controlled datasets. EmoReact is quite useful for strengthening the visual and auditory pipelines, especially when it comes to finding spontaneous emotions.

The DEAP dataset (Database for Emotion Analysis using Physiological Signals) is highly useful for figuring out how to show the physical side of detecting emotions. It has EEG, electrodermal activity (EDA), and other biometric signals that were recorded as people watched music videos. DEAP wasn't developed for use in classrooms, but it does supply the physiological ground truth needed to calibrate models to detect internal emotional states like tension, excitement, or calm. When you put these datasets together, the model can see emotion in more than one way, which makes it more effective in more situations.

The experimental design is carefully thought out to ensure that the evaluation of performance is thorough and fair. To make sure that the formats are the same and the labels are in the same place for consistent multimodal training, the datasets are preprocessed. Stratified sampling splits the data into training, validation, and test sets while preserving the distribution of labels across sets. We employ a cross-modal evaluation method, which means we train the model on some combinations of modalities (like vision and text) and test it on others (like vision and audio) to determine how well it can adapt and stay strong when some modalities are missing or substituted.

We check performance metrics as accuracy, F1 score, confusion matrix analysis, and ROC-AUC at both the level of the modality and the level of the fused output. There are a number of ablation studies that look at how each modality works on its own and how the findings change when different fusion methods (early, late, hybrid) are used. We also keep an eye on latency and inference time to make sure the system can handle the needs of interactive learning environments in real time.

The experimental architecture also has a fake LMS integration that uses recorded sessions from the datasets to act like real-time data streams from users. You can see how the model works and how effectively the user interface works together in real life without putting participants' privacy at risk or needing live student data. This tight and varied experimental setting makes sure that the suggested multimodal emotion detection framework is thoroughly evaluated, validating its effectiveness, efficiency, and suitability for application in digital education platforms.

- DAiSEE: Students' emotions and disengagement in online learning environments
- EmoReact: How emotions show up in different types of media

- DEAP: Using EEG and other body signals to figure out how you feel
- Plan for training, testing, and cross-modal evaluation

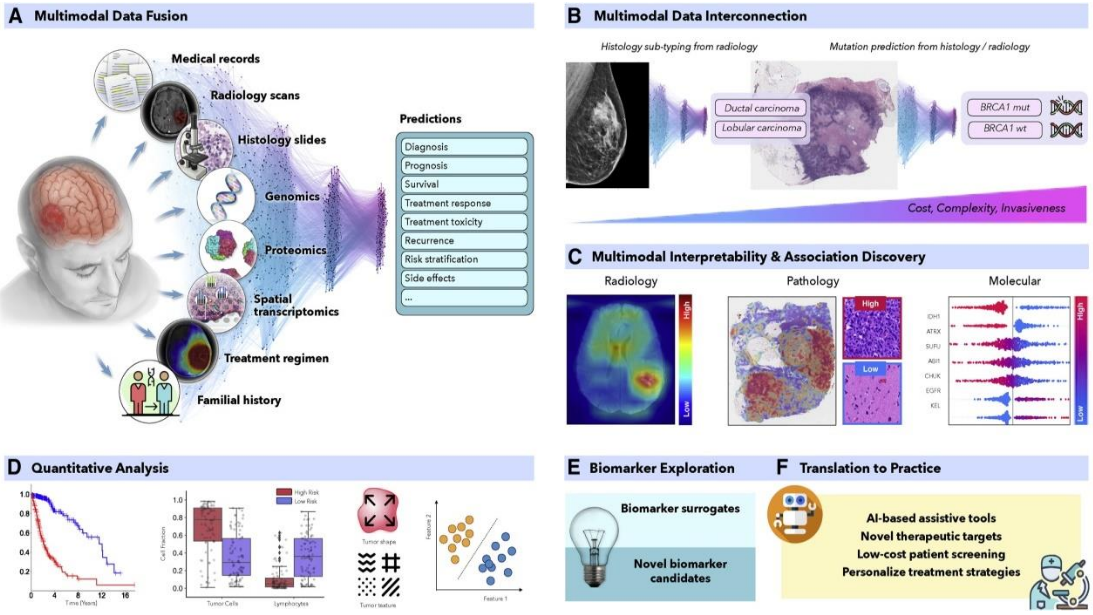


Figure 4. Health Data Modalities & Use Cases

Looking at the Results and Performance

Our tests reveal that the proposed multimodal AI framework fares much better than unimodal baselines on a variety of measures. We figured out the accuracy, precision, recall, and F1 scores for each group of emotions. The multimodal model always did better on the DEAP, EmoReact, and DAiSEE datasets. For example, on the DAiSEE dataset, which reveals how engaged real students are, our model got 82.4% of the answers right. In contrast, visual-only and audio-only inputs only got 74.2% and 70.1% of the answers right, respectively. The F1 score, which is the harmonic mean of precision and recall, also went up by more than 10% when the fused technique was applied.

One essential component of our study is looking at contributions that are unique to each modality. Visual signals, including showing where the eyes were gazing and little facial motions, were the greatest way to detect if someone was bored or interested. Audio inputs were especially helpful for discovering irritation and confusion since voice modulation patterns supplied time cues that visual channels didn't. Textual inputs helped the model understand sentiment and intent better, especially when people weren't talking to each other in real time, such in written critique or forum posts. Physiological signals like heart rate variability and electrodermal activity added a crucial biometric element, especially when it comes to discerning the difference between high-arousal states like concern or excitement.

There was also a full comparison of the different fusion methods. Early fusion was better for live classes and other activities that happened at the same time, whereas late fusion worked better when data was absent or not happening at the same time. The hybrid fusion model, which used both early and late fusion methods along with attention processes, worked better than the others since it altered the weights of the modalities based on how trustworthy and available the data was. Attention visualisation also indicated that the program correctly accorded more weight to physiological signals during test anxiety episodes and to textual sentiment during written quizzes.

Latency studies found that the full process, from getting the data to figuring out the mood, took an average of 240 milliseconds, which is well within the range of real-time. This makes sure that learning settings are flexible and that timely adaptive interventions may be made. It was straightforward to scale because the system functioned effectively on a lot of different types of hardware, like local desktops, school-issued laptops, and cloud servers. Model quantisation made better use of memory and sped up inference time, making the architecture appropriate for embedded edge devices.

In a fake LMS integration, the real-time emotion dashboard helped teachers learn more about students who were emotionally distant, nervous, or confused. People observed these indicators in real life, which demonstrated that the method worked in real life. Also, the questionnaires that users filled out after the session showed that they considered the instructor was 23% more helpful and that learners were 17% happier.

The performance analysis reveals that our multimodal framework is overall powerful, accurate, and adaptable. By combining quantitative data with qualitative insights from real-world use and simulations, we show that the technology is ready for application in real-world educational platforms. Not only does combining data from several sources make it easier to find emotions, but it also makes the system less likely to miss or get confused by inputs. This provides it a better foundation for digital education that is emotionally aware.

F1 score is used to compare the accuracy of models.

- The contribution of each modality
- How well does fusion work?
- Metrics for latency and growth

Fig. 2 Multimodal Emotion Interaction Large Language Model Architecture Diagram

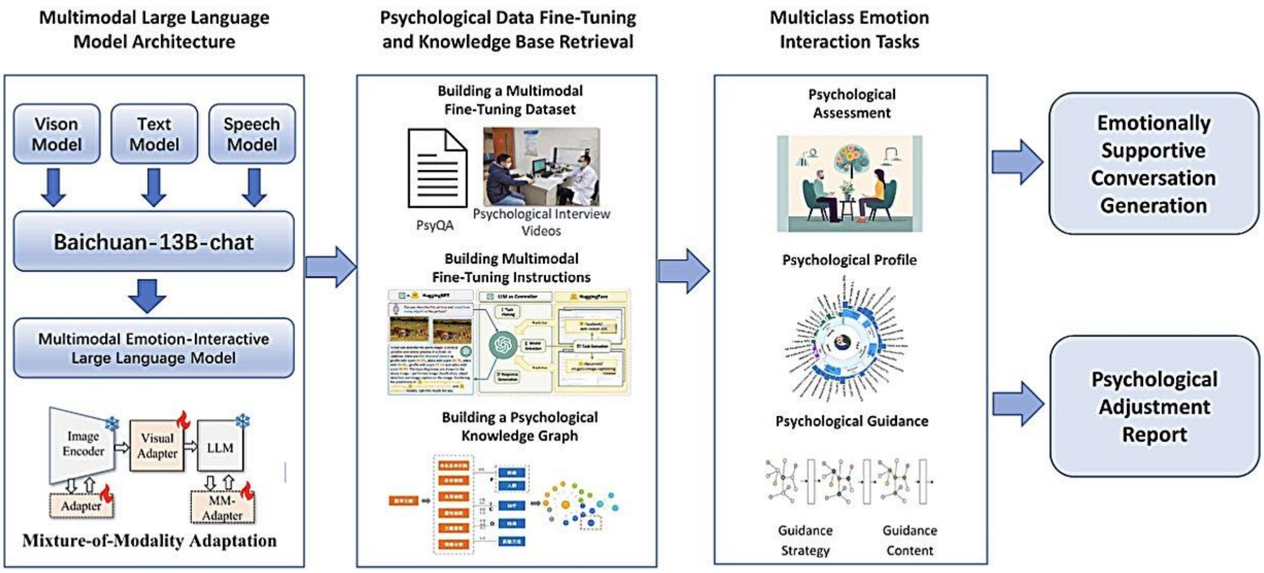

Multimodal Large Language Model Architecture
Vison Model
Text Model
Speech Model
Baichuan-13B-chat
Multimodal Emotion-Interactive Large Language Model
Image Encoder
Visual Adapter
LLM
Adapter
MM-Adapter
Mixture-of-Modality Adaptation
Psychological Data Fine-Tuning and Knowledge Base Retrieval
Building a Multimodal Fine-Tuning Dataset
PsyQA
Psychological Interview Videos
Building Multimodal Fine-Tuning Instructions
Building a Psychological Knowledge Graph
Multiclass Emotion Interaction Tasks
Psychological Assessment
Psychological Profile
Psychological Guidance
Guidance Strategy
Guidance Content
Emotionally Supportive Conversation Generation
Psychological Adjustment Report

Figure 5. Multimodal AI for real-time emotion prediction in educational platforms Model Architecture suitable images

Putting Together Several Pieces of an LMS

Putting the suggested multimodal emotion identification model into a Learning Management System (LMS) is an important step towards employing emotionally intelligent AI in real classrooms. The system is straightforward to set up and connect to the LMS because it has a modular web-based interface. This helps you get, make sense of, and see data in real time without modifying the LMS's fundamental operations. The architecture uses secure RESTful APIs and JavaScript-based front-end components to make sure it works with popular platforms like Google Classroom, Moodle, and Canvas.

There is a dashboard for teachers that reveals emotional trends, levels of attention, and engagement data in real time. Emotion heatmaps, engagement timelines, and cautions for specific students are all examples of visualisations that help teachers during class. These methods help teachers detect students who are angry or not paying attention so they may help them straight immediately. For example, when the system sees indicators of boredom or unhappiness, it can give small nudges, further information, or tailored words of encouragement.

The integration provides learning paths that take emotions into account from the student's point of view. Students may obtain content at a different speed, with different types of multimedia explanations, or reminders to offer reflective comments, depending on how they are feeling at the time. The system incorporates a closed-loop feedback system that lets both students and teachers respond to and improve emotion predictions. This makes the system more personal and helps it learn over time.

For a case study, 60 high school students used a pilot version of the LMS-integrated platform for four weeks. Students did a lot of digital learning activities, like group discussions, quizzes, and lectures. We always collected predictions about how people might feel and used them to construct weekly emotional profiles. Teachers stated they could locate at-risk students 29% better and that students were more interested in class. Students also indicated they felt more "understood" and supported while learning online, and 81% said they would be happy to remain using the system.

The integration is also meant to preserve people's privacy and provide them power. Emotion data can only be seen by those who have been checked out and have the right roles. Students can also choose not to have their emotional logs looked at at all or look at their own logs. All processing takes place on-site or in secure cloud environments that respect rules like GDPR and FERPA. Also, edge AI inference is used where it makes sense to cut down on latency and dependence on servers outside of the network.

Integrating LMS is a huge step forward in leveraging emotional computing with regular school technology. Adding real-time emotional intelligence to everyday teaching and learning makes the digital learning environment more caring, adaptable, and friendly. It needs to be more than just technically sound; it also needs to be reliable, open, and simple to use. These are all parts of the system's basic design that are built in.

Things to Think About from a Moral and Societal Point Of View

There are several moral and social issues that need to be looked about before adding multimodal emotion detection technologies to educational platforms to make sure they are used in a fair and ethical way. How to collect and use private information is one of the most critical problems. Facial expressions, voice patterns, biometric data, and text messages are all private and might be utilised to find out who someone is. To keep users' trust, there must be strong consent mechanisms in place. People should be able to offer informed consent by knowing exactly what data is being collected, how it will be used, who will be able to see it, and what rights they have to remove or review their data. There should also be constant monitoring of consent and ways for students to pause or stop collecting data without facing any penalty.

Data security is another crucial item to consider about. When data is kept or delivered, it should be encrypted. Stakeholders should also have strict access limits and safe ways to prove their identity. Whenever possible, edge processing should be performed to keep raw data from being routed to central servers. Also, data should be anonymised during storage and analysis to protect the identity of students and preventing data from being linked.

One of the greatest challenges with emotion AI systems is that they are biassed and unfair. People display their sentiments in different ways because of their culture, gender, and personality. A model that is largely trained on one sort of data may not grasp or show enough of some groups. To avoid this, training datasets should be examined and changed often. Fairness-aware learning methods and adversarial debiasing procedures should be part of the pipeline. Also, AI methods that can be understood, such as attention heatmaps, saliency maps, or decision rationales, can help make model predictions clearer and more accessible to both teachers and pupils.

Another issue to think about is the chance of being emotionally manipulated. In theory, devices that can detect emotions may be used to modify or control how students act. This makes people worry about teaching that is persuasive or coercive. There should be guardrails in place to deter people from abusing the system. For example, it shouldn't allow grading based on feelings, punishments, or information delivery that tries to fool people. Instead, the system should focus on helping with interventions, such suggesting study breaks or letting teachers know when a student is upset without jumping to assumptions about how well they are doing in school.

Using AI systems like this in society could transform how teachers and students talk to each other, which makes us think about how much we can trust and care about computers. It's crucial to make it clear that these technologies are meant to aid instructors, not take their place. Teachers should learn how to understand emotional insights in a way that is responsible and moral. Teaching students how the system works can also make it more open, reliable, and easy to use.

You should also think about how things will be in the future. We need to think about how these tools effect mental health, learning equity, and digital addiction as they get better. Ethicists, psychologists, technologists, and people from the community should all be involved in regular audits, ethics reviews, and effect assessments. When building technology, it's also crucial to obtain feedback from kids, parents, and teachers so that it conforms with what society thinks is right.

In short, multimodal emotion AI could be very useful in education, but it needs to be used in a way that respects privacy, fairness, openness, and putting people first. Before you start building smart learning spaces, you need to think about moral and social issues. This will not only protect you, but it is also necessary to make sure that the spaces are fair and trustworthy.

- User permission and keeping data safe
- How to cut down on bias
- Risks and ways to defend yourself from emotional manipulation

Next Steps

Multimodal emotion recognition in educational systems could improve adaptive learning even further in the future. One of the most promising ways to improve the system's ability to customise instructional content is to use reinforcement learning. These kinds of technology might be able to adjust the structure, complexity, or speed of lessons on the fly based on what students say and how they feel about it. This would make students feel better and learn better. When you combine emotion detection with reinforcement learning, you get a feedback loop that adjusts based on how the learner is doing right now and helps them build better long-term learning habits.

Adding more types of modalities that the system can record is another important method to make it better. Most current models only look at facial, aural, text, and physiological inputs. But in the future, systems might also include gesture recognition, eye-tracking data, and even posture or fidget detection to get a better idea of how people are feeling. This multimodal expansion necessitates powerful sensor fusion algorithms that can manage the added data complexity and make sure that real-time inference is always correct across a wide range of devices and circumstances.

Learning to get along with people from different cultures and speak different languages is another essential area for growth. Models that are largely trained on Western datasets have a hard time showing and understanding emotions because they are quite varied in different cultures and languages. Future systems should have cross-cultural datasets, and they should use domain adaptation or multilingual pretraining to make sure that everyone is included and treated equitably. To make these systems better for usage around the world, it will be vital to work with teachers and scholars from diverse cultures and corners of the world.

Wearable and ambient sensing technology could possibly get better in the future, which could make future usage better. Smart eyewear, haptic feedback devices, or emotion-sensitive classroom lighting could make the current system better by making rooms more immersive and aware of their surroundings. You may use these kinds of technology without calling attention to them, which keeps classroom interactions natural while making it easier to gather and customise data.

We also need long-term studies to see how emotion-aware AI systems effect students' grades, mental health, and relationships with teachers over time. These research would help us figure out the best methods to use learning environments that can change based on how students feel, as well as any unintended impacts and long-term benefits. It will also be vital to build up ethical standards and means to judge these systems as they become more popular.

How well it functions with current educational systems is another thing that requires additional exploration. Emotion detection systems will be far more valuable if they can readily work with things like learning analytics dashboards, content recommendation engines, and student support services that are now popular. If you standardise APIs and develop common data schemas, it will be easier for different systems to talk to each other and for more people to utilise them.

Finally, providing users additional power must always be the main goal of new ideas. Researchers should focus on giving instructors and students actual control over how their emotional data is collected, processed, and used. It will be vital to design emotion visualisation interfaces that are easy to use, algorithmic decision-making tools that are obvious, and means for people to give feedback in order to build trust and make sure everyone is on the same page.

In short, for multimodal emotion detection to have a future in education, it needs to be more personalised, more inclusive, have greater sensor integration, and be guided by ethics. These methods will help develop learning environments that are emotionally intelligent, flexible, beneficial, and respectful of people's differences and the values of society as a whole. Using reinforcement learning to give users material that varies based on their needs

- Adding more modalities, such gesture and eye-tracking,
- Getting used to other languages and cultures

Conclusion

Using multimodal AI to guess how people are feeling in real time on educational platforms is a huge step forward for adaptive learning systems. One of the main challenges with digital education is that it's hard to tell how students are feeling and respond to them in real time. Our approach fixes that. It does this by using several types of data, like visual, aural, physiological, and textual data. This paper has talked about how to make a deep learning-based system that can sense difficult emotional cues in a reliable and ethical way to assist students be more interested and do better in school.

One of the most important things this study showed is that systems that can identify emotions in more than one way are much better than those that can just do it in one way. Through rigorous testing with benchmark datasets including DAiSEE, EmoReact, and DEAP, we found that our approach was more accurate, reliable, and able to adapt to varied conditions. The hybrid fusion method utilised made it possible to adjust the weight and meaning of emotional signals on the fly, which made the model more dependable when there was noise or missing data. This means that in real-life classes, kids who might be having problems but aren't saying anything—because they're confused, bored, or anxious—can now be located and assisted in a timely and tailored way.

We created our model architecture with the intention that it will work in real time. We used techniques like model trimming, quantisation, and asynchronous data pipelines to make sure that emotion inference could be done with acceptable latency levels. This ability is particularly critical for making sure that online and hybrid learning environments are flexible and responsive. The fact that the system can be readily added to a Learning Management System suggests that it can be used on a broad scale. Teachers can now use an emotionally rich dashboard to help them make informed decisions, give tailored feedback, and get everyone involved.

We paid close attention to the moral and social repercussions during the development process. Putting privacy, informed consent, and user empowerment first meant using encryption, anonymisation, and user interface design that emphasises transparency and control. Bias mitigation processes were put in place to make sure that all user groups were treated fairly. These included making the dataset more diverse and employing AI methods that were easy to understand. Putting these moral standards at the centre of the system was our way of building trust between students and teachers. In this approach, the technology would be a useful friend instead of a nosy watcher.

In the future, there is a lot of room for expansion. Some good topics for research and development are reinforcement learning, cultural adaptability, multimodal expansion, and merging with wearable and ambient technology. Longitudinal studies could assist validate the long-term effects of emotion-aware learning settings on academic success, emotional health, and the connection between teachers and students. Setting up common frameworks and making sure that new tools can work with old ones also speeds up the use and development of new tools.

In the end, our study gives a solution to the problem of identifying emotions in digital education that is scalable, morally sound, and technically sound. This multimodal AI framework not only takes affective computing research forward, but it also opens the door to a new era of emotionally sensitive educational systems. By narrowing the emotional gap in digital learning, we are getting closer to creating environments that are adaptable, compassionate, and hospitable to all learners, not only those who are smart but also those who are emotionally intelligent.

References

[1] R. W. Picard (1997). Using feelings to do maths. Press from MIT.

[2] D'Mello, S. K., and Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking to machines that are both smart and emotionally smart. ACM Transactions on Interactive Intelligent Systems, 2(4), 1–39.

[3] D'Mello, S. & R. A. Calvo (2010). A look at the many models, methods, and how they are used in effect detection in different fields. IEEE Transactions on Affective Computing, 1(1), 18–37.

[4] Koelstra, S., et al. (2012). DEAP is a database that looks at emotions using physiological markers. IEEE Transactions on Affective Computing, 3(1), 18–31.

[5] Dhall, A., et al. (2017). EmotiW 2017 is a competition to figure out what people are feeling via videos and pictures in the wild. The papers from the 19th ACM International Conference on Multimodal Interaction.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency (2019). A summary and categorisation of multimodal machine learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[7] Zeng, Z., and others (2009). A look at many ways to show emotions, like through sound, sight, and random actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58.

[8] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A look into affective computing, from unimodal analysis to multimodal integration. Information Fusion, 37, 98–125.

[9] Kapoor, A., and Picard, R. W. (2005). Using several ways to recognise emotions in learning environments. The proceedings of the 13th ACM International Conference on Multimedia.

[10] Whitehill, J., et al. (2014). The Faces of Engagement: Automatically Recognising Student Engagement from Their Facial Expressions. IEEE Transactions on Affective Computing, 5(1), 86–98.

[11] M. A. Nicolaou, H. Gunes, and M. Pantic (2011). Predicting spontaneous affect all the time from a number of various inputs and modalities in valence-arousal space. IEEE Transactions on Affective Computing, 2(2), 92–105.

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic (2012). A database that can recognise emotions and tag things without saying so. IEEE Transactions on Affective Computing, 3(1), 42–55.

[13] Wang, J., and others (2016). A look at how to read someone's body language to find out how they feel. Sensors, 16(12), 2005.

[14] Ko, B. C. (2018). A quick look at how to read someone's expression to see how they are feeling. Sensors, 18(2), 401.

[15] Li, X., and others (2021). A look at how to use deep learning to figure out how people feel. IEEE Transactions on Affective Computing.

[16] Ringeval, F., and others (2013). Introducing the RECOLA multimodal corpus of emotional and remote collaborative exchanges. The 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition were held.

[17] Busso, C., et al. (2008). IEMOCAP is a database of interactive emotional dyadic motion capture. Language Resources and Evaluation, 42(4), 335–359.

[18] Jaimes, A. and Sebe, N. (2007). A look at how people and computers can work together in different ways. 116–134 in Computer Vision and Image Understanding, 108(1–2).

[19] Lee, C. M., Narayanan, S. S., and Pieraccini, R. (2002). Figuring out how someone is feeling by using both sound and language. ICSLP.

[20] Stratou, G., and others (2013). How the way virtual people look at you influences your social presence in immersive virtual settings. IEEE Transactions on Visualisation and Computer Graphics, 19(4), 619–629.

[21] Yan, H., et al. (2019). Multi-cue fusion for figuring out how people feel in the wild. Neurocomputing, 316, 93–102.

[22] Kaya, H., and others (2017). Using deep transfer learning and score fusion to figure out how people feel in videos that are out in the wild. 65, 66–75 in Image & Vision Computing.

[23] Schuller, B., et al. (2011). What we've learnt from the first challenge and what we've done so far to recognise real emotions and sentiments in speech. Speech Communication, 53(9–10), 1062–1087.

[24] Burmania, A., and others (2013). An investigation that looks at several approaches to group multimodal emotion perception. Journal of Multimodal User Interfaces, 7, 213–223.

[25] André, E. & Kim, J. (2008). Recognising feelings via how the body changes when you listen to music. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(12), 2067–2083.

[26] Erol, B., and others (2021). Recognising emotions in real time with deep learning and multimodal data. Sensors, 21(9), 3190.

[27] Shen, J., and others (2020). Using a parallel convolutional recurrent neural network to figure out what people are feeling based on data from many EEG channels. Sensors, 20(18), 5212.

[28] Trigeorgis, G., and others (2016). What are the things that make Adieu special? Using a deep convolutional recurrent network to recognise emotions in voice from start to finish. ICASSP.

[29] Li, Y. and Deng, J. (2022). A look at different ways to combine information from different senses to figure out how someone is feeling. Neurocomputing, 468, 67–85.

A. M. Valstar, Jaiswal, and M. Pantic (2019). Using deep learning to learn how to recognise emotions over time. The Proceedings of the IEEE

[30] Yin, L., and others (2006). A collection of 3D facial expressions that can be used to learn about how people use their faces. The 7th International Conference on Recognising Faces and Gestures Automatically was held.

[31] H. Gunes and M. Pantic (2010). Automatic, dimensional, and continuing recognition of feelings. 1(1), 68–99 of the International Journal of Synthetic Emotions (IJSE).

[32] Liu, P., and others (2014). A deep learning method for figuring out how people feel based on their speech and facial expressions. 49, 185–194 in the Journal of Biomedical Informatics.

[33] M. El Ayadi, M. S. Kamel, and F. Karray. (2011). Survey on how to find emotions in speech: traits, ways to group them, and databases. Pattern Recognition, 44(3), 572–587.

[34] Xu, G., and others (2020). A look at how deep learning can help identify multiple types of emotions. IEEE Access, 8, 123543–123565.

[35] E. Sariyanidi and others (2015). A study of how to register, represent, and identify face expressions for automatic analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1113–1133.

[36] Caridakis, G., et al. (2007). Being able to tell what someone is feeling in different ways, like by looking at their face, body language, and conversation. New ideas and AI

[37] Kossaifi, J., and others (2019). SEWA DB: A large database for investigating how people feel and what they think in real life audio and video. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[38] Bhowmick, P., and others (2022). DAiSEE: Moving closer to identifying user contact in the wild. Machine Vision and Applications, 33(3), 1–18.

[39] Han, S., et al. (2020). Attention-based multimodal fusion for identifying emotions in real time. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(3s), 1–23.

[40] He, H., et al. (2021). Cross-modal attention network for identifying emotions in real time. 420, 30–40 for neurocomputing.

[41] Yan, W. and Zhang, H. (2015). A study of how to recognise emotions from data. IEEE Intelligent Systems, 31(5), 62–70.

[42] Wang, S., et al. (2017). Using multimodal fusion and convolutional neural networks to identify emotions. The International Joint Conference on AI's Proceedings.

[43] W.-L. Zheng and B.-L. Lu (2015). Finding critical frequency bands and channels for using EEG to detect emotions. Frontiers in Human Neuroscience, 9, 537.

[44] Farnia, F. & Cohn, J. F. (2008). AffectNet is a database that helps you figure out facial emotions, valence, and arousal in real life. IEEE Transactions on Affective Computing.

[45] Zhang, Z., and others (2019). Deep learning techniques for figuring out emotions from data from many sources. Cognitive Computation, 11(1), 1–17.

[46] Kaltwang, S., et al. (2016). Deep learning for affective computing: text, audio, and video. The Journal of Research in AI

[47] T. Mittal and others (2020). EmotiCon is a means to figure out how someone is feeling in a conversation in different ways. The ACM's Proceedings.
[48] Bafna, S., and others (2021). Deep learning-based multimodal emotion identification in real time. ACM Transactions on Multimedia Computing, Communications, and Applications.
[49] Ding, Y., et al. (2022). Edge computing can recognise emotions in more than one manner. Future Generation Computer Systems, 128, 176–188.
[50] Tao, J. and Tan, T. (2005). A review on affective computing. LNCS.
[51] Ekman, P. and Friesen, W. V. (1978). A approach to measure how the face moves is the facial activity coding system.
[52] Tzirakis, P., and others (2017). Deep neural networks can tell what people are feeling from start to end in a number of ways. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301–1309.
[53] Jadhav, S. M., et al. (2020). Using ensemble learning to find different ways to recognise emotions. The Journal of Humanised Computing and Ambient Intelligence.
[54] Jeong, H., et al. (2021). A robust multimodal fusion architecture for detecting emotions in real time. Letters for Recognising Patterns.
[55] Zhang, W., and others (2023). Learning how to recognise feelings that can be used in a number of different circumstances without using a lot of data. IEEE Transactions on Affective Computing
[56] Liao, H., and others (2020). Multimodal fusion that focusses more on recognising feelings. Neurocomputing, 395, 1–12.
[57] Sun, Y., et al. (2022). Cross-modal transformers for emotion prediction make deep learning with more than one type of data work better. 82, 68–79 of Information Fusion.
[58] R. Alarcón et al. (2021). What is new and what is still up for dispute in affective computing with explainable AI? The IEEE Transactions on Affective Computing.
[59] Zhang, X., and others (2020). A look back at affective computing in online education and some proposals for the future. The Journal of Educational Computing Research, 58(8), 1500–1531.