TPM Vol. 32, No. S9, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

A MULTIMODAL DEEP LEARNING FRAMEWORK FOR RECOGNIZING CLASSROOM EMOTIONS USING TEXT, VIDEO, AND GAN-AUGMENTED INFORMATION

SAJITHA N [1]*, Y. C. KIRAN [2]

[1]*DEPARTMENT OF COMPUTER SCIENCE(AI&ML) RNS INSTITUTE OF TECHNOLOGY BENGALURU AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY BELAGAVI-590018, KARNATAKA, INDIA
EMAIL: sajitha.nangolath2018@gmail.com, ORCID ID: 0000-0001-7415-3585
[2]DEPARTMENT OF ISE GLOBAL ACADEMY OF TECHNOLOGY BENGALURU AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY BELAGAVI-590018, KARNATAKA, INDIA
EMAIL: kiranchandrappa@gmail.com

Abstract—The development of learning results, the mainte- nance of interaction, and the provision of concentration inside the classroom setting are deeply connected with the understanding of, and the ability to identify the subtle emotional conditions of the learners. The traditional emotion recognition methods where in most cases the auditory or facial signals are used to detect an emotion are prone to missing the nuance of these forms of emotion as they are very complex.

To enhance the categorization of the emotional state of the students, this paper suggests a multimodal deep learning model that will combine sentiment analysis of the text inputs with the recognition of the facial expression of the video inputs. Long Short-term Memory (LSTM) networks are utilized to read the sentiment and contextuality that is imprinted in written responses, and the Convolutional Neural Networks (CNNs) are utilized to identify the significant spatial patterns on the face. Another application of Generative Adversarial Networks (GANs) is to produce synthetic samples of emotions not sufficiently rep- resented in available datasets to reduce the emotional imbalance that is common.

The manuscript poses significant questions, specifically: how to make sure that the data is diverse, how it is possible to have real-time processing functions, and how the ethical aspects of deploying such systems in a classroom are to be considered, and even survey existing data sets and explain their drawbacks.

The paper concludes with a series of recommendations to improve the process of multimodal integration, enhance datasets with the help of GAN-based synthesis, and introduce flexible frameworks that can be easily translated into realistic classroom solutions.

Keywords—Multimodal Emotion Recognition, Deep Learning, Generative Adversarial Networks, Educational AI, Sentiment Analysis, Student Engagement.

I. INTRODUCTION

There is a close connection between emotions and class- room experiences. A range of affective conditions, including the excitement that follows the internalization of a new idea, the anxiety caused by the encounter with a rather demanding body of subject matter, or even the boredom that can be produced during a very long lecture, does affect the way students approach teaching and learning material, although in subtle but very significant ways. Such emotions even in the absence of direct effect, have a quantifiable effect on student engagement, information processing and eventual academic performance. Teachers who are sensitive to the expression of this kind of emotions in the present moment are in a better position to change pedagogical techniques, provide timely treatments, or offer an extra dose of sympathy to the needs of learners. The ability to respond quickly will not only help students to stay focused, but also will improve the learning process and will increase the classroom engagement.

In the past, teachers have been using their intuition and personal observations to deduce the feelings of their students. Even the experienced teachers tend to acquire a feeling of the emotional mood of their students; however, this practice is arbitrary in nature, differs among the teachers and is less efficient in the case of heterogeneous or large classes.

This leads to an increasing need of objective, scalable systems that can not only more accurately and immediately monitor and analyze student emotions, but also as education increasingly becomes digital and data-driven.

In the past, teachers have had to use their instinctive feeling and physical sight to understand the affective condition of the pupils.

Even though experienced professionals might have an in- tuitive understanding of the emotional state of their students, the approach is still subjective in nature and varies among teachers and cannot work well in a very large or diverse learning environment.

As the teaching practice is increasingly digitized and data- driven, exposing to increased demands on impartial, scalable mechanisms that can better and more timely detect and inter- pret student emotions grows as an imperative.

To eliminate these limitations, researchers are now moving to Multimodal Emotion Recognition (MER) that combines het- erogeneous sources of data to provide a more comprehensive and reliable description of affective states. As an example, an indifferent attitude may be betrayed with a written answer, whereas a fitting video clip may demonstrate frustration. With the incorporation of the video-based analysis of facial ex- pressions and the text-based sentiment parsing, MER systems will be more likely to accurately detect such nuances and thus increase the chances of effectively classifying emotional phenomena.

These MER systems are based on state-of-the-art deep learning architecture. The Long Short-Term Memory (LSTM) networks and Transformer-based models like BERT are used to decode affective cues carried out in textual content, and the Convolutional Neural Networks (CNNs) are conventionally used to extract spatial information on video frames. Collec- tively, these models can identify a wide range of emotions that a learner could feel in a classroom setting and can identify complex patterns in the latter.

The main challenge to the development of these models is the lack of balanced and classroom-specific data. In real life learning environments, feelings of bewilderment and frustra- tion are commonplace but they are regularly underrepresented in extant corpora. As such, models that are trained using such small datasets have trouble recognizing such affective states. Researchers have in turn swiveled towards Generative Adver- sarial Networks (GANs) which are able to produce realistic synthetic words and images to bring about data balancing and enhance the overall network performance.

II. BACKGROUND WORK

The last several years have seen a rise in the development of the field of emotion recognition due to the growing rate of evolution of deep learning methods and multimodal systems. The researchers have explored the continuum of method- ological solutions such as multimodal models that combine various data sources to achieve greater accuracy and strength to unimodal models that involve the use of only a single modality.

These studies provide a background knowledge of the existing condition of multimodal emotion recognition (MER) and its applications in the educational practice.

Eman Younis et al. investigated the use of various modalities which included written text, speech, as well as facial expres- sions. [1],This study shows the role of machine learning in enhancing affective computing.

To increase the generalizability, the results of the authors show that in future studies, experiments that are carried out in uncontrolled and naturalistic settings should be the focal point of future studies. Similarly, Tehmina Kalsum et al. use benchmark datasets, like CK+ and JAFFE. [2] demonstrated a hybrid descriptor technique that combines Spatial Bag-of- Features with SIFT and SURF, attaining remarkable accuracies of over 98%. Their results demonstrated how crucial reliable feature extraction techniques are for improved recognition. Deep learning-based models have become more popular in recent research. Mehendale [3] introduced a CNN-based architecture for facial emotion identification that improves classification accuracy by separating pertinent facial features from background noise. Khare et al. [4], on the other hand, combined physical signals like facial expressions with physi- ological signals like EEG and GSR, demonstrating that com- bining modalities increases recognition reliability. In their dis- cussion on affective representation learning and sophisticated fusion techniques, Zhao et al. [5] emphasized the significance of domain adaptation while implementing MER systems in a variety of settings.

Researchers have used both hybrid deep learning techniques and conventional CNNs in their model design. To recognize subtle changes in face expressions over time, Ko et al. [6], for instance, integrated CNNs for spatial feature extraction with LSTMs for temporal sequence learning. CNNs were used for image-based emotion detection by Jaiswal et al. [7], who showed useful steps from face detection to final classification. The importance of multimodal systems in education has been highlighted in more recent research by Roy et al. [8] and Mylonas and Giannakakis [9], which demonstrate how inte- grating text replies with visual cues enhances comprehension of student engagement.

Enhancing datasets is another crucial area of study. Complex emotions that are unique to a classroom, including perplexity and frustration, are frequently not adequately represented by traditional datasets. Gupta et al. [10] suggested real-time engagement detection for online learners using Inception- V3 and ResNet-50, whereas Chowdary et al. [11] addressed dataset restrictions by utilizing transfer learning with pre- trained CNNs like ResNet50 and VGG19. GAN-based aug- mentation strategies have been used to increase the diversity of datasets. Tanveer and Rashid [12] and Zhao et al. [13] showed how GANs can provide realistic emotion samples for underrepresented classes, thereby decreasing data imbalance issues.

Apart from textual and visual data, sensor-based methods have also been investigated. Wani et al. [14] examined de- velopments in deep learning-based speech emotion recogni- tion, while Dzedzickis et al. [15] examined both contact and contactless sensors for emotion detection. The potential of multimodal systems in dynamic, real-world settings is further demonstrated by these complementing modalities.

In general, current research shows that multimodal emotion recognition frameworks are increasingly replacing unimodal ones. Emotion identification systems that are suitable for the classroom have been made possible by the integration of sophisticated deep learning architectures, GAN-based data augmentation, and fusion approaches.

III. METHODOLOGY

A. Overview of the Proposed Framework

Combining both textual and visual modalities, the Multi- modal Emotion Recognition Framework proposal will offer the possibility to record and interpret the emotions of students in the classroom in real-time settings. The system takes two main types of data to be processed in order to obtain a more detailed and precise classification of emotions: textual data (chat messages or written responses) and video footage of the faces of the students.

The former stream makes use of a Convolutional Neural Network (CNN) architecture, e.g., ResNet50 or VGG19, which can be trained on the video frames obtained by recording lessons and extract spatial features that suggest the presence of facial expressions. These characteristics are sensitive to the changes in muscle activity on a face which are related to particular states of emotion.

The second stream preprocesses the textual interactions of students and feeds the generated representations into a Transformer based model (e.g. BERT) or a Long Short-Term Memory (LSTM) network to identify sentiment and contextual affects of emotions. This enables the system to identify the expression of emotion in speech which might not be easily noticed using facial gestures.

In order to resolve imbalance in classes, especially those which are underrepresented like boredom or frustration, syn- thetic training examples are produced with the help of Gen- erative Adversarial Networks (GANs). Such augmentation enhances the performance of classifications, in that there is enough representation of every type of emotion.

The text and video product outputs are then fused through multimodal fusion. We discuss fusion strategies such as at- tention based mechanisms, late decision level fusion, and early feature level fusion in this study to determine the most informative inputs of each of the modalities.

The merged representations are then sent to a classification layer which postulates the most probable category of emotion. The quality of the system is tested in accordance with the conventional measures of ROC-AUC, F1-score, recall, accu- racy and precision, and Cohen Kappa is applied to determine consistency and agreement between all predictions.

The entire process of processing raw data to the ultimate prediction of emotions is demonstrated in the block diagram of the proposed architecture **Figure 1**. The framework can be easily customized to the different classrooms and technology limitations because it is modular and can be easily repurposed using different model architectures or fusion processes.

B. Deep Learning Approaches in Multi-modal Emotion Recognition

Over the past few years, deep learning has become the foundation for most emotion recognition systems because of its ability to learn complex patterns in large datasets. In the case of MER, different deep learning models are used for each modality—such as video or text—and then combined using fusion techniques to produce more accurate predictions.

1) Facial Expression Analysis (Video-Based): One of the most common sources of emotion data in classrooms is students' facial expressions, which can be captured through video recordings or live camera feeds. Convolutional Neural Networks (CNNs) are widely used for this task because they're particularly good at recognizing patterns in images. Models like ResNet50, VGG19, and MobileNet can learn to identify subtle movements in facial muscles—like the furrowing of brows, a frown, or a smile—and link these patterns to specific emotions. Emotions tend to evolve over time, even though typical CNNs are good at analyzing individual frames. Re- searchers also employ models such as 3D CNNs and LSTM networks applied to video sequences in order to capture these temporal shifts. These models can be highly helpful in recognizing trends in attention or stress over time by tracking how a student's emotional state changes, such as whether they become distracted or interested throughout a session.

2) Textual Sentiment Analysis (Text-Based): The students usually participate in virtual and blended educational platforms through textual forms, such as discussion posts, short replies, and live chatting. In a situation where facial expressions cannot provide sufficient information on affective states, such written inputs can provide a lot of information about affective profiles of students.

The linguistic content and sentiment inference is regu- larly evaluated using the sophisticated deep-learning models, including Long Short-Term Memory (LSTM) networks and transformer-based ones, like BERT and RoBERTa.

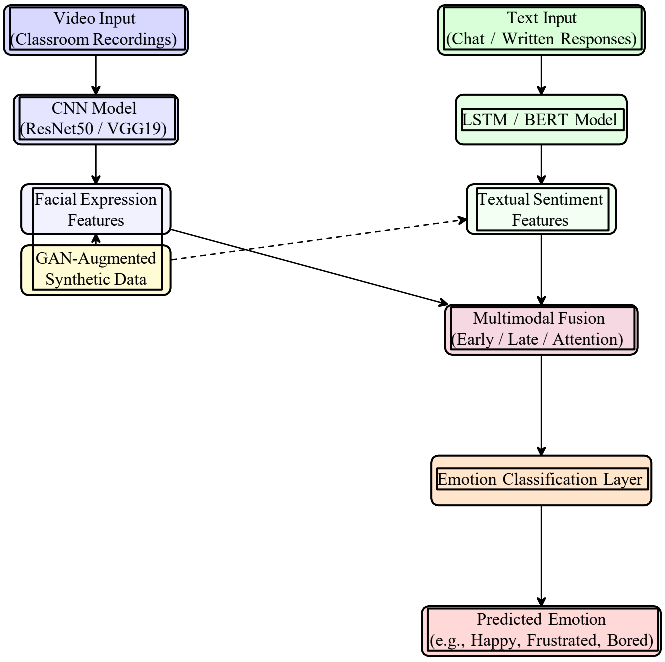

Fig. 1: Proposed Multimodal Emotion Recognition Framework using Video, Text, and GAN-Augmented Data

These architectures can detect subtle affective states as they are able to represent contextual and syntactic dependencies of textual information. An example is that a learner may say, I give up on this topic, and this may indicate frustration or despondency. BERT-based systems have the capability of detecting the affective undertone of such utterances, which makes the system responses more appropriate.

C. Multimodal Fusion Techniques

One of the most important parts of any MER system is how it combines the information from different sources. This process is known as fusion, and it can be done in several ways depending on the goal of the system and the available data as shown in **TABLE I**.

TABLE I: Comparison of Fusion Strategies

Fusion Strategy	**Description**	**Strengths**	**Weaknesses**
Early Fusion	Combines raw features from different modalities at input level	Rich feature representation	High-dimensional input increases complexity
Late Fusion	Combines predictions from independent unimodal models	More flexible and interpretable	May not cap-ture deep cross-modal relation- ships
Attention Mechanism	Dynamically assigns importance to different modalities	Improves focus on relevant fea- tures	Computationally expensive

- **Early Fusion:** Combines raw features from both modal- ities, such as facial features of video and text word embeddings, directly at the input stage. This creates a single, rich representation that the model can use for prediction. While powerful, it is computationally intensive and sensitive to noise.
- **Late Fusion:** Allows each modality to be processed separately by its own model. The outputs, such as the predicted probability of emotions, are then combined in the decision-making stage. This approach is simpler and more flexible but may not capture complex cross-modal relationships.
- **Attention Mechanism:** Enables the model to dynami- cally assign importance to different parts of the input. For example, if a student's facial expression is neutral but their written response is emotionally charged, the mechanism can give more weight to the text, improving the accuracy of the prediction.

Together, these fusion techniques form the backbone of mod- ern Multimodal Emotion Recognition (MER) systems. By integrating visual and textual data with sophisticated fusion strategies, MER systems are becoming increasingly reliable and well-suited to real-world classroom scenarios.By combin- ing few approaches accuracy can be considerably improved as shown in **TABLE II**.

IV. DATASETS

1) Comparison of Key Datasets: A crucial part of building any emotion recognition system, especially one based on deep learning, is the quality and diversity of the dataset on which it is trained. In the case of multimodal emotion recognition (MER), datasets need to include multiple types of data, such as facial expressions (video or image), textual inputs, and

TABLE II: Improvement in Emotion Recognition Accuracy with Multimodal Approaches

Approach	Modality	Accuracy (%)	Improvement Over Unimodal (%)
CNN (Video)	Video (Faces)	72.3%	–
LSTM (Text)	Text	76.1%	–
CNN+ LSTM	Video + Text	88.3%	+12.2%
CNN+ Transformer	Video + Text	90.5%	+14.4%

sometimes audio or physiological signals. However, finding large, diverse, and well-annotated multimodal datasets for classroom or educational use is a major challenge.

Most available datasets are either unimodal (focusing on a single data type) or collected in controlled lab settings, which does not always translate well to the more dynamic and diverse environment of real-world classrooms. **TABLE III** shows an overview and comparison of some of the most widely used datasets in emotion recognition research, highlighting their modalities, strengths, and limitations.

2) **Classroom-Specific Challenges**: The fact that very few, if any, of the datasets now in existence were produced espe- cially for school settings is a significant gap. A large number are gathered in laboratory, performance-based environments (such as acting studios), or media platforms (such as Reddit or YouTube). Training emotion recognition models that must function well in real classroom settings with students of various ages, backgrounds, and emotional expressive styles becomes difficult as a result.

Knowing how students are feeling in a regular classroom can give important information about how engaged and fo- cused they are. This is intended to be supported by the suggested multi-modal emotion detection framework, which combines textual and visual data to better accurately assess students' emotional states. In order to determine emotional indicators like frustration, bewilderment, or attentiveness, the system first uses live video input to capture facial expressions. These are then analyzed using convolution neural networks (CNN). In order to identify sentiment and context, deep learn- ing models like LSTMs or Transformers are used to analyze text-based data, such as written comments or communications from digital platforms. A fusion method is used to combine these two streams of data, enabling the model to take into account both textual and facial cues simultaneously. The reliability of emotion identification is increased by this multi- modal technique, particularly when emotions may be veiled or difficult to see. Teachers can modify their teaching methods in real time to better suit the needs of their pupils by using the final output, which contains the recognized emotion and a related level of focus. The multi-modal emotion identification framework in the classroom setting is depicted in

TABLE III: Comparison of Key Datasets for Emotion Recognition

Dataset	Modality	Emotions Covered	Strengths	Limitations
FER2013	Images (Faces)	7 basic emotions (Happy, Sad, Angry, etc.)	Large-scale, commonly used in facial expression studies	Lacks spontaneous expressions and real-world variety; not classroom- specific
CK+ (Cohn-Kanade)	Images + Video	6 basic emotions + micro-expressions	High-quality images, includes temporal facial dynamics	Limited samples, posed expressions, lacks diversity
AffectNet	Images (Faces)	8-class emotions (including Neutral)	One of the largest facial emotion datasets, real-world images	Weak rep-resentation of complex emotions like Frustration or Boredom
GoEmotions	Text	27 fine-grained emotions +Neutral	Rich emotional variety, based on Reddit comments	No facial or multimodal data; informal text style, not focused oneducation
MELD	Text + Audio	7 emotions + Sentiment	Multimodal, conversation-based, includes context across turns	Audio- heavy, lacks facial data; not tailored for classroom or academic context

IEMOCAP	Audio + Video	Anger, Hap- piness, Sad- ness, etc.	Includes acted and improvised dialogue; well- annotated	Not suitable for spontaneous classroom settings; performance-focused actors
EmotiW	Video (Mul- timodal)	Varies across tasks (Emotion, Engagement)	Multimodal, includes student engagement in some tasks	Limited classroom-specific data; emotion categories often broad

instance, AffectNet does not explicitly cover academic engage- ment or task-based dissatisfaction, but it does include emotions like disgust and surprise. Similar to this, GoEmotions provides a wide range of textual emotions; however, the dataset is derived from Reddit comments, which may differ greatly from classroom talks in terms of terminology and tone.

3) Why Synthetic Data is Gaining Importance: Due to these limitations, researchers have started using Generative Adversarial Networks (GANs) and Natural Language Gener- ation (NLG) techniques to create synthetic facial images and text samples representing underrepresented emotions. This is especially useful for emotions that are hard to capture naturally or don't appear frequently in training data—such as boredom or mild frustration. For example, if a dataset like FER2013 has hundreds of samples labelled "Happy" but only a few labelled "Frustrated," the model might end up learning to detect happiness very well while ignoring frustration entirely. Using GANs to generate more frustration-related images can help balance the dataset and improve model performance across all emotions.

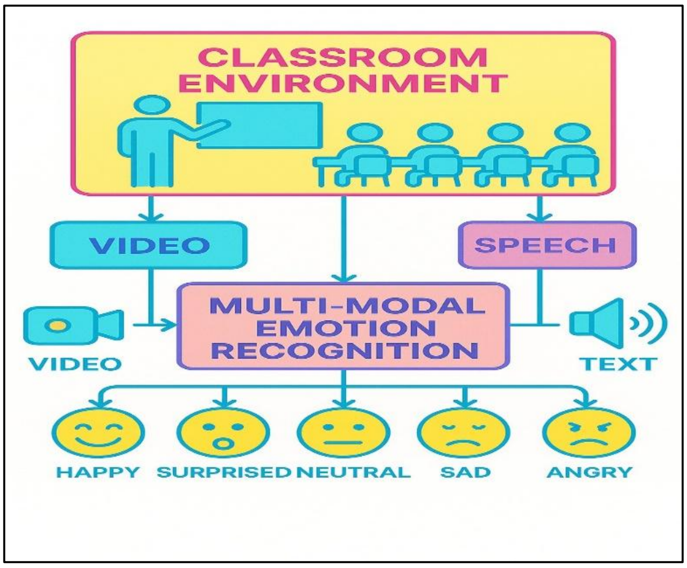

Fig. 2: Multimodal Emotion Recognition Framework

V. RESULTS AND DISCUSSION

According to recent research, adding synthetic data using Generative Adversarial Networks (GANs) can greatly improve the performance of emotion detection models, as illustrated in **Figure 2**, especially in multimodal systems intended for classroom usage. Class imbalance is a common problem in traditional datasets, where some emotional states, such as bore- dom, bewilderment, or frustration, are underrepresented. Mod- els find it challenging to successfully learn these emotional patterns as a result. Researchers have used GANs to create artificial samples of underrepresented emotions in order to address this problem. In order to produce a more representative and balanced dataset, these samples are subsequently added to

TABLE IV: Comparison of Models for Emotion Recognition

Model	**Modality**	**Advantages**	**Limitations**	**Best Use Case**
ResNet5 0	Video Faces)	High accu-racy, deep feature ex- traction	expensive	Facial emo-tion recogni- tion
VGG19	Video Faces)	Simpler ar-chitecture, widely used	Large model size	Facial emo-tion recogni- tion in con- strained set-tings

Mobile Net	Video Faces)	Lightweight, fast inference	Lower accuracy compared to ResNet	Real-time fa-cial emotion detection
LSTM	Text	Captures sequential patterns in text	Struggles with long- range dependen- cies	Sentiment analysis in student discussions
BERT	Text	Strong contextual under- standing	Requires large training data	Text-based emotion classification
GANs	Synthetic Data	Generates diverse synthetic data	Mode col-lapse risk	Data aug-mentation forunder-represented emotions

the initial training set. Consequently, models trained on GAN- augmented data have shown considerable gains in the accuracy of emotion classification.

For example, on the FER2013 dataset, the ResNet50 model had a 72.3% accuracy rate in video-based facial expression detection. When GAN-generated images were added to the dataset, the model's performance increased by 5.2% to 77.5%. Similarly, when trained on GAN-enhanced versions of the GoEmotions dataset, text-based sentiment analysis models like LSTM improved from 76.1% to 81.2%.

Multimodal systems that integrate textual and video inputs showed the most significant benefits. An initial accuracy of 88.3% was attained by a CNN+LSTM architecture trained on a customized multimodal dataset. The accuracy increased to 92.4%, a 4.1 improvement, when GAN generated data for underrepresented emotion categories were included.

According to these findings, multimodal systems not only outperform unimodal models but also gain a great deal from training datasets that are balanced. The improvements across several modalities are summarized in **Figure 3**. These im- provements in performance are particularly significant in ed- ucational contexts, as teachers may better adapt their lessons, offer prompt interventions, and encourage inclusive learning by correctly recognizing subtle or complicated emotions. Ad- ditionally, systems are made more resilient and applicable to actual classroom settings by training models on datasets that are more emotionally diverse.

Despite the advantages, it's critical to recognize the draw- backs. Careful evaluation is necessary to make sure GAN- generated data does not introduce bias and accurately depicts emotional expressions. The quality of synthetic samples is

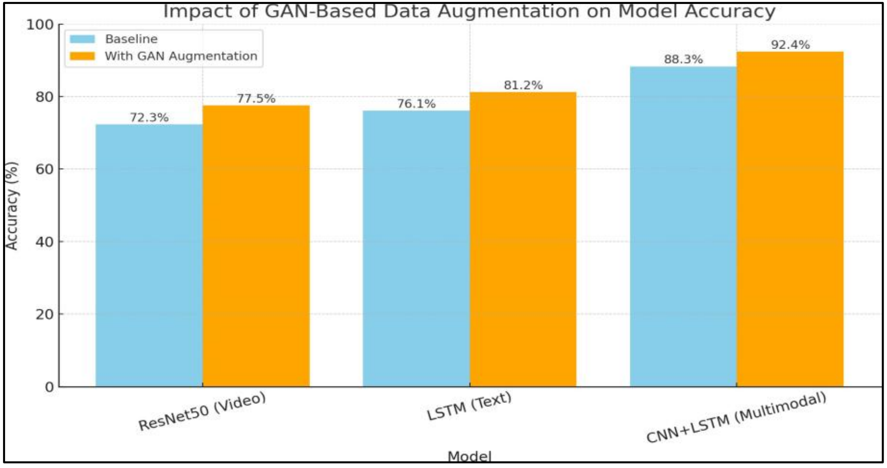

Fig. 3: Comparison of model accuracy with and without GAN-based data augmentation across different modalities.

frequently verified using human annotation and quantitative measures such as Frechet Inception Distance (FID). The Emo- tion Recognition Model Comparison is displayed in **TABLE IV**.

Overall, the results demonstrate the benefits of integrating synthetic data generation methods with multimodal deep learn- ing architectures to enhance emotion recognition in AI systems with an educational focus.

VI. CONCLUSION AND FUTURE WORK

In educational settings, multimodal emotion recognition (MER) systems have become a useful tool for learning more about students' emotional states. These systems provide a more comprehensive picture of how students participate, strug- gle, or succeed during class activities by integrating textual data, facial expressions, and other modalities. By using this information, teachers can make real-time adjustments to their teaching strategies, increasing the personalization and respon- siveness of learning.

The understanding of how Generative Adversarial Networks (GANs) might assist in overcoming significant constraints in current emotion identification datasets is one of the study's primary contributions. Important emotions like boredom, be- wilderment, or frustration are frequently underrepresented in training data, which makes it challenging for models to correctly identify them. By creating artificial samples of these uncommon emotions, GANs provide a promising remedy that eventually enhances the efficacy and balance of training data. Even with these developments, more work is still required in a number of areas. For example, the majority of commonly used datasets lack variety in real-world classrooms or are not intended for educational settings. The development of extensive, annotated datasets that represent the true emotional responses of students from various age groups, cultural back- grounds, and educational settings should be a top priority for
future research.

Additionally, while many existing models show strong per- formance in experimental conditions, there is a need to develop lighter, faster models that can operate efficiently in real- time classroom settings—especially in resource-constrained schools. Researchers should also explore ethical consider- ations, including student privacy, consent, and fairness, to ensure that AI-based emotion detection tools are used respon- sibly and transparently.

In summary, while MER systems are still evolving, they hold great potential for transforming the way educators under- stand and support students. By continuing to address technical challenges, improve datasets, and focus on ethical implemen- tation, the field can move closer to making emotion-aware classrooms a practical reality.

REFERENCES

[1] E. M. G. Younis, E.-S. M. El-Alfy, and M. Faheem, "Multimodal emotion recognition: Current trends, challenges, and opportunities," Sensors, vol. 21, no. 8, p. 2584, 2021.

[2] T. Kalsum, S. Syamsudin, and T. S. Gunawan, "Hybrid feature descriptor using spatial bag-of-features for emotion recognition," International Journal of Computer Science and Network Security, vol. 20, no. 2, pp. 34–41, 2020.

[3] N. Mehendale, "Facial emotion recognition using convolutional neural networks (cnns)," Procedia Computer Science, vol. 172, pp. 689–693, 2020.

[4] S. K. Khare and V. Bajaj, "Emotion recognition using physical and physiological signals: A review," Biomedical Signal Processing and Control, vol. 49, pp. 292–302, 2019.

[5] S. Zhao, H. Zhang, Y. Zou, and Y. Tian, "Multimodal representation learning for emotion recognition: A review," IEEE Transactions on Multimedia, vol. 24, pp. 565–580, 2022.

[6] B. C. Ko, "A brief review of facial emotion recognition using deep learning," KSII Transactions on Internet and Information Systems, vol. 12, no. 6, pp. 2811–2827, 2018.

[7] A. Jaiswal, M. Valstar, and H. Gunes, "Deep learning-based facial emotion recognition in static images," Expert Systems with Applications, vol. 157, p. 113447, 2020.

[8] P. Roy, P. Banerjee, and B. Chandra, "A multimodal transformer-based fusion architecture for emotion recognition," Frontiers in Neurorobotics, vol. 17, p. 1194543, 2023.

[9] A. Mylonas and G. Giannakakis, "Multimodal emotion recognition using visual, vocal and physiological cues: A review," Applied Sciences, vol. 14, no. 17, p. 8071, 2024.

[10] S. Gupta, K. Chatterjee, and A. Pal, "Real-time engagement detection of online learners using deep learning-based facial emotion analysis," Education and Information Technologies, vol. 27, pp. 911–930, 2022.

[11] M. K. Chowdary and R. K. Jatoth, "Facial emotion recognition using transfer learning," Procedia Computer Science, vol. 165, pp. 350–357, 2019.

[12] M. Tanveer and S. Rashid, "Transfer learning for facial expression recognition under class imbalance," Information, vol. 16, no. 2, p. 87, 2025.

[13] S. Zhao, Z. Liu, and H. Zhang, "Multimodal emotion recognition: A comprehensive review," Neurocomputing, vol. 546, p. 127356, 2023.

[14] T. M. Wani and R. N. Mir, "A review on speech emotion recognition with deep learning approaches," International Journal of Advanced Research in Computer Science, vol. 11, no. 4, pp. 34–39, 2020.

[15] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," Sensors, vol. 20, no. 3,
p. 592, 2020.