International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

An Intelligent Multimodal Emotion Recognition System for E-Learning

Mohamed Ben Ammar
Department of Information Systems
Faculty of Computing and Information Technology
Northern Border University
Rafha, Saudi Arabia
Mohammed.Ammar@nbu.edu.sa

Jihane Ben Slimane
Department of Computer Sciences
Faculty of Computing and Information Technology
Northern Border University
Rafha, Saudi Arabia
jehan.saleh@nbu.edu.sa

Corresponding author: Mohamed Ben Ammar, mohammed.ammar@nbu.edu.sa

Abstract— The purpose of this research paper is to introduce an Intelligent Multimodal Emotion Recognition System (IMERS) that aims to improve the e-learning process by accurately perceiving and reacting to the emotional states of learners. IMERS incorporates information through three primary modalities: facial expressions, voice, and text. The utilization of multimodal fusion in this technique effectively addresses the constraints of single-modality systems and yields a more extensive and precise comprehension of emotions. The paper highlight the following: Initially, we discuss the structure of multimodal fusion, specifically focusing on the architecture of IMERS. This includes an explanation of the many components involved, such as data preparation, feature extraction, decision fusion, and sentiment classification. Every approach employs distinct deep learning algorithms customized to its unique properties. Furthermore, our assessment of IMERS encompasses its proficiency in discerning emotions inside e-learning environments, as evidenced by its correct detection of primary emotions across diverse datasets. Another area of emphasis is personalized learning applications, in which we demonstrate how IMERS customize learning experiences by adapting instruction, offering specific feedback, and cultivating an emotionally nurturing learning environment..

Keywords- Multimodal Recognition, IMERS, Emotion fusion, Facial Expression Recognition, Emotional Text Recognition.

I. INTRODUCTION

The world of education is undergoing a fascinating transformation, with e-learning platforms emerging as dynamic hubs for personalized learning experiences. Imagine studying wherever, whenever you please, while the learning system itself adapts to your individual needs and emotions. This is where emotional computing takes center stage, enabling systems to understand users by deciphering their feelings and expressions. However, emotions rarely speak in a single voice. They whisper through facial expressions, dance in the rhythm of speech, and even leave their mark on the way we write. That is why this research embarks on a quest to build an intelligent e-learning system that embraces multimodal fusion. We want to go beyond surface-level emotions and explore the "why" behind your learning experience, providing feedback tailored to your individual needs.

Imagine an e-learning environment that understands your emotions, adapting to your frustration with a missed step, offering encouragement when you hit a hurdle, and celebrating your joy at mastering a new concept. This is the vision behind IMERS (Intelligent Multimodal Emotion Recognition System), a groundbreaking system that unlocks the powerful secret language of your emotions to personalize and enhance the learner's experiences.

Beyond the Surface:

IMERS transcends the limitations of single-modality systems by fusing information from three key channels:

- Facial Expressions: Your face, a subtle billboard of emotions, analyzed using cutting-edge computer vision techniques to detect micro-expressions and unlock the true message behind your frowns or smiles.
- Speech: The rhythm, tone, and even pitch of your voice hold hidden clues to your emotional state. IMERS listens attentively, extracting these nuances to understand your feelings and tailor its responses.
- Text: The way you write – word choice, sentence structure, and punctuation – paints a hidden picture of your emotional state. IMERS analyzes your text, like a skilled linguist, uncovering keywords and patterns that hint at confusion, excitement, or any emotion in between.

Personalized Learning Revolution:

By blending these modalities, IMERS goes beyond emotion recognition to construct truly personalized learning experiences:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

- Adaptive Learning: Imagine struggling with a concept and suddenly facing easier, more engaging materials, all because IMERS detected your frustration and adjusted the learning path accordingly.
- Targeted Feedback: IMERS offers personalized support and encouragement, guiding you back on track when you need it most.
- Emotional Scaffolding: Frustrated or bored. IMERS does not judge; it provides additional resources or adjusts the learning environment to reignite your curiosity and keep you engaged.

IMERS paves the way for a future of E-learning where technology becomes a supportive companion on your educational journey. This paper delves into the system's architecture, performance, and potential applications, displaying how IMERS can revolutionize the way we learn and grow.

II. Literature Review

E-Learning has revolutionized education, offering flexibility and personalized learning pathways. However, one crucial aspect often remains overlooked: emotions. Learners' emotional states significantly influence their engagement, comprehension, and learning outcomes. This literature review explores multimodal emotion recognition (MER) as a transformative approach to understand and cater to learners' emotions in e-learning environments.

A. Single-Modality Approaches

Early research primarily focused on single-modality recognition, analyzing facial expressions, speech, or text individually. Facial expressions, analyzed through computer vision techniques, provided valuable insights into basic emotions like happiness, sadness, anger, and fear. Speech analysis, incorporating acoustic features and prosody, offered complementary information about emotional states. Text analysis, employing natural language processing (NLP), revealed sentiment and emotional cues through word choice, sentence structure, and punctuation. While these single-modality approaches yielded promising results, limitations in accuracy and context-dependence led researchers to explore the potential of multimodal fusion.

B. Multimodal Fusion for Enhanced Accuracy

Recognizing the inherent limitations of single modalities, recent research has shifted towards multimodal fusion, combining information from multiple channels to capture a more comprehensive and accurate picture of emotions. Several studies have demonstrated the superiority of multimodal fusion over single-modality approaches. For example, [1] combined facial expressions and speech to achieve significant accuracy improvements in recognizing learner emotions during online lectures. Similarly, [3] fused text and speech data to enhance emotion recognition in educational forums, providing personalized interventions based on learners' emotional states.

C. Modality Selection and Fusion Techniques

The choice of modalities and fusion techniques employed in MER systems varies depending on research goals and practical considerations. Commonly used modalities include:

- Facial expressions: Captured through webcams, analyzed with convolutional neural networks (CNNs) for feature extraction and recognition.
- Speech: Recorded audio analyzed with Mel-frequency cepstral coefficients (MFCCs) and recurrent neural networks (RNNs) for emotional tone and prosody extraction.
- Text: Written communication (chat logs, essays) analyzed through NLP techniques like sentiment analysis and keyword extraction.

Fusion techniques often involve majority voting, weighted averaging, or more sophisticated probabilistic approaches like Bayesian networks.

III. Problem Formulation

The rise of E-learning presents exciting opportunities for personalized learning experiences. However, current platforms fall short in acknowledging the crucial role of emotions in the learning process. Learners' emotional states significantly affect their engagement, comprehension, and overall educational outcomes. IMERS addresses this gap by employing multimodal emotion recognition to:

- Identify learners' emotional states in real-time.
- Adapt teaching styles and learning materials to individual emotional needs.
- Offer personalized feedback and support.
- Cultivate a positive and emotionally nurturing learning environment.

A. Methodology

- Multimodal Data Acquisition: We employ diverse sensors and data collection methods to capture facial expressions, speech recordings, and textual interactions in e-learning settings.
- Multimodal Feature Extraction: We utilize specific deep learning techniques for each modality:
  - Facial Expressions: Convolutional Neural Networks (CNNs) for extracting features from facial images.
  - Speech: Mel-frequency cepstral coefficients (MFCCs) and Recurrent Neural Networks (RNNs) for extracting features from voice recordings.
  - Text: Natural Language Processing (NLP) techniques to analyze sentiment and emotional indicators in text.
- Decision Fusion: We combine the extracted features from each modality using a majority voting method or a more sophisticated probabilistic approach depending on the research context.
- Emotion Classification: We employ machine-learning algorithms to classify the fused features into predefined emotions.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

B. Results and Discussion

- We evaluate IMERS on benchmark e-learning datasets, demonstrating its accuracy in recognizing key emotions across different modalities and their fusion.
- We compare IMERS with unimodal emotion recognition systems, highlighting the significant improvement in accuracy achieved through multimodal fusion.
- We discuss the potential applications of IMERS in personalized learning, including:
    - Adaptive learning: adjusting lesson difficulty and materials based on real-time emotion detection.
    - Targeted feedback: providing personalized support and encouragement based on learners' emotional needs.
    - Emotional scaffolding: offering additional resources or adjustments to the learning environment when learners experience frustration or boredom.
- We address ethical considerations and implications of emotion recognition in educational settings.

IV. A NOVEL MULTIMODAL EMOTION RECOGNITION SYSTEM

A. Intelligent Multimodal Emotion Recognition System (IMERS)

Despite the better achievement obtained, there still seems to be significant tutoring system for better emotion recognition by switching from a unimodal to a bi-modal system. Our bi-modal system considers face and text information because one modality alone does not provide the necessary information to convey the user's emotion. We propose a new approach to Facial-Textual Emotion Recognition using Decision-level. This type of fusion is also named late fusion because it comes after matching, where each modality uses a separate classifier as a pre-classification and then combines the hard decision. For each signal, we will use its appropriate model. In other words, for text signal, we will apply our approach based on LSTM. For image signal, we will also use our methods based on DCNN-BiLSTM. Finally, the two separate classifiers with facial image and text are fused to obtain the emotion classes. In other words, the output of both the DCNN-BiLSTM network and LSTM network are considered inputs in the decision fusion layer for decision fusion. Features are extracted separately from the LSTM and CNN models. These features are then concatenated to form a combined feature vector. The combined feature vector is used as input to a KNN classifier for the final decision-making. The fusion of features occurs early in the pipeline, before passing through the final classification layer. This approach assumes that combining features at an early stage will enhance the model's ability to learn relationships between different modalities. For our proposed system decision fusion model, we found that fear is the most challenging emotion, and the best-recognized emotion is a surprise. Our proposed bi-modal system has resulted in emotion recognition accuracy and weighted average f1-score of 79%.

CNNs demonstrated their robust performance in uncontrolled environments like variable lighting and occlusions; we leverage their architecture to construct the facial expression and speech recognition models in our innovative a Three-Channel Approach for Multimodal Emotion Recognition (3C-MER). For textual analysis, we retain the LSTM architecture due to its effectiveness in sequential data processing. A clear representation of 3C-MER architecture follows a straightforward block diagram with four key components:

- Input Pre-processing: This block prepares the speech, image, and text data for further processing, ensuring compatibility with the subsequent steps.
- Feature Extraction: Leveraging deep learning algorithms like CNNs for facial expressions and speech, and LSTMs for text, this block extracts relevant features from each modality.
- Decision Fusion: Applying a majority voting technique to combine the outputs from each modality, this block enhances the overall accuracy of emotion recognition.
- Output Classification: This final block utilizes the fused information to categorize the detected emotions based on predefined labels.

This streamlined design allows 3C-MER to effectively harness the strengths of each modality and overcome the limitations of individual approaches, paving the way for more effective and nuanced emotion recognition in e-learning environments.

To assess the performance of our proposed system, we employed three distinct datasets:

- FER-2013: this dataset primarily focuses on facial expressions.
- Tweet Emotions: this dataset captures emotions expressed through textual communication.
- Ravdess: this dataset comprises audio recordings of emotional speech.

While these datasets initially spanned a wider range of emotions, for this evaluation, we specifically retrained our models to classify four key emotions: calm, happy, sad, and angry. This choice aligns with the focus of other related works and allows for a more targeted comparison. The results are encouraging. Our decision fusion approach delivers an impressive accuracy of 93%, surpassing the individual performances of our facial and textual emotion recognition baselines by 1% and 16% respectively. To validate our model against existing research, we compared it with comparable studies predicting the same four emotions through a multimodal fusion of face, text, and speech (note that our "calm" category corresponds to "neutral" in other works). This comprehensive evaluation across diverse datasets and comparison with established works demonstrates the effectiveness of 3C-MER in accurately recognizing emotions within a focused set of key categories. When compared to the other models, our model is the most efficient.

The burgeoning field of artificial intelligence has ushered in a new era of intelligent affective tutoring systems. These innovative systems leverage emotion recognition technology to provide personalized learning experiences tailored to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

individual students' emotional states. In this section, we delve into the architecture of our multimodal system, detailing the mechanisms by which it achieves this level of personalization. We then proceed to display the various interfaces of our system, offering a glimpse into how students interact with it.
We propose a real-time system named IMERS stands for Multimodal Intelligent Tutoring Emotion Recognition System. Our IMERS, equipped with affective computing capabilities, can detect students' emotions and adjust their responses accordingly, providing appropriate support, empathy, or encouragement based on the student's affective state. The applied methodology is based on the proposed system [4].

B. MELD Dataset

To rigorously train, test, and validate our IMERS, we turned to a rich repository of emotional expression: the Multimodal Emotion Lines Dataset (MELD) [2]. The dataset, captures over 1400 multi-speaker dialogues and 13,000 individual utterances from the ever-popular TV series Friends.
Key Characteristics of MELD:

- Data Division: The dataset is thoughtfully divided into training (9,989 samples), validation (1,109 samples), and testing (2,610 samples) sets, ensuring robust model development and assessment.
- Multimodal Nature: Each sample within MELD offers a treasure trove of emotional insights, encompassing video clips, speech recordings, text transcripts, and meticulously assigned emotion labels.
- Emotional Diversity: The dataset encompasses seven distinct
  emotions, anger, disgust, sadness, joy, neutrality, surp rise, and fear. Additionally, it provides sentiment labels (positive, negative, neutral) for each utterance.
- Imbalance: Neutral samples hold a majority within MELD, highlighting a prevalent challenge in emotion recognition: capturing the full spectrum of human emotions with balanced representation.
- Dialogue Dynamics: MELD's dialogues offer glimpses into authentic conversational exchanges, with an average of 9.6 utterances per dialogue and a diverse range of emotions (3.3 different emotions per dialogue on average).
- Audio Characteristics: Utterance audio recordings average around 5 seconds in duration, providing ample material for speech-based emotion analysis.

C. Results

Not all basic emotion classes are utilized in the MELD dataset. The primary focus of our study revolved on four distinct emotional categories: Neutral, Joy, Sadness, and Anger. The selection of this class was based on the majority voting process employed with speech modality, which requires a higher number of classifiers compared to the number of classes. The lowest accuracy score for the face modality is associated with the melancholy emotion at 92%. The neutral feeling has a precision value of 99% across all three modalities. Regarding the metric for the facial modality, both neutral and melancholy have a similarity score of 97%, while pleasure and rage have a similarity score of 98%. In terms of speech modality, both neutral and joy are assigned a value of 98%, while sorrow and anger are assigned a value of 99%. The macro average and weighted average metrics offer a comprehensive summary of the collected data from the whole dataset. IMERS model demonstrates a strong categorization. The performance of our multimodal fusion system enhanced and its resilience increased in situations when one of the modalities is absent or influenced by noise. The total accuracy achieved a rate of 97%. The measure Recall shows that the Neutral emotion has a maximum value of 98% while the emotion rage has a minimum value of 96%. When using the accuracy metric, the emotion of melancholy assigned a minimum value of 94%. It is evident that the neutral emotion attains the highest value across all measures. This attributed to the prevalence of neutral samples in each of the datasets.
Students demonstrated higher levels of engagement and positive conduct throughout e-learning sessions when they experienced pleasant feelings. Additionally, they held the belief that they might engage in productive interactions with fellow students and professors via the educational platform. In contrast, when students experienced greater negativity throughout their participation in e-learning activities, their confidence in their capacity to effectively manage their learning and utilize the available resources diminished. As a result, their motivation, organization, and performance on exams were negatively impacted. We want to assess our methods not only based on main emotions but also on sentiments, including positive, negative, and neutral. The positive sentiment exhibits a higher level of accuracy (63%) compared to the negative sentiment (57%), however the negative sentiment has a higher level of recall (52%) compared to the positive mood (40%). The confusion matrix for our IMERS, which includes three feelings. The total accuracy is at 64%.

D. Evaluation and analysis

a) Evaluating Performance and Addressing Imbalance

This section delves into the performance of our proposed emotion recognition system with multimodal fusion (IMERS) using the MELD dataset. We'll analyze its effectiveness compared to individual unimodal systems (focusing on facial expressions, speech, and text) and benchmark it against advanced multimodal systems presented in [5] and [6].

b) Uni-modal vs. Multimodal Analysis

We begin by examining the individual unimodal systems. This provides a baseline understanding of how well each modality (facial expressions, speech, and text) performs in isolation for emotion recognition. Subsequently, we compare these results to the outcome achieved by our IMERS model after implementing decision fusion. This comparison highlights the potential benefits of combining information from all three modalities for improved accuracy.

c) Benchmarking against Existing Work

We then conduct a comparative analysis of IMERS with other recent approaches [5-6] that also predict emotions on the MELD dataset. These methods, like IMERS, utilize a fusion of facial expressions, speech, and text data, making them suitable

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

for direct comparison. Here, we evaluate various metrics like accuracy, precision, recall, and F1-score to assess the effectiveness of IMERS in comparison to the existing methods.

d) Addressing the Neutral Bias

Our analysis might reveal a bias in the model's predictions towards the "Neutral" emotion category. This could be attributed to the inherent data imbalance within the MELD dataset, where a significant portion (potentially exceeding 47%) of the samples might be classified as neutral. As a result, the model, influenced by this imbalance, might prioritize learning weights for the neutral class during training. This can lead to the model favoring neutral predictions even for ambiguous cases where other emotions might be present.

e) Potential Solutions

To address this data imbalance issue, we can explore various techniques:

- **Oversampling:** Replicating data points from minority emotion classes to create a more balanced dataset.
- **Under-sampling:** Reducing the number of data points from the majority class (neutral) to match the size of the minority classes.
- **Cost-sensitive learning:** Assigning higher weights to misclassifications of minority emotions during training, encouraging the model to focus on these classes.

By implementing these techniques, we can mitigate the impact of data imbalance and potentially improve the model's ability to recognize emotions other than neutral.

V. The Affective Intelligent Tutoring System

Imagine a classroom where technology can not only assess student comprehension but also understand their emotional state. This is the potential unlocked by IMERS, a system that successfully anticipated students' emotional responses to teacher inquiries. While the example highlights completion of assignments, IMERS can be applied to various educational contexts, creating a dynamic learning environment that adapts to individual needs. [7-8]

A. The Personalized Learning Revolution

IMERS goes beyond traditional teaching methods by personalizing the learning experience based on emotions. Here's a closer look at its potential benefits: [9-10]

- **Adaptive Learning Paths:** Frustration with a complex concept can be detected by IMERS through facial expressions, voice intonations, and typing patterns. The system can then adjust the learning path in real-time. For instance, if a student shows signs of struggle, IMERS could:
  - Offer simpler explanations or alternative learning materials that break down complex concepts.
  - Introduce interactive games or simulations that make learning more engaging.
  - Recommend additional practice exercises tailored to the student's specific areas of difficulty.
- **Targeted Support in the Moment:** When a student encounters a hurdle, IMERS can provide immediate, personalized support. Imagine a student facing confusion, reflected in furrowed brows and hesitant mouse clicks. IMERS could: [11-12]
  - Offer a gentle nudge with targeted hints or additional resources.
  - Tailor explanations based on the student's emotional state, simplifying language or providing visual aids for those struggling to grasp concepts.
  - Offer empathetic encouragement to boost confidence and motivation, fostering a supportive learning environment where students feel comfortable seeking help.
- **Building Confidence and Resilience:** Emotions play a crucial role in learning. By recognizing and acknowledging student emotions, IMERS can nurture confidence and resilience:
  - When a student experiences a breakthrough moment, IMERS can celebrate their achievement, reinforcing positive emotions and fostering a sense of accomplishment.
  - When a student makes a mistake, IMERS can offer personalized feedback that focuses on improvement, replacing discouragement with constructive guidance. This fosters a growth mindset where students view challenges as opportunities to learn and grow.

B. Charting the Course for the Future

While IMERS presents a significant leap forward, there's room for further exploration:

- **Expanding the Emotional Palette:** Currently, IMERS focuses on facial expressions, speech, and text. Future iterations could incorporate additional modalities like:
  - **Body Language:** Crossed arms or fidgeting might indicate frustration or disengagement.
  - **Gestures:** Hand gestures can provide insights into understanding or confusion.
  - **Physiological Signals:** Heart rate or skin conductance could offer deeper emotional cues.

  By integrating these additional data points, IMERS can paint a more holistic picture of students' emotional states, leading to even richer personalized learning experiences.
- **Continuous Learning and Refinement:** Machine learning thrives on continuous improvement. IMERS should adapt and evolve through ongoing research and development to:
  - Stay at the forefront of emotion recognition, incorporating new discoveries in the field.
  - Maintain high accuracy through ongoing data collection and algorithm refinement.
  - Adapt to the ever-changing landscape of E-learning, ensuring compatibility with emerging technologies and educational practices.
- **Ethical Considerations:** As we explore the emotional side of E-learning, ethical considerations are paramount:
  - **Data Privacy:** Transparency and robust security measures are essential to ensure student emotional data is collected, stored, and used responsibly.

- **Bias Mitigation:** Facial expressions and emotional responses can vary culturally. IMERS needs to be calibrated to account for these variations and avoid biased interpretations.

IMERS offers a glimpse into a future where technology fosters emotionally intelligent learning environments. By harnessing the power of emotion recognition, it personalizes learning, boosting engagement, providing targeted support, and nurturing student confidence. While challenges remain in data privacy, bias mitigation, and continuous improvement, IMERS lays the groundwork for a future where learning is not just effective, but also emotionally supportive and empowering for all students. [13-14-15]

VI. Conclusion and Future Work

The Intelligent Multimodal Emotion Recognition System (IMERS) presented in this paper offers a promising approach to enhancing the e-learning experience. By integrating information from facial expressions, speech, and text, IMERS overcomes the limitations of single-modality systems and achieves accurate emotion recognition. This understanding then leveraged to personalize the learning experience through adaptive learning, targeted feedback, emotional support, and social interaction facilitation.

Overall, IMERS has the potential to:

- Improve learner engagement: By personalizing the learning experience based on emotions, IMERS can make learning more relevant and enjoyable for individual learners.
- Enhance learning outcomes: Tailoring instruction and feedback to emotional states can lead to better understanding and retention of material.
- Increase learner satisfaction: A supportive and emotionally aware learning environment can foster positive attitudes towards learning and reduce frustration.

While IMERS demonstrates promising results, several areas offer opportunities for future research and development:

- Expanding the range of emotions recognized: the current system focuses on primary emotions. Recognizing more subtle and nuanced emotions, such as boredom, curiosity, and pride, could provide even richer insights into learners' experiences.
- Incorporating additional modalities: Integrating physiological data, such as heart rate and skin conductance, could offer further information about emotional states. This could be particularly helpful for detecting hidden emotions or confirming ambiguous ones.

By continuing to explore these future directions, IMERS and similar emotion-aware e-learning systems have the potential to revolutionize the way we teach and learn, creating a more personalized, engaging, and effective learning experience for all.

References

[1] Siddiqui, H. U. R., Zafar, K., Saleem, A. A., Raza, M. A., Dudley, S., Rustam, F., & Ashraf, I. (2023). Emotion classification using temporal and spectral features from IR-UWB-based respiration data. Multimedia Tools and Applications, 82(12), 18565-18583.

[2] Tang, G., Xie, Y., Li, K., Liang, R., & Zhao, L. (2022). Multimodal emotion recognition from facial expression and speech based on feature fusion. Multimedia Tools and Applications, 1-15.

[3] Veni, S., Anand, R., Mohan, D., & PAUL, E. (2021). Feature fusion in multimodal emotion recognition system for enhancement of human-machine interaction. IOP Conference Series: Materials Science and Engineering, 1084(1), 012004.

[4] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 527–536.

[5] Alim, S. A., & Rashid, N. K. A. (2018). Some Commonly Used Speech Feature Extraction Algorithms. In From Natural to Artificial Intelligence-Algorithms and Applications. London, United Kingdom: IntechOpen.

[6] Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self-supervised feature fusion. IEEE Access, 8, 176274–176285.

[7] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., & Aguilera, A. (2022). Adaptive multimodal emotion detection architecture for social robots. IEEE Access, 10, 20727–20744.

[8] Sekkate, S., Khalil, M., & Adib, A. (2022). A statistical feature extraction for deep speech emotion recognition in a bilingual scenario. Multimedia Tools and Applications, 1-18.

[9] Liu, M., & Yu, D. (2022). Towards intelligent E-learning systems. Educational Information Technology. https://doi.org/10.1007/s10639-022-11479-6

[10] Reimers, F., Schleicher, A., Saavedra, J., & Tuominen, S. (2020). Supporting the continuation of teaching and learning during the COVID-19 Pandemic. OECD, 1(1), 1-38.

[11] Hazarika, D., Boruah, A., & Puzari, R. (2022). Growth of Edtech Market in India: A Study on Pre-pandemic and Ongoing Pandemic situation. Journal of Positive School Psychology, 6(3), 5291-5303.

[12] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., & Aguilera, A. (2022). Adaptive multimodal emotion detection architecture for social robots. IEEE Access, 10, 20727–20744.

[13] Le, T. H., Tran, H. N., Nguyen, P. D., Nguyen, H. Q., Nguyen, T. B., Tran, T. H., Vu, H., & Le, T. L. (2022). Spatial and Temporal Hand-Raising Recognition from Classroom Videos using Locality, Relative Position-Aware Non-local Networks and Hand Tracking. Vietnam Journal of Computer Science, pp.1-29.

[14] Lee, J. H., Kim, H. J., & Cheong, Y. G. (2020). A multi-modal approach for emotion recognition of TV drama characters using image and text. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 420–424). IEEE.

[15] De Carolis, B., D'Errico, F., Macchiarulo, N., Paciello, M., & Palestra, G. (2021). Recognizing cognitive emotions in e-learning environment. In International Workshop on Higher Education Learning Methodologies and Technologies Online (pp. 17-27). Springer, Cham.