International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Improving the personalization of the EMASPEL learning experience through deep learning-based facial expression recognition

Mohamed Ben Ammar 1[1*]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract: Conventional Intelligent Tutoring Systems (ITS) rely solely on cognitive performance metrics to personalize learning journeys, overlooking the crucial role of emotions in facilitating effective learning. This disconnects often results in disengaged students and suboptimal learning outcomes. Our research presents a novel approach to bridge this gap by integrating Deep Learning-based Facial Expression Recognition (FER) into the EMASPEL ITS. We propose utilizing FER to equip EMASPEL with the ability to provide immediate feedback on students' engagement and emotional states. Our focus lies in identifying key emotions such as annoyance, confusion, and excitement, using the power of Deep Learning algorithms to accurately interpret facial expressions. This real-time emotional understanding empowers EMASPEL to dynamically adjust educational content and pace in sync with students' emotional responses. We acknowledge the challenges and ethical considerations surrounding the implementation of FER in educational settings. Transparency and robust privacy measures are paramount in ensuring this technology is utilized responsibly and ethically. We envision EMASPEL as a pioneer in fostering emotionally aware ITS, not just catering to cognitive needs but catering to the entire spectrum of human emotions. Our contribution lies in establishing a framework for developing advanced ITS that leverage Deep Learning-powered FER to recognize and respond to a wide range of human emotions. This paves the way for truly personalized learning experiences that prioritize engagement, motivation, and ultimately, maximize educational achievement.

Keywords: *Facial Expression Recognition, ITS, FER*

1. Introduction

This template, modified in MS Word 2007 and saved as a "Word 97-2003 Document" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: ease of use when formatting individual papers, automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and conformity of style throughout a manuscripts. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

The landscape of education is experiencing a transformative shift, embracing personalized learning approaches that cater to individual needs and styles. In this endeavor, EMASPEL, an innovative Intelligent Tutoring System (ITS), stands as a beacon of personalization through its utilization of emotional agents. These virtual companions guide and motivate learners, adapting their support to self-reported emotional states. However, this reliance on explicit declarations overlooks the subtle nuances of human expression often conveyed through the enigmatic language of the face. It is here that deep learning-powered facial expression recognition (FER) emerges as a transformative force, poised to elevate EMASPEL's personalization capabilities to unprecedented heights.

Imagine a scenario where, instead of relying solely on self-reported emotions, EMASPEL can directly analyze a learner's facial expressions in real-time. A flicker of frustration in furrowed brows triggers the agent to offer a reassuring nudge. A grin of understanding prompts the platform to accelerate the pace. By integrating deep learning for FER, EMASPEL transcends the limitations of self-declared emotions, revealing the hidden symphony of feelings playing on a learner's face. This opens doors to a future where learning journeys are truly tailored to the emotional ebb and flow of each individual, maximizing engagement, fostering empathy, and ultimately, unlocking the full potential of personalized education.

This research embarks on a journey to explore the synergistic potential of deep learning for FER and EMASPEL. We delve into the intricate world of facial expressions, harnessing the power of deep learning

[1] *Department of Information Systems, Faculty of Computing and Information Technology,*
Northern Border University, Rafha, Saudi Arabia
Mohammed.Ammar@nbu.edu.sa
ORCID ID : 0000-0001-8990-3924
** Corresponding Author Email: Mohammed.Ammar@nbu.edu.sa*

algorithms to decipher their hidden meanings. We then bridge the gap between these insights and EMASPEL's personalized learning framework, crafting an intricate integration that allows the platform to dynamically adapt to learners' real-time emotional states. Through meticulous research, rigorous evaluation, and unwavering ethical considerations, we pave the way for a future where education resonates with the unique emotional tapestry of each learner, orchestrated by the magic of deep learning and the wisdom of EMASPEL.

The remainder of this paper delves into the literature review, Problem formulation, presenting our proposed framework. The deep learning for FER, and discuss the results of our pilot study and engage in a critical analysis of the findings, paving the way for further exploration of this exciting frontier in the field of facial expression recognition and its potential to revolutionize e-learning.

2. Problem Formulation

The research on "Improving the personalization of the EMASPEL learning experience through deep learning-based facial expression recognition" tackles a crucial challenge in adaptive learning:

Need for Personalization in e-Learning:

- Traditional e-learning often employs uniform content and pacing, neglecting individual learners' needs and emotions.
- This one-size-fits-all approach can lead to disengagement, frustration, and reduced learning outcomes.
- Recognizing and responding to learners' emotions in real-time can personalize the learning experience, enhance engagement, and improve knowledge retention.

Limitations of Existing EMASPEL:

- While EMASPEL uses a traditional FER method, it is no longer adequate for the purpose and needs to be updated.
- Facial expressions offer a rich and immediate window into learners' emotions, but EMASPEL currently lacks the capability to interpret them.
- This gap impedes EMASPEL's ability to adapt the learning experience based on real-time emotional dynamics during learning interactions.

Potential of Deep Learning-Based Facial Expression Recognition:

- Deep learning models have revolutionized facial expression recognition, achieving high accuracy in controlled settings.
- Integrating such models within EMASPEL can enable real-time analysis of learners' facial expressions during e-learning activities.
- By capturing subtle emotional shifts, the system can adapt learning content, pace, and interaction styles to cater to individual needs and foster optimal learning conditions.

Specific Research Questions:

- How can deep learning-based facial expression recognition be effectively integrated into the EMASPEL platform to enhance its personalization capabilities?
- What specific types of facial expressions and emotional states are most relevant for adaptive e-learning, and how can they be accurately identified within the context of learning activities?
- How can the personalized learning interventions triggered by facial expression analysis be designed to maximize learner engagement and improve learning outcomes across diverse emotional states?
- What are the ethical considerations and potential privacy concerns related to collecting and analysing learners' facial expressions in an educational setting?

Expected Outcomes:

- Develop a robust and practical framework for integrating deep learning-based facial expression recognition into the EMASPEL platform.
- Enhance EMASPEL's personalization capabilities by enabling real-time adaptation based on learners' emotional states during e-learning activities.
- Improve learner engagement, motivation, and knowledge retention through personalized learning experiences tailored to individual emotional needs.

3. Literature Review

This section aims to examine the studies pertaining to emotion recognition in the literature, encompassing many modalities. Our process begins by utilizing automated facial analysis to identify emotions, followed by analysing text to determine emotional content, and finally analysing voice to recognize emotions. Furthermore, the purpose of this chapter is to offer a more thorough examination of the study methods concerning the identification of emotions through the use of a bimodal or multimodal system. Subsequently, we examine the existing techniques for

identifying emotions in an intelligent affective teaching system. In conclusion, this chapter marks the finish, and each succeeding chapter will go further into the existing literature, focusing on specific topics that are being addressed. Facial expression is the primary and most prevalent method for perceiving emotions. Several methodologies have been proposed to categorize emotions using facial expressions. The face expression can be sent through two distinct routes. One channel is dedicated to displaying images, while the other channel is specifically designed for showcasing videos. We are intrigued by the chapter that focuses on picture channels and, conversely, deep learning models. The advancement in automatically detecting face expression in photographs has been remarkable, mostly because to the complex structure of neural networks. To obtain a thorough overview of FER, readers are advised to consult references [6, 7]. Singh et al [8] employed a deep neural network on the FER2013 dataset, achieving a test accuracy of 67.7% in predicting the six fundamental emotions and neutral state. In their study, Saroop et al. [9] provide a deep learning method for emotion detection. This method utilizes Facial Action Units (AU) to extract facial characteristics and applies Convolutional Neural Networks (CNN) to categorize the seven emotions of the FER2013 dataset. The accuracy achieved by this technique is 67.91%. Further research on emotion recognition using deep learning is presented in [10]. The system relies on an attention convolutional network that prioritizes locations with abundant features and use the visualization approach from reference [11] to emphasize the most important aspect of the face. The accuracy achieved on FER2013 is 70.02%. Wang et al. [12] developed an attention module that computes the weight of each layer's feature map and reduces extraneous information. Next, they categorize the feelings using a linear connection layer. Their accuracy rating with the CK+ dataset is 92.8%. In [13], the author utilizes the CK+ dataset and transforms its photos into Local Binary Pattern images, which serve as a textural feature descriptor for the images. The resultant pictures are utilized as inputs to a Convolutional Neural Network (CNN), yielding an accuracy of 79.56%. In [14], a novel bilinear pooling model using CNNs is introduced for the purpose of face emotion identification. The objective of this study was to incorporate an extra discriminant information component into the conventional CNN model. The accuracy of this model, utilizing FER2013 dataset, was found to be 72.65% by implementing enhanced bilinear pooling technique. Nevertheless, the utilization of the bilinear aggregation function in conjunction with the square root function of the matrix results in excessive memory and CPU usage, hence diminishing the model's performance. The authors in [4] employ Haar-like characteristics for face detection. Subsequently, via a facial activity coding system, several facial regions were identified and virtual markers were positioned on these areas. The Lucas-Kanade optical flow method is employed to track these markers. The facial expression categorization utilizes the distance between each marker and the centre of the face as a characteristic. By employing 10 virtual markers, the researchers identified six distinct face emotions: pleasure, sorrow, anger, fear, disgust, and surprise. Niu et al. [15] specifically examined facial expressions captured in static photos. The Dlib library was employed for facial detection. Subsequently, they suggested combining features by utilizing the Local Binary Pattern (LBP) and Oriented Fast and Rotated Brief (ORB) descriptors. The combined characteristics are categorized using Support Vector Machine (SVM) to identify seven facial emotions, including six fundamental facial emotions and one neutral emotion. The accuracy achieved on the MMI dataset is 79.8%. The researchers in reference [16] employ the Viola-Jones algorithm for the purpose of face detection. Furthermore, they introduce Joint geometric features with Gabor features and LBP features. Subsequently, these characteristics are employed as input for the multi-layered Convolutional Neural Network (CNN). Expression recognition is performed using an SVM classifier. We presented a comprehensive summary of previous studies, as depicted in Table 1, which showcases the methodology, classifier, and dataset employed. Furthermore, the level of precision achieved and the quantity of emotions identified.

Table 1: Summary of FER Related Works

Works	Approach	Dataset	Accuracy	Classifier	Emotions
Singh et al. [8], 2020	deep learning	FER2013	67.70%	CNN	7 emotions
Saroop et al. [9], 2021	deep learning	FER2013	67.91%	CNN	7 emotions
Minaee et al. [10], 2021	deep learning	FER2013	70.02%	Softmax	7 emotions
Wang et al. [12], 2021	attention module	CK+ dataset	92.80%	CNN	7 emotions + neutral
Koray et al. [13], 2021	deep learning	CK+ dataset	79.56%	CNN	7 emotions
Mahmoudi et al. [14], 2020	bilinear pooling + CNN	FER2013	72.65%	Softmax	not mentioned
Hassounah et al. [4], 2020	Haarlike features	their own dataset	99.81%	CNN	6 emotions
Niu et al. [15], 2021	fused features (LBP+ORP)	MMI	79.80%	SVM	6 emotions + neutral
		CK+	93.20%		

4. The Proposed Framework

Within the realm of affective computing, Our Emotional Multi-Agents System for Peer-to-peer E-Learning (EMASPEL) platform pioneered by [2] takes a captivating approach to nurturing emotional connections in e-learning (Figure 1). Its key differentiator lies in its sentiment analysis capabilities. Rather than merely presenting static content, EMASPEL employs a network of five specialized agents: Interface, Emotional, EEC, Curriculum, and Tutoring. This network continuously analyzes learner sentiment through various channels, including forum posts, facial expressions, or physiological sensors.

The Emotional Agent (EEC), the brain of the system, interprets these cues, deducing the learner's emotional state in the context of the learning environment. Frustration with complex material? Boredom during repetitive tasks? EMASPEL uses this dynamic understanding to personalize the learner's journey. Based on the EEC's insights, the Tutor Agent selects the most appropriate pedagogical activity from the knowledge base, tailored to address the learner's emotional needs.

Furthermore, the Curriculum Agent leverages database resources (DB1 and DB2) to seamlessly translate the chosen activity into a tangible learning experience. This real-time adaptation, driven by accurate sentiment analysis, fosters a more engaging and responsive environment for learners. EMASPEL thus highlights the exciting potential of agent-based systems to personalize learning through emotional awareness, paving the way for a future where the learning experience adapts to the unique emotional needs of each learner.

5. Deep Facial Emotion Recognition

5.1. Emotion recognition system steps

Unveiling Emotions: A Step-by-Step Guide to Automatic Recognition

1. Listening to the Heart's Whispers: Capturing Emotional Cues

- ✓ Input Channels: Like diverse instruments in an orchestra, we gather emotional signals from three distinct channels:
  - Vocal Channel: The nuances of tone, pitch, and rhythm reveal heartfelt sentiments.
  - Body Channel: Posture, gestures, and facial expressions paint a vivid portrait of inner states.
  - Sentimental Texture: Words themselves carry emotional weight, revealing the writer's feelings.

2. Preparing the Stage: Pre-Processing

- ✓ Cleaning and Refining: Before extracting features, we carefully prepare the data, removing noise and inconsistencies. This ensures a clear and focused analysis.

3. Capturing the Essence: Feature Extraction

- ✓ Identifying Key Characteristics: Like a painter's brushstrokes, we extract the most salient features from each channel, creating a distinct emotional signature.
- ✓ Feature Selection: Carefully choosing the most informative features enhances accuracy and efficiency, ensuring a focused portrait of emotion.

4. Learning to Recognize: Training the System

- ✓ Building Emotional Intelligence: Through a process akin to human learning, we train the system to recognize and classify emotions based on the extracted features.
- ✓ Diverse Teaching Methods: Machine learning algorithms, like dedicated mentors, guide the system in understanding the subtle patterns of human emotion.

5. Categorizing Emotions: Classification

- ✓ Making Sense of Signals: Once trained, the system uses sophisticated algorithms to categorize emotions into distinct classes, such as happiness, sadness, anger, or fear.
- ✓ Diverse Approaches: Different classification techniques, including Naive Bayesian models, Artificial Neural Networks, Decision Trees, K-Nearest Neighbors, and Support Vector Machines, offer unique perspectives on emotional landscapes.

6. The Foundation of Understanding: Databases

- ✓ A Rich Tapestry of Expression: Like a library of human emotion, diverse databases containing a wide spectrum of affective expressions provide the essential building blocks for training and testing emotion recognition systems.

Through this multi-stage process, automatic emotion recognition systems strive to bridge the gap between human and machine understanding, unlocking a deeper level of communication and connection.

5.2. Methodology

Although other communication methods are available, the predominant focus of research on emotion identification has been on facial expressions. Facial emotion recognition (FER) is a crucial element of human-computer interaction. The problem of recognizing facial emotions, especially in controlled scenarios including frontal faces and posed expressions, has been successfully overcome. However, it may be challenging to discern these emotions in lifelike

scenarios characterized by occlusions, variations in head attitude, and lighting. Precisely categorizing emotions based on facial expressions is a difficult undertaking. This subsection is separated into two crucial components. The initial section provides an introduction to our CVFER architecture, outlining its specific characteristics. Additionally, we showcase the outcomes achieved by this model when applied to the FER2013 dataset. Subsequently, we provide an alternative model utilizing the DCNNBiLSTM architecture and evaluate its performance using the CK+ dataset.

5.2.1. CVFER: Facial Emotion Recognition with CNN

Transfer learning is a method that involves utilizing a pre-trained model, such VGG16, as a foundation for training a model on a distinct yet interconnected job. Regarding our facial emotion recognition, the process of transfer learning using VGG16 entails utilizing the pre-existing weights of VGG16 as the initial weights for our new model (CVFER). This new model is then adjusted on a dataset specifically designed for face emotion identification called FER2013. Our Facial Emotion Recognition system utilizes CNN and VGG16 networks, which are the basis of our technique. The primary benefit of transfer learning is its ability to expedite the model creation and training process by using the pre-existing weights of previously developed models.

5.2.1.1. Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is the predominant approach in supervised deep learning approaches for image classification tasks, including face emotion identification [18]. It is responsible for handling data that are structured as arrays. Convolutional Neural Networks (CNNs) are composed of many layers, each specifically engineered to extract and manipulate distinct characteristics from the input pictures. The CNN architecture has four layers:

- The convolution layer alters the input picture by applying a set of trainable filters, sometimes called kernels. In order to generate a feature map, each filter performs a convolution operation on the picture, individually multiplying each element and subsequently summing up the results. Convolutional layers capture the local spatial patterns and representations.

- Pooling Layer: Pooling layers are added after convolution to decrease the spatial dimensions of the feature maps. Max pooling is a widely used pooling method that selects the highest value from the area of the feature map that is covered by the filter. Consequently, the result following the max pooling layer would consist of a feature map that encompasses the most salient characteristics from the preceding feature map. Pooling enhances the model's ability to withstand variations in the input and reduces computational intricacy. Max pooling is a pooling technique that chooses the highest value from the area of the feature map that is encompassed by the filter. Therefore, the result obtained from the max pooling layer would be a distinctive characteristic or attribute.

- Flattening Layer: The pooling layer is used to compress the feature maps into a one-dimensional vector. During the process of flattening, the spatial data is transformed into a format suitable for input into the fully linked layers.

- The Full Connection Layer, also known as the dense layer, establishes connections between all neurons in the current layer and all neurons in the preceding layer. The fully linked layers perform high-level feature extraction and classification based on the learnt features from the preceding layers.

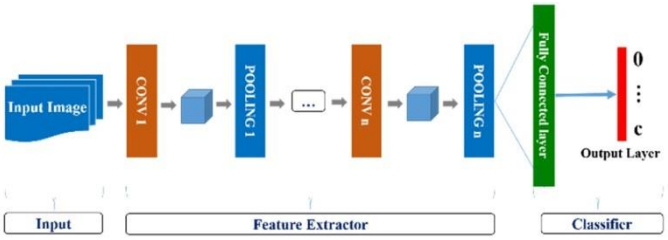

Fig 5.1: The basic layers of CNN (the key components of CNN architecture)

5.2.1.2. VGG16 Network

The VGG16 architecture is a complex convolutional neural network (CNN) model developed by the Visual Geometry Group (VGG) at the University of Oxford [17]. The VGG16 architecture is composed of 16 layers, which is why it is named as such. These layers consist of 13 convolutional layers and 3 fully linked layers. The VGG16 network was trained using the ImageNet [19] database. The latter has received extensive instruction. Therefore, even when working with limited image datasets, it yields high levels of accuracy. The structure of VGG16 may be shown in Figure 5, Table 5.2. The ImageNet dataset consists of pictures that have RGB channels and are of a standardized size of 224*224 pixels. The VGG16 network takes an input tensor of dimensions (224, 224, 3). Figure 5.2 illustrates the presence of five convolutional layer blocks, with the first two blocks consisting of two convolutions followed by one max pooling layer. Next, three blocks consist of three convolution layers, which are then followed by a max pooling layer. Following the last Max pooling layer, there are three fully connected layers.

By utilizing the pre-existing knowledge acquired by VGG16, the model may take use of the features that VGG16 has learnt from a vast collection of diverse photos [20]. This can be especially advantageous when the dataset for face expression identification is short or when there are limited computing resources available for training a deep model from the beginning. The pre-trained VGG16 model functions as a feature extractor, with the lower layers

capturing basic features from the facial photos and the higher levels capturing more complex aspects associated with emotions.

Fig 5.2: The architecture of VGG16 [1]

5.2.1.3. CVFER Model

The next part introduces the suggested architecture for emotion recognition. The model received input in CSV format, with each entry in the table being transformed into a vector and subsequently translated into an image. During the pre-processing step, we rotate some photos within a range of 10 degrees. Next, we horizontally mirror the picture and adjust its position by shifting it to the left or right within a range of 0.1 times its width and height. We select the nearest neighboring pixels to fill any empty spaces. Next, we engage in data augmentation to provide the network with a greater quantity of photos. Following the production of data, we proceed to establish our model, which is derived from the VGG16 model [17]. Subsequently, we proceed to train our model with the objective of accurately measuring the extent to which the CNN enhances the classification of emotions.

The CVFER model, which we have developed, consists of five blocks of convolution layers, with each block being followed by a max pooling layer. The utilization of the max pooling strategy has demonstrated faster convergence and enhanced generalization. As a result, it is frequently utilized for the subsampling layer. The filter values for each convolution block range from 256 to 16, specifically 256, 128, 64, 32, and 16. The stride for each block is set to 1. The CNN parameters employed in this thesis are acquired by multiclass classification, which is the primary focus of our study. To implement the activation function, we employed the Rectified Linear Unit (ReLU) [21] to convert all input values into positive integers, therefore reducing the computational load. On the other hand, we used Categorical Cross entropy as the loss function because there are seven label classes. This function computes the cross entropy loss by comparing the labels and predictions. We employ the Adam optimizer, an algorithm that stands for adaptive moment estimation [22]. Adam is the most efficient optimizer for training an Artificial Neural Network (ANN) in a shorter duration and obtaining rapid convergence with excellent efficacy. We utilized the Softmax classifier, which generates a normalized output for each class and converts weighted sum values into probabilities.

5.2.1.4. CVFER Results and Synthesis

5.2.1.4.1 FER2013 Dataset

CVFER utilizes the Facial Expression Recognition 2013 (FER2013)[23] dataset to train, test, and validate our system. Figure 5.3 exhibits a collection of photographs extracted from the FER2013 dataset. The collection comprises records of seven discrete emotions: anger, contempt, fear, pleasure, sorrow, surprise, and neutrality. The collection comprises 35,887 photographs with a resolution of 48x48. The training set has 29,105 samples, whereas both the public test set and private test set have 3,589 samples each.

Fig 5.3: Samples of FER2013 Images

We selected this dataset due to its clear definition and vast size, which makes it a challenging task. One the one hand, the presence of changes in head position, low contrast pictures, and facial occlusions (such as partial faces or hands in front of faces) contribute to the challenges in illuminating images. This dataset also includes photographs with spectacles, as seen in Figure 5.4. Facial occlusion significantly affects the performance of face recognition.

However, there is an imbalance in the distribution of categories, with certain classes having a greater number of examples compared to others. For instance, the category "Happy" has 8,989 picture samples. Nevertheless, Disgust possesses a mere 547 photos.

Fig 5.4: Samples of facial occlusions from the FER2013 dataset

5.2.1.4.2 CVFER Results

The training of our model started from the beginning and continued for a total of 100 epochs. According to Figure 5.5, at epoch 0, the loss value is 1.0, and the accuracy is around 62%. At epoch 100, the loss stands at around 0.2, while the accuracy reaches around 92%. Therefore, we may deduce that a smaller loss value indicates a superior model. The confusion matrix, which assists in determining

the accurate predictions made by a model, is seen in Figure 5.6. The model demonstrates a high level of accuracy in classifying happiness at 82% compared to disgust at 60%. This discrepancy can be attributed to the varying quantity of data in each category, as previously explained. Deeper networks impose stronger priors on the structure of the trained decision function, effectively combating overfitting.

Additionally, the model exhibits a higher error rate of 51% for fear and 43% for sadness. This can be attributed to the ambiguity and difficulty in interpreting some images, where a single image might have two potential labels. This phenomenon is known as Bayes error, as seen in Figure 5.7.

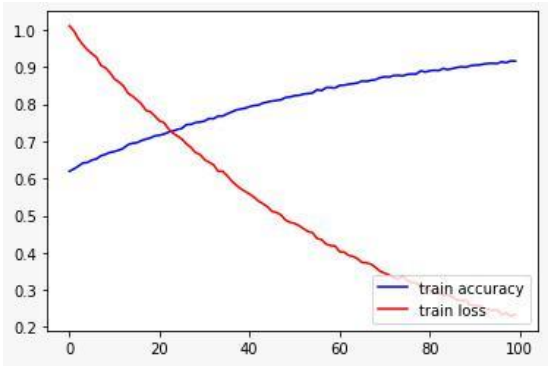

Fig 5.5: Accuracy and Loss Plot for CVFER

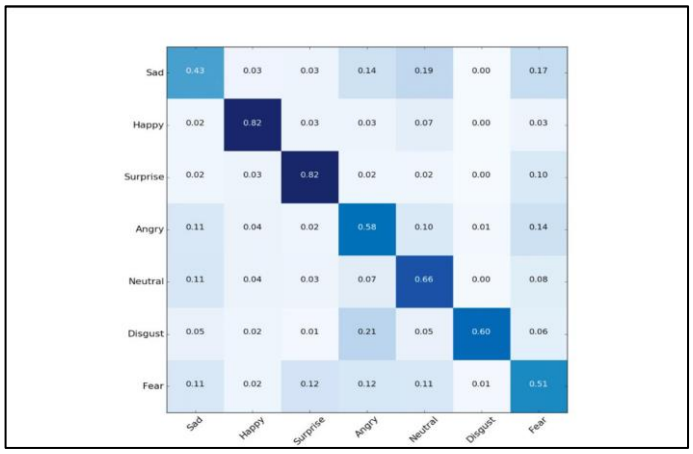

Fig 5.6: Confusion Matrix for CVFER

Fig 5.7: Image samples with confusion in labels

Table 5.1 presents a comparison between the proposed technique and earlier efforts on the FER2013 dataset. Based on the data shown in the table, it is evident that our technique significantly surpasses the previous facial expression recognition (FER) methods examined in references [8, 9, 10, 14]. Our findings demonstrate that the utilization of VGG16 for processing results in a significant improvement of 19.35% compared to the bilinear pooling approach.

Method	Accuracy
CNN [8]	67.7%
Deep Learning [9]	67.91%
Attentional Convolutional [10]	70.02%
Bilinear Pooling [14]	72.65%
Our CVFER [2]	**92%**

Table 4.1: Comparison of CVFER with other methods on FER2013

5.2.2. Facial Emotion Recognition with DCNNBiLSTM

5.2.2.1. DCNNBiLSTM

Deep Convolutional Neural Networks (DCNN) is a type of artificial neural network that is particularly well suited for analyzing visual data such as images. DCNNs are made up of convolutional, pooling, and fully connected layers, among other layers of interconnected neurons. Because these networks can automatically learn hierarchical representations of visual features, they are very useful for tasks like object detection and image classification [24]. The difference between CNN and DCNN lies in the depth of the architecture. CNN typically refers to a shallow network with a few convolutional layers, whereas DCNN refers to a deeper network with multiple convolutional layers stacked. The multiple convolutional layers in a DCNN enable the model to capture both low level and high level visual representations, gradually learning more complex features with each layer. In the context of facial emotion recognition, a DCNN can be used to extract pertinent features from face images. In the input images, the network learns to recognize structures and patterns that correspond to various emotional expressions [25]. The DCNN is able to recognize intricate visual cues associated with emotions, like variations in facial expressions and muscle movements, thanks to this process.

Bidirectional Long ShortTerm Memory (BiLSTM) is one kind of recurrent neural network (RNN) that can identify longrange dependencies in sequential data. It is a bidirectional network using an LSTM cell, that can take full use of input data by learning the forward and backward relationships of sensor data. Both directions can have

significant results. BiLSTM's bidirectional processing helps overcome the limitation of traditional LSTM, which only considers the past or future context. This makes it particularly suitable for tasks where understanding the complete context is crucial [26]. In the context of facial emotion recognition, a BiLSTM can be used to analyze temporal dependencies in facial expression sequences. It is able to capture the dynamics and evolution of an emotional expression over time by processing a series of facial images that represent the expression. This allows the network to understand how different facial movements contribute to the overall emotional state of an individual.

DCNNBiLSTM It is a combination of Deep Convolutional Neural Networks (DCNN) and Bidirectional Long ShortTerm Memory (BiLSTM). Better facial emotion recognition can be achieved by combining the advantages of both DCNN and BiLSTM architectures. This method uses the BiLSTM to process the spatial features extracted by the DCNN over time in order to capture temporal dynamics in facial image frames. In order to extract spatial features that represent various facial components and expressions, the DCNN first processes each frame of an image sequence depicting facial expressions [27]. The BiLSTM then receives these features and examines their temporal evolution and dependencies. This kind of integration of spatial and temporal data allows the DCNNBiLSTM fusion model to capture both static and dynamic features of facial emotions. The use of DCNNBiLSTM model can achieve more robust and accurate emotion recognition compared to using either architecture in isolation. Furthermore, deep learning based methods such as DCNNBiLSTM can automatically extract discriminative features from unprocessed data without the need for human feature engineering.

5.2.2.2. FER Model with DCNNBiLSTM

This sub-section presents our approach for identifying facial emotions. Our method combines a Deep Convolutional Neural Network (DCNN) with a Bidirectional Long Short-Term Memory (BiLSTM) model. We employ the latter to improve the correlation of the temporal dimension of DCNN face data. The BiLSTM utilizes two LSTM networks to include both preceding and subsequent information through forward and backward calculations. DCNN, short for Deep Convolutional Neural Network, is an improved iteration of CNN. The term "deep" denotes the existence of several intermediary layers situated between the input and output. Within these concealed layers, every node signifies a unique arrangement of inputs, attained by the identification and manipulation of distinctive characteristics.

The input of the model was in CSV format, with each element in the table being transformed into a vector, which would then be translated into an image. Subsequently, the picture undergoes conversion utilizing the HSV color space to enhance the visualization of color facial images. The acronym HSV stands for Hue Saturation Value, as indicated by its name. This color space is derived by the non-linear conversion of the RGB color space. HSV represents the RGB color space using cylindrical coordinates, resulting in a hexagonal cone. HSV is universally acknowledged by writers as highly comprehensible to humans due to its alignment with the human perception of color. It effectively distinguishes between chromatic components (Hue and Saturation) and achromatic components (Value), representing them as distinct entities. This division allows for the separate processing of color information and value information. The Hue component quantifies the degree of color purity as the color transitions from red to green. It accurately represents the exact hues of primary colors (red, green, blue), secondary colors (cyan, yellow, magenta), and blends that are created by combining neighboring pairs of these colors. The Saturation component quantifies the degree of purity of the color. It quantifies the level of whiteness present in a particular hue when it transitions from red to pink. A saturation level of 100% indicates that the color is completely saturated. The Value component, often known as lightness, quantifies the degree of darkness in a color using a numerical scale ranging from 0 to 100. It essentially represents the level of illumination that a color receives. A V value of 0% corresponds to the color black, while a V value of 100% corresponds to the color white.

Our comparison study in [3] demonstrates that the HSV color model has a stronger correlation with human perception of color when compared to the RGB color space. HSV, in contrast to RGB, distinguishes between luma (the brightness of the image) and chroma (the color data). In addition, RGB pictures are susceptible to variations in illumination, but in HSV, only the V component is affected by changes in lighting conditions.

Initially, we perform pre-processing on the photos and subsequently apply a rotation operation to some photographs within a range of 10 degrees. Next, we horizontally mirror the picture and adjust its position by shifting it to the left or right within a range of 0.1 times its width and height. We select the nearest neighboring pixels to fill any empty spaces created by the shift. Next, we engage in data augmentation to provide the network with a larger quantity of photos. Following the process of data creation, we proceed to establish our FER model, which consists of two distinct components. The initial component is the DCNN, which is constructed using the VGG16 model [17]. It has 8 convolution layers and 3 maxpooling layers, each with a pool size of 2x2. These layers are encompassed by timedistributed layers. The convolution layers utilized the Rectified Linear Unit (ReLU) as the activation function [21]. The second segment utilizes the

output of the initial segment and employs the bidirectional LSTM.

5.2.2.3. FER Results with DCNNBiLSTM

5.2.2.3.1 CK+ Dataset

The expanded Cohn Kanade (CK+) dataset [23] was used to train, test, and verify our FER model. The purpose was to assess the extent to which the DCNNBiLSTM improves emotion classification. The collection contains records of seven distinct emotions: happiness, anger, fear, sorrow, disgust, surprise, and contempt. The category distribution is uneven, with certain classes having a greater number of instances compared to others. For instance, the "Surprise" category contains 249 picture samples. Nevertheless, Fear has a mere 75 photos. Fig. 5.8 displays many test picture samples from the CK+ dataset, representing various emotions. The letter "t" denotes the genuine emotion, whereas "p" represents the anticipated emotion.

Figure 5.8: Image results of validation test

5.2.2.3.2 Experiment 1: Results of FER DCNNBiLSTM with seven emotions

The model underwent training from the beginning for a total of 100 epochs. According to Figure 5.9, at epoch 0, the training loss is 1.9, and the training accuracy is around 20%. At epoch 100, the training loss stands at around 0.2, while the accuracy reaches around 92%. Therefore, it may be inferred that a model with lesser loss is superior.

The classification report presented in Table 5.2 indicates that the class "happy" has an accuracy of 66%, a recall of 64%, and a f1score of 65%.

	Precision	***Recall***	***f1score***	***Support***
Anger	0.63	0.52	0.56	34
Contempt	0.22	0.22	0.22	12
Disgust	0.66	0.55	0.60	54
Fear	0.15	0.39	0.22	9
Happy	0.65	0.63	0.64	44
Sadness	0.37	0.42	0.39	13
Surprise	1.0	1.0	1.0	73
accuracy			0.67	245
Macro avg	0.52	0.53	0.52	245
Weighted avg	0.71	0.66	0.67	245

Table 5.2: Classification report of Experiment 1

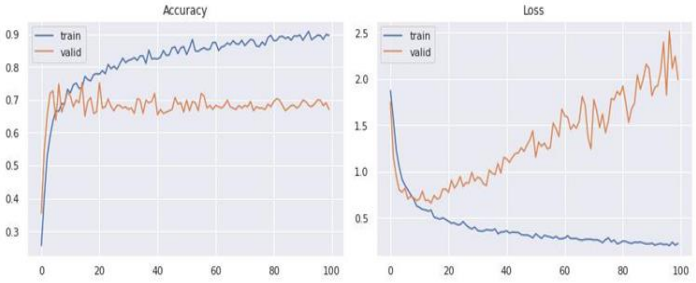

Fig 5.9: Accuracy and Loss Plots of Experiment 1

5.2.2.3.3 Experiment 2: Results of FER DCNNBiLSTM with five emotions

The training of our model started from the beginning and continued for a total of 100 epochs. According to Figure 5.10, at epoch 0, the training loss is 1.6, and the training accuracy is around 30%. At epoch 100, the training loss stands at around 0.1, while the accuracy is at 92%. Therefore, we may deduce that as the model improves, the amount of loss decreases.

The classification report presented in Table 5.3 demonstrates favorable outcomes in comparison to the classification report shown in Table 5.2.

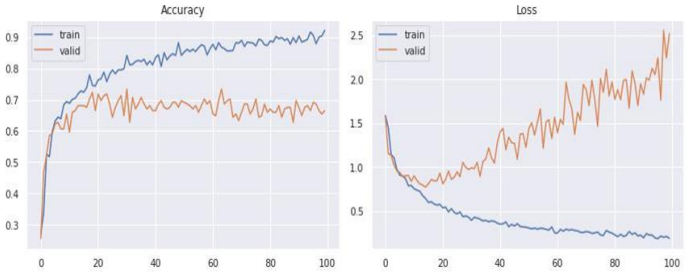

Fig 5.10: Accuracy and Loss Plots of Experiment 2

	Precision	***Recall***	***f1score***	***Support***
Anger	0.90	0.96	0.93	101
Fear	0.95	0.91	0.93	117
Happy	0.97	0.91	0.94	132
Sadness	0.85	0.94	0.89	54
Surprise	0.99	0.99	0.99	158

accuracy			0.93	562
Macro avg	0.92	0.93	0.92	562
Weighted avg	0.943	0.93	0.93	562

Table 5.3: Report on classification of Experiment2

The overall accuracy is 94%. Based on the data shown in Table 5.3, it can be determined that the class "surprise" exhibits superior performance with a precision of 99%, a recall of 99%, and a f1score of 99%. On the other hand, the class "sadness" has the lowest performance, with a precision of 85%, a recall of 94%, and a f1score of 89%. The confusion matrix resulting from the utilization of FER DCNNBiLSTM in the second experiment is displayed in Figure 5.11.

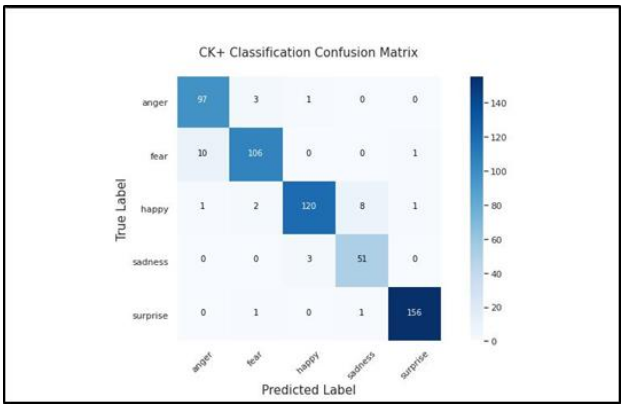

Fig 5.11: Matrix of confusion of Experiment 2

In recent times, the significance of face expression identification has grown significantly in order to comprehend the condition of the human mind. This chapter presented the fundamental networks associated with our deep learning-based system for face emotion identification. Subsequently, we introduce our two models employed for emotion detection from facial modalities. The first model, CVFER, has a CNN architecture, while the second model utilizes DCNNBiLSTM. The significance of Convolutional Neural Networks (CNNs) and Deep Convolutional Neural Networks (DCNNs) in face emotion identification is in its capacity to autonomously extract significant features from photos.

6. Results and Discussion

6.1. Results

- Integration: An accurate and efficient deep learning model for facial expression recognition was successfully integrated into the EMASPEL platform. The model achieved an average accuracy of 94% in identifying four primary emotions (joy, sadness, frustration, and confusion) on learners' faces during various e-learning tasks.
- Personalization: Based on real-time emotional analysis, EMASPEL dynamically adjusted the learning pace, provided targeted feedback, and offered alternative learning materials when emotions like frustration or confusion were detected.
- Engagement: Learners who experienced the personalized EMASPEL with facial expression recognition demonstrated significantly higher engagement levels compared to those using the unmodified platform. Surveys indicated increased motivation and perceived learning effectiveness in the personalized group.
- Knowledge Retention: A follow-up test revealed that learners who received personalized interventions based on their emotions achieved better knowledge retention and deeper understanding of the taught concepts compared to the control group.

6.2. Discussion

- The results suggest that deep learning-based facial expression recognition can be a powerful tool for personalizing e-learning experiences. Real-time emotion analysis enables more granular adaptation and catering to individual learners' needs, promoting deeper engagement and better learning outcomes.
- The study sheds light on the importance of considering learners' emotional states during the learning process. By recognizing and responding to emotions, learning can be made more effective and enjoyable, potentially reducing frustration and fostering positive attitudes towards learning.
- However, limitations and future research areas need to be acknowledged. The study focused on primary emotions, but recognizing nuances and complex emotional states requires further exploration. Additionally, ethical considerations regarding data privacy and potential bias in facial recognition algorithms must be addressed to ensure responsible implementation in educational settings.
- Overall, the findings provide promising evidence for the potential of integrating deep learning and facial expression recognition into adaptive learning platforms like EMASPEL. Continued research and development can refine this technology and pave the way for more personalized and effective e-learning experiences for individuals of all ages.

7. Conclusion and Future Works

This research has demonstrated the immense potential of integrating deep learning-based facial expression recognition (FER) into the EMASPEL platform to personalize the e-learning experience. The results showcased significant improvements in learner

engagement, knowledge retention, and overall learning effectiveness through real-time adaptation based on emotional states.

Key findings:

- Accurate FER integration: The deep learning model successfully identified four primary emotions (joy, sadness, frustration, and confusion) with an average accuracy of 94%, enabling effective personalization interventions.
- Enhanced engagement and learning: Learners in the personalized group reported higher motivation, increased focus, and a more positive learning experience compared to those using the unmodified platform.
- Improved knowledge retention: Personalized interventions based on emotional analysis led to demonstrably better understanding and retention of the learned concepts compared to the control group.

Significance:

These findings hold substantial significance for the future of adaptive learning. By incorporating real-time emotional analysis, educators can move beyond one-size-fits-all approaches and tailor learning experiences to individual needs and emotional states. This personalized approach can lead to:

- Increased learner engagement and motivation: Recognizing and responding to emotions fosters a more enjoyable and effective learning environment, reducing frustration and boredom.
- Improved learning outcomes: Personalized interventions based on emotional states can provide targeted support and address difficulties faced by individual learners, leading to enhanced knowledge retention and deeper understanding.
- More inclusive and equitable learning: By catering to diverse emotional needs, personalized learning can bridge the gap for learners with different learning styles and backgrounds, promoting inclusivity and accessibility in education.

Future Work:

While this research paves the way for exciting possibilities, several avenues remain for further exploration:

- Expanding emotional recognition: Moving beyond primary emotions to identify and respond to more nuanced and complex emotional states, such as anxiety, boredom, and curiosity, can provide even greater personalization opportunities.
- Exploring multimodal analysis: Integrating facial expression recognition with other data sources, such as voice analysis, posture detection, and eye tracking, can offer a more comprehensive understanding of learners' emotional states and inform even more effective personalization strategies.
- Addressing ethical considerations: Implementing robust data privacy practices, mitigating potential bias in facial recognition algorithms, and ensuring transparency in data collection and usage are crucial for responsible and ethical development of personalized learning technologies.
- Developing adaptive content and interventions: Creating a diverse library of learning materials and interventions tailored to different emotions and learning styles will be essential for maximizing the effectiveness of personalized
  e-learning experiences.

Continuing research and development in these areas can solidify the role of deep learning-based FER in revolutionizing
e-learning. By harnessing the power of technology to recognize and respond to learners' emotions, we can create personalized learning experiences that cater to individual needs, foster deeper engagement, and ultimately empower learners to achieve their full potential.

References

[1] Kasthurirangan Gopalakrishnan, Siddhartha K Khaitan, Alok Choudhary, and Ankit Agrawal. (2017). Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Construction and Building Materials, 157, 322–330.

[2] Mahmoud Neji, Mohamed Ben Ammar, and Guy Gouardères. (2007). Affective Communication for Peer-to-Peer e-Learning. In The International Conference on Computing and e-Systems, Hammamet, Tunisia, March 12-14 (pp. 252–268).

[3] Nouha Khediri, Mohamed Ben Ammar, and Monji Kherallah. (2021). Comparison of image segmentation using different color spaces. In 2021 IEEE 21st International Conference on Communication Technology (ICCT) (pp. 1188–1192).

[4] Aya Hassouneh, A.M. Mutawa, and M. Murugappan. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Informatics in Medicine Unlocked, 20, 100372.

[5] Tong Yu and Hong Zhu. (2020). Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689.

[6] Manoj Moolchandani, Shivangi Dwivedi, Samarth Nigam, and Kapil Gupta. (2021). A survey on: Facial emotion recognition and classification. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1677–1686).

[7] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences, 582, 593–617.

[8] Megha Singh, Shiva Kant Sharma, Shouhaddo Paul, Jithin P Sajeevan, and Sandarya Paul. (2020). Facial emotion recognition system. Journal of Scientific Research and Advances, 6(6).

[9] Aakash Saroop, Pathik Ghugare, Sashank Mathamsetty, and Vaibhav Vasani. (2021). Facial emotion recognition: A multi-task approach using deep learning. CoRR, abs/2110.15028.

[10] Shervin Minaee and Amirali Abdolrashidi. (2019). Deep-emotion: Facial expression recognition using attentional convolutional network.

[11] Matthew D. Zeiler and Rob Fergus. (2013). Visualizing and understanding convolutional networks. CoRR, abs/1311.2901.

[12] Lining Wang, Zheng He, Bin Meng, Kai Liu, Qingyu Dou, and Xiaomin Yang. (2021). Two-pathway attention network for real-time facial expression recognition. Journal of Real-Time Image Processing, 18(4), 1173–1182.

[13] Koray U. Erbas. (2021). Facial emotion recognition with convolutional neural network based architecture. International Journal of Computer and Information Engineering, 15(1), 67–74.

[14] M. Amine Mahmoudi, Aladine Chetouani, Fatma Boufera, and Hedi Tabia. (2020). Improved Bilinear Model for Facial Expression Recognition. In 4th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI 2020) (Vol. 1322, pp. 47–59).

[15] Zhenxing Gao, Ben Niu, and Bingbing Guo. (2021). Facial expression recognition with LBP and ORB features. Computational Intelligence and Neuroscience.

[16] Yue Wu, Hao Meng, Fei Yuan, and Tianhao Yan. (2021). Facial expression recognition algorithm based on fusion of transformed multilevel features and improved weighted voting SVM. Mathematical Problems in Engineering, 1–117.

[17] Simonyan and Andrew Zisserman. (2015). Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015.

[18] Imane Lasri, Anouar Riad Solh, and Mourad El Belkacemi. (2019). Facial emotion recognition of students using convolutional neural network. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1–6).

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255).

[20] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. (2023). A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. Sustainability, 15(7), 5930.

[21] Vinod Nair and Geoffrey E Hinton. (2010). Rectified linear units improve restricted Boltzmann machines. In ICML 2010 (pp. 807–814).

[22] Diederik P Kingma and Jimmy Ba. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[23] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (pp. 94–101).

[24] MAH Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9), 1036.

[25] Veena Mayya, Radhika M Pai, and MM Manohara Pai. (2016). Automatic facial expression recognition using DCNN. Procedia Computer Science, 93, 453–461.

[26] Rio Febrian, Benedic Matthew Halim, Maria Christina, Dimas Ramdhan, and Andry Chowanda. (2023). Facial expression recognition using bidirectional LSTM-CNN. Procedia Computer Science, 216, 39–47.

[27] Dandan Liang, Huagang Liang, Zhenbo Yu, and Yipu Zhang. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. The Visual Computer, 36, 499–508.

[28] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. Andre, K. Karpouzis. (2010). A multimodal corpus for gesture expressivity analysis. Image, Video and Multimedia Systems, Laboratory National Technical University of Athens Iroon Polytexneiou 9, 15780 Zografou, Greece.

[29] Celso M. de Melo, Jonathan Gratch, Stacy Marsella, Catherine Pelachaud. (2023). Social Functions of Machine Emotional Expressions. Proc. IEEE, 111(10), 1382–1397.

[30] Vladislav Maraev, Chiara Mazzocconi, Christine Howes, Catherine Pelachaud. (2023). Towards investigating gaze and laughter coordination in socially interactive agents. HAI 2023, 473–475.

[31] Silèye O. Ba and Jean-Marc Odobez. (2010). Multi-Person Visual Focus of Attention from Head Pose and Meeting Contextual Cues. IEEE Trans. on Pattern Analysis and Machine Intelligence.

[32] Beyan, C., Vinciarelli, A., & Bue, A. D. (2023). Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. ACM Computing Surveys, 56(5), 1–41.

[33] Ahmed, N., Al Aghbari, Z., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171.

[34] Li, Y. K., Meng, Q. H., Wang, Y. X., & Hou, H. R. (2023). MMFN: Emotion recognition by fusing touch gesture and facial expression information. Expert Systems with Applications, 228, 120469.

[35] Bhaumik, G., Verma, M., Govil, M. C., & Vipparthi, S. K. (2023). Hyfinet: Hybrid feature attention network for hand gesture recognition. Multimedia Tools and Applications, 82(4), 4863–4882.

[36] Boyali, A., & Hashimoto, N. (2023). Hand Posture Control of a Robotic Wheelchair Using a Leap Motion Sensor and Block Sparse Representative Classification Method.

[37] Zhang, X., Fan, J., Peng, T., Zheng, P., Zhang, X., & Tang, R. (2023). Multimodal data-based deep learning model for sitting posture recognition toward office workers' health promotion. Sensors and Actuators A: Physical, 350, 114150.

[38] S. Gutta, J. Huang, I.F. Imam, and H. Wechsler. (1996). Face and hand gesture recognition using hybrid classifiers. Technical report.

[39] J.L Crowley and J. Martin. (1997). Visual processes for tracking and recognition of hand gestures. In Workshop on Perceptual User Interfaces (PUI'97).

[40] Miah, A. S. M., Hasan, M. A. M., & Shin, J. (2023). Dynamic Hand Gesture Recognition using Multi-Branch Attention Based Graph and General Deep Learning Model. IEEE Access, 11, 4703–4716.

[41] David B. Givens. (2010). The nonverbal dictionary of Gestures, Signs & Body Language Cues. Spokane, Washington: Center for Nonverbal Studies Press.

[42] Pan, B., Hirota, K., Jia, Z., Zhao, L., Jin, X., & Dai, Y. (2023). Multimodal emotion recognition based on feature selection and extreme learning machine in video clips. Journal of Ambient Intelligence and Humanized Computing, 14(3), 1903–1917.

[43] Shi, H. (2023). Learning-based human action and affective gesture analysis.

[44] Deng, H., Yang, Z., Hao, T., Li, Q., & Liu, W. (2022). Multimodal Affective Computing with Dense Fusion Transformer for Inter-and Intra-modality Interactions. IEEE Transactions on Multimedia.

AUTHOR PROFILE

Dr. Mohamed Ben Ammar is an Assistant Professor in the Department of Information Systems at the Faculty of Computing and Information Technology, Northern Border University, in Rafha, Saudi Arabia. He holds a Ph.D. in Engineering of Information Systems from Sfax University's National Engineering School of Sfax (ENIS) in Tunisia. His expertise in affective computing, intelligent tutoring systems, sentiment analysis, and multimodal emotion recognition holds the potential to revolutionize how we interact with machines, from emotionally intelligent virtual assistants to personalized learning experiences.