Human Emotion Recognition Based on Facial Expression Using Convolution Neural Network

Gheed T. Waleed* and Shaimaa H. Shaker

Department of Computer Science, University of Technology, Baghdad, Iraq
Email: cs.21.03@grad.uotechnology.edu.iq (G.T.W.); Shaimaa.h.shaker@uotechnology.edu.iq (S.H.S.)
*Corresponding author

***Abstract*—Emotion recognition has become an essential aspect of human-computer interaction, encompassing a wide range of applications such as virtual reality, cognitive science, and digital health. Identifying human emotions through facial expressions is challenging due to the intricate and constantly changing nature of facial movements. However, the advancements that have been made in deep learning methods, specifically Convolutional Neural Networks, have shown promising results in this field. This study aims to present a Convolutional Neural Network-based deep learning model for precise identification of emotions from facial expressions. The proposed system underwent training and evaluation using the Multimodal EmotionLines Dataset (MELD) and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The experimental results validate the efficacy of the suggested methodology, attaining a remarkable accuracy rate of 96.3% in the identification of emotions for the MELD dataset and 95.86% accuracy rate in RAVDESS dataset. The system effectively tackles the challenge of distinguishing between fear and surprise emotions, which can exhibit considerable similarity, making differentiation more challenging.**

***Keywords*—emotion recognition, facial expressions, human-computer interaction, face detection, deep learning, convolution neural network**

I. Introduction

Recognizing emotions through facial expressions is essential to human connection and communication. To grasp social cues, empathize with others, and make wise decisions in a variety of interpersonal circumstances [1]. One of the main ways people express their emotions non-verbally is through facial expression, in which they reflect a wide range of emotions, including joy, sadness, anger, fear, surprise and disgust, among others, though minor, variations in their facial muscle movements [2]. These behaviors work as potent indicators of an individual's affective state, conveying important details about their intentions, emotions, and responses to inputs from the outside world [3]. The applications of this capacity span a wide range of disciplines, from psychology and neuroscience to artificial intelligence and human-computer interaction.

Manuscript received April 30, 2024; revised August 5, 2024; accepted August 12, 2024; published December 12, 2024.

The origins of facial emotion recognition can be traced back to the later part of the 20th century when psychologists and researchers initiated investigations into the correlation between facial expressions and emotions [4]. Paul Ekman and his colleagues made a significant breakthrough in identifing and classifying universal facial expressions [5]. This work led to create multiple theories and models for recognizing emotions based on facial expressions. Theoretically, there is an unlimited amount of different facial expressions, making it impossible to recognize every possible expression in a system with finite memory. By designating Action Units (AUs), such as upper lid raiser, outer brow raiser and many more, Ekman and Friesen made it possible. These AUs were created using the altered facial elements [6, 7].

Technological progress has facilitated the development of advanced algorithms and machine learning models that can automatically recognize facial emotions. Data-driven systems that use machine learning and deep learning techniques to automatically identify emotions from face photos have been made possible by the advent of digital imaging technologies and the accessibility of large-scale facial expression datasets [8]. Deep learning, particularly Convolutional Neural Networks (CNNs), has enabled the recent extraction and learning of numerous features [9, 10] for an excellent facial emotion detection system [11]. Regarding facial expressions, it is worth mentioning that the lips and the eyes are the primary sources of cues, while other facial characteristics like the ears and hair have a minor impact [12]. This means that the machine learning framework should focus primarily on crucial facial characteristics and exhibit reduced sensitivity towards other facial regions. Challenges in the realm of emotion recognition include uneven lighting, varied positions, and other facial accessories. The drawback of simultaneously optimising feature extraction and classification in emotion detection using conventional methods [13, 14]. This paper will utilize a proposed model of Convolutional Neural Network (CNN) to carry out facial emotion recognition on both the Ryerson Audio-Visual Database of Emotional Speech and Song RAVDESS dataset [15] and the Multimodal Emotion Lines Dataset (MELD) [16], which is considered state-of-the-art as it is the first research to extract facial features from MELD while simultaneously achieving a high accuracy rate, as demonstrated in the results section. The approach effectively resolves the

doi: 10.12720/jait.15.12.1366-1373

challenge of distinguishing between fear and surprise emotions, which often exhibit notable similarities, making the differentiation more complex.

The subsequent sections of this paper are outlined below. Section II provides an overview of the current research and algorithms in expression recognition. Section III provides a comprehensive description of the proposed method. Section IV offers a detailed account of the experimental methodology and presents the findings obtained from analyzing the Multimodal Emotion Lines Dataset MELD dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The entirety of the paper is summarized in Section V.

II. Literature Review

Researchers and scientists have investigated diverse techniques for the identification of facial emotions. Commonly employed techniques encompass the utilization of deep learning algorithms, such as convolutional neural networks and recurrent neural networks. These algorithms have demonstrated encouraging outcomes in accurately discerning emotions from facial expressions. Wang *et al*. [17] proposed the recognition of emotions based on the auxiliary model. This study integrates information from three primary sub-regions, namely eyes, nose, and mouth, with the entire facial image using a weighting procedure to capture the utmost amount of information. The model is assessed by the utilization of four databases, namely The Extended Cohn-Kanade Dataset (CK+), Facial Expression Recognition 2013 Dataset (FER2013), Static Facial Expressions in the Wild (SFEW), and Japanese Female Facial Expression (JAFFE). Norden *et al.* [18] suggested an innovative method for recognizing facial expressions, by employing transfer learning using pre-trained Visual Geometry Group (VGG16) and ResNet50 models. The databases used in this analysis were the Japanese Female Facial Expression (JAFFE) and FER2013 datasets. Experimental results demonstrated that ResNet50 outperformed other prominent approaches in terms of accuracy, as reported in the literature. While Bodapati *et al.* [19] proposed that deep Convolutional Neural Networks can extract features for the purpose of emotion recognition. This work utilizes the VGG16 model for feature extraction and a multi-class Support Vector Machine (SVM) for classification, in contrast to their method which used CNNs for emotion recognition. Nithya Roopa *et al.* [20] conducted a study where they utilized Inception V3 for emotion recognition. They achieved a test accuracy of 39% on the Karolinska Directed Emotional Faces (KDEF) database. Shaees [21] and his colleagues introduced a transfer learning technique that utilizes a support vector machine classifier on characteristics retrieved by convolutional neural networks such as AlexNet. Based on the present testing results, Resnet50 has demonstrated the highest level of classification accuracy when compared to other cutting-edge methods for this particular task. These extracted features are subsequently utilized as input for Support Vector Machines (SVMs) to accomplish classification. The study employed two databases, CK+ and The Natural Visible and Infrared Facial Expression Database (NVIE), to conduct the research and achieved acceptable levels of accuracy. Singh and Nasoz [22] introduced a study on facial emotion identification with convolutional neural networks. The study employed various models, namely VGG 19, VGG 16, and ResNet50, in conjunction with the fer2013 dataset to execute the experiment. Among the three models considered, VGG 16 exhibited the best level of accuracy, reaching 63.07%. In another study, Ozdemir *et al.* [23] introduced an emotion identification system based on the LeNet architecture. This study utilizes a combined dataset consisting of the JAFFE and KDEF datasets, as well as additional proprietary data collected by the researchers. The Haar cascade library is employed in this study to eliminate extraneous pixels that do not contribute to the detection of facial expressions.

III. Materials and Methods

This paper proposes a system to detect and recognize emotions through facial expressions, an overview of the proposed system is shown in Fig. 1.

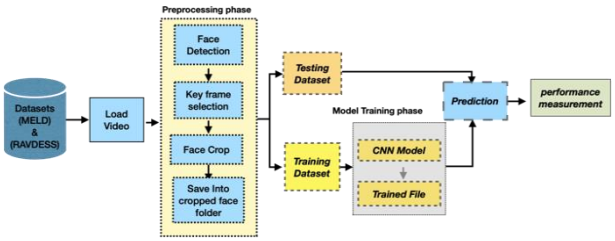


Fig. 1. Overview of the proposed recognition system.

A. Data Collection

The initial stage of the research process involves obtaining a diverse and accurately labelled dataset of facial photographs that portray a wide range of emotions. We employed two distinct datasets, with the first one being Multimodal Emotion- Lines Dataset (MELD) [15], which is an expansion and refinement of Emotion Lines Dataset [24]. The MELD dataset comprises approximately 13,000 utterances extracted from 1,433 dialogues gathered from the television series Friends. Every verbal utterance is marked with emotion and sentiment labels, and includes audio, visual, and textual elements. The total number of videos in MELD dataset that will be used is 9,990 labeled videos of 7 different emotions, with each of them being in MP4 file format and having a duration of 10 seconds or less. Table I provides a comprehensive analysis of the Multimodal Emotion Lines Dataset (MELD) statistics, including the number of samples and the percentage across different categories.

The distribution of samples across each emotion exhibits heterogeneity, with a significant fraction leaning towards neutral and joy, whereas fewer examples are observed for fear and disgust. The wide range of training examples guarantees the ability to apply learned knowledge to diverse emotional expressions, varying lengths of discourse or speaker turns, and realistic variations in conversational tone. On the other hand, the

RAVDESS database has 4,904 videos featuring actors portraying eight different emotions: happy (752 films), sorrow (752 videos), anger (752 videos), fear (752 videos), disgust (384 movies), surprise (384 videos), neutral (376 videos), and tranquility (752 videos). The video clips have a size of 1,280×720 pixels and a frame rate of 30 frames per second.

TABLE I. EMOTIONS DISTRIBUTION IN MELD DATASET

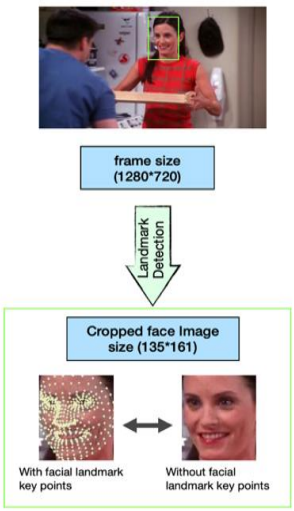
Emotion	No. of samples	Percentage (%)
Anger	1,627	13 %
Disgust	359	2 %
Fear	242	2 %
joy	2,209	17 %
Natural	6,469	43 %
Sad	1,400	13 %
Surprise	1,001	10 %

B. Preprocessing

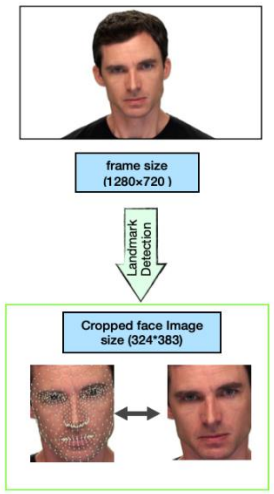
Prior to any further analysis, it is imperative to preprocess the image data. After acquiring MELD and RAVDESS datasets, we employed a range of preprocessing techniques to improve the quality. These tasks encompassed operations such as detecting, scaling, aligning faces [25] and normalizing the data [26]. Preprocessing is applied to the labeled datasets so that we reduce the time cost of the system and improve the overall performance. Initially, facial detection was accomplished using MediaPipe Face Mesh, a powerful yet efficient tool for real-time identification of facial landmarks.

Facial landmark detection algorithms utilize the BlazeFace prototype, which facilitates the automated identification of the precise positions of key facial landmark points on an image or video [27]. The process commences by capturing video frames using OpenCV, converting these frames from Blue, Green, Red (BGR) to Red, Green, Blue (RGB) format, as MediaPipe operates with RGB images. The key landmarks typically encompass the facial regions such as the tip of the nose, the corners of the eyes, the eyebrows, the mouth, and other facial contours, estimated to be around 468 3D facial landmarks. The system comprises two concurrent deep neural network models: a detector that analyzes the entire image and identifies the positions of faces, and a 3D face landmark model that uses these positions to estimate the approximate 3D surface through regression. Accurately cropping the face significantly reduces the necessity for typical data augmentations such as affine transformations involving rotations, translations, and scale adjustments. Instead, it enables the network to allocate most of its capacity towards enhancing the accuracy of coordinate prediction. Furthermore, facial landmark detection (Mediapipe mesh) has the capability to generate crops using the face landmarks detected in the previous frame. The face detector is only activated to re-localize the face when the landmark model can no longer identify the presence of a face.

Preprocessing operations involve scaling and standardization to ensure uniformity and enhance the performance of the model. Images are reduced to a standard resolution of 224×224 pixels to maintain consistency across the dataset. This helps reduce computational complexity and ensures that the model can handle images of the same size. Standardization involves rescaling pixel values to a range between 0 and 1 by dividing the pixel intensities by 255. The scaling stage in this approach standardizes the provision of information, resulting in faster preparation and more stable convergence during model training. This preprocessing pipeline ensures accurate and standardized detection of the face. Fig. 2 illustrates how landmark face detection can identify a face in the background on (a) a frame sample of the MELD dataset, and (b) a frame sample of the Ravdess Dataset.



(a)



(b)

Fig. 2. Example of face detection using Mediapip mesh. (a)MELD frame, (b) RAVDESS frame.

C. Convolution Neural Network

Defining the architecture of the neural network is essential for the development of the CNN model. This entails determining the number of convolutional layers, pooling layers, fully connected layers, and the activation functions to be employed.

Furthermore, it is crucial to take into consideration parameters such as the learning rate, batch size, and number of epochs when training the model. We developed

a customized architecture for Convolutional Neural Network specifically designed for the purpose of recognizing facial emotions as shown in Fig. 3. The CNN model employed in this study achieves an ideal balance between complexity, efficiency, and scalability. The utilization of two densely connected layers, each consisting of 128 and 64 units respectively, along with the Rectified Linear Unit (ReLU) activation function, empowers the model to effectively identify complex relationships and intricate patterns within the data provided. Rectified linear units (ReLU) are used to prevent the problem of vanishing gradients and ensure fast convergence. This is because sigmoid or hyperbolic tangent functions can experience diminishing gradients and slower adjustment over time. The feature extraction section consists of three convolutional layers, which are subsequently followed by max-pooling layers. This architecture is designed to extract hierarchical features from facial images. The convolutional layer in a neural network utilizes a convolution kernel, which is responsible for the convolution operation. The parameters of this kernel are learned during the training process. During the process of convolution, each kernel is applied to the input image to compute a feature map. Put simply, the convolution kernel moves across the input image and calculates the dot product between the input and the convolution kernel at each spatial position. Next, the feature map of each kernel is overlaid in the depth dimension to generate the output image of the convolutional layer.

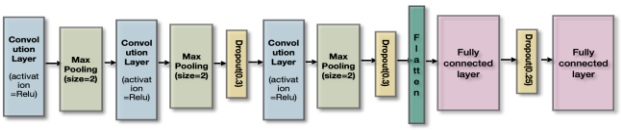


Fig. 3. The proposed CNN model.

Ultimately, the entire input matrix is traversed, resulting in the resultant matrix. The dimensions of each kernel are smaller than the dimensions of the input. The connectivity between each neuron in the activation map and the input image is limited to a small local area. This implies that the receptive field of each neuron is equivalent to the size of the kernel. The pooling layer, also known as the down-sampling layer, is designed to decrease the dimensions of the matrix produced by a convolutional layer. Pooling reduces the number of features and parameters, thereby simplifying the computational complexity of convolutional networks. Pooling additionally minimizes overfitting to a certain degree, thereby enhancing the convenience of optimization.

Max pooling is utilized in our CNN model to decrease the spatial dimensions of the input data and extract the most significant features. This methodology enables the model to concentrate on crucial facial expression patterns and enhance its capacity to precisely categorize emotions. Furthermore, our system integrates an Adam optimizer during the training procedure, thereby augmenting the model's capacity to acquire and adjust to various facial expressions.

The system also incorporates a dropout operation, which serves to prevent overfitting and enhance the model's generalization capability. This is accomplished by stochastically eliminating a specific percentage of neurons during the training process, compelling the network to depend on various combinations of features and diminishing its dependence on neurons. Setting the dropout rate to 0.5 randomly removes units during training, thereby improving robustness and flexibility. In comparison to conventional architectures such as VGG or ResNet, this model is less intricate, reducing computational requirements while maintaining sufficient complexity to benefit from the dataset. Flatten layers is also used to facilitate the transformation of multi-dimensional feature maps into a singular, one-dimensional vector, allowing for compatibility with fully connected layers used in classification. The design of our architecture draws inspiration from cutting-edge CNN models and has been tailored to efficiently capture subtle facial expressions. The Adam optimizer is utilized for optimization due to its ability to dynamically adjust the learning rate for each parameter, resulting in quicker convergence and improved generalization performance. We utilized Softmax function in the fully connected layer as showed in Eq. (1), and employed categorical cross-entropy as the loss function as demonstrated in Eq. (2), for the compiling of the system:

$$\sigma(z)_i = \frac{e^{z_i}}{\Sigma_{j=1}^{k} e^{z_j}} \tag{1}$$

where, z_i is the i-th element of the input vector z, $\sigma(z)_i$ is the *i*-th element of the output vector $\sigma(z)$, which indicates the probability of class *i*.

$$l(y, \hat{y}) = -\Sigma_{i=1}^{C}\, y_i \cdot \log(\hat{y}) \tag{2}$$

where, C is the number of classes, y_i is the true probability of the *i-th* class (1 for the true class, 0 for others), $\hat{y}_i$ is the predicted probability of the *i-th* class (output of the softmax function). This tailored framework ensures effective training and precise accuracy for tasks involving emotion recognition, making it well-suited for the specified application.

In addition, we conducted a thorough evaluation of the model's performance using metrics such as accuracy, which is measured as Eq. (3):

$$Accuracy = \frac{no.correct\ predictions}{total\ no.prediction} \tag{3}$$

IV. Result and Discussion

This section presents the empirical analysis of the suggested methodology to assess the proposed system. Prior to delving into the details of the model's performance on the datasets. The model underwent training for 90 epochs on the RAVDESS dataset and 125 epochs on MELD dataset from the beginning, using a DELL computer, Processor Intel® Core i7-8550U CPU 1.80 Hz, 1.99 GHz, RAM 16GB, 64bit operating processor.

A. Face Detection Results

The face detection process using facial landmark (Mediapipe Mesh) achieved high accuracy rates on both datasets. Table II displays the accuracy rates of face detection achieved through the implementation of the Facial Landmark detection. According to the accuracy rates presented in the table, it is clear that the Facial Landmark detection has exhibited impressive performance in detecting faces. The method's high accuracy rates underscore its potential for precisely identifying and localizing faces in diverse environments.

TABLE II. NUMBER OF DETECTED FACES USING FACIAL LANDMARK DETECTION ON MELD AND RAVDESS

Faces	MELD Dataset			RAVDESS Dataset	
	Face Image	Not Face	Accuracy(%)	Face Image	Accuracy(%)
Anger	37,877	2742	92.76%	19,415	100%
Disgust	9,722	816	91.61%	19,770	100%
Fear	9,727	785	91.93%	18,002	100%
Joy	4,284	352	91.78%	18,634	100%
Neutral	92,803	5027	94.58%	27,177	100%
Sadness	20,844	1223	94.13%	18,198	100%
Surprised	40,852	2907	92.88%	17,356	100%

B. Model Training and Testing Results

The dataset was divided into training and testing sets using scikit-learn's model selection module. 80% of the data were designated as the training data, while 20% constituted the testing data in RAVDESS dataset. This ensures that the progress of the model is guided by continuous evaluation of evidence data, thus preventing overfitting. The validation part of the training method provides a reliable evaluation of model performance on a subset of data that was not used for training. To mitigate the problem of overfitting in our experiment, we randomly selected 4,000 facial images from each of the categorized labeled folders with different emotions. This number was chosen because it corresponds to the minimum number of facial images detected in the (Joy) folder. We conducted experiments with three distinct batch sizes and three different data splits. The model's performance was assessed using the accuracy metric.

Our experimental findings indicate that the complexity and distribution of data significantly influence the efficacy of the proposed convolutional neural network architecture for emotion classification. The batch size determines the number of training examples used in each iteration, affecting the speed at which the model reaches a solution and its overall efficiency. Through experimentation with various batch sizes, it was discovered that batch sizes of 64,100, and 150 had the greatest impact on speed, error rate, and training duration, depending on the specific style. Occasionally reducing the batch size to 64 frequently improved accuracy and reduced loss, but it also increased the duration of training. Increasing the batch sizes, such as to 150, resulted in faster convergence but slightly lower precision. The batch size of 64 achieved optimal balance between precision and training efficiency, leading to a high accuracy rate within a reasonable training duration. Partitioning the data into distinct training and testing sets significantly impacts the model's ability to generalize. Three distinct ratios of training to testing data were examined: 70:30, 80:20, and 90:10. The performance of the distributions was evaluated using accuracy metric. The highest level of accuracy was achieved using a 90:10 split for MELD and 80:20 split for RAVDESS, indicating sufficient training data for effective learning and enough testing data for reliable evaluation.

An exhaustive scan was conducted to thoroughly investigate the specified ranges of batch sizes and distributions. Each blending was meticulously evaluated to determine the most effective solutions. Within each data split, cross-validation calculated the average execution estimates across different folds. The model was trained using different combinations of batch sizes and learning rates for a specific number of iterations using the Adam optimization algorithm. Early halting was implemented to alleviate problems caused by overfitting. Table III summarizes the hyperparameters and their effect on the accuracy rate and loss on the MELD dataset. Table IV, on the other hand, demonstrates the same hyperparameters and their effect on the accuracy rate and loss for the RAVDESS dataset. Considering that human emotion recognition involves classifying emotions into multiple categories, the confusion matrix serves as a comprehensive metric for assessing and illustrating the effectiveness of the proposed method.

TABLE III. ACCURACY AND LOSS RATE ON DIFFERENT BATCH SIZES AND DATA SPLITS FOR MELD

Items	Batch size	loss	Accuracy rate (%)
Data split	64	12.9%	95.62%
Training Data 70%	100	14.22%	95.32%
Testing Data 30%	150	14.91%	94.99%
Data split	64	16.26%	94.38%
Training Data 80%	100	12.2%	95.59%
Testing Data 20%	150	17.64%	94.13%
Data split	64	11.23%	96.31%
Training Data 90%	100	12.16%	95.75%
Testing Data 10%	150	11.76%	95.28%

TABLE IV. ACCURACY AND LOSS RATE ON DIFFERENT BATCH SIZES AND DATA SPLITS FOR RAVDESS

Items	Batch size	Loss	Accuracy rate (%)
Data split	64	27.21%	93.54%
Training Data 70%	100	34.03%	93.73%
Testing Data 30%	150	14.91%	93.26%
Data split	64	32.21%	93.28%
Training Data 80%	100	21.01%	95.73%
Testing Data 20%	150	45.19%	92.28%

The confusion matrix which is used to evaluate the effectiveness of the proposed method for recognizing human emotions is shown in Fig. 4 displays the true classes in rows and the predicted classes in columns showing that the emotion (Disgust) achieved the highest rate among other emotions.

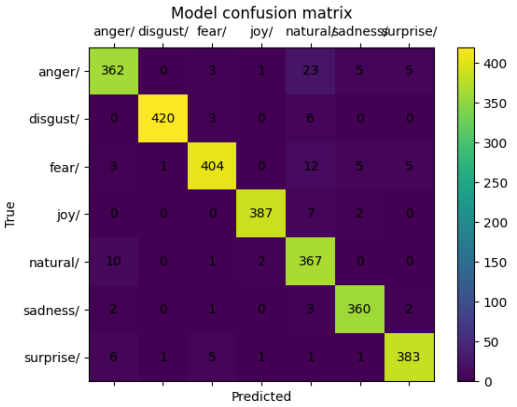


Fig. 4. Confusion matrix of Seven basic classified emotions on MELD dataset.

Fig. 5 illustrates the model accuracy with 120 epoch.

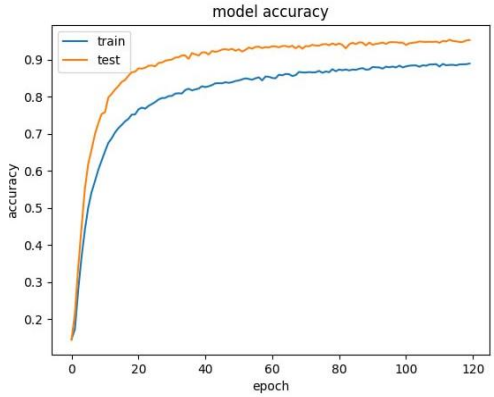


Fig. 5. Model training versus testing accuracy for MELD using proposed model.

Fig. 6 shows the model loss with 120 epochs using the MELD dataset.

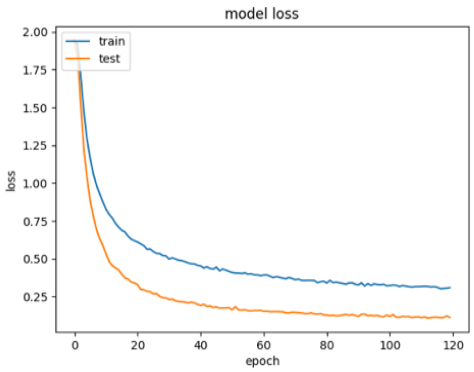


Fig. 6. Model training versus testing Loss for MELD using proposed model.

Fig. 7 displays the true classes in rows and the predicted classes in columns showing that the emotion (Disgust) achieved the highest rate among other emotions.

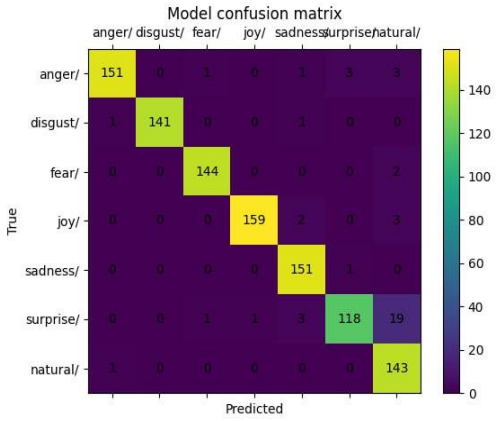


Fig. 7. Confusion matrix of Seven basic classified emotions on RAVDESS dataset.

Fig. 8 demonstrates the model accuracy with 80 epoch by use the RAVDESS dataset.

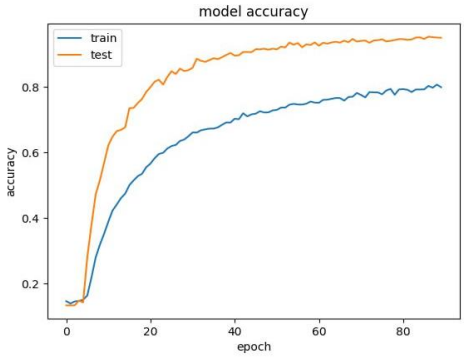


Fig. 8. Model training versus testing accuracy for RAVDESS using proposed model.

Fig. 9 illustrates the model loss rate with 100 epoch by using the RAVDESS dataset.

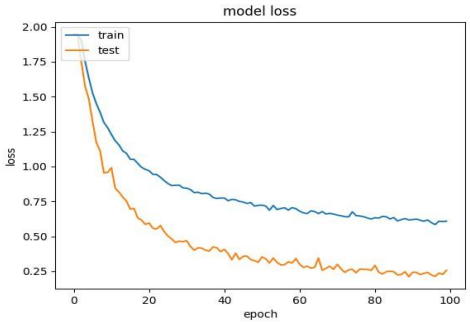


Fig. 9. Model training versus testing Loss for RAVDESS using proposed model.

The model has achieved great results when comparing between fear and surprise, two emotions with similar facial expressions. The deep neural network employs sequential convolutional layers to detect subtle variations in facial expressions. Fear and surprise can be differentiated by focusing on nuanced differences in the shape and stiffness

of the eyebrows, lips, and eyes. For example, fear typically involves more pronounced wrinkling of the forehead and a tight mouth, while surprise is characterized by raised eyebrows and a relaxed, open mouth. Moreover, data augmentation techniques offer different approaches that assist the model in accurately recognizing these subtle differences. The model's visualizations of extracted features, such as Gradient-weighted Class Activation Mapping, demonstrate that it focuses on specific facial regions when distinguishing between fear and surprise. This highlights the model's ability to effectively differentiate between these two emotions.

The CNN-based facial emotion recognition model achieved state-of-the-art performance on the MELD dataset, and a high accuracy rate on RAVDESS dataset in comparison with other models as shown in Table V, which shows a comparison with other existing studies. and Table VI, which compare our models with other generated models on the same tested samples of RAVDESS and MELD datasets using different evaluation metrics.

TABLE V. A COMPARISON BETWEEN OUR MODEL AND OTHER MODELS ON RAVDESS DATASET

Model	Accuracy rate
The proposed model	95.73%
C. Jiménez [28]	80.08%
C. Jiménez [29]	86.70%

TABLE VI. EVALUATION METRICS FOR DIFFERENT MODELS ON THE MELD AND RAVDESS DATASETS

Model	Accuracy		Precision		Recall		F1-Score	
	M	R	M	R	M	R	M	R
Proposed CNN	96.3%	95.8%	95.2%	95.1%	96.1%	95.8%	95.6%	95.4%
SVM	78.5%	76.4%	77.0%	75.2%	78.0%	76.0%	77.5%	75.6%
Random forest	81.2%	79.7%	80.5%	78.9%	81%	79.5%	80.7%	79.2%
VGG 16	93.6%	92.4%	92.8%	92%	93.2%	92.3%	93%	92.2%
ResNet 50	94.8%	94.2%	94.1%	93.9%	94.5%	94.1%	94.3%	94%

Note: M stand for MELD; R stand for RAVDESS

The comprehensive comparison results highlight the superior performance of the proposed CNN model across both data sets, surpassing traditional machine learning models such as SVM and random forest, as well as deep learning architectures like VGG16 and ResNet50. This achievement highlights the effectiveness of the proposed model in capturing complex facial features and subtle variations in expressions that are crucial for accurate emotional recognition. The innovative strategy of our CNN, incorporating modified convolutional and pooling layers alongside advanced techniques such as drop out and Adam optimisation, has proven crucial in achieving these results. The impressive performance of the proposed model demonstrates its potential in real-time emotional recognition systems, enhancing human-computer interaction by providing accurate and prompt responses.

Convolutional Neural Networks are required for efficient emotion recognition, as they allow for quick analysis while minimising resource consumption. When adequately trained, a Convolutional Neural Network (CNN) can rapidly assess sentiments within milliseconds, which is crucial for immediate response in real-time scenarios. It is crucial to note that all processing requires time. Graphics processors efficiently process large inputs within a short period of time, while central processors may take seconds to handle complex designs. When deploying models to physical devices, it is necessary to strike a balance between accuracy and computational efficiency due to their limited capabilities. Quantisation can reduce the size of a model and increase its speed without sacrificing much accuracy. In order for Convolutional Neural Networks to function effectively in real-world scenarios, it is crucial to achieve superior performance while also considering practical limitations, in order to maintain user satisfaction.

The CNN model used for facial emotion recognition frequently misclassifies emotions due to variations in muscle movements, lighting issues, obstructions, and head orientations. These difficulties highlight the vulnerability of the model to environmental factors and its inability to capture nuanced expressions. To make progress, the dataset needs to be expanded with a variety of diverse illustrations, different lighting and pose variations, and advanced data augmentation techniques. In addition, it is better to add another recognition techniques for more accurate results like voice or body. It is essential to overcome these challenges in order to develop facial emotion recognition frameworks that are more reliable.

V. CONCLUSION

In conclusion, the study on emotion recognition utilizing Convolutional Neural Networks (CNN) has produced promising results, achieving a state-of-the-art with an accuracy rate of 96.31% on MELD and 95.73% on RAVDESS datasets. Ultimately, the achievement of this study sets the stage for ongoing progress in the field of emotion recognition and its wide-ranging implementations. Potential future work for this study may encompass integrating upper body movements into the facial emotion recognition system. Empirical data indicates that movements of the upper body can convey significant emotional signals that enhance facial expressions. Integrating analysis of upper body movement can enhance the accuracy and resilience of the emotion recognition system, particularly in real-life situations where individuals commonly employ both facial expressions and body language to communicate their emotions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Gheed T. Waleed and Shaimaa H. Shaker conceptualized, carried out methodology, and conducted formal analysis. Gheed T. Waleed performed experimental tasks, drafted the original manuscript, and Shaimaa H. Shaker supervised, reviewed, and edited the manuscript. Both authors contributed equally to the final version of the manuscript. All authors had approved the final version.

REFERENCES

[1] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, Nov. 2021. doi: 10.1109/MSP.2021

[2] D. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," *Journal of Physics: Conference Series*, vol. 1193, 012004, 2019.

[3] S. A. Al-agha, H. H. Saleh, and R. F. Ghani, "Multi-model emotion expression recognition for lip synchronization," in *Proc. 2020 1st. Information Technology to Enhance E-learning and Other Application (IT-ELA)*, Baghdad, Iraq, 2020, pp. 171–177.

[4] S. Modi and M. H. Bohara, "Facial emotion recognition using convolution neural network," in *Proc. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2021, pp. 1339–1344. doi: 10.1109/ICICCS51141.2021

[5] P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System*, Consulting Psychologists Press, 1978.

[6] J. Harrigan, R. Rosenthal, K. Scherer, *New Handbook of Methods in Nonverbal Behavior Research*, Oxford University Press, 2008.

[7] V. Bettadapura, "Face expression recognition and analysis: the state of the art," arXiv preprint, arXiv:1203.6722, 2012.

[8] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, 3046, 2021.

[9] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Applied Sciences*, vol. 2, no. 3, 446, 2020.

[10] A. Altalbi, S. Shaker, and A. Ali, "Anomaly detection from crowded video by convolutional neural network and descriptors algorithm: Survey," *International Journal of Online & Biomedical Engineering*, vol. 19, no. 7, July 2023.

[11] A. Ali and M. Abdulmunem, "Image classification with Deep convolutional neural network using tensorflow and transfer of learning," *Journal of the College of Education for Women*, vol. 31, no. 2, pp. 156–171, 2020.

[12] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 500–508, 2006.

[13] S. Gupta, "Facial emotion recognition in real-time and static images," in *Proc. 2018 2nd International Conference on Inventive Systems and Control (ICISC)* 2018, pp. 553–560.

[14] A. Mollahosseini, D. Chan, M. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[15] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS One*, vol. 13, no. 5, 2018.

[16] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," arXiv preprint, arXiv:1810.02508. 2018.

[17] Y. Wang, Y. Li, Y. Song, X. Rong, "Facial expression recognition based on auxiliary models," *Algorithms*, vol. 12, no. 11, 227, 2019.

[18] F. Norden and F. Marlevi, "A comparative analysis of machine learning algorithms in binary facial expression recognition," Master dissertation, School of Electrical Engineering and Computer Science (EECS), 2019.

[19] J. Bodapati and N. Veeranjaneyulu, "Facial emotion recognition using deep CNN based features," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7, 2019.

[20] N. Roopa, "Emotion recognition from facial expression using deep learning," *International Journal of Engineering and Advanced Technology (IJEAT)*, pp. 2249–8958, 2019.

[21] S. Shaees, H. Naeem, M. Arslan, M. Naeem, S. Ali, and H. Aldabbas, "Facial emotion recognition using transfer learning," in *Proc. 2020 International Conference on Computing and Information Technology (ICCIT-1441)*, 2020, pp. 1–5.

[22] S. Singh and F. Nasoz, "Facial expression recognition with convolutional neural networks," in *Proc. 2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0324–0328.

[23] M. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using CNN architecture," in *Proc. 2019 Medical Technologies Congress (tiptekno)*, 2019, pp. 1–4.

[24] S. Chen, C. Hsu, C. Kuo, and L. Ku, "Emotionlines: An emotion corpus of multi-party conversations," arXiv preprint, arXiv:1802.08379, 2018.

[25] F. Q. Abdulalla, A. S. Abduljabar, and S. H. Shaker, "A survey of human face detection methods," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 10, no. 2, 2018. doi: 10.29304/jqcm.2018.10.2.392

[26] H. Al-Dabbas, R. Azeez, and A. Ali, "Machine learning approach for facial image detection system," *Iraqi Journal of Science*, pp. 6328–6341, 2023.

[27] S. Gupta, P. Kumar, and R. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 8, no. 8, pp. 11365–11394, 2023.

[28] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," *Sensors*, vol. 21, no. 22, 7665, 2021.

[29] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. Montero, F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," *Applied Sciences*, vol. 12, no. 1, 327, 2021.