Deep Neural Network and Transformer Models for Emotion Recognition

Sabah Anwer Abdulkareem[1*]
[1]Department of Computer Engineering, College of Engineering, University of Diyala, Diyala,32001, Iraq.
sbh_anwar@uodiyala.edu.iq

Abstract

Recent decades have seen significant progress in the growing intersection between artificial intelligence and the way we understand and manage emotions, commonly referred to as emotional intelligence. Although understanding and predicting emotions is one of the greatest human abilities, it has now become an important research area in computer science and information technology. Emotion recognition is a fundamental task in affective computing and a key development area for improving human computer interaction. Early efforts to perform emotion recognition relied primarily on traditional techniques of signal analysis and were unable to provide definitive results. The limitations of conventional signal analysis methods in emotion recognition, the requirement for effective and scalable deep learning models, and the incorporation of these models into practical applications are the main issues this study attempts to address. Deep learning models have become vital due to the breakthroughs in artificial intelligence over the past decade. To this end, in this study, we trained two models, Vision Transformer (ViT) and ResNet50V2, to recognize emotions from image datasets. Among these models, ViT stood out with an impressive accuracy of 78% and an AUC of 0.86 for emotion detection and identification on the FER2013 dataset, demonstrating its strength in understanding facial expressions. This work establishes to develop scalable, efficient, and adaptable models that integrate seamlessly into real-world applications while continuing to push the boundaries of human-computer interaction. This work enhances the ongoing advancements in efficient computing by showcasing the effectiveness of Vision Transformer in emotion detection and offering insights into its use in various interactive systems. In the future, researchers can investigate sophisticated data augmentation approaches. Including other modalities, such as speech, physiological cues, and facial expressions, aids in capturing emotions in greater detail.

Keywords: Emotion Recognition, Deep Learning, CNN, Human-Computer Interaction, Vision Transformer

Article history: Received: *11 Nov 2024****; Accepted:*** *12 Jan 2025****; Published: 15 Mar*** *2025*

1. Introduction

Human emotions play a powerful role in shaping our actions, influencing how we interact with others and make decisions. Being able to predict emotions is becoming increasingly important not just for improving how humans and computers work together but also for applications in areas like mental health care and customer service. With the advancement of technology, People expect interactive machines to interpret, understand, and cross emotions in human-like ways. This creates an increasing need for systems of emotion recognition that can precisely sense and respond to human feelings, making interactions more natural and efficient Body language. Systems of emotion recognition work by an emotional person's

[*] Corresponding author: sbh_anwar@uodiyala.edu.iq

identifying state depending on verbal and non-verbal signals, such as facial expressions, voice tone, and body language. In 2020, 19.87 million USD was valued in the global market for emotion recognition. It is expected to grow significantly, arriving at 52.86 USD million by 2026, paid by an interesting Compound Annual Growth Rate (CAGR) of 18.01% between 2021 and 2026. This fast growth confirms the increasing significance of this technology in a wide range of areas [1].

The technology of emotion recognition, supported by Artificial Intelligence (AI), collects auditory and visual cues, such as voice expressions and facial expressions, to achieve more accurate results. The technology can better explain human emotions by integrating these sensory inputs with AI. This progress makes easier the way for the AI robots' creation and responding machines capable to human actions and voices, enabling them to perform tasks in ways that conform more closely with human behavior and expectations [2].

While discussing the related work, it is evident that while numerous studies have advanced research in facial expression recognition using deep learning, there are critical gaps. Existing methods, particularly CNN-based methods, have a tendency to suffer from the selection of layers for feature extraction, leading to compromised performance. However, these limitations indicate the necessity for innovative architectural designs that can capture fine facial details. Previous studies on facial expression recognition using methods of deep learning often confront challenges, especially in improving the convolutional neural networks (CNNs) architecture. An important problem was the appropriate layers selection within CNNs, which directly affected the model's ability to efficiently extract features of relevance from facial images. selection of insufficient layers led to less-than-optimal performance, as critical details in facial expressions were either overgeneralized, underrepresented, or hampering the recognition accuracy.

These restrictions highlighted the need for more accuracy and flexible approaches to CNN architecture design for got better emotion recognition. Basic emotions such as happiness, sadness, anger, disgust, fear, and surprise are fundamental and cross-cultural emotions that influence human behavior. Researchers have developed classification and dimensional models [3]. suggest an attentional convolutional network–based deep learning method that outperforms previous models on different datasets, such as CK+, FER-2013, JAFFE and FERG, by concentrating on the face's critical areas. The findings revealed that certain facial features are responsive to various emotions [4]. suggested a model that contains 10 fundamental and 56 secondary emotions. These large emotions numbers are a challenge to predict and test, where increasing complexity to emotion recognition whenever increases positive and negative emotions coexistence. Few research studies take contradiction, but [5], authors show better in algorithms of sentiment detection to handle contradiction.

Transformers, which were first applied in natural language processing, have in the recent past claimed to operate with success in computer vision tasks like facial emotion recognition. However, there is no yet established research that investigates and knows the comparability of ViT-based models to CNN for this purpose. Moreover, most known research fails to thoroughly analyze how scalable and efficient transformer-based approaches are in real-time emotion recognition. Closely associated with the above limitations, our study examines the performance of ViT vs. ResNet50V2 in facial emotion recognition. We compare the efficiency of Vision Transformer (ViT) and ResNet50V2 models on facial emotion recognition from the FER2013 dataset in this work. Our contributions are summarized as follows:

1. We provide a comparative analysis of CNN-based (ResNet50V2) and transformer-based (ViT) architectures for emotion recognition, highlighting their strengths and limitations.
2. We demonstrate that ViT achieves superior performance, with an accuracy of 78% and an AUC of 0.86, showcasing its potential for improved emotion detection.
3. We analyze the scalability and adaptability of transformer-based approaches for real-world emotion recognition applications, addressing

challenges such as computational efficiency and model robustness.

Our hypothesis is that transformer models like ViT are more suited for facial emotion recognition compared to standard CNN architecture due to the fact that they are more efficient at handling long-range dependencies within facial expressions.

2. Related work

Using advanced computer vision and AI to read and understand people's feelings through their facial expressions is based on facial emotion recognition [5]. While current developments make use of deep learning, specifically Convolutional Neural Networks, traditional approaches concentrated on rule-based systems and manually generated features. CNNs automatically create hierarchical representations from raw visual input, enabling more complex and accurate emotion recognition [6]. Anger, joy, sadness, fear, surprise, and disgust are common face expressions associated with emotions. Every emotion has a corresponding facial muscle shape and movement, such as a smile for enjoyment or a widening of the eyes and lifted eyebrows for astonishment. FER algorithms use these expressions to determine an individual's emotional state [7].

Robust FER models in supervised learning require training data. To train algorithms to characterize patterns and correlations between emotions and visual attributes, we need labelled datasets of faces with comparable emotions. Increasing dataset quantity, variety, and quality improves FER approaches. Eye position, lip movement, and facial muscle movements are extracted from photos or videos using FER. Machine learning algorithms is typically trained on labeled data to recognize patterns in face expressions. FER has several applications, including social robotics. With FER, machines can effortlessly communicate and respond to human emotions. By allowing systems to respond to emotions, we can improve user-computer interaction. The strategy can help educational programs discover changes in disgruntled students. FER also interacts with entertainment and e-commerce recommendation systems. These technologies can increase emotional relevance and personalization by recognizing users' emotional responses and suggesting products that fit their moods [8].

In healthcare, FER facilitates the monitoring of mental health. Facial expression analysis can facilitate the early detection of illnesses and the monitoring of therapy outcomes. It is beneficial in autism research, enabling carers and experts to interpret individuals' facial expressions with autism spectrum disorders. Because of the diversity of facial emotion expressions across diverse cultural backgrounds, ages, and genders, creating universal emotion models is challenging. Therefore, maintaining a balance with technological advances is the primary focus in developing FER. The future of FER is promising, which includes gesture and speech recognition along with facial analysis [9].

Bota et al. [10] conducted a thorough examination of emotion identification by extracting physiological features and passing them into machine learning models. However, the review study that has been conducted did not utilize the strategies of systematic review, which should adhere with the PRISMA guidelines. Furthermore, their review work provided minimal analysis, and lack of improved future suggestions and directions. Previous studies have also used dual-modal joint learning [11]. The LXMERT framework uses transformers to obtain multimodal representations of images and language using self-learning and joint learning mechanisms [12]. In the study of Odhiambo and Mwashita [13], the authors defined a joint attention network for audio-visual synchronization.

In the previous work of Adyapady [14], an extensive analysis of emotion identification from facial images utilizing machine learning and deep learning methodologies. Their method for reviewing emotion detection does not adhere to PRISMA principles. The authors examined several methodologies, datasets, and several applications of emotion recognition. In the work of Guo et al. [15], the authors introduced an innovative deep convolutional neural network (DCoT) model using transformer encoders for EEG-based emotion recognition, examining the reliance of emotion recognition on individual EEG channels and the interpretation of extracted features, achieving an average accuracy of 93.83% across three classification tests. The mean accuracy across three subject-independent categorization tasks is 83.03%.

In e-learning, individual user face detection uses facial expressions to assess emotions. To highlight the influence of emotion on human attention mechanisms, reasoning, learning, perception, and decision-making, this interdisciplinary field integrates psychology, computer science, and cognitive science. An example of affective applied computing is the automated-driving assistance systems, which typically adopt several types of signals, including physiological, to help monitor the drivers, issue alerts on risks, or perform certain actions such as considering the safety by reducing the speed. Another group of researchers in [16], [17], and [18]. conducted a comprehensive analysis of the state-of-the-art research in emotion identification. In their work, they highlighted the available data and machine learning models adopted for sentiment analysis and the fusion of extracted features. Their research does not consider the theories of emotion identification and does not address the reported low evaluation metrics performance of multimodal identification systems. In the works of Debnath et al. [18] and Hassan et al. [19] presented a thorough analysis of emotion identification and methods related to word embedding, architectural frameworks, and training processes, and the integration of various modalities to express emotions by Kanjo et al. [20] the authors focused on advanced emotion recognition in conversations through natural language processing, without examining multimodal emotion recognition [21]. Conducted a survey of multimodal analysis, addressing challenges and trends in text and video, but did not investigate additional body language such as gait, gesture, or interaction between various emotions. Reviewing audio elements that are associated with emotions and musical emotion recognition, Mehendale [22] focuses on EEG-based emotion recognition while acknowledging that their approaches are restricted to EEG methods. In the study of Minaee et al. [23], the authors focus on facial expression recognition techniques emphasizes postural differences, excluding other patterns utilized in wearable and portable devices for feeling sensing by Schoneveld et al. [24] the authors shown that the methodology employed is the main focus of research on automated emotion recognition in clinical settings; however, the larger MER trends and issues taken into account in this study are not. Valuable areas are highlighted while omitting discussions of trends, challenges, and future directions. provided a PRISMA-compliant systematic review on speech signal emotion recognition. Although their approach addressed the use of ML and DL techniques, it did not address comprehensive research objectives or research obstacles [25]. Table 1 shown Researches Summaries

Table 1: The summary of the related work

Ref.	Approach	Future Directions	Accuracy
Dixit and Satapathy [26]	Vision Transformer for emotion classification	Enhancing the interpretability of the model and adding multimodal data	84.5%
Karani et al. [27]	Lightweight CNN + efficient data processing algorithms	Creating smaller models and increasing accuracy while using less resources	79.9%
Aggarwal et al. [28]	Pretrained models + transfer learning for emotion recognition	Enhancing generalization in various contexts	83.8%
Zakieldin et al. [29]	Cross-attention with hybrid feature networks	Improving attentional processes and reaching more intricate emotional states	87.0%
Manalu and Rifai [30]	CNN + LSTM for temporal feature extraction	lowering latency and integrating with more reliable temporal models	82.1%
Abad and Gholamy [31]	Transformer + CNN fusion for multimodal data	Adding more modalities and cutting down on computation costs	85.2%
Aina et al. [32]	Hybrid CNN + attention mechanisms for real-time analysis	Improving accuracy in dynamic circumstances while lowering computing needs	80.2%

Therefore, the largest challenge to FER is facial expression variability across cultures, ages, and genders, which restricts developing a universal model with good generalizability to different populations. Changes in illumination conditions, occlusions, and head poses also influence the accuracy of recognition. Addressing all these challenges remains one of the main research challenges. Despite the progress in FER, current methods have some limitations, such as the absence of good generalization over diverse datasets, inefficient use of computational resources, and limited multimodal fusion methods. The majority of current work has focused on CNN-based models, which are excellent at feature extraction but not at modeling long-distance dependencies in facial expressions. Certain of the recent advancements in transformers, such as Vision Transformers (ViT), have exhibited tremendous potential in image processing but are not explored in FER.

In this work, we bridge this gap by suggesting two classification approaches using ResNet50V2 and ViT models to identify emotions from facial image data. The proposed model combines the strengths of convolutional-based and fully attention-based systems to enable more efficient information exchange among different layers. Through the integration of transformers into FER, we aim to push the boundaries of emotion recognition and improve the development of stronger and more adaptable FER systems for practical applications.

3. Methods and Materials

3.1 Datasets

The FER2013 dataset, which stands for face Expression identification 2013, is extensively utilized in the domain of face expression identification and is frequently employed in machine learning and computer vision applications. It includes labeled data of facial expressions for training and testing models that can automatically identify human emotions from facial expressions. The dataset includes 35,887 greyscale 48 × 48 pixel pictures of faces. Each image represents one of seven unique facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. Each image is labeled with the corresponding emotion, and the data is split into training, validation, and test sets. Each image is represented as a flat vector of 2,304 dimensions (48 × 48 pixels), and labels are provided as integers corresponding to emotion classes. FER2013 was submitted as part of the Kaggle Facial Expression Recognition Challenge in 2013 [33]. The FER-2013 dataset includes the following number of images for training and testing. The whole data description is below in Fig. 1.

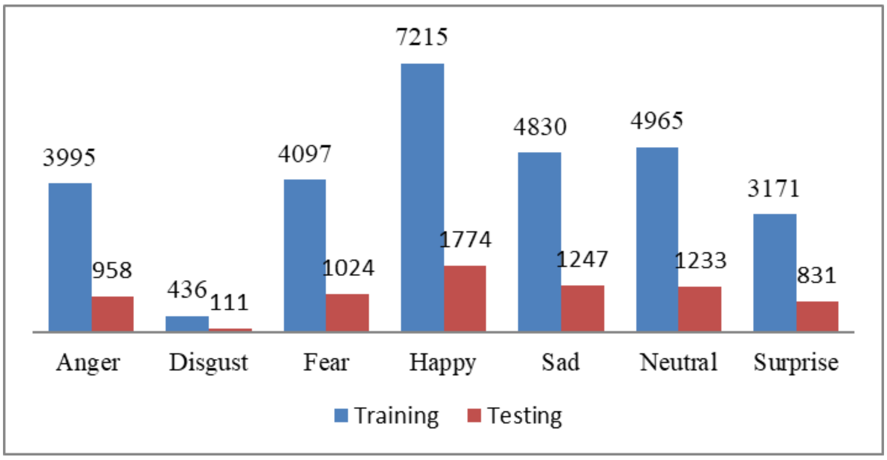

Fig. 1 Data description.

3.2 Methodology

This study utilises innovative deep learning and vision algorithms within the realm of computer vision approaches. This study presents a classification methodology utilizing ResNet50V2 and LLaVa, an effective vision transformer model. This study employs ResNet50V2, an enhanced iteration of the ResNet architecture, which utilises residual learning to address the vanishing gradient problem in emotion detection, facilitating the extraction of deep hierarchical features from photos that convey nuanced facial expressions related to emotions. Moreover, LLaVa utilises extensive vision-language models for picture analysis. The LLaVa model incorporates visual elements to improve the understanding of emotional context, rendering it appropriate for emotion recognition tasks. This methodology seeks to recognize expressions from facial picture data to ascertain emotions. The objective of multi-emotion recognition is to ascertain the existence of emotions. The overall architecture of the proposed method for multi-class emotion detection uses images as input to identify emotions. Models will be assessed based on accuracy, precision, and the AUC-ROC curve. Fig. 2 illustrates the Block diagram of the suggested technique.

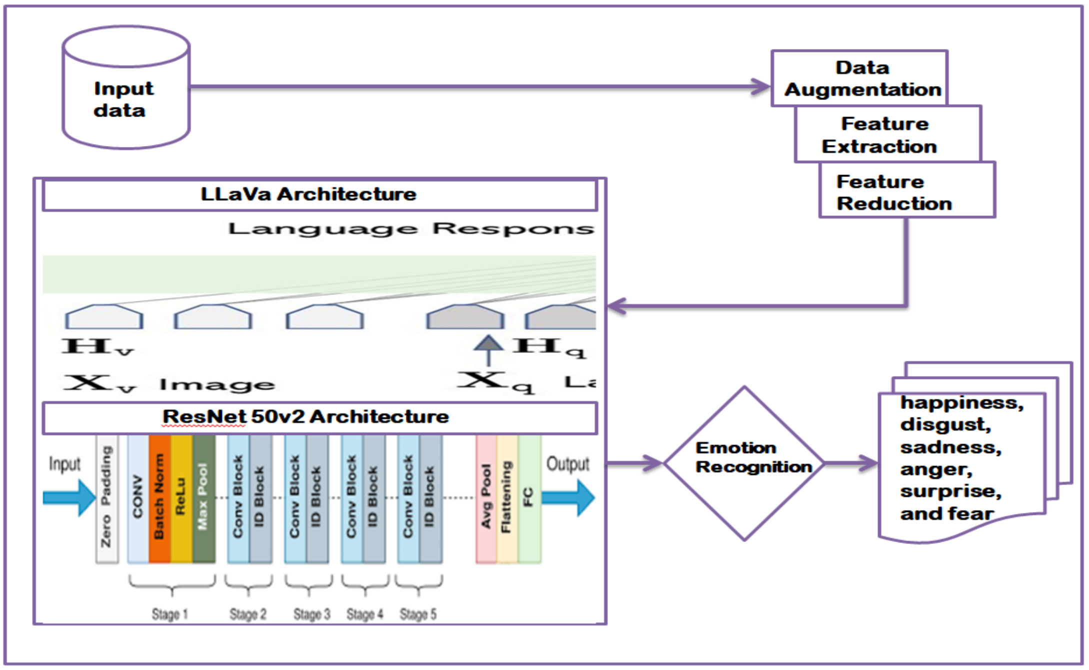

Fig. 2 Block diagram of the suggested emotion recognition method

3.2.1 Preprocessing Stage

The initial phase involves processing photos and converting them into a standardized format following preprocessing procedures. Image processing streamlines the ensuing procedural processes. The converted greyscale image was supplied for analysis following the standardization of the data. The computer vision algorithm will be implemented in the subsequent phase.

3.2.2 Feature Extraction and Classification

One of the critical models for multi emotion recognition is feature extraction and reduction that include deep neural networks and transformer models. Feature extraction extracts the most appropriate information from photographs, which aids in determining emotional predictions. This technique reduces data of raw image to a critical set and useful features, such as textures, edges, and major facial landmarks, which are required for emotion recognition [34]. After feature extraction, feature reduction is used to optimize the data by less informative or removing redundant features. The removal of superfluous features enhances the model's efficacy and performance. The model mitigates overfitting. The reduction of features expedites training and inference.

The integration of these processes enhances the accuracy of emotion detection by focusing on critical components and eliminating extraneous information. This facilitates the comprehension of affective expressions through models. The multi-emotion detection extraction method employs the foreground to enhance the model's performance. This method distinguishes the subject matter from the context. In order to facilitate the model's learning, we accentuate the image's critical components for emotion identification by focusing on the face at the front.

While the image is accessible, eliminate any external commotion or distractions. Background data that contains irrelevant information may compromise the model's accuracy and reliability. To ensure that your model can accurately identify emotions, it is necessary to partition the features [35]. Deep neural networks and transformation models are able to identify subtle facial expression changes as a result of enhanced input data [36].

ResNet50V2 has been used to efficiently recognize many faces in complex scenarios, as its deep architecture enables extracting intrinsic features from images, which is essential for distinguishing between different objects and recognizing emotional expressions [37]. LLaVA incorporates vision processing to improve its capabilities, and it uses deep learning to identify numerous emotions. LLaVA is an effective tool for applications that analyze visual data because of its ability to comprehend and generate textual descriptions of visual content. More accurate and context-aware predictions of emotional expressions are made possible by LLaVA's capacity to handle both visual and textual data. LLaVA has a deeper understanding of complicated emotional signals. It can improve the model's capacity to detect nuanced emotions and translate visual patterns to verbal descriptions. LLaVA can decipher the feelings conveyed by a picture with more precision than before. ResNet50V2 utilizes identity mapping, batch normalization, and weight scaling to strengthen the gradient flow so that it augments stability and the rate of convergence during training.

Hyperparameters used in ResNet50V2 include an optimizer as weight decay with Adam, learning rate = 0.001, batch size = 32, epochs = 50, and loss function = categorical Cross-Entropy. LLaVa uses vision processing for capability strengthening and utilizes deep learning to detect a vast number of emotions. Unlike CNNs, LLaVa employs vision-language alignment to recognize emotions using contextual cues. LLaVa is a very effective tool in visual data-processing tasks because of its ability to comprehend and generate textual descriptions of visual content. The fact that LLaVa can process visual and textual data makes more accurate and contextual predictions of emotional expressions possible. Hyperparameters of LLaVa are as follows: transformer layers are 12, embedding size is 768, attention heads are 8, Dropout rate is 0.1, optimizer is AdamW, learning rate is 2e-5 (fine-tuned with warm-up strategy), and batch size is 16. LLaVa can read the emotions depicted by an image more accurately than ever before, owing to its multimodal learning approach, as it can read emotions in a human-like manner.

The models are trained using transfer learning, beginning with pre-trained weights of ImageNet models and fine-tuning them on the emotion recognition dataset. Stratified K-fold cross-validation (K=5) is employed to make the models generalizable and avoid bias. Training is performed using NVIDIA A100 GPUs and TensorFlow 2.0 to enable optimized execution. Early stopping and learning rate annealing techniques are employed to avoid overfitting.

4. Results and Discussion

This section contains a discussion and analysis of the results. The dataset is partitioned into three sets: training, testing, and validation. The training accuracy is determined at the conclusion of each epoch to assess the model's ability to accurately predict the results of the training dataset. The validation accuracy is determined by utilizing the validation dataset to evaluate the model's ability to generalize to unobserved data. The training loss is computed after each block during the epoch and subsequently averaged over the entire epoch. The total time required for all training sessions is ultimately determined by calculating the validation loss after each epoch on the validation dataset.

An epoch is a complete cycle in which the model goes through the entire training dataset, as shown in Tables 2 and 3. During each epoch, the model reviews all the data, makes predictions, and checks how far these predictions deviate from the results. It then adjusts its internal blocks to improve them. Without this process on multiple images, the model becomes better at hiding data from the data, leading to improvements in accuracy, precision, recall, and other metrics. The objective during training is to minimize the loss function, which quantifies the divergence between the model's predictions and the actual values. Over time, the model adjusts itself to ensure it is correct.

For large datasets, where not all the data is a single integral, the data is split into smaller chunks to retrieve the batches. The complaints of this model are that the batches are one-time batches, and when it has gone through all the batches, an epoch is completed. Choosing the right number of epochs is important. If it is less, the model has not learned enough from the data (underfitting). If it is more, the model has started to memorize the training data, including its noise, including generalizing well to the new data (overfitting).

Table 2: Loss and execution of LLaVa (ViT)

Epoch	Training Loss	Validation Loss	Execution Time (s)
1	2.30	2.29	55.2
2	2.15	2.12	55.4
3	1.99	1.95	55.6
4	1.85	1.80	55.8
5	1.7	1.65	56
6	1.55	1.52	56.2
7	1.4	1.35	56.3
8	1.25	1.22	56.5
9	1.09	1.05	56.7107
10	0.95	0.91	56.8964
11	0.80	0.75	57.0821
12	0.65	0.59	57.2679
13	0.51	0.44	57.4536
14	0.35	0.29	57.6393
15	0.20	0.14	57.825
16	**0.05**	**0.01**	58.0107

To implement a model for emotion recognition from facial expressions using the FER2013 dataset, focusing on seven major emotions: happiness, sadness, anger, fear, neutral, disgust, and surprise. The system uses ResNet50V2 and LLaVA and compares it with other common deep neural architectures, which is the Alex CNN model, and evaluates model performance with accuracy, precision, recall, AUC-ROC curve, and confusion matrix.

Table 3: Loss and execution of ResNet50V2

Epoch	Training Loss	Validation Loss	Execution Time (s)
1	2.1	2.05	55.3
2	1.95	1.9	55.6
3	1.8	1.75	55.7
4	1.6	1.5	55.9
5	1.3	1.25	56.1
6	1.17	1.09	56.29
7	0.97	0.89	56.48
8	0.78	0.69	56.67
9	0.58	0.49	56.86
10	0.39	0.29	57.05
11	0.19	**0.09**	57.24
12	**0.01**	0.11	57.43
13	0.2	0.31	57.62
14	0.39	0.51	57.81
15	0.59	0.71	58
16	0.79	0.91	58.19

After implementing an emotion recognition system from facial expressions, we observed better statistics with the second model (LlAVA Vision Transformer (ViT)), as shown in Table 4. The system compares the performance of ResNet50V2, LLaVA, and Alex CNN models and evaluates the model's performance with accuracy, precision, recall, AUC-ROC curve, and confusion matrix as displayed in Figs. 3 and 4.

Table 4: Model performance in accuracy, precision, recall and AUC.

Model	Accuracy	Precision	Recall	AUC
LLava	0.78	0.77	0.76	0.86
ResNet50V2	0.76	0.75	0.74	0.84
Alex CNN	0.75	0.74	0.73	0.83

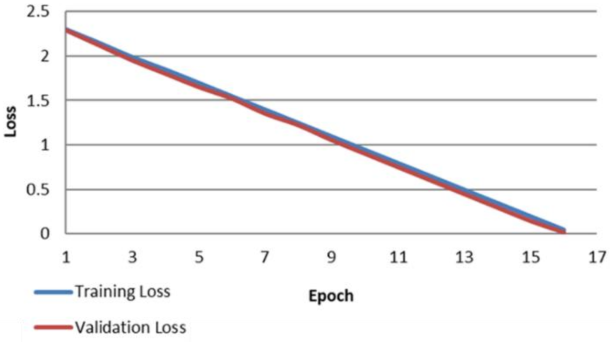

Fig. 3 Learning curve for LLaVa

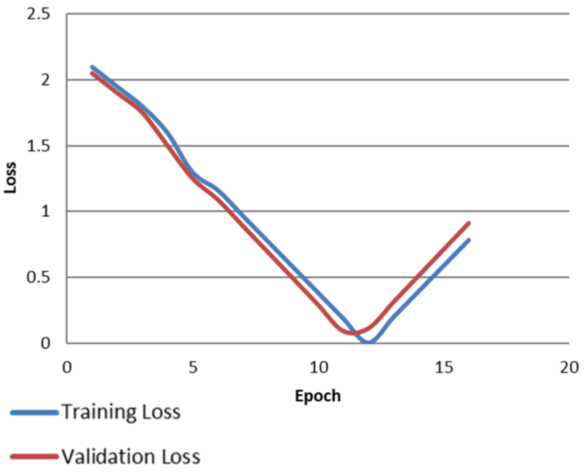

Fig. 4 Learning curve for ResNet50V2

Using transformer forwarding can improve the prediction accuracy of the emotion recognition model; however, the overall efficiency of training a deep learning model is a challenge when deploying models in a practical environment due to the huge training time and computing space requirements. Future research will require studying data management and developing more powerful and efficient machine learning algorithms to form a positive study on deploying machine learning algorithms for practical machine learning problems. The accuracy results indicate that Vision Transformer (ViT) achieved the highest accuracy of 78%, followed by ResNet50V2 at 76%, and Alex CNN at 75%, as shown in Fig. 5.

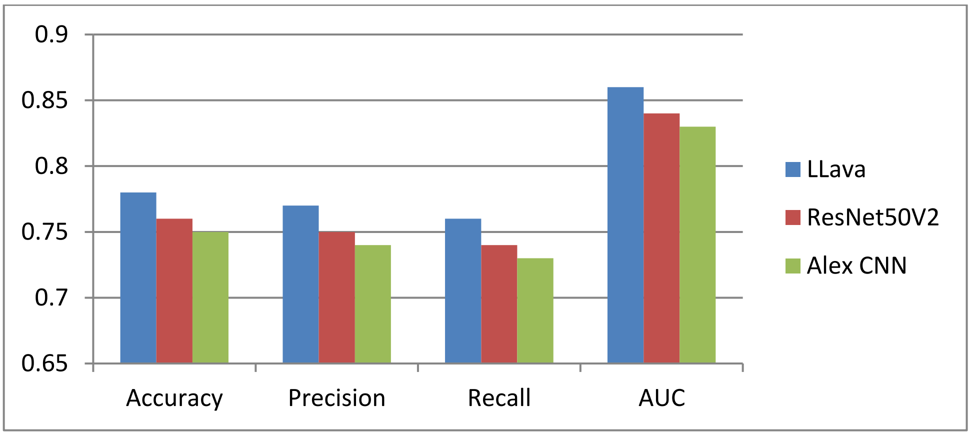

Fig. 5 Accuracy, precision, recall, and AUC factor results among model

Confusion matrices in Tables 5 and 6 showed each model’s strengths and weaknesses in predicting specific emotions. For example, Llava ViT showed better discrimination between emotions like “anger” and “sadness,” while Alex CNN had difficulty distinguishing between “neutral” and “fear.” Misclassifications were more pronounced for ResNet50V2 for overlapping emotions like as “happiness” and “surprise”. Finally, Llava ViT and ResNet50V2 showed slightly longer execution times compared to Alex CNN, which maintained its efficiency.

Table 5: Confusion Matrix for Llava.

True \ Predicted	**Happiness**	**Sadness**	**Anger**	**Fear**	**Neutral**	**Disgust**	**Surprise**
Happiness	**81.54%**	2.77%	1.85%	2.46%	5.23%	4.31%	1.85%
Sadness	4.03%	**80.54%**	2.68%	2.01%	6.71%	2.01%	2.01%
Anger	3.51%	1.75%	**76.02%**	5.26%	8.19%	2.34%	2.92%
Fear	4.05%	1.69%	6.08%	**79.39%**	4.39%	2.03%	2.36%
Neutral	5.82%	5.09%	8.73%	6.55%	**69.09%**	2.91%	1.82%
Disgust	9.15%	3.92%	3.92%	3.27%	5.88%	**71.90%**	1.96%
Surprise	6.06%	3.37%	4.04%	3.03%	2.02%	4.04%	**77.44%**

Table 6: Confusion Matrix for ResNet50V2.

True \ Predicted	Happiness	Sadness	Anger	Fear	Neutral	Disgust	Surprise
Happiness	**83.08%**	3.08%	1.54%	3.08%	4.62%	3.08%	1.54%
Sadness	3.33%	**83.33%**	3.33%	1.67%	5.00%	1.67%	1.67%
Anger	4.41%	1.47%	**73.53%**	5.88%	8.82%	2.94%	2.94%
Fear	3.51%	1.75%	5.26%	**80.70%**	5.26%	1.75%	1.75%
Neutral	5.17%	3.45%	10.34%	6.90%	**68.97%**	3.45%	1.72%
Disgust	9.38%	3.13%	3.13%	4.69%	6.25%	**71.88%**	1.56%
Surprise	5.08%	3.39%	3.39%	1.69%	1.69%	3.39%	**81.36%**

5. Conclusion

The study evaluates deep learning models' capacity to predict seven emotion classes on the FER-2013 dataset, therefore addressing the significant difficulty of emotion recognition from visual data. Their top layers were rebuilt to meet the particular needs of the dataset, therefore enabling the pre-trained models to be adapted to this work. With 78% accuracy, Llava ViT came out among the models tested as having the best performance. Although the 2–3% accuracy difference from other models would not seem significant, perceptual confusion matrix analysis verified the dependability of this method and so became the recommended alternative for additional improvement. Furthermore, strengthening its resilience, qualitative analysis revealed consistent model results over the FER-2013 dataset. We want to enhance the model going forward to reach better accuracy without overfitting and cut computing costs. Once these objectives are met, the method will be modified for more challenging, multi-task models able to concurrently manage chores like age prediction and emotional recognition. This vision aims to inspire the research community to investigate the synergy between transformers and deep neural network models for next uses and use the suggested method to meet various difficulties. The architecture presents various issues that demand more research nevertheless its potential. First, it depends on already-existing data; most photos have a range in size from 86 to 256 pixels.

This restriction begs questions regarding whether the method would work equally effectively on higher-resolution datasets. Second, the size estimate of found objects could be erroneous or non-intuitive without a strong recognition process. Incomplete or overlapping parameter descriptions might hamper the optimization of deep CNNs employing transformer block frameworks and impede transformer stacking. To implement a dependable model from end to end, more steady depth levels and validation procedures allowing input spaces of at least average width are required. Finally, the model must incorporate techniques to handle ill-defined and nonlinear aspects while normalizing feature distributions to avoid sensitivity problems or output collapse. This work opens the path for more sophisticated emotion identification systems by resolving constraints and investigating creative ideas.

Nevertheless, some limitations need to be noted in order to enhance future research endeavors. Firstly, the use of the FER-2013 dataset, which is made up of fairly low-resolution images (86 to 256 pixels), brings questions regarding the generalizability of the model to higher-resolution datasets and real-world scenarios where image quality is highly variable. Future research needs to investigate the effect of image resolution on model performance and modify architectures accordingly. Second, the model may lack the ability to accurately estimate the intensity of detected facial expressions without robust object recognition processes. Occlusions, poor lighting, or partially visible faces may affect classification performance. Therefore, additional pre-processing techniques such as attention mechanisms or self-supervised learning may provide robustness.

Conflict of interest

The author declares that the publishing of this article does not include any conflicts of interest. Furthermore, the author has strictly adhered to ethical problems such as plagiarism, informed consent, misconduct, data fabrication and falsification, multiple publishing and submission, and redundancy.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

References

[1] Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. (2024). Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions. Information Fusion, 105, 102218. https://doi.org/10.1016/j.inffus.2023.102218

[2] Siam, A. I., Soliman, N. F., Algarni, A. D., Abd El-Samie, F. E., & Sedik, A. (2022). Deploying machine learning techniques for human emotion detection. Computational intelligence and neuroscience, 2022(1), 8032673. https://doi.org/10.1155/2022/8032673

[3] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046. **https://doi.org/10.3390/s21093046**

[4] Suárez-Cabal, M. J., Suárez-Otero, P., de la Riva, C., & Tuya, J. (2023). MDICA: Maintenance of data integrity in column-oriented database applications. Computer Standards & Interfaces, 83, 103642. https://doi.org/10.1016/j.csi.2022.103642

[5] Saranya, G. (2024). Integrated Vision and Sensor Based Analysis for Sleep Apnea Using FeatFaceNet Deep Learning. Journal of Electrical Engineering & Technology, 19(1), 655-664. https://doi.org/10.1007/s42835-023-01549-1

[6] Abdulrahim, A. (2024, March). The Impact of Key Determinants of BIM Technology Adoption on Organizational Performance within the UAE Construction Industry. In BUiD Doctoral Research Conference 2023: Multidisciplinary Studies (pp. 472-480). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56121-4_45

[7] Alanazi, M. F., Ali, M. U., Hussain, S. J., Zafar, A., Mohatram, M., Irfan, M., ... & Albarrak, A. M. (2022). Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model. *Sensors*, *22*(1), 372. https://doi.org/10.3390/s22010372

[8] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83, 19-52. https://doi.org/10.1016/j.inffus.2022.03.009

[9] Anzabi, N., & Umemuro, H. (2023). Effect of different listening behaviors of social robots on perceived trust in human-robot interactions. *International Journal of Social Robotics*, *15*(6), 931-951. https://doi.org/10.1007/s12369-023-01008-x

[10] Bota, P. J., Wang, C., Fred, A. L., & Da Silva, H. P. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE access*, *7*, 140990-141020. https://doi.org/10.1109/ACCESS.2019.2944001

[11] Cheng, J., Liu, R., Li, J., Song, R., Liu, Y., & Chen, X. (2023). Motion-robust respiratory rate estimation from camera videos via fusing pixel movement and pixel intensity information. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/TIM.2023.3291770

[12] Li, Bin, and Dimas Lima. "Facial expression recognition via ResNet-50." International Journal of Cognitive Computing in Engineering 2 (2021): 57-64. https://doi.org/10.1016/j.ijcce.2021.02.002

[13] Odhiambo, M. O., & Mwashita, W. (2024). Security Provision for the Internet of Intelligence Using Autonomous Mobile Agents. In *From Internet of Things to Internet of Intelligence* (pp. 147-156). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55718-7_8

[14] Adyapady, R. R., & Annappa, B. (2023). A comprehensive review of facial expression recognition techniques. *Multimedia Systems*, *29*(1), 73-103. https://doi.org/10.1007/s00530-022-00984-w

[15] Guo, J. Y., Cai, Q., An, J. P., Chen, P. Y., Ma, C., Wan, J. H., & Gao, Z. K. (2022). A transformer based neural network for emotion recognition and visualizations of crucial EEG channels. Physica A: Statistical Mechanics and its Applications, 603, 127700. https://doi.org/10.1016/j.physa.2022.127700

[16] de Mendonça, L. J. C., Ferrari, R. J., & Alzheimer's Disease Neuroimaging Initiative. (2023). Alzheimer's disease classification based on graph kernel SVMs constructed with 3D texture features extracted from MR images. Expert Systems with Applications, 211, 118633. https://doi.org/10.1016/j.eswa.2022.118633

[17] Alnaggar, M., Siam, A. I., Handosa, M., Medhat, T., & Rashad, M. Z. (2023). Video-based real-time monitoring for heart rate and respiration rate. Expert Systems with Applications, 225, 120135. https://doi.org/10.1016/j.eswa.2023.120135

[18] Debnath, T., Reza, M. M., Rahman, A., Beheshti, A., Band, S. S., & Alinejad-Rokny, H. (2022). Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. Scientific Reports, 12(1), 6991. https://doi.org/10.1038/s41598-022-11173-0

[19] Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. Information Fusion, 51, 10-18. https://doi.org/10.1016/j.inffus.2018.10.009

[20] Kanjo, E., Younis, E. M., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. Information Fusion, 49, 46-56. https://doi.org/10.1016/j.inffus.2018.09.001

[21] Lian, Z., Liu, B., & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 985-1000.https://doi.org/10.1109/TASLP.2021.3049898

[22] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences*, *2*(3), 446. https://doi.org/10.1007/s42452-020-2234-1

[23] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, *21*(9), 3046.https://doi.org/10.3390/s21093046

[24] Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146, 1-7. https://doi.org/10.1016/j.patrec.2021.03.007

[25] Kamble, K., & Sengupta, J. (2023). A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals. *Multimedia Tools and Applications*, *82*(18), 27269-27304. https://doi.org/10.1007/s11042-023-14489-9

[26] Dixit, C., & Satapathy, S. M. (2024). Deep CNN with late fusion for real time multimodal emotion recognition. Expert Systems with Applications, 240, 122579. https://doi.org/10.1016/j.eswa.2023.122579

[27] Karani, R., Jani, J., & Desai, S. (2024). FER-BHARAT: a lightweight deep learning network for efficient unimodal facial emotion recognition in Indian context. Discover Artificial Intelligence, 4(1), 35. https://doi.org/10.1007/s44163-024-00131-6

[28] Aggarwal, A., Keserwani, K., & Chauhan, A. (2023). Systematic literature review of current trends in artificial intelligence-based intrusion detection systems. International Journal of Computing and Digital Systems, 14(1), 1-xx. https://doi.org/10.1515/jisys-2023-0248

[29] Zakieldin, K., Khattab, R., Ibrahim, E., Arafat, E., Ahmed, N., & Hemayed, E. (2024). Vitcn: Hybrid vision transformer with temporal convolution for multi-emotion recognition. International Journal of

Computational Intelligence Systems, 17(1), 64. https://doi.org/10.1007/s44196-024-00436-5

[30] Manalu, H. V., & Rifai, A. P. (2024). Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. Intelligent Systems with Applications, 21, 200339. https://doi.org/10.1016/j.iswa.2024.200339

[31] Abad, S., & Gholamy, H. (2023). Evaluation of machine learning models for classifying malicious URLs.https://www.diva-portal.org/smash/get/diva2:1766378/FULLTEXT01.pdf

[32] Aina, J., Akinniyi, O., Rahman, M. M., Odero-Marah, V., & Khalifa, F. (2024). A hybrid Learning-Architecture for mental disorder detection using emotion recognition. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3421376

[33] FER-2013: A facial expression recognition dataset" by I. Goodfellow, Y. Bengio, and A. Courville. This is part of the Kaggle Facial Expression Recognition Challenge, published in 2013(https://www.kaggle.com/datasets/msambare/fer2013)

[34] Anand, M., & Babu, S. (2024). Multi-class facial emotion expression identification using dl-based feature extraction with classification models. International Journal of Computational Intelligence Systems, 17(1), 25. https://doi.org/10.1007/s44196-024-00406-x

[35] Chen, D., Wang, P., Yue, L., Zhang, Y., & Jia, T. (2020). Anomaly detection in surveillance video based on bidirectional prediction. Image and Vision Computing, 98, 103915. https://doi.org/10.1016/j.imavis.2020.103915

[36] Mohammed, H. I., & Waleed, J. (2023, March). Hand gesture recognition using a convolutional neural network for arabic sign language. In AIP Conference Proceedings (Vol. 2475, No. 1). AIP Publishing. https://doi.org/10.1063/5.0104256

[37] Sarada, B., Narasimha Reddy, K., Babu, R., & Ramesh Babu, B. S. S. V. (2024). Brain tumor classification using modified ResNet50V2 deep learning model. *International Journal of Computing and Digital Systems*, *16*(1), 1-10.