Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

The Implementation of Transfer Learning by Convolution Neural Network (CNN) for Recognizing Facial Emotions

Anbananthan Pillai Munanday[1], Norazlianie Sazali[1,*], Arjun Asogan[1], Devarajan Ramasamy[2], Ahmad Shahir Jamaludin[1]

[1] Faculty of Manufacturing and Mechatronic Engineering Technology, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia
[2] Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia

ARTICLE INFO

Article history:
Received 10 May 2023
Received in revised form 17 August 2023
Accepted 23 August 2023
Available online 14 September 2023

Keywords:
CNN; FER-2013; JAFFE; CK+; Transfer Learning; Deep Learning

ABSTRACT

The primary objective of this study is to develop a real-time system that can predict the emotional states of an individual who commonly undergoes various experiences. The primary methodology suggested in this research for detecting facial expressions involves the integration of transfer learning techniques that incorporate convolutional neural networks (CNNs), along with a parameterization approach that minimizes the number of parameters. The FER-2013, JAFFE, and CK+ datasets were jointly used to train the CNN architecture for real-time detection, which broadened the range of emotional expressions that may be recognized. The proposed model has the capability to identify various emotions, including but not limited to happiness, fear, surprise, anger, contempt, sadness, and neutrality. Several methods were employed to assess the efficacy of the model's performance in this study. The experimental results indicate that the proposed approach surpasses previous studies in terms of both speed and accuracy.

1. Introduction

1.1 Manifestations of Emotions

One of the most important parts of successful two-way communication between people is being able to read their facial expressions of emotion. Facial expressions are a potent form of communication as they convey one's emotions and genuineness, providing meaning to what is being communicated. Automated Facial Expression Recognition (AFER) is an interdisciplinary field that intersects different domains, including Artificial Intelligence, Psychology, Neuroscience, and Behavioural Research. It is commonly acknowledged that the progress in computing technology has significantly accelerated research in the fields of artificial intelligence (AI) and pattern recognition. Human and machine communication requires a good rapport to establish a genuine connection. However, only 38% of information is transmitted through voice and language, whereas 55% is conveyed through facial expressions. It is possible to deduce someone's emotional state just by

* *Corresponding author.*
E-mail address: azlianie@ump.edu.my

https://doi.org/10.37934/araset.32.2.255276

observing their face. There is no universally accepted definition of emotion, despite the abundance of research on this topic [1]. Emotional display is one of several ways a feeling may manifest itself, and this could be one approach. Unlike feelings, emotions can be easily created.

Accurately predicting emotional states in real time holds significant importance and has the potential to bring about transformative impacts across various domains. This capability has profound implications for mental health, as it enables timely interventions and personalized treatment plans based on individuals' emotional states. Additionally, in the realm of human-computer interaction, accurate real-time emotion prediction allows interactive systems to adapt their responses and interfaces, creating more intuitive and engaging user experiences. Moreover, in the context of personalized services, understanding customer's emotional responses in real-time empowers businesses to tailor their offerings and marketing strategies to enhance customer satisfaction and loyalty. Accurate real-time emotion prediction also fosters healthier social relationships, as it enhances individuals' ability to understand and empathize with others' emotions, leading to improved communication and social bonding. Furthermore, this capability contributes to the advancement of artificial intelligence and robotics, enabling more emotionally aware systems and robots that can interact more naturally and effectively with humans. Thus, accurate real-time prediction of emotional states holds immense potential for improving mental health support, human-computer interaction, personalized services, social relationships, and the development of intelligent systems.

Despite its intricacy, the majority of tasks related to the complex structure can be accomplished by utilizing the frontal face as input. Therefore, significant resources have been allocated to the development of automated methods for encoding expressions.

Under controlled conditions, it is feasible to identify basic facial emotions, such as those expressed on the front faces or through side postures, which can contribute to the effective completion of the assignment. The study of emotion encompasses several vital and complex fields of study. Emotion detection systems have the ability to identify a range of common emotions [2], including neutral, happy, sad, surprised, angry, fearful, and disgusted, as shown in Figure 1.

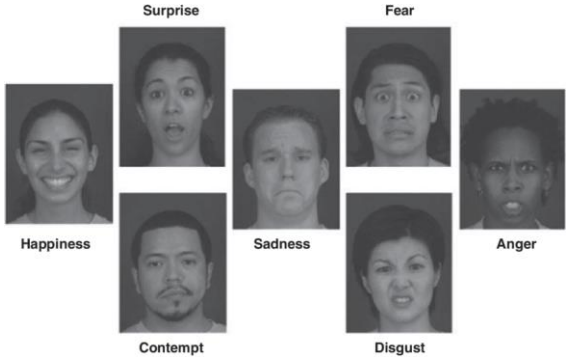

Fig. 1. Universal facial emotions [3]

The presence of 21 different emotional states is due to compound emotions rather than universal emotions. A compound emotion, such as a combination of happiness and surprise, is created when both emotions are combined [2]. This is due to the fact that the physiological sensations of surprise expressions are mixed with happiness, as shown in Figure 2.

There are several possible uses for technology that can recognize facial expressions and determine an individual's emotional state. Applications of emotional and social awareness technology include enhancing gaming experiences, identifying drowsy drivers, and detecting signs of pain or distress in patients [4]. There are also devices that make it easier to spot drivers who are too

tired to drive. Most previous work on FER has focused on enhancing the process of feature extraction stages. Features extracted from appearance-based, geometric-based, or hybrid-based feature extractors as shown in Figure 3, are then fed into various classifiers to make decisions using different methodologies.

Fig. 2. Combination of happy and surprised emotions

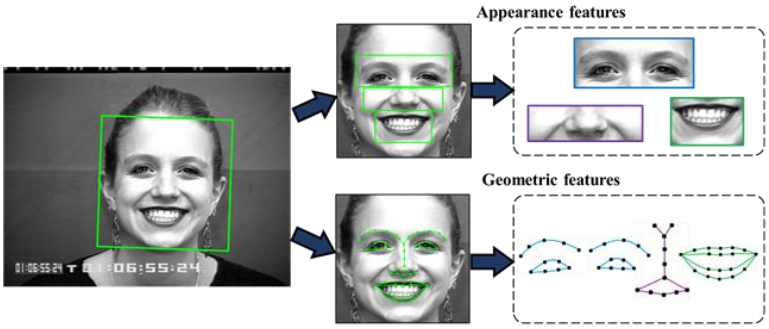

Fig. 3. Appearance and geometric-based feature extractors [5]

However, the computational difficulty of conventional approaches makes it challenging to achieve high recognition rates. Although the basic model can capture some aspects of emotional expressions, it fails to capture the complexity and subtlety of our emotions in real-life situations. Therefore, a more sophisticated approach that combines the action coding system and the continuous approach is required to capture expressions accurately. Variations in head orientation, illumination, and occlusion can introduce additional levels of complexity.

As individuals may exhibit diverse ways of conveying their emotions, the interpretation of an image may differ based on various factors, including background luminosity, background hue, image orientation, and other characteristics, as illustrated in Figure 4, which depicts an instance of illumination impact. Another aspect that plays a role in determining interpreted pictures is the common practice of minimizing the significance of real, unposed expressions. Therefore, it is essential to have a reliable and automated Facial Expression Recognition (FER) system that can adapt to different situations.

Several approaches to automatically recognizing facial expressions have been suggested and reviewed. A considerable proportion of the initial studies utilized graphical representations in geometric form. These studies involved the use of active contours to recover the shape of the lips and eyes, vector descriptors for tracking facial movements, and variable mesh models in 2D. Other techniques employed Gabor filters, local binary patterns, and other appearance-based representations [4]. These feature extraction methods were commonly combined with various

classifiers to identify action units or categorize emotions, ensuring the accuracy of the findings. In this research, random forests and support vector machines (SVMs) were the two classifiers that were most often utilized.

Fig. 4. Effect of illumination changes in face images [6]

Convolutional neural networks (CNNs) are artificial neural networks comprising multiple layers designed to mimic the structure of the human brain and address complex problems. These artificial neurons, equipped with a bias and an activation function, derive an output from an image, thereby assigning greater significance to the picture. Artificial neurons have a wide range of applications, including image classification, recognition, and segmentation. Additionally, artificial neurons can perform simple convolutions [7]. Increasing the amount of data provided to the convolutional neural network can lead to the development of a more reliable and accurate deep learning model. One approach to achieve this is through deep learning-based facial expression recognition, which can recognize a broad spectrum of human emotions (such as anger, fear, neutrality, happiness, disgust, sadness, and surprise) from single images.

This technique aims to automatically recognize facial expressions to accurately reflect an individual's emotional state. When training a CNN, images of faces from an emotional expression dataset are labelled and used. After this stage, the suggested CNN model will determine the appropriate facial expression. Previously established CNN-based models have demonstrated a high level of accuracy in recognizing emotions in FER testing. However, these models still require a large amount of memory and computational power. Due to these limitations, they cannot be deployed in resource-constrained environments [8]. A lightweight CNN, on the other hand, requires less memory and processing power, making it suitable for real-time applications.

Transfer learning in CNN is a technique that extracts features from the FER2013, CK+, and JAFFE datasets [10–13] by using pre-trained convolutional neural network models on a large dataset. Instead of training a new CNN model from scratch, transfer learning allows us to fine-tune an existing model by adjusting its parameters to improve performance [13]. This approach is particularly useful when there is a limited amount of training data available, as it saves time and computational resources. By utilizing transfer learning, we can leverage the knowledge learned from a large dataset and apply it to a new dataset FER2013, CK+, and JAFFE [10–14]**,** resulting in improved accuracy and efficiency in training CNN models.

To enable real-time detection on the combined dataset, transfer learning (TL) was employed from the FER-2013 dataset during the training process. This approach was adopted because previous studies have demonstrated that using TL with the CK dataset leads to improved precision [15]. Moreover, time is saved as TL eliminates the need for initial training. In a nutshell, the following factors were taken into consideration as a direct consequence of this research:

i. Developing a CNN with a simpler structure to facilitate quicker detection.
ii. The goal of this project is to accelerate the process of detection by utilizing transfer learning while maintaining the reliability of the result.
iii. A comprehensive dataset was compiled from various sources, including natural surroundings and laboratory environments, to achieve timely target detection for real-time purposes.
iv. Creating an application that can identify facial expressions in real time with a reduced runtime cost compared to existing works in the current research literature.
v. Emphasizing the distinct variations that arise among various convolutional neural network techniques under similar conditions.
vi. Utilized CNN visualization techniques were utilized to enhance understanding of the model generated through the application of the most up-to-date and advanced FER techniques to various datasets. This enables us to make more accurate predictions based on analysis outcomes.
vii. Demonstrating that the network for detecting emotions can generalize over a broad variety of datasets and FER.

1.2 An Analysis of Prior Research in the Relevant Field

Deep learning-based facial expression recognition is a sophisticated technology utilized for this purpose. It has achieved exceptional success in image classification and can accurately describe emotions based on single images, including anger, fear, neutrality, happiness, disgust, sadness, and surprise.

The objective of this method is to automate the detection of facial expressions, which can aid in accurately evaluating someone's emotional condition. To apply this approach, a Convolutional Neural Network (CNN) was trained with a set of labelled facial expression images extracted from benchmark facial expression datasets. A dataset consisting of facial images curated for the purpose of analysing facial expressions is commonly known as a facial expression dataset, which includes a variety of facial images. Next, the proposed CNN model predicts the intended facial expression. The CNN technique is often used to achieve this specific goal. The term "facial expression analysis" can be traced back to the work of Reddi *et al.,* [16]. Since then, the wide range of applications for FER has fuelled an increasing number of research initiatives in this field. Applications in healthcare, nonverbal communication, and even the study of human behaviour can all benefit from the implementation of an autonomous facial expression detection system. The applications of such a system are not limited to the ones listed above.

Despite the fact that the field of emotion detection has been around for a few decades, it remains an interesting field for research due to its numerous beneficial applications. Researchers have invested considerable time and effort in developing algorithms capable of accurately identifying human emotions. However, despite notable advancements in this area, most research conducted on still images has been unable to address the challenges posed by variations in lighting, changes in facial position, and obscured facial expressions. The usual automated approach for facial expression recognition (FER) comprises three crucial stages: detecting the face, extracting its features, and identifying the emotions of the subject.

Feature extraction techniques primarily rely on facial appearance and geometrical properties. Appearance-based algorithms are employed to extract intensity, gradient, and textural variations from the designated facial region as shown in Figure 5.

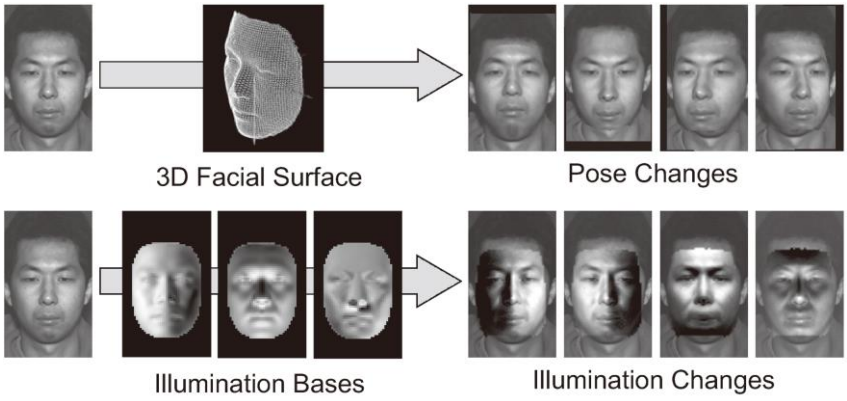

Fig. 5. Conceptual generation of various facial Images using appearance-based algorithms [17]

Conversely, geometric facial features are extracted based on the underlying anatomy of the face, as shown in Figure 6. To recognize different emotions, for instance, it is essential to calculate the distances between various facial landmarks. These landmarks are located on the face. Moreover, the hybrid technique integrates both geometric and visual methodologies for feature extraction. To recognize emotions, it is crucial to input the gathered handcrafted characteristics into established classifiers.

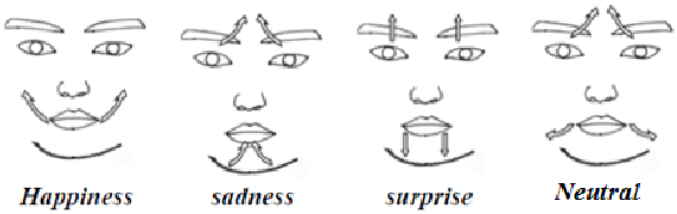

Fig. 6. Geometric-based algorithms facial muscle movements [18]

Zhang *et al.,* [19] developed a highly effective approach for classifying halftone pictures and processing images, which can be used to analyse key elements of both still and motion images. Unsupervised learning and stacked sparse autoencoders (SAE) were used to extract grayscale image features. The FER-2013 and CK+ datasets were utilized to train a CNN model based on the ResNet architecture, and the model was used to extract characteristics from the data. In addition to the complexity perception classification (CPC) technique, other classifiers such as Softmax, linear SVM, and random forest were employed. The combination of CNN+Softmax with CPC resulted in recognition rates of 71.35 percent for the FER2013 dataset and 98.78 percent for the CK+ dataset.

Lekdioui *et al.,* [20] proposed an approach to identifying facial expressions based on the texture and shape descriptors of the face. CNN can yield good results when trained to analyse a face and recognize the features that influence its predictions [21]. This factor is crucial in obtaining positive outcomes from CNN. Similarly, Happy introduced a method for extracting facial patches using landmark points and LDA features. Baffour *et al.,* [22] developed a flexible hypothesis pooling strategy to address the issue of image multi-categorization. With this method, the model can take any assumptions about the object segment as input. Each concept is then connected to a single CNN through a chain of other concepts and events. In summary, the use of maximum pooling is a viable option for generating standard predictors in multi-label predictions by combining the outputs of various hypotheses within the model.

Several issues related to facial emotion recognition applications may arise, such as alignment, face detection, and face recognition. Chowdary *et al.,* [23] proposed a multi-level CNN with 18 layers, similar to VGG. The proposed model not only utilizes the existing high-level features but also incorporates information from the underlying layers. By implementing a multi-level CNN, improvements in the accuracy of the core CNN model were achieved, reaching 69.21%.

Liu *et al.,* [24] employed an original optical flow technique called MDMO. An affine transformation was then used to enhance illumination, address its complications, and account for the subject's head movement in order to optimize texture information extraction. Training a support vector machine (SVM) classifier on the data of regions of interest (ROIs) in the face helped in identifying genuine emotions. The model achieved an accuracy rate of 73.03 percent on the FER dataset.

In order to investigate visually observed features, a part of the research community developed an autoencoder. Zeng *et al.,* [25] designed deep sparse autoencoders (DSAE) to train appearance-based features that can be utilized for recognizing facial emotions. On the other hand, Naumann *et al.,* [26] developed a deep learning model for facial alignment and correction based on landmark features and recurrent recognition. Chen *et al.,* [27] demonstrate the ability to identify genuine emotions in images using deep learning.

In contrast to earlier studies that utilized manual feature extraction from facial images and employed a two-step approach for detecting smiles by training a classifier, deep learning seamlessly integrates feature learning and classification into a single model, thereby increasing efficiency. The procedure of identifying the feature points is without a doubt the most difficult part of the geometric approach. In their research, Ezerli *et al.,* [28] proposed a geometric method that utilized feature points to isolate the eye and mouth areas. These areas were then mapped quadrilateral to obtain input for a fuzzy membership algorithm, ensuring precise categorization of facial emotions. Pang *et al.,* [29] developed a system that accurately applies the CNN DL algorithm to identify visual objects.

Ongoing research efforts are currently focused on the demanding area of video analysis called "human activity identification," which is garnering considerable interest. Ronao *et al.,* [30] proposed an efficient and effective system for detecting human activity, utilizing sensors that are typically present in smartphones. The suggested approach was evaluated on diverse experimental datasets, producing outstanding results. Christou *et al.,* [31] introduced a 13-layer CNN model that achieved an accuracy of 91.12% when tested on the validation subset of the Fer2013 dataset. The model was employed to obtain these results.

In addition to the CK+, JAFFE, and FACES datasets, Sajjanhar *et al.,* [32] also used them. The results indicated that the VGG19 model achieved the highest accuracy on the FACES dataset, showcasing its best performance to attain this level of accuracy.

Chen *et al.,* [27] developed a two-stage approach that includes training separate convolutional neural networks (CNNs) using the CK+ and BU-4DFE datasets. On the other hand, Dosovitskiy *et al.,* [33] utilized Flownet 2.0, a novel automated micro-expression analysis technique , to improve the performance of their dual-template CNN model [33–39]. Although the proposed model did not perform as well as traditional methods, it still achieved accuracy rates of 95.4% and 77.4% on the CK+ and BU-4DFE datasets [27], respectively. Kumar *et al.,* [38] employed a frequency-domain-based approach to remove low-intensity expression frames. Their analysis revealed that low-intensity frames lack textural variation. The emotional snapshot created from the last set of high-intensity frames is improved by the presence of strong motion in those frames. Once all these intense frames have been evaluated, each is assigned an emotion using the appropriate CNN model.

Spatial Pyramid Pooling (SPP), introduced by He *et al.,* [41] has been widely used in computer vision automated systems. These systems encompass various applications such as expression analysis, anti-spoofing technologies, and semantic segmentation. ASPP, developed by [42–47], is extensively employed in several applications, including object identification, image classification, and image segmentation. These are just a few examples of the numerous applications that have utilized ASPP in these domains. Similarly, there is a growing trend of using CNN-based algorithms as a replacement for conventional feature extraction methods due to their self-learning capabilities and the rapid advancement of deep learning models. Consequently, many researchers have become increasingly interested in exploring this area [48], primarily because CNN can learn on its own.

One prominent study by Zhang *et al.,* proposed a novel deep learning architecture called "Facial Expression Transformer" (FET) for FER [49]. The FET model leveraged self-attention mechanisms to capture fine-grained spatial dependencies within facial images, leading to improved expression recognition accuracy. Their results demonstrated superior performance on benchmark datasets such as CK+ and RAF-DB.

In another study, Li *et al.,* explored the use of multimodal approaches for FER by integrating both facial and physiological signals [50]. Their work demonstrated that combining facial features extracted from deep convolutional neural networks with physiological signals, such as heart rate and electrodermal activity, resulted in enhanced emotion recognition accuracy. This fusion of multiple modalities has shown promise in capturing a more comprehensive representation of emotional states.

Additionally, Mollahosseini *et al.,* proposed a novel dataset named "AffectNet" that contains a large-scale collection of diverse facial expressions labelled with valence and arousal values [51]. This dataset enabled researchers to train and evaluate deep learning models for continuous emotion prediction. The availability of such datasets has facilitated advancements in continuous emotion recognition, opening new possibilities for real-world applications.

Furthermore, recent studies have explored the use of generative models, such as Generative Adversarial Networks (GANs), for FER. For instance, the work of Xu *et al.,* introduced a GAN-based approach for facial expression synthesis, allowing for data augmentation and addressing the scarcity of labelled training data [52]. Their method generated realistic facial expressions that facilitated the improved performance of deep learning models for FER.

Facial emotion recognition is crucial in various domains [53], but low-resolution images can hinder accurate classification [54]. To address this, Ullah *et al.,* have explored the use of 2-D canonical correlation analysis (2-D CCA) for image super-resolution in facial emotion recognition [55]. 2-D CCA finds a correlation between low-resolution and high-resolution image pairs and learns a transformation function. This technique has shown improved accuracy compared to traditional methods. Researchers have also introduced variations and enhancements, such as incorporating constraints or combining 2-D CCA with other techniques like deep learning. These approaches hold promise for enhancing resolution and improving emotion recognition. Further research is needed to optimize these schemes for real-world applications.

Facial emotion recognition using deep learning techniques has been an active area of research in computer vision and affective computing. Ullah *et al.,* [56] have explored the use of deep convolutional neural networks (CNNs) such as VGGNet, ResNet, and InceptionNet, trained on large-scale datasets like FER-2013, CK+, and JAFFE, [9-12,57] to achieve robust and accurate emotion recognition. Transfer learning techniques have been employed to fine-tune pre-trained models and leverage learned features from general image recognition tasks. Temporal modelling with recurrent neural networks (RNNs), particularly LSTM networks, has been utilized to capture dynamic facial expressions over time. The fusion of multiple modalities and the integration of generative adversarial

networks (GANs) for facial emotion synthesis and augmentation have also been explored. These approaches have advanced the field, improving accuracy and robustness in facial emotion recognition. However, challenges such as occlusion and pose variation still require further research for real-world applications.

These recent advancements and studies highlight the dynamic nature of FER research, with a focus on developing more accurate models, exploring multimodal approaches, introducing new datasets, and leveraging generative models. Despite these contributions, there remains a need for further research to address challenges such as handling non-frontal and partially obscured facial expressions, improving model generalization across different demographics, and enhancing real-time performance in challenging scenarios.

According to our research, significant advancements have been made in the field of FER. However, current practical applications require solutions that can achieve desired accuracy without incurring high computational costs. Conventional approaches may be faster in identifying FER, but they are less accurate when considering FER in its natural context. The high computational costs associated with using most CNN-based algorithms, which are parameter-intensive, are the main cause of this issue. Despite the success of CNN-based strategies, this challenge persists. For instance, the VGG-16 architecture, widely used in research, requires over 138 million parameters, making it difficult to implement in scenarios with limited resources.

In our proposed method, we incorporate images taken both outdoors and in controlled laboratory settings to improve real-time recognition. This approach differs from previous studies that primarily focused on utilizing frontal faces in laboratory environments for real-time analysis. To address these limitations, we have developed a CNN specifically designed for FER in practical applications, with lower computational demands and fewer parameters (less than 1.3 million). Additionally, transfer learning (TL) can be employed to reduce training time and enhance the accuracy of the model. Moreover, this CNN model can be utilized with images captured in various settings, including controlled laboratory environments and outdoor conditions, to improve real-time recognition. This departure from previous studies, which predominantly focused on frontal faces in laboratory settings, contributes to a more comprehensive analysis. By incorporating transfer learning techniques and fine-tuning pre-trained models, this study seeks to improve the accuracy and efficiency of emotion recognition, thus bridging the gap between existing approaches and the requirements of real-world applications.

2. Methodology

2.1 Introduction

The recommended approach is presented in Figure 7 at a high-level overview. Our research introduces a novel framework that involves the utilization of a CNN within a simulated environment for expression analysis. To ensure the impartiality of our proposed model, we start by acquiring fresh raw images from various datasets (referred to as image acquisition) and eliminating any potential dataset bias. Next, we employ facial detection and resizing methods to isolate the region of interest and preprocess the resulting images to prepare them for training our algorithm in facial emotion identification. Initially, we train the images from the merged dataset using the training data. Our approach incorporates transfer learning to generate a trained model. Once the training phase is completed, our suggested model proceeds to analyse the selected image to detect the presence of a facial structure. When a face is detected by a cascade classifier, the image progresses to the next stage of pre-processing.

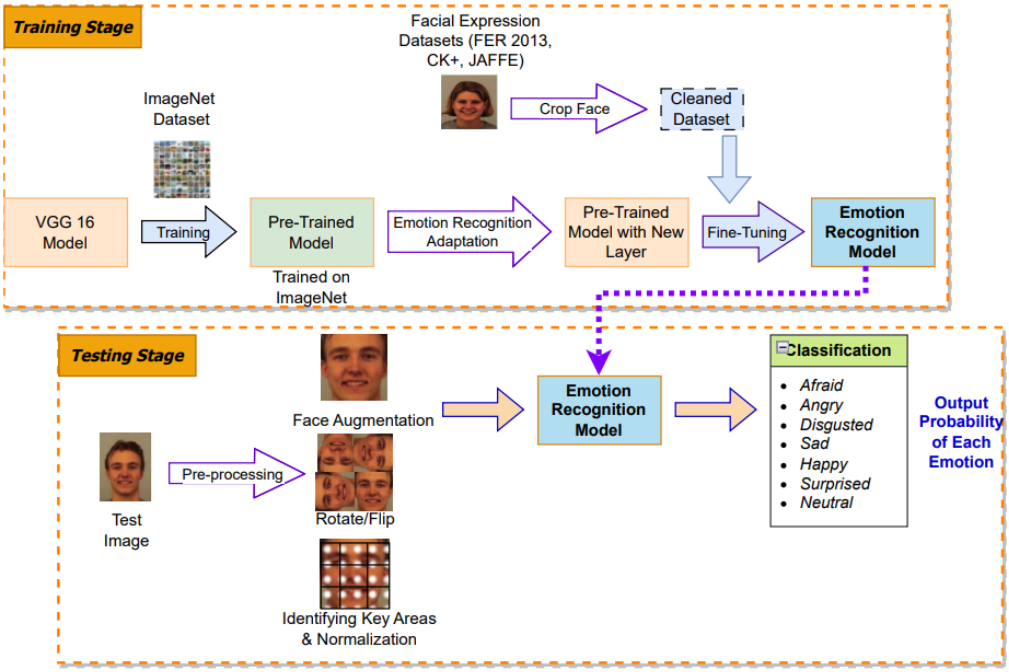

Fig. 7. Proposed system architecture [10–13], [53]

The image pre-processing stage consists of several distinct steps, as illustrated in the proposed system architecture. Various methods and technologies are employed to improve facial expression recognition, including cropping, rotating, flipping, and stretching the detected facial area. Subsequently, normalization and magnification techniques are applied to detect landmarks and align the designated facial expressions. Adjabi *et al.,* utilized the process flow for each phase as depicted in Figure 8, choosing it for inclusion in this study due to its proven effectiveness [58]. In the final stage, supplementary data is incorporated to enhance the accuracy of the model's predictions.

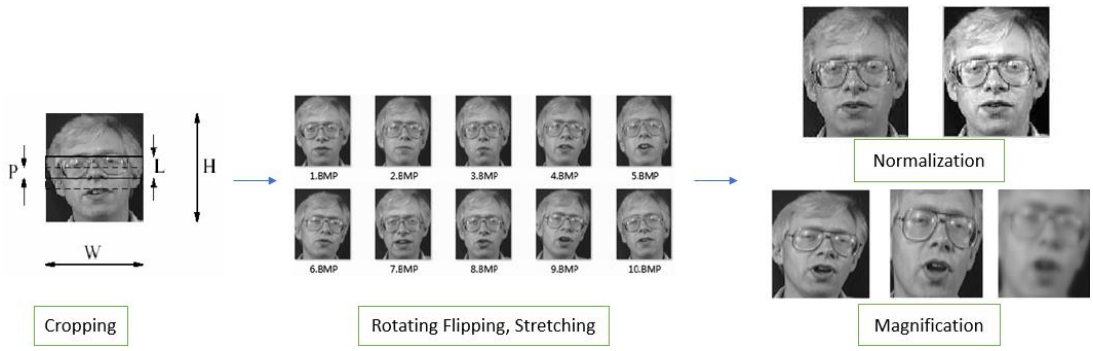

Fig. 8. Process flow

2.2 Pre-Processing

Several factors, such as the capturing device and lighting conditions, influence the need for image pre-processing. Typically, data pre-processing is performed to enhance the image quality before utilization. Standard pre-processing techniques include resizing, histogram equalization, noise reduction, and normalization. Extensive pre-processing tasks may result in longer runtimes, which can hinder real-time detection capabilities. Therefore, our proposed approach utilizes a minimal number of pre-processing steps that do not compromise accuracy. In our study, we employ a two-step pre-processing phase consisting of face detection and data augmentation.

For facial recognition, we utilize the Haar cascade classifier, which provides excellent detection efficiency and accuracy. By adjusting the settings such as minimum neighbours and scale factor, the classifier can detect multiple objects in real-time [59]. We convert the colour image to grayscale before feeding it to the classifier. The classifier then generates four coordinates, which can be used to create a rectangle around the face. To ensure consistency, we scale the detected faces to 48 by 48 pixels.

In real-world scenarios, various factors such as lighting, sound levels, and audience positions can vary significantly. Additionally, background elements may not accurately reflect an individual's emotional state. To train the FER model using CNN, it was necessary to pre-process the visual semantic input to ensure alignment and standardization. This pre-processing step involves the following methods:

i. The initial stage in facial identification within computer vision is to detect faces, which involves recognizing the facial area within an image. Localization determines the face boundaries while facial coordinate detection involves locating the face within the image. The Viola-Jones (V&J) face detector is one of the popular methods used for facial recognition.
ii. Deep learning-based FER systems heavily rely on data augmentation. However, in order to train the CNN model and ensure its adaptability for recognizing various emotions, a large amount of data is required. Pre-processing and cropping of input images are required before they can be utilized in the machine-learning pipeline.
iii. Facial registration is a widely used initial step in face recognition tasks. It involves the adjusting of a sample face to match a known reference face. The process of facial registration aims to align the sample face with the reference face.
iv. The eyes, mouth, nose, and eyebrows are some of the most recognizable facial features that act as landmarks. In our approach, we begin by detecting the individual's head and neck in the image and paying attention to any distinctive attributes of their facial region of interest (ROI).
v. Changes in lighting and head orientation can significantly affect performance and lead to apparent changes in image quality. As a result, we discuss two typical methods for normalizing faces to minimize these variations: standardizing head position and adjusting brightness. Our experiment consisted of three databases with comparable numbers of face images but varying resolutions, resulting in distinct facial expressions.

Thus, to identify the facial boundary, we initially utilized images from the Haar Cascade Library. The identified facial expressions were then cropped and resized to the same dimensions, resulting in rectangular images. To reduce the sparsity of the neural networks, the pixel values of the images were transformed into 48x48 grayscale images before being fed into them.

Researchers often employ data augmentation techniques to artificially expand their datasets and enhance the performance of deep learning models. One effective technique for generating additional data for deep learning models is to use the ImageDataGenerator class from Keras's preprocessing package. This class can generate tensor images in batches on the fly.

2.3 CNN Architecture

The goal of the CNN model proposed in this study is to effectively train the pixel values in the rectangular facial expression region for efficient processing. This enables rapid responses using the

deep artificial neural network model developed in this research. We designed our CNN architecture with the goal of effectively and efficiently training the pixel values within the rectangular area of the facial expression region, as depicted in Figure 9. This is because the image size of FER-2013 is notably smaller (48x48) than the typical input size (224x224 or 299x299) of deep learning models, which influenced our architectural choices. Resizing an image result in additional redundant pixels, leading to duplicate information and diminished feature learning. Additionally, converting grayscale images to colour images requires extra computation, since the FER-2013 dataset only provides grayscale images. Therefore, it is advisable to create a CNN architecture with a lower number of parameters to reduce computational time and memory usage. The input to the CNN is a grayscale image of 48x48 pixels, which undergoes convolution through the convolutional layer (CL).

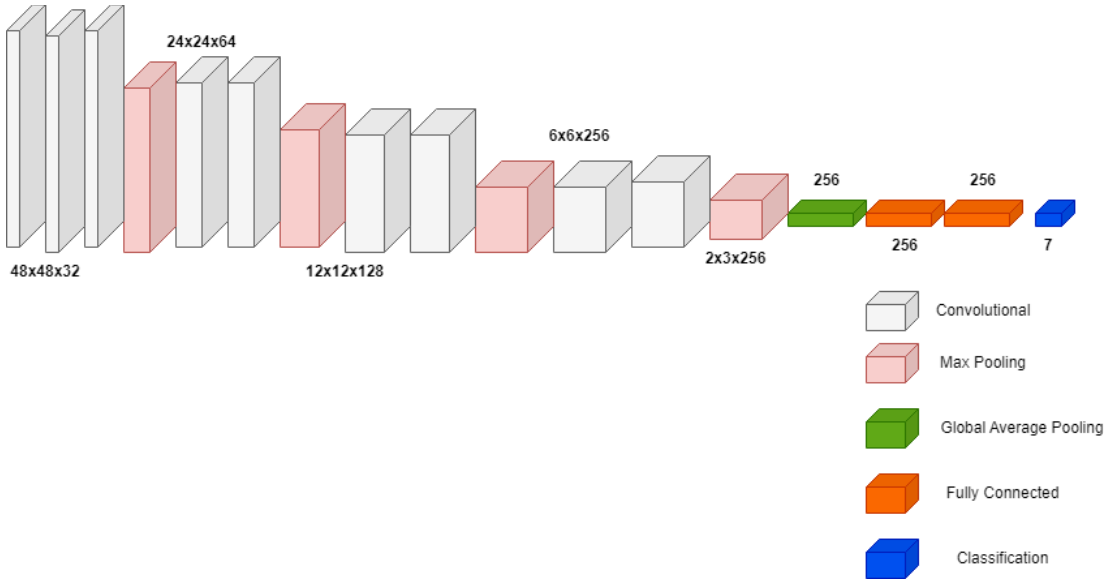

Fig. 9. Convolution Neural Network (CNN) model architecture [16]

The process starts with the application of convolutional layers, which use filters to extract features from image patches.

Upon receiving a 48x48 input image, the convolutional layer initially produces 32 feature maps by convolving with 32 3x3 kernels. Following the initial convolutional layer, we employed seven additional convolutional layers with 3x3 filters and a stride of 1. These layers were used to extract features at different levels, including 32, 64, 128, 256, and 256 features. This is presented as Eq. (1).

$$A_j^i = m(\sum_{t=1}^{N-1}. A_i^{i-1} * w_{ij} + wt_b) \quad (1)$$

The convolutional operation is indicated by the * symbol, where the feature maps are represented by Ai and the filter by w. Through the use of suitable filters, the convolutional layer (CL) can capture spatial and temporal dependencies within an image. Nonlinear characteristics and interaction effects are captured by the ReLU activation function, which always follows the CL. The function returns 0 when a negative value is passed in, but any positive x value will be returned. One can utilize Eq. (2) to determine the input neuron's value, represented as x.

$$f(x) = \max(0, x) \quad (2)$$

After each pair of convolutional layers (CLs), a Max Pooling Layer (MPL) is employed to downsample the output feature maps and reduce their dimensionality. The MPL involves using 2x2 filters with a stride of 1. The MPL downsamples the feature maps to eliminate redundant information. The following formula is used to compute the MPL as Eq. (3).

$$A_j^i = F(MPA_i^{i-1} + w_b) \tag{3}$$

To reduce the computation cost and prevent overfitting, a simplified representation of the data needs to be created by downsampling the dimensions, which allows the system to learn about features within binned sub-regions. This process involves producing a representation of the data and removing any non-overlapping parts. Furthermore, incorporating Max Pooling Layers (MPLs) in the network architecture provides some degree of translation invariance to the features, while also reducing the number of parameters that need to be trained.

In order to train the model, a grayscale image of size 48 x 48 is inputted and various optimization and regularization techniques are applied. The final output class represents only one emotion from the set of seven emotions. The architecture of the CNN includes four layers of convolution and four layers of MaxPooling.

2.4 Network Training

During the training phase of the network, a test size of 25% was selected. To ensure parameter convergence, a batch size of 32 and 500 epochs were utilized. The defined learning rate was 10^{-3}. A stride of 2 was used along with a kernel size of 2x2 for each convolutional and max-pooling layer.

2.4.1 Hyper parameters

We have utilized various hyperparameters in our model to analyse the facial expression database. Instead of exploring the effectiveness of the model on different databases, we have provided a brief overview of our training approach. We trained our model on all the data from the experiment and made efforts to standardize the architecture and hyperparameters to the best of our ability.

Each model underwent a total of fifty training epochs. The initial weights of the network were generated by sampling from a Gaussian distribution with a mean of zero and a certain standard deviation. To prevent overfitting, we applied a regularization or shrinkage technique by setting coefficients to 0. We tested dropout values ranging from 0.1 to 0.4 before making a decision. Finally, we chose to use Softmax activation for multiclassification in the dense layer.

2.5 Testing in Real-Time

The CNN architecture proposed in this study was trained and tested using real-world data. To detect human faces from a computer's camera at a rate of 30 frames per second, the Haar Cascade Library was utilized. The detected images were then passed to the model for classification. The resulting predictions were displayed on a secondary screen, showing the probability of each facial expression belonging to a certain class. The system displayed the emotion with the highest probability on top of the Haar cascade frame, and this process was repeated for every 30 frames per second of the live camera stream.

2.6 Transfer Learning

Transfer learning (TL) is a widely used technique in deep learning, where pre-trained model weights are utilized to start training for a new task. TL aims to enhance the performance of the model across various problems, resulting in faster training and better overall efficiency.

Transfer learning (TL) is especially beneficial in the context of facial expression recognition (FER) as the available datasets are often limited in size, making it challenging to train CNN-based models effectively. While millions of images would be ideal for training CNNs, FER datasets generally consist of only a few hundred or thousand images, resulting in difficulties in training CNN-based models and the potential for overfitting due to the lack of data.

To address the issue of limited datasets, we employ the inductive transfer learning approach. This involves utilizing weights obtained from training on a larger dataset to initialize the training process for a smaller dataset. In our case, we utilize transfer learning by using a pre-trained model from the larger FER-2013 dataset [12] to train on the CK+ and JAFFE datasets [10–12], [53]. This allows us to leverage the knowledge and feature representations learned from the larger dataset to improve the performance of our model on the smaller datasets.

2.6.1 Fine-tuning the Convolution Neural Network (CNN)

Fine-tuning is an essential step in transfer learning that allows us to adapt a pre-trained model to a specific task, such as emotion recognition. When fine-tuning a model, we selectively update the weights of certain layers while keeping the weights of earlier layers fixed. This process enables the model to specialize and learn task-specific features while leveraging the general knowledge acquired from pre-training on large-scale image datasets.

In our study, we utilize various pre-trained models, including ResNet, AlexNet, VGG16, VGG19, Inception, and our proposed models. These models have been trained on extensive image datasets, such as ImageNet, to learn rich and general representations of visual patterns. They have demonstrated strong performance in image classification tasks and are widely used in transfer learning scenarios.

When fine-tuning these models for emotion recognition, we typically freeze the early layers, which capture low-level features like edges and textures, since these features are already well-learned and transferable across tasks. The later layers, which capture more abstract and high-level features, are fine-tuned to adapt to the specific emotional patterns we want to recognize.

By fine-tuning the pre-trained models, we allow them to specialize and learn emotional features that are relevant to our target task. This approach saves significant training time and computational resources compared to training a model from scratch. Furthermore, the pre-trained models provide a strong starting point with learned representations that can capture meaningful patterns in images, improving the overall performance and generalization ability of the model for emotion recognition.

2.7 Datasets

In order to assess the performance of our model, we chose two widely used datasets: CK+ and JAFFE [10–12, 53]. Additionally, we included the FER-2013 dataset due to its comprehensive nature, versatility, and free accessibility. The expressions captured in the FER-2013 dataset were obtained from an uncontrolled setting, while the CK+ and JAFFE datasets were obtained in a highly regulated laboratory environment.

The FER-2013 dataset [12] was created as part of the ICML 2013 Workshop on Representation Learning, which included a challenge to create a dataset. This dataset consists of 35,887 grayscale images depicting seven distinct emotional states. The images were obtained through Google's image search API. It is widely used for training deep learning models, and each image has a size of 48x48 pixels.

In 2010, the CK+ dataset [11], which is an extended version of the CK dataset, was released to researchers. It contains facial expressions from 213 individuals displaying emotions ranging from neutral to intense. The dataset includes photographs of six universal expressions of emotions, including contempt, and the images are labelled using FACS codes. Compared to the original CK dataset, this dataset has a greater number of participants and sequences.

The JAFFE dataset comprises 213 grayscale images depicting six universal emotions. The images were captured in a controlled setting using ten Japanese female models and have a resolution of 256x256 pixels [10, 11, 53].

3. Results

3.1 Introduction

In this section, we will provide an overview of the offline and online experiments conducted and present a detailed analysis of the evaluation results obtained from our model. Our evaluation methodology is based on experiments carried out on two important facial expression recognition datasets, which will also be discussed. We will address the challenges and issues encountered when working with databases in this section as well. Furthermore, we will evaluate the performance of our model using different hyperparameters and conduct experiments on both the CK+ and JAFFE datasets to assess its effectiveness. The number of records used from each dataset is presented in Table 1.

Table 1
Total Number of records in the study dataset [10-13,53]

Emotions	CK+	JAFFE	FER-2013
Surprise	123	52	2137
Sad	379	43	1292
Neutral	369	47	3107
Happy	416	49	2391
Fear	276	43	2316
Disgust	39	46	789
Angry	247	47	3215

Researchers in facial expression recognition (FER) are increasingly turning to deep learning as a powerful approach to overcome challenges such as variations in lighting, obstructed views, identification bias, and low-intensity expression recognition, especially in challenging environmental conditions. However, deep learning models require a large dataset for training to accurately capture subtle shifts in expression. The availability of sufficient and high-quality training data is therefore crucial for the development of effective deep FER systems.

3.2 Data Analysis

To enable both testing and training, the dataset was divided into training and testing sets with 70% and 30%, respectively. A detailed description of the entire data preparation process is provided in the approach section. To ensure unbiased testing on the CK+, FER, and JAFFE datasets, we applied

five-fold and ten-fold cross-validation techniques. Data augmentation was exclusively applied to the training data, while normalization was applied to the validation and testing data. Training the proposed model on the FER-2013 dataset proved challenging due to its diverse nature. One of the most difficult tasks was developing a concise CNN architecture with minimal pre-processing steps.

Our proposed approach achieved a substantial accuracy improvement of 2.51% compared to several state-of-the-art CNN-based models. To ensure accuracy despite its small size, we evaluated the JAFFE dataset using the ten-fold cross-validation method. Using the pre-trained weight on FER-2013 and without any augmentation, our proposed method achieved an accuracy of 97.66% on the JAFFE dataset, using the same experimental setup. Additionally, by incorporating data augmentation, we achieved an improved accuracy of 93.92% on the JAFFE dataset. In addition to the inclusion of a contempt class, the CK+ dataset's range of emotions from neutral to extremely compelled researchers to investigate novel evaluation approaches. In this research, we used the first frame as a control and analysed emotions using the sixth through twelfth peak frames.

In order to overcome any potential bias that may be introduced by suboptimal training-test division, whether done randomly or by the user, a five-fold cross-validation method, which is a well-established method, is employed. The combination of two datasets not only resulted in a larger dataset but also provided an opportunity to include both laboratory-controlled and natural photos from the FER-2013 dataset. After obtaining the pre-trained weight from the combined dataset, we utilized it in the real-time application. Using this approach, we achieved an accuracy rate of 71.45% for transfer learning from FER-2013 to the new dataset (FER-2013 plus CK+). Compared to other state-of-the-art methods, our proposed approach outperforms in augmentation, CNN structure, and transfer learning.

The FER-2013 dataset includes images captured in natural settings, which poses challenges for accurate facial expression recognition. The proposed augmentation procedure increases the model's accuracy by approximately 4% and enhances its adaptability during training. Table 2, Table 3, and Table 4 present a comparison of our proposed model with previous studies.

Table 2
Comparison of results for the FER-2013 dataset

Methodology	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)	Neutral (%)	Efficiency (%)
ResNet	64	66	61	69	61	64	73	65.4
AlexNet	58	55	53	57	50	54	64	55.9
VGG16	69	64	67	73	59	65	75	67.4
VGG19	73	70	74	81	64	66	81	72.7
Inception	78	71	75	80	72	71	74	74.4
Proposed	77	75	80	85	70	76	81	77.7

Table 3
Comparison of performance with existing literature on the JAFFE dataset

Methodology	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)	Neutral (%)	Efficiency (%)
ResNet	63	66	65	69	64	70	82	68.7
AlexNet	59	55	57	57	53	60	73	59.1
VGG16	70	63	71	73	65	71	84	70.7
VGG19	72	70	78	81	67	72	90	76.0
Inception	79	71	79	80	75	77	83	77.7
Proposed	77	74	83	84	72	81	89	80.0

Table 4
Comparison of performance with existing literature on the CK+ dataset

Methodology	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)	Neutral (%)	Efficiency (%)
ResNet	74	74	70	78	77	85	97	79.3
AlexNet	68	63	62	66	66	75	88	69.7
VGG16	79	72	76	82	75	86	99	81.3
VGG19	83	78	83	90	80	87	95	85.1
Inception	88	79	84	89	88	92	98	88.3
Proposed	86	82	88	93	85	96	94	89.1

Facial expression recognition (FER) involves detecting facial muscle activity, which is typically captured in grayscale photographs. To enhance the efficiency and accuracy of the training process, the proposed method incorporates transfer learning (TL). Fine-tuning the weights to minimize the error rate is a time-consuming task when starting from raw data. To achieve the highest degree of accuracy in practical applications, transfer learning is utilized to fine-tune the final model.

The choice of an appropriate optimization technique is crucial for achieving higher accuracy by facilitating the identification of optimal input parameters through the minimization of error rates.

3.2 Performance Evaluation of Transfer Learning for Facial Expression Recognition with Previous Research

The recognition of facial expressions has gained significant attention in recent years due to its potential applications in a range of fields, including healthcare, security, and entertainment. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown remarkable outcomes in facial expression recognition. Nevertheless, developing a CNN model from scratch requires a significant amount of labelled data and computing power, which can be expensive and time-consuming. In contrast, transfer learning allows the use of pre-trained CNNs to extract features from images and fine-tune the model for a specific task with a smaller amount of labelled data. The aim of this study is to assess the effectiveness of transfer learning in facial expression recognition by utilizing three benchmark datasets - CK+, JAFFE, and FER-2013 [10–13, 53]. These datasets are widely employed for evaluating the performance of facial expression recognition methods.

The objective is to compare the performance of transfer learning with traditional approaches used in previous research. The objective of this study is to conduct a comprehensive performance evaluation of transfer learning for facial expression recognition by analysing the results of various studies. This analysis aims to identify areas for future research and provide a deeper understanding of the effectiveness of this approach. This experiment's findings will contribute to the advancement of facial expression recognition technology and inform researchers and practitioners about the effectiveness of transfer learning for this task.

Greco *et al.,* [60] utilized transfer learning with a pre-trained deep convolutional neural network and attained the highest performance in the EmotiW 2016 facial expression recognition challenge, achieving a weighted accuracy of 62.45%. This was significantly better than the second-best approach which achieved a weighted accuracy of 53.77%. Li *et al.,* [61] also concluded that the transfer learning based model for the JAFFE and CK+ datasets yielded an overall accuracy of 90.2%. This performance is higher than the accuracy achieved by several other approaches.

Aguilera *et al.,* [50] achieved an accuracy of 63.85% on the Emotion Recognition in the Wild dataset using a transfer learning approach with a pre-trained deep neural network. This was higher than the accuracy achieved by several other approaches.

Fernandez *et al.,* [62] reported achieving an accuracy of 66.5% on the FER-2013 dataset by utilizing a transfer learning method with pre-trained deep convolutional neural networks. This was also higher than the accuracy achieved by some other approaches.

Overall, it is difficult to directly compare the performance of these studies due to differences in the datasets, pre-trained models used, and evaluation metrics. However, all these studies demonstrate the effectiveness of transfer learning for facial emotion recognition, achieving high accuracy compared to other approaches.

It is important to highlight that the effectiveness of transfer learning approaches can be influenced by various factors, including the volume and quality of the target dataset, the selection of the pre-trained model, and the specific objective of the task at hand. Hence, it is crucial to conduct a thorough evaluation of various transfer learning techniques within the context of a specific application to determine their performance.

3.3 Novelty of Real-Time Deep Learning for Accurate and Dynamic Emotional State Prediction

The primary novelty of this study lies in the development of a cutting-edge system that revolutionizes the field of emotion recognition. Our approach is distinguished from existing methods by its unique combination of advanced technologies and ability to accurately predict emotional states in real-time. By leveraging the power of deep learning, specifically Convolutional Neural Networks (CNNs), our system achieves exceptional performance and surpasses many state-of-the-art algorithms.

One key contribution of our study is the integration of real-time processing capabilities into the emotion recognition process. Traditional approaches often rely on offline analysis, which limits their applicability in dynamic and time-sensitive contexts. Our system addresses this limitation by enabling on-the-fly emotion recognition from live video streams or real-time data inputs. This opens up exciting possibilities for a wide range of applications, including mental health monitoring, human-computer interaction, and personalized services.

Another important aspect that sets our study apart is the innovative use of CNNs, which eliminates the need for manual feature extraction. Traditional methods often require domain experts to manually define and extract relevant features, a labour-intensive and time-consuming process. In contrast, our CNN-based approach automatically learns and extracts discriminative features directly from the data, enabling more accurate and efficient emotion recognition.

Furthermore, our system incorporates advanced techniques such as transfer learning and data augmentation. Transfer learning allows us to leverage pre-trained models trained on large datasets, such as the FER-2013 dataset, to initialize the training process for smaller datasets like CK+ and JAFFE. This approach effectively addresses the challenge of limited training data, improving the generalization and performance of our system. Data augmentation techniques further enhance the training process by artificially expanding the dataset, increasing its diversity and robustness.

The unique value of our proposed system lies in its ability to deliver timely and precise emotional insights. By accurately predicting emotional states in real time, our system can provide valuable information for mental health professionals, enabling early detection and intervention in cases of emotional distress. In human-computer interaction, our system can enhance the user experience by adapting interfaces and interactions based on the user's emotional state, leading to more personalized and engaging experiences. Moreover, in personalized services and social relationships, our system can provide valuable feedback and recommendations based on the emotional states of individuals, promoting improved communication and well-being.

In summary, our study represents a significant advancement in the field of emotion recognition by combining real-time processing capabilities, deep learning techniques, and innovative methodologies such as transfer learning and data augmentation. The unique value and innovation of our proposed system holds tremendous potential for a wide range of applications, impacting fields such as mental health, human-computer interaction, and personalized services. Through our research, we push the boundaries of possible and open new avenues for future research and practical implementation in the exciting field of emotion recognition.

4. Conclusions

The objective of this research is to develop a real-time system that can accurately recognize various emotional expressions of a group of individuals from a single video frame. The proposed method not only achieves high accuracy on the FER-2013 dataset but also surpasses many state-of-the-art algorithms on the JAFFE and CK+ datasets. In contrast to conventional classifiers and CNN-based methods, the suggested software can efficiently detect a broad spectrum of emotional states with relatively lower computational complexity. The utilization of CNN-based techniques eliminates the requirement for manual feature extraction, which can be a demanding and time-consuming task. Unlike other machine learning techniques that require extensive pre-processing, CNNs can produce state-of-the-art results with minimal pre-processing. One of the major difficulties for current research methodologies, including the proposed strategy, is recognizing facial expressions that are strongly non-frontal and partially obscured. This issue needs to be addressed and improved to achieve better results. In order to address the challenges, a more diverse and comprehensive dataset is needed that can handle extreme scenarios. Moreover, there is a need to improve the accuracy of the face detection system to reliably recognize faces in challenging scenarios, such as low illumination and occlusion, which would lead to better real-time performance of the proposed system. Although the proposed approach utilizes static images for FER, it may be worthwhile to explore the use of video data in the future to leverage time-domain information.

Acknowledgement
Authors would like to thank Ministry of Higher Education Malaysia and Universiti Malaysia Pahang Al Sultan Abdullah for funding under grant PGRS230344 and RDU220124.

References
[1] Liu, Weifeng, Caifeng Song, and Yanjiang Wang. "Facial expression analysis using a sparse representation based space model." In *2012 IEEE 11th International Conference on Signal Processing*, vol. 3, pp. 1659-1662. IEEE, 2012. https://doi.org/10.1109/ICoSP.2012.6491899

[2] Kandhro, Irfan Ali, Mueen Uddin, Saddam Hussain, Touseef Javed Chaudhery, Mohammad Shorfuzzaman, Hossam Meshref, Maha Albalhaq, Raed Alsaqour, and Osamah Ibrahim Khalaf. "Impact of Activation, Optimization, and Regularization Methods on the Facial Expression Model Using CNN." *Computational Intelligence and Neuroscience* 2022 (2022). https://doi.org/10.1155/2022/3098604

[3] Krajinović, Karla. "Neverbalna komunikacija: mikroekspresije." PhD diss., University of Zagreb. Department of Croatian Studies. Division of Communication Sciences, 2017.

[4] Ozdemir, Mehmet Akif, Berkay Elagoz, Aysegul Alaybeyoglu, Reza Sadighzadeh, and Aydin Akan. "Real time emotion recognition from facial expressions using CNN architecture." In *2019 medical technologies congress (tiptekno)*, pp. 1-4. IEEE, 2019. https://doi.org/10.1109/TIPTEKNO.2019.8895215

[5] Benitez-Garcia, Gibran, Tomoaki Nakamura, and Masahide Kaneko. "Facial expression recognition based on local fourier coefficients and facial fourier descriptors." *Journal of Signal and Information Processing* 8, no. 3 (2017): 132-151. https://doi.org/10.4236/jsip.2017.83009

[6] Anwarul, Shahina, and Susheela Dahiya. "A comprehensive review on face recognition methods and factors affecting facial recognition accuracy." *Proceedings of ICRIC 2019: Recent Innovations in Computing* (2020): 495-514. https://doi.org/10.1007/978-3-030-29407-6_36

[7] Podder, Tanusree, Diptendu Bhattacharya, and Abhishek Majumdar. "Time efficient real time facial expression recognition with CNN and transfer learning." *Sādhanā* 47, no. 3 (2022): 177. https://doi.org/10.1007/s12046-022-01943-x

[8] Banala, Rajesh, Vicky Nair, and P. Nagaraj. "Performance of Secure Data Deduplication Framework in Cloud Services." In *International Conference on Artificial Intelligence and Data Science*, pp. 392-403. Cham: Springer Nature Switzerland, 2021. https://doi.org/10.1007/978-3-031-21385-4_32

[9] Lyons, Michael, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. "Coding facial expressions with gabor wavelets." In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pp. 200-205. IEEE, 1998. https://doi.org/10.1109/AFGR.1998.670949

[10] Lyons, Michael J. "" Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset." *arXiv preprint arXiv:2107.13998* (2021). https://doi.org/10.2139/ssrn.3900990

[11] Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "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. IEEE, 2010. https://doi.org/10.1109/CVPRW.2010.5543262

[12] Goodfellow, Ian J., Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski et al. "Challenges in representation learning: A report on three machine learning contests." In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pp. 117-124. Springer berlin heidelberg, 2013. https://doi.org/10.1007/978-3-642-42051-1_16

[13] Wikanningrum, Anggit, Reza Fuad Rachmadi, and Kohichi Ogata. "Improving lightweight convolutional neural network for facial expression recognition via transfer learning." In *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/CENIM48368.2019.8973312

[14] Bhatti, Yusra Khalid, Afshan Jamil, Nudrat Nida, Muhammad Haroon Yousaf, Serestina Viriri, and Sergio A. Velastin. "Facial expression recognition of instructor using deep features and extreme learning machine." *Computational Intelligence and Neuroscience* 2021 (2021): 1-17. https://doi.org/10.1155/2021/5570870

[15] Nesakumar, A. Darwin, R. Priyadharshini, K. Pavithra, L. Sherin, K. N. Pavithra, and P. Mugilan. "Emotional Based Playback System for Autism." In *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, pp. 1328-1332. IEEE, 2022. https://doi.org/10.1109/IIHC55949.2022.10060348

[16] Reddi, Palasatti Srinivasa, and A. Sri Krishna. "CNN Implementing Transfer Learning for Facial Emotion Recognition." *International Journal of Intelligent Systems and Applications in Engineering* 11, no. 4s (2023): 35-45.

[17] Sato, Atsuhshi. "Advances in face detection and recognition technologies." *NEC J Adv Technol* 2 (2005): 28-34.

[18] Avabhrith, Ramya, and Jharna Majumdar. "Human Face Expression Recognition."

[19] Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE signal processing letters* 23, no. 10 (2016): 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[20] Lekdioui, Khadija, Yassine Ruichek, Rochdi Messoussi, Youness Chaabi, and Raja Touahni. "Facial expression recognition using face-regions." In *2017 international conference on advanced technologies for signal and image processing (ATSIP)*, pp. 1-6. IEEE, 2017. https://doi.org/10.1109/ATSIP.2017.8075517

[21] Hajarolasvadi, Noushin, and Hasan Demirel. "Deep facial emotion recognition in video using eigenframes." *IET Image Processing* 14, no. 14 (2020): 3536-3546. https://doi.org/10.1049/iet-ipr.2019.1566

[22] Baffour, Prince Awuah, Henry Nunoo-Mensah, Eliel Keelson, and Benjamin Kommey. "A Survey on Deep Learning Algorithms in Facial Emotion Detection and Recognition." *Inform: Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi* 7, no. 1 (2022): 24-32. https://doi.org/10.25139/inform.v7i1.4563

[23] Chowdary, M. Kalpana, Tu N. Nguyen, and D. Jude Hemanth. "Deep learning-based facial emotion recognition for human–computer interaction applications." *Neural Computing and Applications* (2021): 1-18. https://doi.org/10.1007/s00521-021-06012-8

[24] Liu, Yong-Jin, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. "A main directional mean optical flow feature for spontaneous micro-expression recognition." *IEEE Transactions on Affective Computing* 7, no. 4 (2015): 299-310. https://doi.org/10.1109/TAFFC.2015.2485205

[25] Zeng, Yun, Xilin Liu, and Lehua Cheng. "Facial emotion perceptual tendency in violent and non-violent offenders." *Journal of interpersonal violence* 37, no. 17-18 (2022): NP15058-NP15074. https://doi.org/10.1177/0886260521989848

[26] Naumann, Sandra, Mareike Bayer, and Isabel Dziobek. "Preschoolers' sensitivity to negative and positive emotional facial expressions: an ERP study." *Frontiers in Psychology* 13 (2022): 828066. https://doi.org/10.3389/fpsyg.2022.828066

[27] Chen, Jingying, Yongqiang Lv, Ruyi Xu, and Can Xu. "Automatic social signal analysis: Facial expression recognition using difference convolution neural network." *Journal of Parallel and Distributed Computing* 131 (2019): 97-102. https://doi.org/10.1016/j.jpdc.2019.04.017

[28] Ezerceli, Özay, and M. Taner Eskil. "Convolutional Neural Network (CNN) Algorithm Based Facial Emotion Recognition (FER) System for FER-2013 Dataset." In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1-6. IEEE, 2022. https://doi.org/10.1109/ICECCME55909.2022.9988371

[29] Pang, Shuchao, Juan José del Coz, Zhezhou Yu, Oscar Luaces, and Jorge Díez. "Deep learning to frame objects for visual target tracking." *Engineering Applications of Artificial Intelligence* 65 (2017): 406-420. https://doi.org/10.1016/j.engappai.2017.08.010

[30] Ronao, Charissa Ann, and Sung-Bae Cho. "Human activity recognition with smartphone sensors using deep learning neural networks." *Expert systems with applications* 59 (2016): 235-244. https://doi.org/10.1016/j.eswa.2016.04.032

[31] Christou, Nikolaos, and Nilam Kanojiya. "Human facial expression recognition with convolution neural networks." In *Third International Congress on Information and Communication Technology: ICICT 2018, London*, pp. 539-545. Springer Singapore, 2019. https://doi.org/10.1007/978-981-13-1165-9_49

[32] Sajjanhar, Atul, ZhaoQi Wu, and Quan Wen. "Deep learning models for facial expression recognition." In *2018 digital image computing: Techniques and applications (dicta)*, pp. 1-6. IEEE, 2018. https://doi.org/10.1109/DICTA.2018.8615843

[33] Dosovitskiy, Alexey, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. "Flownet: Learning optical flow with convolutional networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2758-2766. 2015. https://doi.org/10.1109/ICCV.2015.316

[34] Burada, Sreedhar, BE Manjunath Swamy, and M. Sunil Kumar. "Computer-Aided Diagnosis Mechanism for Melanoma Skin Cancer Detection Using Radial Basis Function Network." In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1*, pp. 619-628. Singapore: Springer Nature Singapore, 2022. https://doi.org/10.1007/978-981-19-2350-0_60

[35] Balaji, K., P. Sai Kiran, and M. Sunil Kumar. "Power aware virtual machine placement in IaaS cloud using discrete firefly algorithm." *Applied Nanoscience* 13, no. 3 (2023): 2003-2011. https://doi.org/10.1007/s13204-021-02337-x

[36] Kumar, M. Sunil, B. Siddardha, A. Hitesh Reddy, Ch V. Sainath Reddy, Abdul Bari Shaik, and D. Ganesh. "Applying The Modular Encryption Standard To Mobile Cloud Computing To Improve The Safety Of Health Data." *Journal of Pharmaceutical Negative Results* (2022): 1911-1917. https://doi.org/10.47750/pnr.2022.13.S08.231

[37] Wei, Heng, and Zhi Zhang. "A survey of facial expression recognition based on deep learning." In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 90-94. IEEE, 2020. https://doi.org/10.1109/ICIEA48937.2020.9248180

[38] Jain, Neha, Shishir Kumar, Amit Kumar, Pourya Shamsolmoali, and Masoumeh Zareapoor. "Hybrid deep neural networks for face emotion recognition." *Pattern Recognition Letters* 115 (2018): 101-106. https://doi.org/10.1016/j.patrec.2018.04.010

[39] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016. https://doi.org/10.1109/CVPR.2016.90

[40] Zhang, Yongmian, and Qiang Ji. "Active and dynamic information fusion for facial expression understanding from image sequences." *IEEE Transactions on pattern analysis and machine intelligence* 27, no. 5 (2005): 699-714. https://doi.org/10.1109/TPAMI.2005.93

[41] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 9 (2015): 1904-1916. https://doi.org/10.1007/978-3-319-10578-9_23

[42] Sanjay, S., SS Nithish Soorya, R. Vengatesh, and KC Sri Hari Priya. "Security Access Control System Enhanced with Face Mask Detection and Temperature Monitoring for Pandemic Trauma." In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pp. 1-6. IEEE, 2022. https://doi.org/10.1109/CONIT55038.2022.9848266

[43] Venkateswara Reddy, L., Ganesh Davanam, T. Pavan Kumar, M. Sunil Kumar, and Mekala Narendar. "Bio-Inspired Firefly Algorithm for Polygonal Approximation on Various Shapes." In *Intelligent Computing and Applications: Proceedings of ICDIC 2020*, pp. 95-107. Singapore: Springer Nature Singapore, 2022. https://doi.org/10.1007/978-981-19-4162-7_10

[44] Ganesh, D., K. Jayanth Rao, M. Sunil Kumar, M. Vinitha, M. Anitha, S. Sai Likith, and Racheal Taralitha. "Implementation of Novel Machine Learning Methods for Analysis and Detection of Fake Reviews in Social Media." In *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 243-250. IEEE, 2023. https://doi.org/10.1109/ICSCDS56580.2023.10104856

[45] Ahmed, Shamim, Robiul Alam, Md Rasel Hossain, Md Mominul Islam, Md Imran Hossain, and Tamim Tabassum. "An IoT based Smart Robot that Aids in the Prevention of COVID19 Spread." In *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1-6. IEEE, 2022. https://doi.org/10.1109/STI56238.2022.10103253

[46] Srivastava, Devesh Kumar, and Dharmendra Narayan Jha. "Hibiscus Flower Health Detection to Produce Oil Using Convolution Neural Network." In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, pp. 1-5. IEEE, 2022. https://doi.org/10.1109/ASSIC55218.2022.10088339

[47] Chang, Yaohui, Chunhua Gu, and Fei Luo. "A novel energy-aware and resource efficient virtual resource allocation strategy in IaaS cloud." In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1283-1288. IEEE, 2016. https://doi.org/10.1109/CompComm.2016.7924911

[48] De Luna, Robert G., Elmer P. Dadios, and Argel A. Bandala. "Automated image capturing system for deep learning-based tomato plant leaf disease detection and recognition." In *TENCON 2018-2018 IEEE Region 10 Conference*, pp. 1414-1419. IEEE, 2018. https://doi.org/10.1109/TENCON.2018.8650088

[49] Sun, Ning, Yao Song, Jixin Liu, Lei Chai, and Haian Sun. "Appearance and geometry transformer for facial expression recognition in the wild." *Computers and Electrical Engineering* 107 (2023): 108583. https://doi.org/10.1016/j.compeleceng.2023.108583

[50] Aguilera, Ana, Diego Mellado, and Felipe Rojas. "An Assessment of In-the-Wild Datasets for Multimodal Emotion Recognition." *Sensors* 23, no. 11 (2023): 5184. https://doi.org/10.3390/s23115184

[51] Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild." *IEEE Transactions on Affective Computing* 10, no. 1 (2017): 18-31. https://doi.org/10.1109/TAFFC.2017.2740923

[52] Liu, Leyuan, Rubin Jiang, Jiao Huo, and Jingying Chen. "Self-difference convolutional neural network for facial expression recognition." *Sensors* 21, no. 6 (2021): 2250. https://doi.org/10.3390/s21062250

[53] Du, Shichuan, Yong Tao, and Aleix M. Martinez. "Compound facial expressions of emotion." *Proceedings of the national academy of sciences* 111, no. 15 (2014): E1454-E1462. https://doi.org/10.1073/pnas.1322355111

[54] Liu, Tai-Ling, Peng-Wei Wang, Yi-Hsin Connie Yang, Gary Chon-Wen Shyi, and Cheng-Fang Yen. "Association between facial emotion recognition and bullying involvement among adolescents with high-functioning autism spectrum disorder." *International journal of environmental research and public health* 16, no. 24 (2019): 5125. https://doi.org/10.3390/ijerph16245125

[55] Ullah, Zia, Lin Qi, D. Binu, B. R. Rajakumar, and B. Mohammed Ismail. "2-D canonical correlation analysis based image super-resolution scheme for facial emotion recognition." *Multimedia Tools and Applications* 81, no. 10 (2022): 13911-13934. https://doi.org/10.1007/s11042-022-11922-3

[56] Ullah, Zia, Lin Qi, Asif Hasan, and Muhammad Asim. "Improved Deep CNN-based Two Stream Super Resolution and Hybrid Deep Model-based Facial Emotion Recognition." *Engineering Applications of Artificial Intelligence* 116 (2022): 105486. https://doi.org/10.1016/j.engappai.2022.105486

[57] Lyons, Michael, Miyuki Kamachi, and Jiro Gyoba. "The Japanese female facial expression (JAFFE) dataset." *The images are provided at no cost for non-commercial scientific research only. If you agree to the conditions listed below, you may request access to download* (1998). https://doi.org/10.5281/ZENODO.3451524

[58] Adjabi, Insaf, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. "Past, present, and future of face recognition: A review." *Electronics* 9, no. 8 (2020): 1188. https://doi.org/10.3390/electronics9081188

[59] Babenko, Artem, and Victor Lempitsky. "Efficient indexing of billion-scale datasets of deep descriptors." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2055-2063. 2016. https://doi.org/10.1109/CVPR.2016.226

[60] Greco, Antonio, Nicola Strisciuglio, Mario Vento, and Vincenzo Vigilante. "Benchmarking deep networks for facial emotion recognition in the wild." *Multimedia tools and applications* 82, no. 8 (2023): 11189-11220. https://doi.org/10.1007/s11042-022-12790-7

[61] Li, Kuan, Yi Jin, Muhammad Waqar Akram, Ruize Han, and Jiongwei Chen. "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy." *The visual computer* 36 (2020): 391-404. https://doi.org/10.1007/s00371-019-01627-4

[62] Fernández, Dennis Núnez. "Multi-subject Continuous Emotional States Monitoring by Using Convolutional Neural Networks." In *2019 International Conference on Control of Dynamical and Aerospace Systems (XPOTRON)*, pp. 1-4. IEEE, 2019. https://doi.org/10.1109/XPOTRON.2019.8705963