International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

CNN Implementing Transfer Learning for Facial Emotion Recognition

Palasatti Srinivasa Reddi[1], Dr. A. Sri krishna[2]

Submitted: 01/11/2022 **Accepted:** 04/02/2023

Abstract: The study's primary objective is the development of a framework that operates in real time and can determine the emotional state that most people in a group are experiencing on average. This research article suggests a basic method for recognising facial expressions by combining transfer learning with convolutional neural networks (CNNs) that have few parameters (TL). The suggested CNN architecture was jointly trained on the FER-2013, JAFFE, and CK+ datasets for real-time detection, which expanded the scope of what can be detected emotionally in terms of expressions of emotion. The model could notify when a person was happy, sad, surprised, scared, angry, disgusted, or neutral. Several methods were used to sort out how well the model worked. Experimental findings support the claim that the proposed approach is more effective than other research in terms of both precision and speed.

***Keywords**: CNN, FER-2013, CF+, JAFFE, Transfer Learning, Deep Learning*

1. Introduction

Expressions of emotion that can be read from a person's face are an essential component of successful two-way communication between individuals. Expressions are a potent way of communicating due to the fact that they divulge one's feelings, reflect one's level of sincerity, and provide context for what is being said. When it comes to automatic facial expression recognition (AFER), you will find that you are at the crossroads of a number of different domains, such as artificial intelligence, psychology, neurology, and behavioural research. It is common knowledge that advancements in computing have sped research across a variety of sectors, including artificial intelligence (AI) and pattern recognition. For there to be a genuine connection between human and machine, there must first be a cordial rapport between the two. However, sound and language transmit only 38% of information, while 55% is sent through facial expressions. You may learn a lot about a person's mood just by looking at their face. Despite a mountain of literature on the topic, there is not one single definition of an emotion that is accepted by everyone. [1] One of the many ways in which a feeling might make itself known is through the display of emotion. This is one of the ways. Emotions, in contrast to feelings, can be fabricated.

Despite the complexity of the structure, the majority of the work can still be done using the word "face" as an input. As a result, a significant amount of research has been focused on automating the encoding of facial expressions. It is feasible to recognise fundamental expressions, such as frontal faces and the emotions conveyed by side postures, in a controlled environment, which makes it possible to carry out the work. The study of emotion covers a vast, important, and sophisticated range of topics. "Neutral," "happy," "sad," "surprised," "angry," and "disgusted" are some of the most common emotions that emotion detection systems seek for and identify. "Fear" is another common emotion. Compound emotions, as opposed to universal emotions, are responsible for the existence of 21 distinct emotional states. When joy and surprise are combined, a compound feeling such as "happy surprise" is produced. This is because the bodily symptoms of surprise are combined with joy.

The ability of technology to read an individual's facial expressions and automatically calculate their emotional state has a number of potential applications. Some examples of emotionally and socially aware technologies are ones that can make games more fun, find drivers who are too tired to drive, and tell when a patient is in pain or suffering.[2] Other examples include devices that help improve the identification of drowsy drivers. Up until this point, the majority of the research that was done on FER was focused on improving the feature extraction step. The features that were acquired using one of the several approaches that can broadly be classified as appearance, geometric, or hybrid feature extractors are then delivered to various classifiers so that they can make their determinations. On the other hand, traditional methods are computationally challenging, making it difficult to attain

[1]*Research Scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University College of Engineering, Guntur, AP State, India, Email:psrinvasareddi@gmail.com*
[2]*Professor & HoD, Department of Information Technology, RVR & JC College of Engineering, AP State, India, Email: ask@rvrjc.ac.in*

high recognition rates. Both the action coding system and the continuous technique have some kinks to work out before they can adequately describe emotions in the real world. The effect model, on the other hand, disregards the nuanced and delicate character of our successful displays. The intricacy of the situation is going to shift, and that includes things like head posture, illumination, and occlusion.

Because people have a variety of ways of expressing themselves, the same image can have multiple meanings depending on factors such as the brightness of the backdrop, the colour of the backdrop, the position of the picture, and many other features. The fact that genuine, unposed expressions are often downplayed is another factor that contributes to how the image is understood. As a result, a dependable automated FER in a variety of contexts is critical. The process of automatically recognising facial expressions has been attempted and tested using a variety of different methods. A significant number of the earlier works made use of geometric representations. These included vector descriptors for facial motion, active contours for retrieving the shape of the lips and eyes, and 2D changeable mesh models. Others turned to methods that were based on the representation of an appearance, such as the use of Gabor filters and local binary patterns. [3] These feature extraction techniques were often linked with one of several repressors in order to transfer these feature vectors to emotion classification or action unit identification. This is done to ensure that the results are as precise as possible. The most commonly used repressors in this study were support vector machines, also known as SVMs, and random forests.

Convolutional neural networks are multi-layered, artificial neural networks that replicate the structure of the human brain in an effort to solve complex problems. These artificial neurons increase the weight of a picture by adding a bias and an activation function in order to generate an output from the image. Image categorization, recognition, and segmentation are all examples of potential applications for artificial neurons, which can also be used to do basic convolutions. [4] By feeding additional information into the convolutional neural network, a deep learning model can be created that is more robust and accurate. One of these methods is called deep learning-based facial expression identification, and it is able to distinguish between a variety of human feelings (including anger, fear, neutral, pleased, disgust, sadness, and surprise) based on a single photograph.

This strategy seeks to achieve the objective of automatically recognising facial expressions in order to accurately characterise a person's emotional state. During the training process for a CNN, tagged facial photos from a dataset of facial expressions are used. The proposed CNN model will then determine which phrase is supposed to be carried out after this step. For the FER test, previously created CNN-based models have demonstrated a level of recognition accuracy that is satisfactory. The model still has high computational costs, and a significant amount of memory is needed to run it. Because of these limits, they cannot be implemented in an environment with limited resources. [5] A lightweight CNN is one that requires little in the way of memory and computing power, allowing it to be deployed in real time.

Transfer learning (TL) from FER-2013 was utilised to train the merged dataset for real-time detection. This was done since TL improves accuracy when applied to the CK dataset. In addition, time has been saved because TL does away with the necessity of performing initial training. In a nutshell, the following were mostly included as a result of this research:

(a) Creating a CNN structure that is more straightforward in order to promote faster detection

(b) Accelerating the detection process through the application of transfer learning while preserving the integrity of the final product.

(c) Carrying out real-time detection using a comprehensive dataset that was gathered from a variety of sources, including natural environments as well as laboratory situations.

(d) The creation of a real-time facial expression identification application that has a lower runtime cost than other works in the existing body of research in the literature.

(e) We place an emphasis on the empirical differences that exist between various convolutional neural network algorithms in contexts that are consistent.

f) We use CNN visualisation methods in order to improve our understanding of the model that was created by applying the most recent and cutting-edge techniques in FER to a variety of datasets.

g) We show that the emotion-detection network can generalise across a wide range of datasets and FER.

2. Related Works

One of these technologies is known as deep learning-based facial expression identification. It has demonstrated outstanding performance in image classification. As a result, it can be used to distinguish human emotional states such as anger, fear, neutrality, happiness, disgust, sadness, and surprise.

This technique makes an effort toward automatic recognition of facial emotions in order to determine an individual's emotional state with more precision. In this method, a CNN is trained using a dataset of tagged facial images that are collected from a dataset of facial expressions. The dataset of facial expressions contains the facial images. The CNN model that has been proposed then determines which expression was meant to be used. The CNN method is frequently used for this particular function. The work that Darwin did in 1872 laid the groundwork for the field of study known as "facial expression analysis." Since that time, the numerous applications that may be found for FER have been the driving force behind an ever-increasing number of research initiatives. An automatic facial expression recognition system could be useful in many different areas, including human behaviour analysis, nonverbal communication, and healthcare applications. These are just a few of the many fields that could potentially reap the benefits of such a system. Even though it has been around for a few decades, the area of emotion recognition is still very much active in modern times. This is due, in large part, to the numerous useful applications that the field has. Recently, researchers have been focusing their efforts on developing an algorithm that can artificially recognise human emotions. Even though the findings have improved, the vast bulk of the research that has been done on static photographs has not been successful in resolving the concerns of illumination, position change, and masked facial expression.

Face detection, feature extraction, and emotion recognition are the three fundamental stages that comprise the conventional automatic FER method. FER stands for face emotion recognition. The Haar-cascade classifier is a method that is frequently used to locate the important parts of the face.

The processes that are based on the look and geometric features of the face are the ones that are employed most frequently for feature extraction. The intensity, gradient, and textural variation of the specified face region are all retrieved using algorithms that are based on appearance.

On the other hand, geometric feature extraction is founded on the anatomy that lies beneath the surface of the face. It is helpful to compute the distances between facial spots in order to recognise emotions, for example. These spots are located on the face. In addition, the hybrid method incorporates visual as well as geometric strategies for the purpose of feature extraction. It is important to feed the collected handcrafted features into conventional classifiers in order to be able to identify emotions.

Zhang et al. [6] developed an extremely efficient system for the categorization of halftone images and the processing of images, which can be used to analyse important aspects of movies and still photos. The characteristics of grayscale images were extracted using unsupervised learning and stacked sparse autoencoders (SAE). The Fer2013 and CK+ datasets were utilised in the training of a CNN model that was based on ResNet. This model was then used to extract features from the data. Softmax, linear SVM, and random forest were among the classifiers used, in addition to the proposed complexity perception classification algorithm (CPC). When CNN+Softmax is used with CPC, recognition rates for FER2013 come in at 71.35 percent, while recognition rates for CK+ come in at 98.78 percent. Davidson et al. [7] recognised micro-expressions using a histogram-oriented gradient. Lekdioui et al. [8] proposed a texture and shape descriptor-based method to detect facial emotion. Good results from CNN can be obtained, according to Khorramiet al. [9], if CNN is trained to analyse a face and identify the characteristics that influence CNN's predictions. This is the key to achieving good results from CNN. In a similar vein, Happy et al. [10] used a salient facial patch extraction method based on landmark points and LDA as a feature. To solve the problem of picture multi classification, Wei et al. developed a flexible hypothesis pooling technique. With this approach, the model can accept as input any number of hypotheses regarding the object segment. Each notion was ultimately linked to a single CNN through a series of connections. In conclusion, traditional predictors for multilabel predictions can be obtained by employing maximum pooling to take the results of the model's many hypotheses and averaging them. In applications that are connected to FER, some of the issues that can arise include alignment, the detection of faces, and the recognition of those faces. Nguyen et al. [11] came up with the idea for a model of a multi-level, 18-layer CNN that is comparable to VGG. This model incorporates information from the level below it in addition to the high-level elements that are already present in it. There has been an improvement in the accuracy of a fundamental CNN model to 69.21%, and a multi-level CNN has been shown. A unique method of optical flow known as MDMO was utilised by Liu et al. [12] To improve texture information extraction, an affine transformation was then utilised as a means of simplifying the lighting system and avoiding its complexities as well as the subject's head movement. The data from the facial ROIs were used to train an SVM classifier to recognise real emotions, and the resulting model achieved an accuracy of 73.03 percent on the FER dataset.

A portion of the research community came up with an autoencoder so they could do research on visually observed characteristics. Deep sparse autoencoders were built by Zeng et al. [13] in order to train appearance-based attributes for the purpose of face emotion recognition (DSAE). The deep learning model that Parka et al. present for face alignment and correction is based on landmark characteristics and repeated recognition. Chen et al. [14] demonstrate that they are capable, with the help of deep learning, of recognising genuine smiles in photographs.

Deep learning has the potential to quickly combine feature learning and classification into a single model, in contrast to previous research that collected handcrafted features from face photographs and trained a classifier to carry out grin recognition in a two-step procedure. The most challenging aspect of the geometric technique is undoubtedly the process of locating the feature point. Priya et al. [15] proposed a geometric strategy for isolating the eye and mouth regions by making use of feature points in their study. Following this, a quadrilateral mapping of the regions was carried out in order to provide input to a fuzzy membership algorithm for the purpose of accurate facial emotion classification. A system for tracking visual targets was created by Pang et al. [16], using the CNN DL algorithm with very high accuracy.

The hard field of video analysis known as "human activity detection" is currently the focus of a large number of ongoing research efforts. Ronao et al. [17] demonstrated a human activity detection system that is efficient and effective, and it is based on the sensors found in smartphones. The approach produced results that were considered state-of-the-art when applied to a variety of experimental databases. When applied to the Fer2013 dataset, the 13-layer CNN model that Christou et al. [18] suggested yielded an accuracy of 91.12% on the validation dataset. This result was obtained by using the model.

Sajjanhar et al. [19] made use of not only CK+ but also JAFFE and FACES. According to the findings, the VGG-19 model achieved the highest level of accuracy on the FACES dataset. It achieved this level of precision by performing at its best.

Chen et al. [20] proposed a two-stage process based on difference convolutional neural networks (DCNNs) trained on the CK+ and BU-4DFE datasets, respectively. On the CK+ dataset, it was found that the suggested model has an accuracy of 95.4%, whereas on the BU-4DFE dataset, it has an accuracy of 77.4%. Li et al. [21] applied the new automated micro-expression analysis methodology Flownet 2.0 [22] [30-35] to enhance the efficiency of a dual-template CNN model, despite the fact that the results were still mediocre in comparison to the more traditional methods. Next, Kumar et al. [33] eliminated expression frames that had a low intensity by utilising an approach that was based on the frequency domain. According to the findings of their research, low-intensity frames are those that have the least degree of variety in their textures. The presence of significant motion in the final set of high-intensity frames will serve to enhance the emotional snapshot that has been produced from those frames. The appropriate CNN model is then used to assign an emotion to each of these intense frames after they have all been processed. Since it was first proposed by He et al. [23], SPP has been successfully implemented in a wide variety of automated systems throughout the computer vision literature. These systems include those for semantic segmentation, anti-spoofing applications, expression analysis, and a great deal of other applications. Since its first suggestion by Chen et al. [24-29], ASPP has been effectively implemented in a variety of domains, including object recognition, picture segmentation, and image classification, to name just a few of these domains' applications. In a similar manner, CNN-based algorithms are becoming increasingly popular as a viable alternative to more conventional methods of feature extraction. Researchers have shown a growing interest in investigating this field as a result of CNN's innate capability for self-learning as well as the rapid development of deep learning models[23].

Our investigation has led us to the conclusion that, despite the significant progress that has been made in the FER problem, strategies that achieve acceptable precision with fewer computing costs are not currently viable for application in a real-world setting. This is the finding that we came to after conducting our research. Traditional methods require less time to recognise FER, but they are not very precise when taking into consideration FER in its natural environment. The high computational costs associated with most CNN-based approaches are due to their parameter-intensive nature. This is despite the fact that CNN-based approaches are effective. For instance, the VGG-16 design, which has been used in a variety of studies, needs around 138 M parameters, making it difficult to implement in an environment with restricted access to resources.

In addition, the suggested method combines images taken in the wild and in controlled lab environments to improve real-time detection. This is in contrast to previous research, which relied primarily on lab-controlled frontal faces for real-time analysis. Because of these limitations, a CNN that is more lightweight and uses fewer than 1.3 million parameters has been developed for use with FER in a real-world environment. In addition to this, the TL approach can be used to cut down on the total amount of time spent training while also improving accuracy.

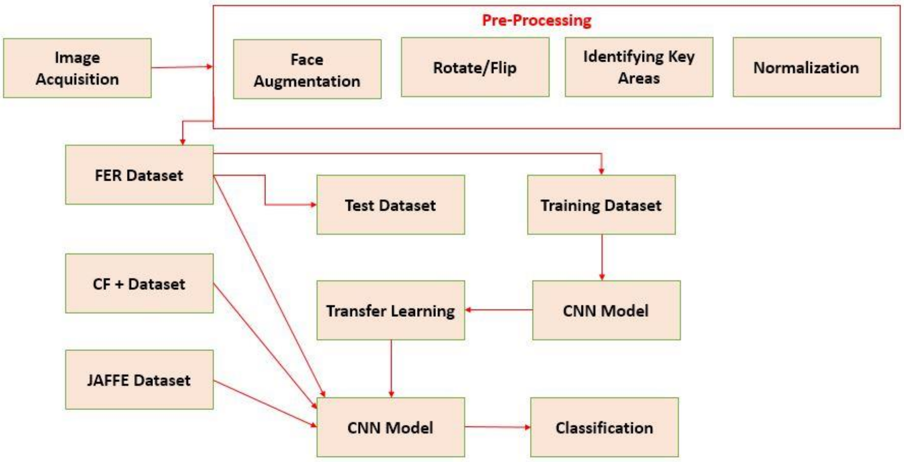

Fig. 1: Proposed System Architecture

3. Methodology

Figure 1 depicts a high-level overview of the suggested strategy. In this study, we propose an innovative framework for analysing expressions by using a CNN in a simulated setting. At the outset of our suggested model, we acquire a new raw image from multiple datasets (also called image acquisition) to rule out the possibility that our model is biased in favour of any one particular dataset. Face detection and resizing were used to isolate the target area, and then the resulting images were augmented to get them ready for training in facial emotion identification. First, the pictures from the combined dataset are trained using the training data. By applying the transfer learning method developed, we produce a trained model. Once training is complete, our recommended model starts analysing the chosen image to see if a face is present. If a face is found in the selected image using a cascade classifier, the image moves on to the next phase of pre-processing.

This stage involves image pre-processing, which entails numerous unique steps as shown in the proposed system architecture. For example, by using various methods and technology, we can improve the facial expressions that are identified. Such as cropping, rotating, flipping, and stretching the detected face. After that, normalisation and magnification methods are used to identify landmarks and register the specified facial emotions. The last step is to provide the model with more information for making accurate predictions.

3.1 Pre-processing

Whether or not pre-processing of images is necessary is determined by a number of variables, including the capturing instrument and the range of lighting conditions. Pre-processing the data usually improves the image quality before it is used. Resizing, histogram equalisation, noise reduction, and normalising are all examples of frequent pre-processing procedures. However, real-time detection capacity may be jeopardised by the additional runtime costs associated with substantial pre-processing operations. Therefore, the suggested method uses a minimal number of pre-processing steps without sacrificing precision. As part of the pre-processing phase, our work has used a two-step process: face detection followed by data augmentation.

As a result of its superior detection efficiency and precision, the Haar cascade classifier has been used for face detection in this case. Many objects can be detected in real-time in this case by adjusting the scale factor and minimum neighbours settings. Before feeding the colour image to the classifier, the programme converts it to grayscale, at which point the classifier outputs four coordinates that can be used to draw a rectangle around the face. In addition, in order to maintain uniformity Scaling the identified faces down to 48 by 48 pixels

In natural settings, things like lighting, sound levels, audience positions, and other factors tend to change a lot. background art that doesn't really relate to how someone is feeling. Before CNN was used to train the FER model, the visual semantic input had to be pre-processed to align and standardise the input. The following methods are included in the pre-processing step:

(1) The first stage of face recognition in computer vision is face detection, which identifies a face region in an image. Localization means defining the bounds of the face, whereas finding the face coordinates in the image is what this phrase is referring to. When it comes to facial recognition, the tried-and-true Viola-Jones (V&J) face detector is a common choice.

(2) Data augmentation is crucial in a deep learning-based FER system. However, in order to train the CNN model and ensure that it can be used to solve any recognition problem, a huge amount of data is required. Before an input image can be used in the machine learning process, it must be processed and cropped.

(3) One common preliminary method for the face recognition task is face registration. Facial registration is the process of adjusting a sample face so that it looks like a known good reference face.

(4) Four of the most prominent features of the face that serve as landmarks include the eyes, mouth, nose, and eyebrows. We locate the subject's head and neck in the image, and we make a mental note of any distinguishing aspects of their facial ROI.

(5) Variations in lighting and head position can have a significant impact on performance, leading to noticeable shifts in the resulting image quality. Therefore, we describe two standard approaches to normalising faces in an effort to reduce these differences: adjusting the brightness and the position of the head to be standard. There were three databases, each with roughly the same number of face images but different resolutions (and thus distinct facial emotions).

Consequently, the images from the Haar Cascade library were initially used to detect the facial circumference. After that, the identified facial emotions were clipped and recorded to the same size, making them all rectangles. Pictures' pixel values were also transformed to 48x48 grayscale images before being fed into neural networks. This was done to keep the neural networks as sparse as possible.

Data enhancement The performance of deep learning models can be improved with a larger dataset. So, most researchers now use augmentation to artificially expand the size of their dataset. The ImageDataGenerator class in Keras's preprocessing package has been used to add data in real time. This class makes a batch-wise tensor image.

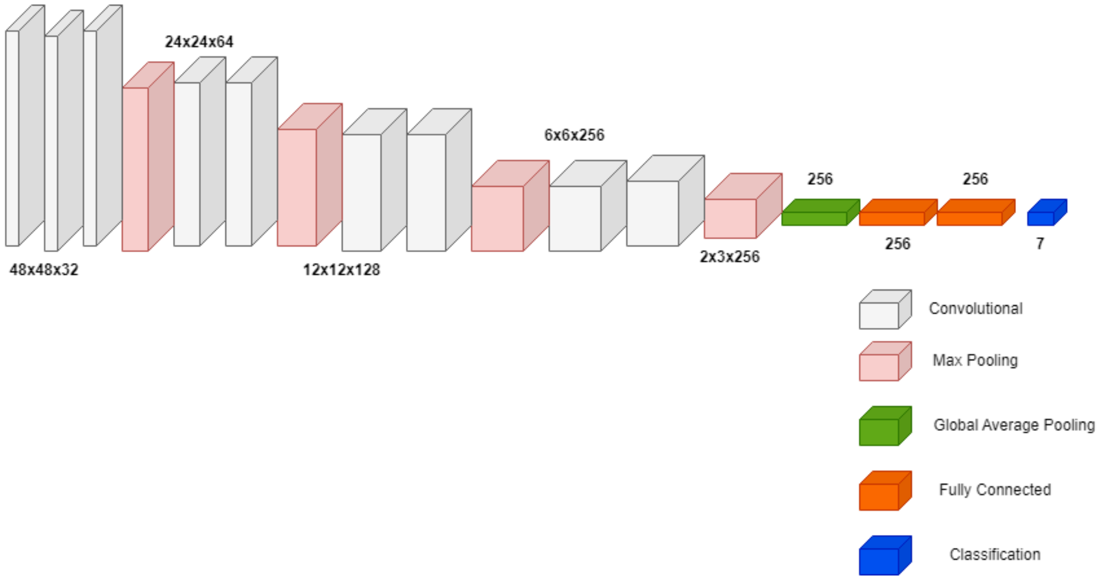

Fig. 2: Model Architecture

3.2 CNN architecture

The suggested CNN architecture aims to efficiently and quickly train the pixel values in the rectangle region that encompasses facial expressions in order to permit speedy responses with the deep artificial neural network model that was produced as a result of the research. Our motivation for developing CNN's architectural framework can be summed up as follows: To begin, the image size (48x48) of FER-2013 is significantly lower than the input size (224x224 or 299x299) of most deep learning models. When a picture is resized, it retains extra pixels that are never used, leading to duplicate information and stale feature learning. Second, if the input image is a colour image, Because FER-2013 only provides grayscale images, KA must perform an additional computation with redundant values to convert a greyscale image to a colour image. Finally, it is recommended that a CNN structure with fewer parameters be developed in order to reduce memory and computational time requirements. The input to a CNN is a 48x48 grayscale image, which is sent on to the convolution layer (CL).

The process starts with convolution layers, which are in charge of pulling out features from image patches by applying filters to them.

When presented with a 48x48 input image, the CL first generates 32 feature maps using 32 3x3 kernels. After that, 7 more CLs are convolved with 3 X 3 filters with stride 1 to extract 32, 64, 128, 256, and 256 features, respectively.

$$\dot{A}_j^i = m\left(\Sigma_{t=1}^{N-1} \cdot \dot{A}_i^{i-1} * w_{ij} \ + wt_b\right)$$
(1)

Here, the convolutional operation is given by the * operator. A_i represents the feature maps, and w is the filter. By employing the proper filters, CL is able to accurately capture the temporal and geographical details of interdependence present in an image. Due to its interaction impact and nonlinear features, the ReLU activation function always comes after the CL. If a negative value is passed into the function, it will return 0. However, that value is what you get back for any positive x. Input neuron x can be used in equation (2) to derive the value.

$$\mathrm{f(x)} = \max(0, \mathrm{x})$$
(2)

To condense the output feature maps from the CLs, the MPL, or Max Pooling Layer, is a layer of 2x2 filters with stride 1, which is placed after every pair of CLs. By down sampling the feature maps, MPL is able to get rid of superfluous data. The formula for determining MPL is:

$$\dot{\mathrm{A}}_{\mathrm{j}}^{\mathrm{i}} = \mathrm{F}(\mathrm{MP}\dot{\mathrm{A}}_{\mathrm{i}}^{\mathrm{i}-1} + \mathrm{w}_{\mathrm{b}})$$
(3)

Through the process of down-sampling, dimensions can be reduced, which then enables inferences to be made about features that are included inside binned sub-regions. In order for the system to learn, it must first produce a representation of the data and then get rid of any parts of that representation that do not overlap. It aids overfitting by giving a simplified representation to work with. It delivers the addition of basic translation invariance to the explicit representation, which reduces the computation cost by reducing the number of parameters to learn.

To train our model, we feed it a 48 x 48 grayscale image and apply several regularisation and optimization techniques to it. From a set of seven feelings, only one is displayed in the final output class. The CNN architecture includes four convolutional layers and four MaxPooling layers.

3.3 Network Training

The test size was set at 25% in the training phase of the network. In order to get the network parameters to converge, a batch size of 32 and an epoch number of 500 were chosen. Definition of the learning rate: 10–3. Every convolutional and max-pooling layer's kernel size is set to 2x2, with a stride of 2.

3.3.3.1 Hyper parameters

Our proposed model makes use of a number of hyper parameters to examine the facial expression database. We briefly discussed our training method instead of delving into the model's efficiency in more depth across different databases. Our model was trained using all of the data from our experiment; despite this, we tried our best to standardise the architecture and hyper parameters. There were fifty total training epochs for each model. The initial network weights are generated using a zero- and standard-deviation-based random Gaussian. To avoid overfitting, a regularisation or shrining approach was used by setting the coe2cients to 0 (a dropout value of 0.1 to 0.4 was used). After experimenting with various permutations, Softmax activation is used to select the dense layer's multiclass cation.

3.4 Real-Time Testing

The suggested CNN architecture was trained, and then the trained model was subjected to real-world testing. To begin, the Haar Cascade library could identify human faces in as few as 30 frames per second from a computer's camera. The photos that had been discovered were then forwarded to the model so that questions about the categories to which they belonged could be asked. The predictions resulted in a secondary screen displaying the facial expression's likelihood of belonging to a particular class, with the higher-class emotion being a frame of the Haar cascade that has been overwritten. This operation was carried out on 30 consecutive frames of the live camera feed every second.

3.5 Transfer learning

In deep learning, transfer learning (TL) is a common technique in which the weights from one trained model are used to commence training for a subsequent task. In this type of learning, the goal is to improve performance on a variety of problems. Method quickens the training while also increasing the model's overall effectiveness. Recently, it has been put to use in fields outside of FER, such as the identification of languages, the detection of cancer, the diagnosis of diseases, and the investigation of the causes of Parkinson's disease. TL is also helpful because the majority of the currently available datasets are too small for CNN-based FER to function effectively. While millions of images would be ideal for training CNNs, the vast majority of FER datasets only contain hundreds or thousands of images. Because there is so little data available, CNN-based model training is getting more and more challenging. In certain circumstances, there is a risk of overfitting.

By utilising the inductive TL approach, this problem of having insufficiently large data sets has been resolved. Weights for inductive TL gained during training on a large dataset are initially stored in the model. After that, the model applies those weights, in the form of a pre-trained weight, to a smaller dataset. When compared to FER-2013, both CK+ and JAFFE have a significantly lower total number of images in their collections. Because of this, the training for the CK+ and JAFFE datasets has been done using the pre-trained model that is saved in the FER-2013 database.

3.6 Datasets

For the purpose of evaluation, the two standard datasets that are used most frequently, CK+ and JAFFE, have been selected. Another dataset, FER-2013, was chosen for this research because it is the most comprehensive, can be used in a variety of contexts, and is freely accessible. Expressions found in FER2013 were collected from an uncontrolled environment, whereas those found in the CK+ and JAFFE All data was collected in a carefully monitored laboratory. Setting. FER-2013: The ICML 2013 Workshop [45] on Representation Learning is hosting three challenges, one of which is the building of a dataset. It collected 35,887 photos to represent the seven different emotional levels by using the image search API that Google provides. It is regarded as one of the most useful datasets for the purpose of instructing deep learning models. The number 48948 refers to the size of the grayscale photos.

The CK+ dataset was made available to researchers in 2010. A greater number of participants and sequences have been included in this expanded version of the CK dataset. It is made up of the facial expressions of 213 individuals who displayed feelings ranging from neutral to extreme. The dataset contains photos of six different universal expressions of emotion, including contempt. The images included in a dataset are categorised by code using FACS.

Jaffe: There are 213 black-and-white images in the archive, depicting six different feelings that are experienced by everyone. The posed black-and-white photographs have a resolution of 2569 x 256 pixels and were done with 10 female Japanese models.

4.0 Experimental Results

This part describes the overall scope of the offline and online experiments. This section presents the evaluation findings of our model as well as the outcomes of experiments conducted on two significant facial expression recognition datasets. Following a brief introduction to database issues and difficulties, the FER model's performance is analysed across a range of hyper parameters applied to two real-world datasets (CK+ and JAFFE). Table 1 gives the number of records selected from the various datasets for the study

To address issues such as changing light, obscured views, a perspective other than the front of the head position identification bias, and low-intensity expression recognition, many researchers are turning to deep learning equipment as the FER study shifts its focus to tough environmental situations. To accurately capture subtle shifts in expression through deep learning, a sizable dataset for training purposes is essential. The lack of sufficient and high-quality training data presents the biggest obstacle for deep FER systems.

In order to test and train, we have divided the dataset into 70% and 30%, respectively. The procedure for preparing the data is described in full in the approach. Only training data undergoes the augmentation procedure; validation and testing data are normalised instead. In order to conduct a fair test on the CK+, FER, and JAFFE datasets, the five-fold and ten-fold cross-validation methods were implemented. The FER-2013 dataset was used to train the proposed model. Due to the varied nature of the FER-2013 dataset, constructing a compact CNN architecture with few pre-processing steps is one of the most difficult tasks. However, the enhancement procedure contributes to a 2.51% boost in precision. Multiple state-of-the-art CNN-based models are surpassed by the suggested technique. Due to its small size, the JAFFE dataset is evaluated using the ten-fold cross-validation method to ensure correctness. Using the augmented accuracy of 93.92% on the JAFFE dataset and the pre-trained weight on FER-2013, The proposed method, on the other hand, was able to reach 97.66% accuracy using the identical experimental setup but without the use of any augmentation. The range of emotions represented in the CK+ dataset, from neutral to extreme, plus the addition of a contempt class, forces researchers to experiment with new approaches to evaluation. In the research, the first frame serves as a control, while the sixth through twelfth peak frames are used to examine the emotions themselves. And to counteract any bias introduced by a non-optimal training-test split (whether chosen at random or by the user), it employs the tried-and-true five-fold cross-validation method. Combining two sets of data allows for a larger dataset as well as the inclusion of both natural (FER-2013) and laboratory-controlled photos. The pre-trained weight from the combined dataset is then used in the real-time application. Transfer learning from FER-2013 to the new dataset (FER-2013 plus CK+) is 71.45% accurate. When compared to other state-of-the-art approaches, the proposed method performs better when it comes to augmentation, CNN structure, and TL. Since FER-2013's photos were acquired in the wild, the augmentation procedure boosted accuracy by about 4% and provided greater flexibility during model training. table 2, table 3 and table 4 give the comparison of the proposed model with the existing literatures

Table 1. Number of records for the study dataset.

Emotion	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

Table 2: FER Dataset result comparison

Methodology	Anger %	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: Performance comparison with literature on the JAFFE dataset

Approach	Anger %	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: Performance comparison with literature on the CF+ dataset

Methodology	Anger %	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 muscle activity, which can be seen in black and white photographs, is the primary basis for FER. The suggested method incorporates TL to boost training velocity and precision. When we start with raw data to train, adjusting the weights to lower the error rate is more labour intensive. To ensure the highest degree of precision in practical implementations, TL is employed to refine the final model.

Selecting an appropriate optimization method is critical for improved precision as it aids in locating the optimal input parameter by lowering the error rate.

5. Conclusion

The goal of the research is to develop practical real time framework that is capable of recognising the range of emotional expressions displayed by a group of people based on a single frame of video. The proposed method not only achieves a high level of accuracy when applied to FER-2013, but it also surpasses the vast majority of state-of-the-art algorithms when applied to the JAFFE and CK+ datasets. The software is superior to typical classifiers and even work based on CNN in its ability to recognise a broad range of emotional states with significantly less computational effort. The CNN-based algorithms eliminate the time-consuming and difficult requirement for manual labour in the process of extracting complicated features. They do not require large pre-processing modules as other machine learning approaches do, yet they nevertheless produce cutting-edge results, hence they are superior to those techniques. While there are benefits to the proposed approach, there are also some potential downsides. be improved upon by further investigation. The identification of facial expressions that are extremely non frontal and occluded is a big difficulty for the research methodologies that are now in uses well as the strategy which is being offered. So that we can solve the aforementioned dilemma in the future, a more comprehensive dataset that takes into account dire circumstances is required. Additionally, additional work needs to be done on the face detection system so that it can recognise people's faces even when they are in odd scenarios. This would improve the system's ability to recognise faces in real time. Even though the method that is being offered uses still photographs for FER, it is possible that in the future, moving ones may be explored in order to make better use of time-domain data.

Conflicts of interest

The authors declare no conflicts of interest.

Reference

[1] R. Breuer and R. Kimmel, “A Deep Learning Perspective on the Origin of Facial Expressions.” arXiv :1705.01842.

[2] Kandhro, “Impact of Activation, Optimization, and Regularization Methods on the Facial Expression Model Using CNN”, Computational Intelligence and Neuroscience, vol. 2022, pp. 1–9, Jun. 2022, doi: 10.1155/2022/3098604.

[3] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, “Real Time Emotion Recognition from Facial Expressions Using CNN Architecture,” in 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, Oct. 2019, pp. 1–4. doi: 10.1109/TIPTEKNO.2019.8895215.

[4] T. Podder, D. Bhattacharya, and A. Majumdar, “Time efficient real time facial expression recognition with CNN and transfer learning,” Sādhanā, vol. 47, no. 3, p. 177, June. 2022, doi: 10.1007/s12046-022-01943-x.

[5] Banala, R., Nair, V., Nagaraj, P. (2022). Performance of Secure Data Deduplication Framework in Cloud Services. In: Kumar, A., Fister Jr., I., Gupta, P.K., Debayle, J., Zhang, Z.J., Usman, M. (eds) Artificial Intelligence and Data Science. ICAIDS 2021. Communications in Computer and Information Science, vol 1673. Springer, Cham. https://doi.org/10.1007/978-3-031-21385-4_32

[6] Zhang K, Zhang Z, Li Z and Qiao Y 2016 Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters. 23(10): 1499–1503

[7] Fox A S, Lapate R C, Shackman A J and Davidson R J 2018 The nature of emotion: Fundamental questions. eds. Oxford University Press

[8] Lekdioui K, Messoussi R, Ruichek Y, Chaabi Y and Touahni R 2017 Facial decomposition for expression recognition using texture/shape descriptors and SVM classifier. Signal Processing: Image Communication. 58: 300–312

[9] P. Khorrami, T. Paine, and T. Huang, ‘Do deep neural networks learn facial action units when doing expression recognition’ in Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 19–27, Santiago, Chile, December 2015.

[10] Happy S L and Routray A 2014 Automatic facial expression recognition using features of salient facial patches. IEEE Transactions on Affective Computing. 6(1): 1–2

[11] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I. Na, and S. H.Kim, "Facial Emotion Recognition Using an Ensemble of MultiLevel Convolutional Neural Networks," International Journal of Pattern Recognition and Artificial Intelligence, 2018.

[12] Liu P, Han S, Meng Z and Tong Y 2014 Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 1805-1812

[13] Zeng N, Zhang H, Song B, Liu W, Li Y and Dobaie A M 2018 Facial expression recognition via learning deep sparse autoencoders. Neurocomputing. 273: 643–649

[14] J. Chen, Q. Ou, Z. Chi, and H. Fu, "Smile detection in the wild with deep convolutional neural networks," Machine Vision and Applications, vol. 28, no. 1-2, pp. 173–183, 2017.

[15] Priya R V 2019 Emotion recognition from geometric fuzzy membership functions. Multimedia Tools and Applications. 78(13): 17847–17878

[16] S. Pang, J. J. del Coz, Z. Yu, O. Luaces, and J. D´ıez, "Deep learning to frame objects for visual target tracking," Engineering Applications of Artificial Intelligence, vol. 65, pp. 406–420, 2017.

[17] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," Expert Systems with Applications, vol. 59, pp. 235–244, 2016.

[18] N. Christou and N. Kanojiya, "Human Facial Expression Recognition with Convolution Neural Networks," Singapore, 2019,pp. 539-545: Springer Singapore.

[19] Sajjanhar, Z. Wu, and Q. Wen, "Deep learning models for facial expression recognition," in 2018 Digital Image Computing:Techniques and Applications (DICTA), 2018, pp. 1-6: IEEE.

[20] J. Chen, Y. Lv, R. Xu, and C. Xu, "Automatic social signal analysis: Facial expression recognition using difference convolution neural network," Journal of Parallel and Distributed Computing, vol. 131, pp. 97-102, 2019.

[21] B. E. Manjunath Swamy. "Personalized Ranking Mechanism Using Yandex Dataset on Machine Learning Approaches." Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1. Singapore: Springer Nature Singapore, 2022.

[22] Burada, Sreedhar,"Computer-Aided Diagnosis Mechanism for Melanoma Skin Cancer Detection Using Radial Basis Function Network." Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1. Singapore: Springer Nature Singapore, 2022.

[23] Kumar, M. S, et al. "Deep Convolution Neural Network Based solution for Detecting Plant Diseases." Journal of Pharmaceutical Negative Results (2022): 464-471.

[24] Prasad, Tvs Gowtham, et al. "Cnn Based Pathway Control To Prevent Covid Spread Using Face Mask And Body Temperature Detection." Journal of Pharmaceutical Negative Results (2022): 1374-1381.

[25] Venkateswara R, L., et al. "Bio-Inspired Firefly Algorithm for Polygonal Approximation on Various Shapes." Intelligent Computing and Applications: Proceedings of ICDIC 2020. Singapore: Springer Nature Singapore, 2022. 95-107.

[26] Kumar, A., Kumar, M. S., Vikhar, P. A., Ghodke, V., Waghulde, R. R., & Rathod, M. (2022). Improving the Software Privacy in the OFDM 5G Communication Integrated with License Key in the Hardware Communication Parameters. International Journal of Intelligent Systems and Applications in Engineering, 10(2s), 236-240.

[27] Natarajan, V. A., Kumar, M. S., Tamizhazhagan, V., & Chevdumoi, R. M. (2022). Prediction Of Soil Ph From Remote Sensing Data Using Gradient Boosted Regression Analysis. Journal of Pharmaceutical Negative Results, 29-36.

[28] Kumar, M. S, et al. "Deep Convolution Neural Network Based solution for Detecting Plant Diseases." Journal of Pharmaceutical Negative Results (2022): 464-471.

[29] P. Sai Kiran,"Resource aware virtual machine placement in IaaS cloud using bio-inspired firefly algorithm." Journal of Green Engineering 10 (2020): 9315-9327.

[30] Balaji, K., P. Sai Kiran. "Power aware virtual machine placement in IaaS cloud using discrete firefly algorithm." Applied Nanoscience (2022): 1-9.

[31] Kumar, M. S, et al. "Applying The Modular Encryption Standard To Mobile Cloud Computing To Improve The Safety Of Health Data." Journal of Pharmaceutical Negative Results (2022): 1911-1917.

[32] S. Li and W. Deng, "Deep facial expression recognition: a survey," IEEE Transactions on Affective Computing, vol. 13, p. 1, 2020.

[33] Jain N, Kumar S, Kumar A, Shamsolmoali P and Zareapoor M 2018 Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters. 115: 101–106

[34] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 770-778

[35] Lin M, Chen Q and Yan S 2013 Network in network. arXiv preprint. arXiv:1312.4400