Journal of Computer Science

Original Research Paper

PM-ViT a Framework for the Recognition of Emotions and Proclivity toward Mental Illness Using Facial Expressions

[1,2]Priti Rai Jain, [1]Syed Mohammad Khurshaid Quadri and [1,2]Anuradha Khattar

[1]*Department of Computer Science, Jamia Millia Islamia, India*
[2]*Department of Computer Science, Miranda House, University of Delhi, India*

Article history
Received: 20-09-2024
Revised: 18-11-2024
Accepted: 09-12-2024

Corresponding Author:
Priti Rai Jain
Department of Computer Science, Jamia Millia Islamia, India
Email: pritirai.jain@mirandahosue.ac.in

Abstract: Automated emotion recognition is being used as a powerful technology in various fields in the present times. Facial Emotion Recognition (FER) aims to identify the emotional states of individuals based on their facial expressions. While in recent years, Convolutional Neural Networks (CNNs) have shown noteworthy performance in image classification tasks, however, the latest adoption of transformers for computer vision tasks has become really influential. This study proposes a novel ViT-based framework PM-ViT to explore the performance of Visual Transformer (ViT) based models on emotion recognition tasks and compare its performance with a CNN-based approach and other existing ViT-based models that recognize emotions from images. The proposed model PM-ViT takes facial images as input. It recognizes the expression and does a binary classification into two classes negative emotions and positive emotions. In addition to emotion recognition, in case the emotions found are negative PM-ViT does a further classification in three classes mild, moderate, and severe basis the perceived strength of negative emotion and hence the proclivity that the person may be having a mental illness. The experimental findings demonstrate that the model using CNN achieves an F1-score of 81.0% on AffectNet and 97.8% on the CK+ augmented dataset whereas the proposed PM-ViT achieves an F1-score of 84.0% on AffectNet and 99.7% on CK+ augmented dataset. The performance of PM-ViT surpasses the performance of the existing ViT-based techniques that determine emotions from images.

Keywords: Computer Vision, Vision Transformers, Emotion Detection, Affectnet, CK+

Introduction

Recognizing emotions helps a human being to understand the intents and psychological states of other human beings. Analysis of facial expressions is a prominent research area for emotion recognition (Lili *et al*., 2022; Pennycook, 1985; Rai Jain *et al*., 2021).

Depression is known to affect a wide range of nonverbal behaviors. Studies claim that the face conveys a lot of nonverbal cues; the face could be a valuable tool for diagnosing depression (Jiang *et al*., 2021). Facial nonverbal behavior is critical for assessing depression severity (Ellgring, 2007; Niu *et al*., 2021). Expressions of anger were rated more strongly by individuals with Major Depressive Disorder (MDD) (Branco *et al*., 2018).

Further, there is a substantial shortage of mental health experts, including psychologists, psychiatrists, physicians, and nurses. India has only 0.75 psychiatrists per hundred thousand people, compared to approximately 6 psychiatrists per hundred thousand people in nations with greater per capita incomes (Garg and Glick, 2018; Jain and Quadri, 2021). Automated facial expression recognition has potential applications in many areas, more so in the aforesaid scenario. One could extract crucial information from real-time facial expressions using a computer. The system could be programmed to raise an alarm when a person displays a negative facial expression, such as that of fear, disgust, anxiety, etc. This information could assist medical personnel to monitor a patient's mental health.

The current work proposes a visual transformer-based model that takes facial images as input. It recognizes the expression and does a binary classification into two classes negative emotions and positive emotions. When the emotions found are negative it is further classified into three classes mild, moderate, and severe, based on the

perceived strength of negative emotion and hence the chance of the person having a mental illness.

The proposed model has the potential to supplement existing mental health assessments and enable regular, low-cost, automated use in situations when a specialist is not easily accessible or the patient is incapable or unwilling to participate. A similar framework could also be used to identify other psychiatric conditions. Studies have found links between psychological and somatic problems using facial expressions (Simcock *et al*., 2020). This research may contribute to the creation of fresh approaches for the early identification or diagnosis of mental illnesses like anxiety and depression through the use of facial emotion.

Recent studies in facial emotion detection point out that visual transformers demonstrate superior performance in both accuracy and efficiency in modern-day image categorization when compared to CNNs (Chaudhari *et al*., 2022; Bobojanov *et al*., 2023; Zakieldin *et al*., 2024). A ViT effectively retains the global image features, which allows it to classify natural images with high precision (Kim *et al*., 2022).

The major objectives of the present study are:

- OBJ1: To understand and describe the working of a ViT
- OBJ2: To develop a fine-tuned framework using ViT architecture for emotion recognition. These models take facial images as input and recognize the expression and its perceived strength in order to classify the image at two levels. At the first level, the model does a binary classification and categorizes facial images into positive emotions and negative emotions. Thereafter, when the emotions found are negative, the model does a 3-way classification into mild, moderate, or severe depending on the strength of the negative emotions.
- OBJ3: To compare the efficiency of CNN based model (henceforth referred to as Model 1) against the proposed ViT-based model (henceforth referred to as PM-ViT Model)

The Primary Contributions of the Current Work Are

The current study proposes a novel model PM-ViT that takes facial images as input. It recognizes the expression and does a binary classification into two classes negative emotions and positive emotions.

In case the emotions found in the image are negative PM-ViT does a further classification into three classes mild, moderate, and severe based on the perceived strength of negative emotion and hence the proclivity that the person may be having a mental illness.

The experimental findings prove the superiority of the proposed model PM-ViT over the CNN-based model on images of both CK+ (Lucey *et al*., 2010) and AffectNet (Mollahosseini *et al*., 2019) datasets. It also demonstrates the superior performance of the PM-ViT over the existing state-of-the-art (SOTA) ViT-based models that recognize emotions from images.

Related Work

Most of the experiments done in this area use three types of methods: (A) machine learning-based techniques like 'Support Vector Machines (SVM), 'Principal Component Analysis (PCA), 'K-Nearest Neighbors (KNN), and ensemble methods, etc., (B) CNN based models and (C) ViT based models. A few related studies based on these techniques are discussed in the following subsections.

Machine Learning Based Models

Several studies such as (Al Jazaery and Guo, 2021; He *et al*., 2019; Wen *et al*., 2015; Zhu *et al*., 2018) have used machine learning methodologies to study the relationship between facial nonverbal behavior and Beck Depression Inventory-II (BDI-II) scores (Beck *et al*., 1996).

The researchers in Lili *et al*. (2022) used information gathered from facial expressions to develop and test classifiers that can determine if a user is 'depressed or not depressed'. The user's face will be photographed in order to extract the user's facial traits using Gabor filters, which are then used to forecast depression. These facial traits were classified using PCA and cascade classifiers. The number of negative feelings in the collected image was used to calculate the level of depression. The result was an F-Score of 76.7%. It is claimed to be moderate due to a low number of samples and features.

The model by Georgescu *et al.* (2019) uses a KNN algorithm for local learning and a ‘Support Vector Machine’ (SVM). The SVM classifier predicts the class label. The experiment achieves an accuracy of 63.31% on the AffectNet 7-emotions.

CNN Based Models

A CNN-based model EM-UDA by Jain *et al*. (2024) that categorizes facial expressions to interpret human feelings uses the technique of unsupervised domain adaptation. Th model achieves an accuracy of 83.9% when AffectNet is used in the source domain and CK+ in the target domain. It attains an accuracy of 74.55% FER 2013 images are used in the target domain.

The local ‘Sliding Window Attention Network’ (SWA-Net) model proposed in Qiu *et al*. (2023) for facial emotion recognition, uses feature-level cropping to avoid complex preprocessing and to preserve the integrity of local features at the same time. SWA-Net achieves an accuracy of 90.03% on RAF-DB, 89.22% on FER+, and 63.97% on AffectNet.

Researchers Liu *et al*. (2022) proposed a multi-modal DCNN to evaluate in real-time the severity of depression based on the facial expressions and body movements of the patient. In order to evaluate the severity of depression experienced by MDD patients, the researchers developed a metric, the "Behavioral Depression Degree" (BDD). BDD combines the action and expression entropies. The results show a Pearson similarity of 74% between BDD and "Self-rating Depression Scale", "Self-rating Anxiety Scale" and “Hamilton Depression Scale”.

The study by Jiang *et al*. (2021) proposed a 'Deep Neural Network' (DNN) using three hundred and sixty-five video-based interviews (total of 88 h) from a group of twelve depression patients both pre and post "Deep Brain Stimulation" (DBS) treatment. A Regional CNN detector and an ImageNet pre-trained CNN, both pre-trained on a very large public dataset, were used to extract 7 basic emotions. The OpenFace toolkit was used to extract facial activity units. Using leave-one-subject-out cross-validation, an Area Under the Curve of 0.72 was obtained for the categorization of remission and 0.75 for response to treatment.

The study (Huang *et al*., 2019) uses an analysis of the individual's Instagram posts to predict the likelihood of depression. In order to predict depressive propensity, it combined and assessed three features text, visuals, and behavior using a DL classifier. Using a five-layered CNN, the suggested model forecasts these users' propensity for depression with an F1 score of 82.3%. According to the study, the suggested model is more robust than models that just use text or visuals because it incorporates both.

A multi-modal DCNN was developed by researchers Liu *et al*. (2022) to assess the degree of depression in real time based on the patient's body language and facial expressions. The researchers created a measure called the "Behavioural Depression Degree" (BDD) to assess the degree of depression that MDD patients endure. The action and expression entropies are combined in BDD. According to the findings, BDD and the "Self-rating Depression Scale," "Self-rating Anxiety Scale," and "Hamilton Depression Scale" have a 74% Pearson similarity.

Visual Transformer Based Models

Recent studies indicate that ViTs demonstrate superior performance in modern-day image categorization techniques in contrast to CNNs (Chaudhari *et al*., 2022; Bobojanov *et al*., 2023; Zakieldin *et al*., 2024; Raghu *et al*., 2021; Bousaid *et al*., 2022). A ViT effectively retains the global image features, which allows it to classify natural images with high precision (Kim *et al*., 2022). The Squeeze ViT model in Kim *et al*. (2022), combines local and global features to improve FER performance while lowering computing complexity through a reduction in the number of features. A CNN is used to process an input image to extract the global and local features as landmark tokens and visual tokens. The squeeze module preserves strong discriminative aspects while modifying the feature dimensions. The Squeeze ViT receives concatenated tokens that are fed into its several encoders and squeeze module stacks. The model achieved an accuracy of 99.54 for CK+.

A Mask Vision Transformer (MVT) proposed in [29] comprises 2 modules: (i) A dynamic relabeling module to correct the wrong labels in facial emotion recognition datasets in the wild and (ii) Mask Generation Network (MGN), built using transformers, to generate a mask that can remove the backgrounds and occlusion of facial images. Results show that the MVT obtains an accuracy of 88.62% with RAF-DB, 89.22% with FER Plus, and 64.57% with AffectNet-7 respectively.

The study (Chaudhari *et al*., 2022) uses a fine-tuned ViT and applies it to an amalgamation dataset of 'FER2013, AffectNet, and CK+48 (AVFER) for image recognition. The models used were fine-tuned ResNet-18, 'ViT-B/16/S, ViT-B/16/SG and ViT-B/16/SAM’ attaining an accuracy of 50.1, 52.3, 52.4 and 53.1% respectively for 8 emotions. The study looks at how the training and validation losses alter for the ViTs in comparison to ReSNet-18 and concludes that ResNet-18 has a high training loss as compared to ViT-based models and that 'ViT-B/16/S, ViT-B/16/SG' have a higher validation loss because of inadequate data.

Two 'Attentive Pooling' (AP) modules are proposed in the study by Xue *et al*. (2023). In order to prevent noisy patches from affecting identification performance in the modules, an 'Attentive Patch Pooling' (APP) module is used to pick the distinct local patches from CNN feature maps. The proposed AP modules with the Vision Transformer (APViT) model combine APP and 'Attentive Token Pooling' (ATP). APViT achieves an accuracy of 66.91% on AffectNet images and 91.98% on RAF-DB images.

In order to address FER in the wild, the study (Ma *et al*., 2023) proposed the 'Visual Transformers with Feature Fusion' (VTFF) model, with two basic phases. To start with, 'Attentional Selective Fusion' (ASF) is used to leverage two distinct feature maps created from CNNs with two branches. ASF obtains discriminatory information by integrating several features with local-global attention. This model achieves an accuracy of 88.14, 88.81, and 61.85, on RAF-DB, FERPlus, and AffectNet respectively.

'Few-Shot Facial Expression Recognition' with a 'Self-Supervised Vision Transformer' (SSF-ViT) is proposed by Chen *et al*. (2023) to train a DL model with less labeled data. This was achieved by mixing 'Self-Supervised Learning' (SSL) with 'Few-Shot Learning' (FSL). The ViT encoder is pretrained with 4 self-supervised tasks: jigsaw puzzle, masked patch prediction, image rotation prediction,

and image denoising and reconstruction. According to the results, SSF-ViT claims to have an accuracy of 90.98, 74.95, 66.04, and 63.69%, on RAF-DB, FER2013, AffectNet, and SFEW 2.0 respectively.

The evolutionary approach using Particle Swarm Optimization (PSO) and the Vision Transformer are combined in the (Fliss and Zemzem, 2024) hybrid intelligent model. The majority of emotions in photos were correctly detected by the suggested model, which achieved an accuracy of 100% for the CK+ dataset and 95.93% for the FER2013 Plus dataset.

CNN and VIT

Both CNN and ViT are two important techniques used in Computer Vision (CV) having different architectures and approaches. Convolutional Neural Networks (CNN) have been the predominant network for quite some time, but ViT demonstrated that they can outperform CNN-based models in recognition, detection, segmentation, and other tasks. ViTs leverage transformer architectures to process images as sequences of tokens, allowing them to capture global dependencies and learn effective representations. The major benefit of ViTs lies in their capacity to efficiently gather global contextual data via the self-attention mechanism. In comparison to CNNs, ViTs are a recent approach to computer vision tasks. ViTs are based on transformer architectures that were initially developed in the context of Natural Language Processing (NLP) tasks. They treat images as a sequence of tokens after demarcating the image into a sequence of patches. They then use transformer blocks to process these token sequences thus enabling interactions among different parts of the image. ViTs use self-attention methods to extract global dependencies in the input image. This enables them to effectively model long-range interactions as compared to CNNs. They capture subtle spatial dependencies between facial features and expressions. Although it takes a huge amount of data to pre-train a ViT model, in comparison to CNNs, ViTs need less data to train effectively and they give more promising results. This makes them suitable for applications where there is a scarcity of labeled data. However, the specific performance of these architectures varies due to factors like dataset quality, dataset size, the complexity of the task, model architecture and design, etc.

Convolutional Neural Network (CNN)

A CNN has (i) a Convolution layer, (ii) a Batch normalization layer, (iii) a Rectified Linear Unit (ReLU) or non-linearity activation layer, (iv) a Max-pooling layer, and (v) a Fully connected layer. Convolutional layers are the first layers in a standard CNN architecture. They extract computationally interpretable characteristics from the picture by passing through the kernels or filters from left to right. Low-level characteristics (such as colors, gradient direction, edges, etc.,) are extracted by the first layer, while high-level features are extracted by the lower layers that follow. Then, the pooling layers minimize the data obtained by the convolutional layers, yet preserve the salient characteristics. Ultimately, the flattened output of the convolutional and pooling layers is passed into the fully connected layers, which carry out the classification. The convolutional layers are interspaced by non-linear activation functions and max-pooling layers (Figs. 1-2). CNNs adaptively learn spatial hierarchies of features from images through their convolutional layers in this manner. They process data via a 2D-grid manner by sliding a set of filters over the input image to extract features. This makes them proficient in focusing on the spatial hierarchies of the features such as edges, textures, and object parts. CNNs have been widely used to efficiently capture facial landmarks, expressions, and emotions. Some well-known CNN architectures that have been successfully fine-tuned for facial emotion recognition tasks include VGG, ResNet, Inception (Szegedy *et al*., 2015), DeepFace (Taigman *et al*., 2014), etc. For effective training, CNNs need large volumes of data. Deep CNNs quite often suffer from issues such as vanishing gradients and overfitting.

Vision Transformer (ViT)

A transformer is a high-capacity network architecture that can approximate complex functions. It uses attention to understand the sequence of information. It is a general-purpose architecture that can process a variety of data formats such as text, audio, image, and video. This makes them highly suited to multi-modal deep learning tasks. It consists of two main parts (i) Encoder and (ii) Decoder (Vaswani *et al*., 2017).

Fig. 1: Blocks in a convolutional neural network

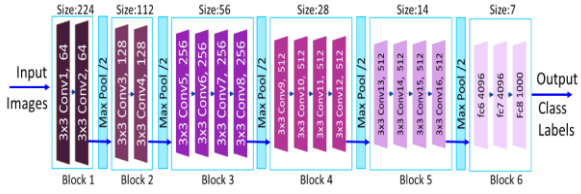


Fig. 2: Block Architecture of VGG-19 having five blocks of convolutional layers (total 16) interspersed with a max-pool layer. The last block has three fully connected layers. Adapted from (Simonyan and Zisserman, 2014)

A ViT is an encoder-only transformer (Dosovitskiy *et al*., 2020). It does not have a decoder. The block diagram of a ViT is shown in Fig. (3). The major benefit of ViTs lies in their capacity to efficiently gather global contextual data via the self-attention mechanism. This allows them to represent long-range dependencies and contextual linkages, hence enhancing their resilience in tasks that need comprehending global context. Additionally, based on input data, ViTs can adaptively modify the self-attention mechanism's receptive fields, improving their ability to record local and global characteristics and strengthening their resistance to changes in object perspective, rotation, or size.

The operation of a vision transformer is detailed as follows.

Image to Patches

The ViT breaks down the image into smaller non-overlapping sections called *patches*. Each patch is typically a square region of the image as shown in Fig. (3). For e.g., in the case of the PM-ViT each image has a size of 224×224 pixels and the patch size is 16 pixels, so, the image is divided into 196 patches, as, [(224*224)/(16*16)] = 196 patches:

$$x \in \mathbb{R}^{W*H*C} \rightarrow x_p^i \in \mathbb{R}^{P*P*C} \; where \; i = 1 \; to \; 196$$

The stride determines how many pixels the sliding window will move in each step. In the case of the PM-ViT, the stride is 16 pixels. Since the patch size is equal to the stride there is no overlap between the patches.

These patches are then flattened/ reshaped from a 2D vector to a 1D vector. Each of these patches is divided into smaller units called tokens. Since most images input to the PM-ViT are colored (i.e. have 3 channels) and the patch size is 16, every patch has 16*16*3 = 768 tokens. Each patch is treated like a distinct input sequence of tokens where each token represents a specific part of the patch.

Linear Projection of Flattened Patches

The Linear Projection Layer works on each flattened patch by transforming each 1D vector into a lower dimensional vector while preserving the important features and relationships. The process of linear projection involves two main operations (a) Weight matrix multiplication and (b) Bias addition. This is very similar to the CNN models where we multiply weights with input and add the bias. It involves multiplying each element of the flattened sequence by a weight and adding a bias term. The bias may also be zero. Both these weights and biases are learned during the training process.

Linear projection can be done for all patches in one step. The result of these steps is a transformed vector of lower dimensionality i.e., a vector that has fewer elements than the original vector. It leads to a reduction in the number of input features used to represent a particular object. The idea behind reducing the dimensionality is that the lower dimensional vectors require less memory and less computational resources thus ensuring faster and effectual operations. Moreover, dimensionality reduction makes the vector representations more robust and focused on essential features because it extracts the most essential features and the most essential information while discarding the less significant details, eliminating noise and irrelevant variations in the data.

Position Embeddings

It is important to understand that in the case of CNN, the operations of convolution translation and scaling are equivariant. This equivariant property is very important for object recognition, detection, and segmentation. The pooling operation in CNN is translation invariant and scale invariant. This means if an object is translated in scale the output of pooling is not affected. This invariant property is important in object recognition. Thus, CNNs have both equivariance and invariance.

In contrast, transformers do not have a notion of equivariance. This issue is resolved by position embedding. Position Embedding Layer is added to each patch of the image. It indicates the location of each patch in the image. Positional encoding is needed because all the data is fed into the transformer together, so, the transformer has no knowledge about the sequence of patches or their position viz-a-viz the original image. Thus, positional embedding provides the position information to ViT.

Position embedding means adding a unique position identifier to the linear projection of each patch. This enables the vision transformer to know the arrangement of the sequence of patches in the original image during training. This is the only inductive bias in a vision transformer. Everything else is learned.

Further, the vision transformer has a learnable class embedding called class token. This is represented by the * sign at position 0 in the Fig. (3).

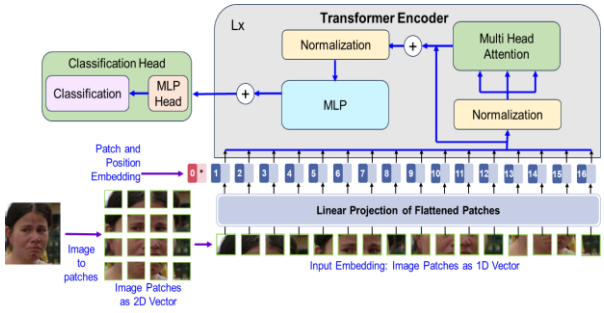


Fig. 3: Architecture of vision transformer adapted from (Dosovitskiy *et al*., 2020)

Transformer Encoder

The encoder consists of a stack of transformer blocks one on top of the other. A block diagram of a single transformer encoder block is shown in Fig. (4). It has 5 key elements that enable its functioning:

- Multi-Head Attention (MHA)
- Multi-Layer Perceptron (MLP)
- Layer Normalization (LN)
- Residual Connections (RC)
- Positional Embeddings (PE)

In Fig. (4), the flow of information is upwards. The MHA lets the different patch embeddings communicate with each other while the MLP lets each embedding think independently about what it has just learned from its neighbors. There is no information sharing between patches. Both MHA and MLP are compute-heavy blocks and most of the processing of the transformer logic happens here. The RC and LN are the optimizations that are included to help the optimization process. Both RC and LN appear twice. RCs help gradients flow while LN aims to stabilize learning. Positional embeddings are added to the embedded patches at the input and they help the transformer identify which patch embedding comes from which location. The following subsection discusses Single-Head Attention (SHA) followed by a subsection that discusses MHA.

Single-Head Attention

An image is divided into p patches. Then $p + 1$ (that accounts for the class token) = N, which is the length of embedding. Let D be the dimension of each embedding. Presume X is the input to the self-attention head of the vision transformer, then, X is given by the relation $X \in \mathbb{R}^{N \times D}$.

The input X is a sequence of embeddings, that is projected thrice to produce 3 matrices of the same shape - queries, keys, and values as follows:

$$Query = Q = W^Q X$$
$$Key = K = W^K X$$
$$Value = V = W^V X$$

In ViT, the attention mechanism operates on pairs of patches in the image grid using the *Query-Key-Value* concept. These new projections can be thought of as having particular meanings:

- The query may be interpreted as the features of interest
- Keys are features relevant to the Query
- Values are what actually gets communicated, it is a matrix of the original features that are to be scaled by the probabilities computed by the attention function

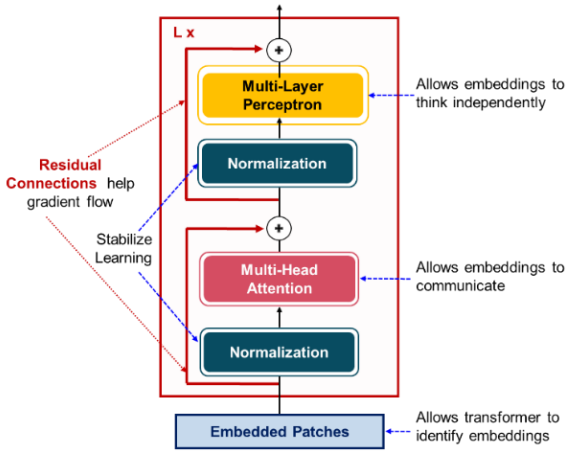


Fig. 4: The transformer encoder (Vaswani *et al*., 2017; Dosovitskiy *et al*., 2020)

When Q is the matrix of queries with dimension d_k, K is the matrix of keys with dimension d_k, and V is the matrix of values of the original features with dimension d_v. Then, the matrix of outputs is given by the 'scaled dot product attention'. It is expressed by the function- $Attention(Q, K, V)$ as shown in Eq. (1):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The dot products of the $query$ are computed with all $keys$ and each of them is divided by $\sqrt{d_k}$, the normalizing factor. Then the normalized dot product is changed to probabilities by the softmax function. These probabilities determine the weights of the $values$ and are used to scale the original input features. The results are added to the residual connections and normalized.

Upon multiplying Q by K^T (K transpose) we get an $N \times N$ matrix in which each row has similarities from the query of one embedding (what we are interested in) to the keys of all the other embeddings (what is relevant to the query). The softmax function is applied along the rows of the matrix to normalize them to probability vectors. Since the softmax is sensitive, when one input is much bigger than the other, it does not give useful gradients. Therefore, to avoid peaky affinities when the inner products are large, the product QK^T is divided by $\sqrt{d_k}$, the square root of the dimension of keys. The justification for this normalization constant is that if we assume we have a query and key vectors as Q and K respectively and that Q and K are independent random variables with mean zero and variance one. Then, their inner product has a variance that is the same as the key dimension that can cause the magnitudes to get very big for high dimensions which in turn saturates the softmax and ruins the gradients. If this assumption is approximately correct then normalizing by $\sqrt{d_k}$ will keep the variance around one. Finally, V is used

to compute a weighted sum of values effectively. Every embedding gets a weighted sum of the values of all other embeddings as per query-key likeness.

Multi-Head Attention (MHA)

The first layer of the transformer encoder in a ViT is the MHA Layer. It has multiple attention blocks in an encoder module (Fig. 5).

Queries, keys, and values are represented by Q, K, and V respectively and h represents the number of attention heads. All these attention blocks have the same network structure. This makes a difference because different learned representations from the multiple heads can improve the overall network performance. Rather than using the same attention function on $_{model}$-dimensional keys, values, and queries, better results are obtained by linearly projecting queries, keys, and values h times using different learned linear projections to d_k, d_k, and d_v dimensions, in that order. (Here h corresponds to the number of heads in the multi-head attention module). The attention function is applied simultaneously to all projected copies of queries, keys, and values to provide d_v dimensional outputs. They are concatenated and projected again to find the concluding results as shown in Fig. (5). MHA permits the model to cooperatively work on information from various representation subspaces at various positions. If a single attention head is used, this is inhibited by averaging.

$$MultiHead(Q,K,V) = Concat(head_1,..,head_h)W^O \qquad (2)$$

where, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and the projections are the parameter matrices:

$W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$

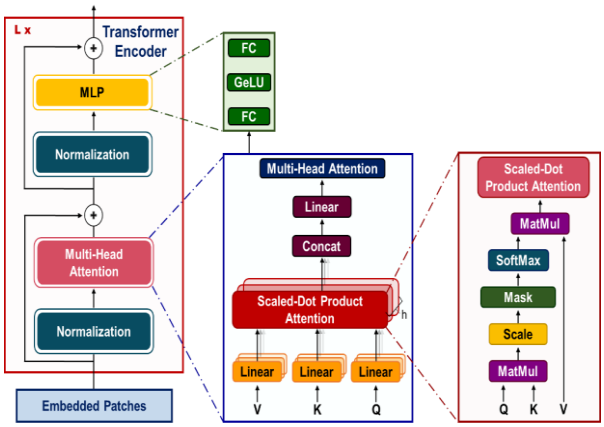


Fig. 5: The MHA module consists of many layers running in parallel (Vaswani *et al*., 2017)

The Attention function is a mapping between a query and a set of key-value pairs to output. Attention aids each patch to attend to and also gather information from all other patches. It captures dependencies between the patches and also enables the model to consider the global context. Each patch of the image is converted to a high-dimensional feature vector and the attention is calculated as the dot product of two feature vectors. It is the length of the projection between the two vectors. MHA concatenates all the attention outputs linearly to the correct dimensions. The multiple attention heads focus on the local and global dependencies in the image.

Every patch of the image is treated as both a "query" and a "key" for attention calculation. At a point in time, the query represents the patch (q) whose representation is being updated based on its relationship with other patches. The key represents all the patches that the query patch (q) attends to. It determines how much attention each key patch should receive from the query patch (q). The value component represents the content of each patch and it is used to compute the weighted sum of values. This weighted sum becomes the output of the attention mechanism.

The attention mechanism computes attention scores between the query and key patches. These attention scores determine to what extent each key patch contributes toward the updated representation of the query patch. The attention scores are typically calculated using a dot product between the query and key embeddings, followed by a *softmax* operation to get attention weights. Thereafter, the attention weights determine the weighted sum of the value embeddings. This weighted sum of value embeddings is the updated representation of the query patch. It captures information from all relevant key patches. This helps the deep learning model to understand the relationship between the different parts of the image by assigning importance scores to the patches and focusing on the most relevant information which helps the model to make better sense of the image.

Each embedding gets a weighted sum of the values of other embeddings based on the query, and key similarity. The attention mechanism is useful but could become very slow if the patches want to send multiple messages. The solution to this problem as proposed by the original authors of ViT is to perform multiple attention operations in parallel using a collection of h separate attention heads. For each of the h heads there is a separate learned query matrix to produce the queries, another learned key matrix to produce the keys, and yet another learned value matrix to produce the values. The scaled dot product attention is applied just as before to obtain an output once this is done for all h heads.

Asymptotic Complexity of Multi-Head Attention

The asymptotic complexity of MHA (ignoring the projections) is $O(n^2 * D)$ which means that the cost of

attention is quadratic in the sequence length. Many researchers have provided ways to reduce this complexity but no model to the best of our knowledge, outperforms the vanilla transformers.

Fully Connected MLP

Now that the embeddings have shared their thoughts with each other in the MHA, they may be allowed to do some thinking alone about what they have learned. This is achieved via a two-layer MLP that is applied independently on each embedding. Therefore, the self-attention layer is followed by a "feed-forward network layer". This layer consists of a couple of FC layers with non-linear activation functions. The output from each patch is passed to this layer. The layer captures the complex non-linear relationships amongst the various patches. The original transformer used ReLU for its nonlinearity. More recently Gaussian Error Linear Unit (GELU) has become more popular for this task. Both these nonlinearities are fairly similar in shape except that the GELU is a little smoother near the origin. The operation of the MLP may be mathematically expressed as:

$$MLP(x) = W_2\sigma(W_1x + b_1) + b_2 \quad (3)$$

Residual Connections

In the transformer encoder block in Figs. (4-5) and the residual connections are depicted by arrows going around the side. Residual connections help with optimization. The residual corresponds to simply adding the input back to the output. The original authors suggest that optimizing the residual mapping is simpler when compared to optimizing the original unreferenced mapping (Vaswani *et al*., 2017). Empirically residual connections make a big difference to optimization He *et al*., 2016).

Normalization Layer

The normalization layer reduces training time and stabilizes training. The two-layer Norm blocks appear before the MHA and the MLP layers in a transformer encoder. Norm is very similar to batch Norm. In both cases, inputs are normalized by subtracting their mean and dividing by the square root of variance plus some positive constant offset for numerical stability. To avoid losing expressiveness the results are scaled by a learned gain parameter and a learned bias parameter is added. Mathematically:

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{Var[x] + \epsilon}} . \gamma + \beta$$

Where, γ is the learned gain and β is the learned bias, ϵ for numerical stability.

Classification Head

The last layer of the transformer encoder is the Classification Head (aka Prediction Head). It maps the output of the transformer into the desired output format. The Multi Linear Perceptron Head is an extra linear layer for the final classification.

The proposed model PM-ViT uses a ViT large architecture as the backbone. It is trained end-to-end using labeled data and the Cross-Entropy loss function. During training, parameters of the model (including those of the transformer encoder layers, positional encodings, and embedding layers) are optimized using the Adam (Adaptive Moment Estimation) optimizer. Adam is based on two main techniques: adaptive learning rates and momentum optimizer.

Datasets Used

The experiments were conducted using images from the two benchmarked datasets AffectNet (Mollahosseini *et al*., 2019) and CK+ (Lucey *et al*., 2010). The details of the images used from these datasets are in Table (1).

AffectNet Dataset

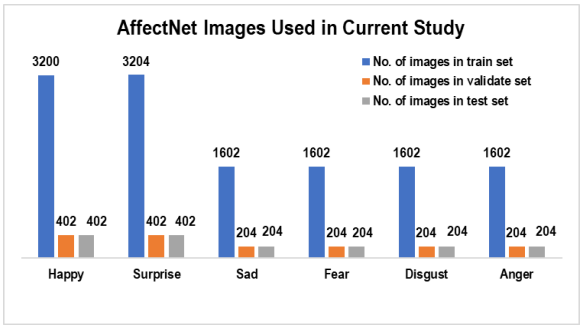
AffectNet is a vast collection of nearly 1000,000 spontaneous pictures in the wild. Spontaneous expressions occur naturally and are not controlled by anyone. Consequently, the feelings are better captured in natural evocations of expressions. The pictures in this dataset were gathered from the Internet using 3 search engines and 1250 emotion-specific keywords in 6 languages. In light of the resources available to carry out the proposed study, a subset of these images for six basic emotions, namely- 'happy (joy), surprise, sad, fear, disgust, and anger' was extracted as a stratified random sample as depicted in Fig. 6(a). Thereafter the images were combined as 2 classes happy and surprise taken in one class as positive emotions and sad, fearful, disgust, and anger taken in the class as negative emotions (Fig. 7(a)).

Extended Cohn-Kanade Dataset (CK+)

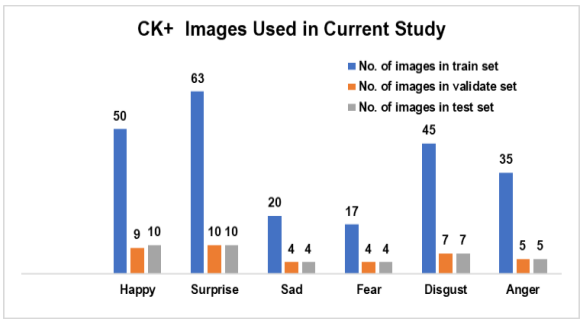
The dataset CK+ has posed images featuring 123 different individuals using 593 videos. These videos have been flattened into 5876 images. For the creation of posed datasets, the participant is requested to 'pose the emotion'. Posed datasets are relatively easy to collect and thus more common. In CK+, 327 images have associated emotion labels. Of these 327, 309 pictures have the 6 emotions considered in this study (Fig. 6(b)). The number of images in CK+ is small, therefore they were augmented 7 times before training the model (Fig. 7(b)).

Table 1: Images used in this study

Dataset	Emotion type	Train	Validate	Test	Total images
AffectNet	Positive	6404	804	804	8012
	Negative	6408	816	816	8040
Total # images		12812	1620	1620	16052
CK+	Positive	113	19	20	152
	Negative	117	20	20	157
Total # images		230	39	40	309



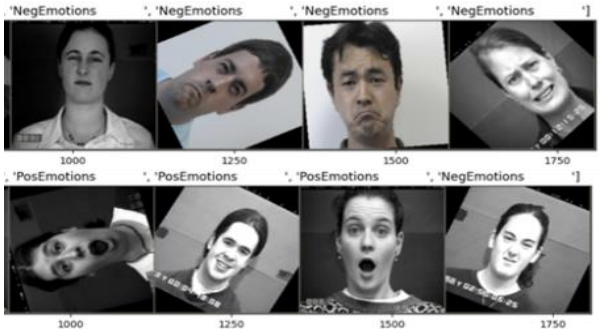
(a)



(b)

Fig. 6: (a) Count of images extracted from AffectNet (Mollahosseini *et al*., 2019) dataset and used in the current study. (b) Count of Images Extracted from CK+ (Lucey *et al*., 2010) dataset, augmented 7 times and then used in the current study

(a)



(b)

Fig. 7: Sample Images from AffectNet and CK+ datasets; (a) AffectNet images; (b) Augmented CK+ images

Materials and Methods

Implementation Environment

The experiments were conducted using a virtual machine at Google Collaboratory Pro on an ASUS RYZEN 7 laptop with 16 GB RAM. These models were built using PyTorch and run on an NVIDIA Ampere Architecture (A100 GPU). The virtual machine had 40 GB GPU RAM and 80 GB RAM in the execution environment. All experiments were conducted 3 times, executing 100 epochs each time and this study reports the average outcomes.

CNN Based Fine-Tuned Model (Model 1)

Model 1 uses the VGG-19 as backbone architecture and is fine-tuned to process facial images from AffectNet and augmented CK+ datasets. It takes a facial image as input and does a two-step classification. In the first step, it recognizes the emotions in the input image and does a binary classification into two classes – negative emotions and positive emotions. In case the emotions found are negative it does a further ternary classification into three classes mild, moderate, and severe based on the perceived strength of the negative emotion and hence the chance of the individual having a mental illness. The thresholds used for classifying the negative emotions into corresponding strengths is shown in Table (2).

The block diagram of Model 1 is shown in Fig. (8). Model 1 is trained using labelled data. The image size used was 224×224. It uses weighted "cross entropy loss" as loss function, weighted "Stochastic Gradient Descent (SGD)" as optimizer and has a learning rate of 1.00e-03, weight decay of 5.00e-04 and momentum of 0.9, as its hyperparameters.

The PM-ViT Model

The PM-ViT is a fine-tuned vision transformer based on ViT Large as backbone architecture (Fig. 9). It is trained end-to-end using labelled data. Exactly like model 1, the PM-ViT recognizes the emotions in the input image and does a binary classification into two classes negative emotions and positive emotions in step 1. In case the emotions found in the image are negative it does a further classification into 3 classes mild, moderate and severe based on the strength of negative emotion and predicts the chance of the individual having a mental illness. The logic of this experiment is based on the Algorithm 1 below. The image size used by PM-ViT is 224×224, the patch size is 16×16 and the stride is also 16. The model uses “Cross-Entropy Loss" as loss function, a learning rate of 1.04e-4 and weight decay of 1.00e-4. During training, the parameters for this model were optimized using grid search and Adam (Adaptive Moment Estimation) optimizer. Even though grid search is time consuming, it was used to tune the hyperparameters because it is easy to implement and when the search space is not very large it

guarantees finding the best combination. Adam is based on two main techniques: Adaptive learning rates and momentum optimizer. Adam was used as it is known to work well with ViT large. The tuning details of learning rates and weight decay used for optimizing PM-ViT using AffectNet images is as shown in Table (3). GELU was used for non-linearity.

Table 2: Classification of images with negative emotions vis-a-vis the intensity (mild, moderate or severe) that the person could be suffering from a mental illness

Extent of negativity	Class
0.75 > prob >= 0.50	mild
0.90 > prob >= 0.75	moderate
>= 0.90	severe

Table 3: Performance of PM-ViT for various learning rates and weight decay values

Parameter	Value	Accuracy (%)
Learning Rate	1.04e-03	82.15
	1.04e-04	82.80
	5.04e-04	82.63
Weight Decay	1.00e-04	82.38
	5.00e-04	82.59
	1.00e-03	82.17

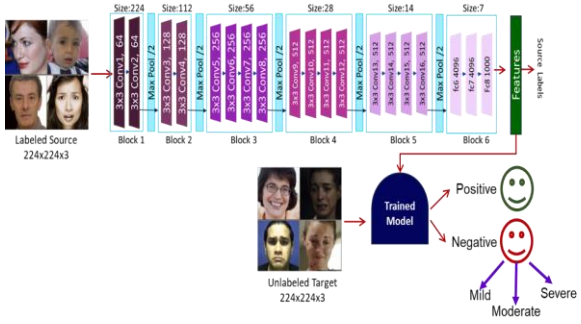


Fig. 8: Model 1, the fine-tuned model CNN using VGG-19 as backbone architecture

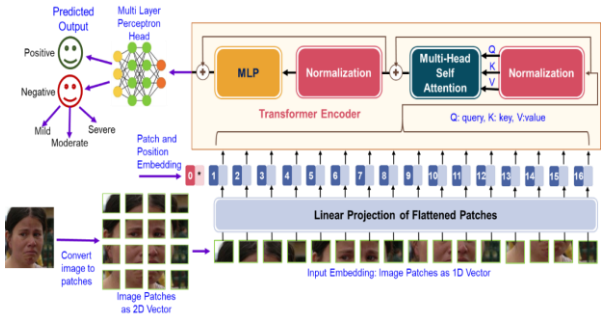


Fig. 9: Proposed PM-ViT Model, using ViT large as its backbone architecture, fine-tuned to classify AffectNet and CK+ augmented dataset image.

Results and Discussion

Experiments were conducted using images extracted form AffectNet and images from augmented CK+ datasets. The results demonstrate that using the CNN based Model 1 the experiment achieves an accuracy of 81.48% and a F1-score of 81.0% on AffectNet dataset. Model 1 achieves an accuracy of 97.8% and a F1-score of 97.8% on CK+ augmented dataset. These results are presented in Fig. (10) and Table (4).

The ViT based proposed PM-ViT framework surpasses the performance of the CNN based Model 1 by achieving an accuracy of 83.78% and a F1-score of 84.0% on AffectNet and an accuracy of 99.7% and an F1-score of 99.7% on CK+ augmented dataset (Table 4).

The Figs. (11-12) show the performance of PM-ViT in comparison to the various ViT based SOTA models that have been used to recognize emotions using the AffectNet and CK+ datasets respectively. The findings therein clearly demonstrate that the proposed PM-ViT's superiority over the SOTA ViT based models used for emotion recognition using the same datasets as PM-ViT (Table 4).

Algorithm 1: Logic of PM-Vit Model

Input: An image of a face
Output: Class Label, Intensity Label

1. Cut the input image into non overlapping square patches and arrange them into a sequence from top left to bottom right
2. Patch Embedding: Convert each patch into a vector of lower dimensions by applying linear projection.
3. Generate positional embeddings for each patch and tokenize- combine patch embeddings with positional encoding.
4. Class token- include a learnable embedding at the position zero to represent the entire image.
5. Pass the tokenized embeddings through multiple transformer encoder layers.
6. Global Representation Extraction: The patch embeddings are aggregated to obtain a global representation of the image. This is achieved using mean pooling.
7. Obtain Classification Head output- a linear layer that uses the global representation to do a binary classification followed by a three-way classification if required:
    a. Classify the image as depicting positive emotions or negative emotions.
    b. If the image depicts negative emotions, categorize the chance of the person having a mental illness as mild, moderate or severe using the thresholds in Table (2).

End

Table 4: Performance of proposed study vs state-of-the-art

Study	Method used/ technique	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Georgescu *et al*. (2019)	KNN (Supervised Learning)	AffectNet	63.31			
Handrich *et al*. (2020)	CNN (Supervised Learning)	AffectNet	79			
Georgescu *et al*. (2019)	CNN and BOVW+ global SVM (supervised learning)	AffectNet	63.2			
Ma *et al*. (2021)	CNN based MTCNN (supervised learning)	AffectNet	61.85			
Hasani *et al*. (2022)	BReG-NeXt (supervised learning)	AffectNet	68.50			
Aina *et al*. (2024)	DNet (a CNN), ViT and ResNet50.	AffectNet, FER 2013	78			
Gao *et al*. (2023)	SSA-ResNet18 (supervised learning)	AffectNet	65.04			
Gao *et al*. (2023)	SSA-ICL-ResNet18 (Supervised Learning)	AffectNet	65.78			
Hossain *et al*. (2023)	CNN Fusion (transfer learning)	SFEW 2.0	80.34			
Kurian and Tripathi (2022)	UDA - Generative adversarial model (GAN)	FER-2013, Kaggle Autism dataset	71.53			
Kurian and Tripathi (2022)	UDA - Generative adversarial model (GAN)	CK+, Kaggle Autism dataset	71.53			
Jain *et al*. (2024)	CNN based EM-UDA Transfer Learning	AffectNet, FER 2013	74.55			74.87
Jain *et al*. (2024)	CNN based EM-UDA Transfer Learning	AffectNet, CK+	83.9			82.8
Chaudhari *et al*. (2022)	ResNet-18	AVFER (FER2013, AffectNet, CK+48)	52.77	50.90	60.05	49.43
Chaudhari *et al*. (2022)	ViT-B/16/S	AVFER (FER2013, AffectNet, CK+48)	54.89	54.85	62.25	51.84
Chaudhari *et al*. (2022)	ViT-B/16/Sharpness Aware Minimizer	AVFER (FER2013, AffectNet, CK+48)	56.94	54.70	63.10	62.20
Ma *et al*. (2023)	Visual Transformers with Feature Fusion (VTFF)	AffectNet	61.85			
Li *et al*. (2021)	Mask Vision Transformer (MVT)	AffectNet	64.57			
Chen *et al*. (2023)	SSF-ViT	AffectNet	66.04			
Xue *et al*. (2023)	APViT	AffectNet	66.91			
Kim *et al*. (2022)	Squeeze ViT	CK+	99.54			
Present Study Model 1	CNN-VGG	AffectNet	81.48	82.0	82.0	81.0
Present Study Model 1	CNN-VGG	CK+ Augmented	97.8	96.4	99.2	97.8
Proposed Model PM-ViT	ViT	AffectNet	83.78	84.0	84.0	84.0
Proposed Model PM-ViT	ViT	CK+ Augmented	99.7	99.4	100	99.7

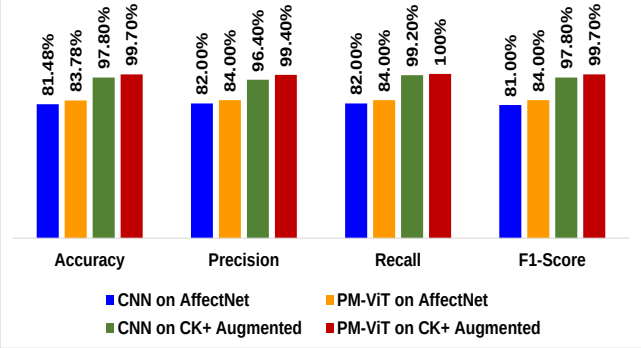


Fig. 10: Performance of CNN based Model 1 Vs PM-ViT

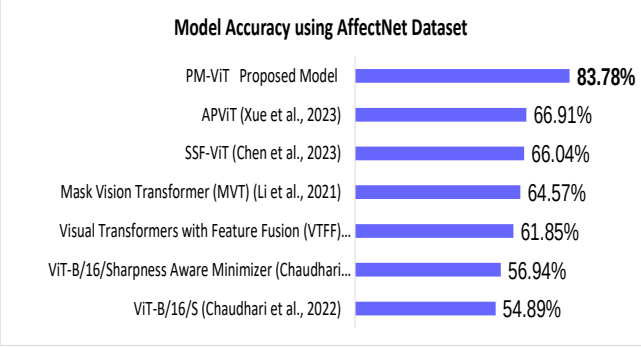


Fig. 11: Performance of PM-ViT Vs other ViT based SOTA models Using AffectNet Dataset

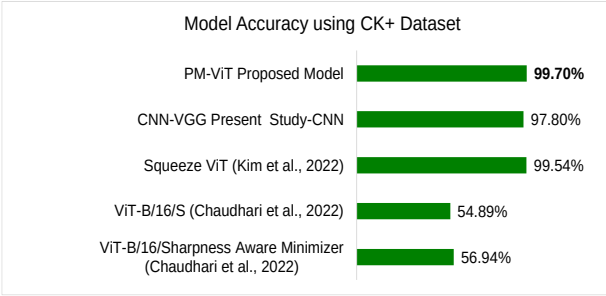


Fig. 12: Performance of PM-ViT Vs other ViT based SOTA models Using CK+ Dataset

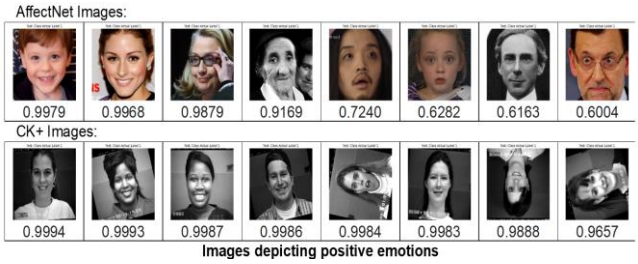


Fig. 13: Sample images from AffectNet and CK+ augmented, after the first step binary classification, showing positive emotions along with the strength of the positive emotions

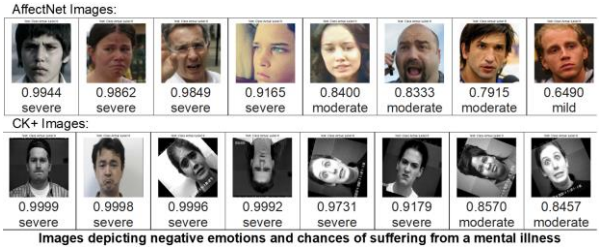


Fig. 14: Sample output of PM-ViT, showing images from AffectNet and CK+ augmented, after the 2nd step, the 3-way classification, showing negative emotions along with the perceived strength of the negative emotions

The outputs of the second stage of classification by PM-ViT to determine the proclivity towards mental illness in case the emotion in the image is found to be negative. The results of the images processed at this stage are as depicted in Figs. (13-14).

Conclusion

The current study proposes a novel framework PM-ViT that takes facial images as input, recognizes the emotions in the input image. It does a binary classification of these images into classes negative emotions and positive emotions as step 1. In case the emotions found in the image are negative it does a further classification, as step 2, whereby it classifies the negative emotion images into three classes mild, moderate and severe basis the perceived strength of negative emotion in that image. This is to predict the proclivity of the individual towards a mental illness.

The experimental findings demonstrate that using the CNN based model 1 the experiment achieves an accuracy of 81.48% and a F1-score of 81.0% on AffectNet dataset. It achieves an accuracy of 97.8% and a F1-score of 97.8% on CK+ augmented dataset, whereas the PM-ViT framework achieves an accuracy of 83.78% and a F1-score of 84.0% on AffectNet and an accuracy of 99.7% and an F1-score of 99.7% on CK+ augmented dataset. This, is indicative of the superiority of the vision transformers over CNN for image processing tasks.

The experimental findings also show that the proposed PM-ViT model has a better performance over the other SOTA vision transformer-based models proposed in the recent studies.

The proposed model has the potential to supplement existing mental health assessments, allow regular low-cost evaluation, automated evaluation in situations where a specialist is not easily accessible or the patient is unwilling or unable to engage. Similar framework could also be used to recognizing early signs of mental health deterioration or prodrome detection.

This research may contribute to the creation of fresh approaches for the early identification or diagnosis of

mental illnesses like anxiety and depression through the use of facial emotion.

Future Directions

The work can be extended for finer categorizations of emotions, for a better assessment of the emotional states vis-à-vis mental ailment prediction using the facial images. Researchers would also like to extend this study for use with multimodal dataset.

Translational applications require real-time emotion detection and hence focused research towards making efficient architectures that can detect emotions without hampering a user's experience of the device may be helpful. Developing light models with smaller architectures and few trainable parameters that are capable of performing well in real-time emotion recognition tasks for handheld computational devices can be an application area of the proposed model.

Acknowledgment

The authors express their heartfelt gratitude to the reviewers of this study for their insightful and helpful comments; to the editors of this study for their prompt efforts in managing the manuscript that have considerably helped to enhance the original submission.

Funding Information

The authors did not receive any funding for this study.

Author's Contributions

Priti Rai Jain: Conceived and designed the experiments, performed the experiments, performed the computation work, analyzed the results, prepared all figures and/or tables, prepared all the drafts and the final manuscript.

Syed Mohammad Khurshaid Quadri: Supervised the entire work and approved the final draft.

Anuradha Khattar: Conceived the experiment, analyzed the results, approved the final draft.

Ethics

The authors confirm that this manuscript has not been published elsewhere and that no ethical issues are involved.

References

Aina, J., Akinniyi, O., Rahman, Md. M., Odero-Marah, V., & Khalifa, F. (2024). A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition. *IEEE Access*, *12*, 91410–91425. https://doi.org/10.1109/access.2024.3421376

Al Jazaery, M., & Guo, G. (2021). Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features. *IEEE Transactions on Affective Computing*, *12*(1), 262–268. https://doi.org/10.1109/taffc.2018.2870884

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory–II. *Psychological Assessment*. https://doi.org/10.1037/t00742-000

Bobojanov, S., Kim, B. M., Arabboev, M., & Begmatov, S. (2023). Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. *Applied Sciences*, *13*(22), 12271. https://doi.org/10.3390/app132212271

Bousaid, R., El Hajji, M., & Es-Saady, Y. (2022). Facial Expression Recognition Using a Hybrid ViT-CNN Aggregator. *Business Intelligence*, 61–70. https://doi.org/10.1007/978-3-031-06458-6_5

Branco, L. D., Cotrena, C., Ponsoni, A., Salvador-Silva, R., Vasconcellos, S. J. L., & Fonseca, R. P. (2018). Identification and Perceived Intensity of Facial Expressions of Emotion in Bipolar Disorder and Major Depression. *Archives of Clinical Neuropsychology*, *33*(4), 491–501. https://doi.org/10.1093/arclin/acx080

Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation*, *5*(4), 80. https://doi.org/10.3390/asi5040080

Chen, X., Zheng, X., Sun, K., Liu, W., & Zhang, Y. (2023). Self-supervised vision transformer-based few-shot learning for facial expression recognition. *Information Sciences*, *634*, 206–226. https://doi.org/10.1016/j.ins.2023.03.105

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16×6 Words: Transformers for Image Recognition at Scale. In *arXiv:2010.11929v2*. https://doi.org/10.48550/arXiv.2010.11929

Ellgring, H. (2007). *Non-verbal communication in depression*.

Fliss, I., & Zemzem, W. (2024). A novel PSO-ViT approach for facial emotion recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *11*(7), 2297016. https://doi.org/10.1080/21681163.2023.2297016

Gao, H., Wu, M., Chen, Z., Li, Y., Wang, X., An, S., Li, J., & Liu, C. (2023). SSA-ICL: Multi-domain adaptive attention with intra-dataset continual learning for Facial expression recognition. *Neural Networks*, *158*, 228–238. https://doi.org/10.1016/j.neunet.2022.11.025

Garg, Parie, & Glick, S. (2018). AI's Potential to Diagnose and Treat Mental Illness. *Harvard Business Review*, *22*.

Georgescu, M.-I., Ionescu, R. T., & Popescu, M. (2019). Local Learning with Deep and Handcrafted Features for Facial Expression Recognition. *IEEE Access*, *7*, 64827–64836. https://doi.org/10.1109/access.2019.2917266

Handrich, S., Dinges, L., Al-Hamadi, A., Werner, P., & Aghbari, Z. A. (2020). Simultaneous Prediction of Valence/Arousal and Emotions on AffectNet, Aff-Wild and AFEW-VA. *Procedia Computer Science*, *170*, 634–641. https://doi.org/10.1016/j.procs.2020.03.134

Hasani, B., Negi, P. S., & Mahoor, M. H. (2022). BReG-NeXt: Facial Affect Computing Using Adaptive Residual Networks with Bounded Gradient. *IEEE Transactions on Affective Computing*, *13*(2), 1023–1036. https://doi.org/10.1109/taffc.2020.2986440

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.1109/cvpr.2016.90

He, L., Jiang, D., & Sahli, H. (2019). Automatic Depression Analysis Using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding. *IEEE Transactions on Multimedia*, *21*(6), 1476–1486. https://doi.org/10.1109/tmm.2018.2877129

Hossain, S., Umer, S., Rout, R. Kr., & Tanveer, M. (2023). Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. *Applied Soft Computing*, *134*, 109997. https://doi.org/10.1016/j.asoc.2023.109997

Huang, Y., Chiang, C.-F., & Chen, A. (2019). Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram. *Proceedings of the 8th International Conference on Data Science, Technology and Applications DATA*, 32–40. https://doi.org/10.5220/0007833600320040

Jain, P. R., & Quadri, S. M. K. (2021). Emerging Role of Intelligent Techniques for Effective Detection and Prediction of Mental Disorders. *Intelligent Data Communication Technologies and Internet of Things*, 185–198. https://doi.org/10.1007/978-981-15-9509-7_16

Jain, P. R., Quadri, S. M. K., & Khattar, A. (2024). EM-UDA: Emotion Detection Using Unsupervised Domain Adaptation for Classification of Facial Images. *IEEE Access*, *12*, 140262–140276. https://doi.org/10.1109/access.2024.3467990

Jiang, Z., Harati, S., Crowell, A., Mayberg, H. S., Nemati, S., & Clifford, G. D. (2021). Classifying Major Depressive Disorder and Response to Deep Brain Stimulation over Time by Analyzing Facial Expressions. *IEEE Transactions on Biomedical Engineering*, *68*(2), 664–672. https://doi.org/10.1109/tbme.2020.3010472

Kim, S., Nam, J., & Ko, B. C. (2022). Facial Expression Recognition Based on Squeeze Vision Transformer. *Sensors*, *22*(10), 3729. https://doi.org/10.3390/s22103729

Kurian, A., & Tripathi, S. (2022). Unsupervised Domain Adaptation for Facial Emotion Recognition in Autistic Children. *Workshops at 18th International Conference on Intelligent Environments (IE2022)*, 66–75. https://doi.org/10.3233/aise220022

Li, H., Sui, Mingzhe, Zhao, F., Zha, Z., & Wu, F. (2021). MVT: Mask Vision Transformer for Facial Expression Recognition in the wild. *ArXiv:2106.04520v2*. https://doi.org/10.48550/arXiv.2106.04520

Lili, N. A., Nurul Amiraa, M. R., MasRina, M., & Nurul Amelina, N. (2022). Depression Level Detection from Facial Emotion Recognition Using Image Processing. *Proceedings of the 8th International Conference on Computational Science and Technology*, 739–750. https://doi.org/10.1007/978-981-16-8515-6_56

Liu, D., Liu, B., Lin, T., Liu, G., Yang, G., Qi, D., Qiu, Y., Lu, Y., Yuan, Q., Shuai, S. C., Li, X., Liu, O., Tang, X., Shuai, J., Cao, Y., & Lin, H. (2022). Measuring depression severity based on facial expression and body movement using deep convolutional neural network. *Frontiers in Psychiatry*, *13*, 1017064. https://doi.org/10.3389/fpsyt.2022.1017064

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. https://doi.org/10.1109/cvprw.2010.5543262

Ma, F., Sun, B., & Li, S. (2021). Robust Facial Expression Recognition with Convolutional Visual Transformers. *ArXiv Preprint ArXiv:2103.16854*, *2*(6), 7.

Ma, F., Sun, B., & Li, S. (2023). Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing*, *14*(2), 1236–1248. https://doi.org/10.1109/taffc.2021.3122146

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. https://doi.org/10.1109/taffc.2017.2740923

Niu, M., Tao, J., & Liu, B. (2021). Multi-Scale and Multi-Region Facial Discriminative Representation for Automatic Depression Level Prediction. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1325–1329. https://doi.org/10.1109/icassp39728.2021.9413504

Pennycook, A. (1985). Actions Speak Louder Than Words: Paralanguage, Communication and Education. *TESOL Quarterly*, *19*(2), 259. https://doi.org/10.2307/3586829

Qiu, S., Zhao, G., Li, X., & Wang, X. (2023). Facial Expression Recognition Using Local Sliding Window Attention. *Sensors*, *23*(7), 3424. https://doi.org/10.3390/s23073424

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do Vision Transformers See Like Convolutional Neural Networks? *Advances in Neural Information Processing Systems*, 12116–12128.

Rai Jain, P., Quadri, S. M. K., & Lalit, M. (2021). Recent Trends in Artificial Intelligence for Emotion Detection using Facial Image Analysis. *Proceedings of the 2021 Thirteenth International Conference on Contemporary Computing*, 18–36. https://doi.org/10.1145/3474124.3474205

Simcock, G., McLoughlin, L. T., De Regt, T., Broadhouse, K. M., Beaudequin, D., Lagopoulos, J., & Hermens, D. F. (2020). Associations between Facial Emotion Recognition and Mental Health in Early Adolescence. *International Journal of Environmental Research and Public Health*, *17*(1), 330. https://doi.org/10.3390/ijerph17010330

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *arXiv:1409.1556v6*. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). https://doi.org/10.1109/cvpr.2015.7298594

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/cvpr.2014.220

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*. Advances in Neural Information Processing Systems.

Wen, L., Li, X., Guo, G., & Zhu, Y. (2015). Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding. *IEEE Transactions on Information Forensics and Security*, *10*(7), 1432–1441. https://doi.org/10.1109/tifs.2015.2414392

Xue, F., Wang, Q., Tan, Z., Ma, Z., & Guo, G. (2023). Vision Transformer with Attentive Pooling for Robust Facial Expression Recognition. *IEEE Transactions on Affective Computing*, *14*(4), 3244–3256. https://doi.org/10.1109/taffc.2022.3226473

Zakieldin, K., Khattab, R., Ibrahim, E., Arafat, E., Ahmed, N., & Hemayed, E. (2024). ViTCN: Hybrid Vision Transformer with Temporal Convolution for Multi-Emotion Recognition. *International Journal of Computational Intelligence Systems*, *17*(1), 64. https://doi.org/10.1007/s44196-024-00436-5

Zhu, Y., Shang, Y., Shao, Z., & Guo, G. (2018). Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing*, *9*(4), 578–584. https://doi.org/10.1109/taffc.2017.2650899