A Study on Emotional Analysis Based on Deep Learning

Sudhanshu Raghuwanshi (Dept. of Computer Science and Engineering), Research Scholar, Glocal University, Saharanpur, Uttar Pradesh
Dr. Geetu Soni, Professor (Dept. of Computer Science and Engineering), Glocal University, Saharanpur, Uttar Pradesh

ABSTRACT

This research develops EmoNet, an emotion identification system using Deep Neural Networks, to accurately detect various facial emotions. EmoNet, with 40 layers, outperforms conventional models, achieving an 8% accuracy improvement on FER2013 and a 0.2% improvement on JAFFE, and can classify 3589 images into seven categories in under 2.77 seconds. The model handles variations in facial size, illumination, and angles, utilizing residual layers to enhance classification and engagement detection. Tested with three Indian datasets, it achieves an 86.87% accuracy rate. EmoNet's integration of Orthogonal High Pass filters further improves its real-world classification performance.

Keywords: EmoNet, JAFFE, Residual Layers, Conventional Models

1. INTRODUCTION

1.1 Recognising Expressions on the Face

A person's facial expressions reveal a lot about their emotional condition right now. When you want to know how someone is feeling right now, go no farther than their emotions. Developing the capacity to convey feelings via one's face begins at a young age and continues to mature throughout a person's life. Every person's facial expression is very consistent, and it's incredibly intuitive for each individual to decipher the meaning behind a person's expression, even though we encounter millions of faces during our lives. One may tell how someone is feeling by looking at their face; expressions like happiness, sadness, rage, surprise, etc. During the Aristotelian period, people started to analyse facial movements and research physiognomy, two extremely old fields of study. A person's expressions are a product of their genes coordinating a set of predetermined, disjointed facial muscle actions. Human facial expressions have been studied since before the Darwinian era. It is challenging to transfer emotional traits onto computer-based systems that can learn and understand human facial expressions, despite the fact that emotions are intrinsic properties that can be changed via peer influence and experience. In order to comprehend people's present emotional states, researchers have developed machine learning algorithms, thanks to recent technical and scientific advancements. It is possible to associate references to face emotions with computationally intelligent systems by using automated methods of facial expression recognition. The ability to detect and respond to human emotions is a hallmark of artificial intelligence (AI) systems. Behavioural sciences, psychology, telecommunications, educational technology, automotive safety, and human-computer interaction are some of the areas that researchers are presently concentrating on while creating emotionally intelligent computer models. Training and recognising a group of people's facial expressions is a challenging undertaking, and recognition is a data-driven process. For FER to be successful, it must be able to compensate for a wide range of conditions, including but not limited to: age, cultural background, low-intensity representation, occlusion, non-frontal image location, lighting, and occlusion. It takes a lot of properly labelled data to train and evaluate the aforementioned categories. While precise categorization is critical, recognition accuracy is affected by numerous variables. Here we focus on training a system that can,

- Facial expression analysis should be taught using a framework that learns from its mistakes.
- Identify arbitrary body language and facial expressions across databases.
- The FER2013, IMFDB, CK+, and JAFFE datasets are used for accurate facial expression recognition.

1.2 The Advancement of Deep Learning Networks

It is possible to separate the architecture of a Deep Neural Network (DNN) into two separate components. The initial step is to acquire the image's attributes and include them into a Feature Map (F_m). Each feature is fed into the network via a neuron that is linked to the local receptive fields of the layer below it. Collecting these features and qualities allows one to ascertain the spatial relationship between different local traits and other attributes. The second stage is to collect all of the feature maps. The cell size remains constant throughout all feature maps, which are flat. The feature map, which is the network's activation function, is constructed using a sigmoid function. Sharing weight on the same mapping plane further restricts the number of available network parameters. A calculation layer follows every neural network convolutional layer. The processing of the secondary derivative and the local average both fall within the purview of this layer. The dimensional complexity is reduced by this unique two-function extraction method. A Convolutional Neural Network (CNN) can have its design improved by increasing both the number of neurons and the thickness of its layers. As a result, the amount of processing power needed typically rises. Advanced graphics processing unit (GPU) computers have made previously impossible computations doable. Overfitting and underfitting are common problems with learning complexity during validation and training. To get around this problem, it would be best to build a sparse network that takes cues from the biological networking system in humans. The proposed method employs Convolutional Neural Networks for the purpose of classifying emotive facial expressions. The procedure is divided into the following three parts:

- The effort put into Facial Emotion Recognition is highlighted through the use of pre-processing.
- Education, Evaluation, and Verification.
- Develop the EmoNet architecture.

Prior to the training and testing phases, the images undergo a preliminary processing step. Histogram equalisation is a tool for reducing the impact of lighting, contrast, illumination, and brightness variations. Face recognition is essential since emotions can only be expressed through facial expressions. Every single image in the dataset underwent the affine transformation, with a variation angle ranging from -45 degrees to 45 degrees. A single face is included in each photograph in the dataset. This model employs a method known as Progressive Resizing (PR) to accommodate a range of facial sizes. Using this method, reducing the time needed to resize all of the photographs to a specific, predefined dimension is a breeze. The standard procedure for FER is shown in Figure 1.1.

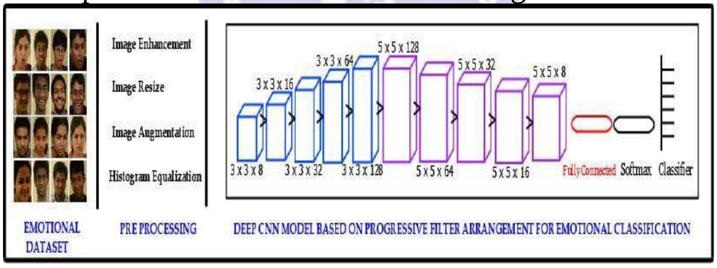


Fig. 1.1: Facial Emotion Recognition Architecture

Due to the wide variety of human appearances, face registration methods vary between datasets. Research and a reliable method of comparison are essential in image processing. Repetition of huge feature maps and network over fitting are inevitable outcomes of the challenging process of learning the significance of each pixel. There is heavy reliance on each dataset's data for Learning and Validation. If a network wants to improve its cross-database efficiency and eliminate reliance barriers, it needs to be able to interact well with external data. One of the most popular computer vision algorithms, CNN, has been the subject of numerous investigations of its effectiveness and resilience. In order to consistently categorise emotions, this system identifies, accumulates, and interprets data from images. Improving its functionality is the primary focus of the application. A fully integrated seven-class network forms the basis of the design. There are a total of ten convolutional layers in the CNN, along with four max-pooling layers and two average-pooling layers. We do this so that we can collect neurons in the greatest possible quantity. A completely linked layer follows the convolutions, and a classification layer and a SoftMax layer sort the feelings into several categories. The network's window size (Ws) is initially set to 3x3 and subsequently increased to 5x5 for the layers that follow. Because of this, more individuals can now benefit from a single training programme. Due to the Rectified Linear Unit (ReLU) leak, CNN is vulnerable to nonlinearity. It makes use of convolutional filters that have 8,16,32,64, and 128 steps. With Max-Pooling, translation invariance is improved and deeper level work is reduced. The model uses numerous filters to extract features from the same image, hence convolution is done individually for each filter and layered. Fig. 1.2 shows the strata. I, image size, p padding, r filter, n_c channels, f_n filters, and s stride are shown in Eq. 1.

$$[I,I,I_c]*[r,r,C]=\left[\left[\frac{I+2p-r}{s}+1\right]\left[\frac{I+2p-r}{s}-1\right],f_n\right] \tag{1.1}$$

Based on Eq. 1.1 convolution is carried out on each image based on I , C and r , pre- serving the relation between pixels and creating a matrix of feature maps. s is used to shift the filter over image pixels using p

$$R_{lf}(\theta)=R_{lf}(\theta)+\lambda\Omega(w_v) \qquad (1.2)$$

$R_{lf}(\theta)$, is the regularization loss function. where, w_v is the weight vector, λ is the factor for regularization (coefficient), and $\Omega(w_v)$. The Regularization Function is given by

$$\Omega(w_v)=l_2 w^T{}_v w_v \qquad (1.3)$$

This model employs l_2 Regularisation to create a unique and stable solution. The l_2-norm squares error dynamically adjusts the model to minimise errors. SGD was utilised in the network. The stochastic gradient descent approach accelerates learning towards the optimal value by oscillating along the steepest path. This may cause fit issues. A momentum controller prevents oscillation and enhances learning curve smoothness. The equation for Stochastic Gradient Descent with Momentum (SGDM) is:

$$\epsilon_{l+1}=\epsilon_l-\alpha\Delta R_{lf}(\epsilon_l)+\gamma(\epsilon_l-(\epsilon_{l-1}))$$
(1.4)

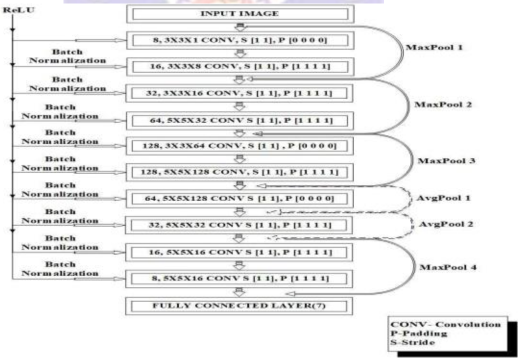

ReLU
INPUT IMAGE
8, 3X3X1 CONV, S [1 1], P [0 0 0 0]
Batch Normalization
16, 3X3X8 CONV, S [1 1], P [1 1 1 1]
Batch Normalization
32, 3X3X16 CONV, S [1 1], P [1 1 1 1]
Batch Normalization
64, 5X5X32 CONV S [1 1], P [1 1 1 1]
Batch Normalization
128, 3X3X64 CONV, S [1 1] , P [0 0 0 0]
Batch Normalization
128, 5X5X128 CONV, S [1 1], P [1 1 1 1]
Batch Normalization
64, 5X5X128 CONV S [1 1], P [0 0 0 0]
Batch Normalization
32, 5X5X32 CONV S [1 1], P [1 1 1 1]
Batch Normalization
16, 5X5X16 CONV S [1 1], P [1 1 1 1]
Batch Normalization
8, 5X5X16 CONV S [1 1], P [1 1 1 1]
FULLY CONNECTED LAYER(7)
MaxPool 1
MaxPool 2
MaxPool 3
AvgPool 1
AvgPool 2
MaxPool 4
CONV- Convolution
P-Padding
S-Stride

Fig. 1.2: EmoNet's Architecture

$$u_i=u_i-l_r\frac{\partial u}{\partial u_i} \tag{1.5}$$

1.3 Recurrent Neural Networks with a Small Residue

The primary goal of this study is to build a domain that can distinguish between different human emotions in order to enhance feature extraction. Achieving this goal is made possible by expanding emotions into six separate groups and developing a model that can reduce the number of hyperparameters needed and the detection bias. The network's convolutional layers are on the left side of the picture, while the Skip network is on the right. Connected to the layers below them, the convolution layers are called intermediate layers. The symbol SIU1CONV1 represents a single convolutional unit, while the symbol S1U1BN1 represents the Batch normalisation layer. Two convolutional layers and one batch normalisation layer comprise each group in the overall structure of these layers. After the second set of layers, the remaining connections are added to the third set, and the second connection is added to the fifth set from the fourth set. The ReLu layers provide connectivity and are in charge of activation passing between sets of layers. Furthermore, the images fed into the network are passed through a convolution layer on the input layer. Three further layers—one for classification, one for pooling, and one for using the SoftMax algorithm—make up the final architecture. In order to aid in dimensionality reduction, the weights provided by the different distributions utilised in the previous levels are aggregated in the Pooling layer. In order to classify the image according to its distribution, the classification layer supplies the class. a manageable burden In order to extract facial traits, the ResNet algorithm was employed. Extracting and differentiating specific features is the first step in accurately analysing facial expressions. It is possible to divide this assignment into three parts:

- Data Preparations
- Emotional Understanding through Residual Network.
- Results Validation

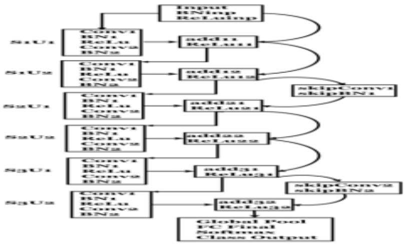


Fig. 1.3: Custom Shallow ResNet Model

The model that was displayed in Figure 1.4 was utilised to derive significant inferences from the image. Every residual block is built with a batch normalisation layer, a 3x3 convolution layer, and a ReLU activation feature. The enhancement also incorporates a batch normalisation layer and a 3x3 convolution layer. Bypassing these two levels, the skip connection instead forms a connection right before the ReLU activation capability. By recycling these leftover bricks, a residual network is finally formed. Three sections (s_1, s_2, s_3) and three subunits (u_1, u_2, u_3) make up the network. Layers such as Batch Normalisation (BN), Rectified Linear Units (ReLu), and Convolutional (Conv) are present inside each section. The extra layer is then notified of each segment. Then, following the first two sections, you'll find two skipnet layers (SkipConv). Degradation issues can lead to serious training mistakes in deep neural networks. Reason being: fitting and disappearing gradients are sources of concern. It is not necessarily the case that deeper networks are "more difficult" to match intuitively. If there are indeed N layers with maximum data accuracy, then the network's performance at N layer would be efficient since the identity-mapping layers M (x) that follow N can only be learnt by mapping the layer that has to be taught. The opposite is true: driving weights in a way that precisely achieves identity mapping is no easy feat. The concept of residual learning is multiplied by the residual function, which is represented by the notation R(x) − x.

As shown in fig.1.3, this is interpreted as a stack of layers that computes the mapping as y= R(x)+ x. The learning of y=R(x) may occur immediately via L(x), where

$$y = (x_i , P_i) + x \qquad (1.6)$$

where, L can be multiple layers. A shortcut operation and element-wise addition are done by the ‘+′ operation as seen in Fig. 1.4.

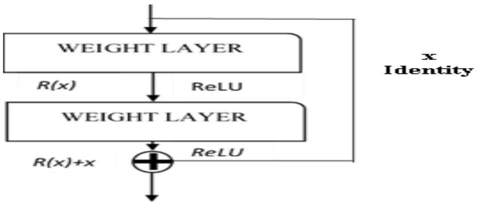


Fig. 1.4: A Residual Function

Although the network is deep, the training time stays the same even after adding additional skip connections; this is because they do not add parameters. As the number of network layers increases, the number of parameters does not grow monotonically. Projection matrices Ps on a space L(x) can give an approximation of the dimensions. Therefore, after adding the projection matrix, the equation for the residual network is as follows:

$$y = L(x, \{P_i\}) + P_s x \qquad (1.7)$$

P_s is utilised for matching the dimensions of the previous layer with the next layer, and the addition of the identical mapping coefficients controls the degradation of gradients. The L(x, Pi) function represents a network with many convolutional layers. On each of the feature maps, the element-wise addition is performed channel-by-channel. The eq 1.7 layer is an intermediary step between the first two convolutional sections that transfers data from the lower layers to the higher ones. Each successive layer takes as input the prediction values computed by the layers before it, and the residual connections help link those values to the layer to which they belong. It is the job of the residual function to determine the real value and compare it to the predicted value. If x is equal to or greater than the real value, the residual function will be zero, and the derivative will be larger. Batch normalisation is also performed in the block, along with the residual connections. The goal is to bring the numbers down to a level where the derivatives aren't so little that we can just ignore them, even though they don't add much. Everything that has been said about the layers and parameters is what the Optimised Hyper Parametric (OHP) method produces. The tuning aids in building the required number of layers after each iteration to derive meaningful interpretations from intermediate and high-level inputs, which in turn form a pool of weighted probabilities. In order to classify the images during testing and validation, these probabilities are utilised. In the section that follows, we will discuss how a surrogate model is used to achieve the OHP.

1.4 Bayesian Network Model

Building on the foundation laid by the convolutional neural networks (CNNs) discussed in Section 1.3, the EmoNet improved the efficacy of emotional interpretation. Each hyperparameter must be manually adjusted when using manual parameter selection. A sophisticated system for selection is produced by this procedure. By identifying parameters based on the True objective function, the Bayesian Optimisation Model was used to find the network hyper parameters. This model chooses the hyperparameter to assess the actual objective function and constructs the network based on the probability of determining the objective function (Of). Given an infinite amount of resources, it would be impossible to compute the objective function at every single point. Consequently, a surrogate model (S_m) was trained on the hyper parameter and real objective function pair to represent the objective function.

$$S_m = P(H_p | O_f) \qquad (1.8)$$

Table 1.1: Hyper Parameter Table

Hyper Parameter	Value
Image size	128 x 128 x 3
Networks Layers	74
Initial LearnRate(Lr)	1.00E-03
Gradient Threshold	’l2norm’
Image size	128 x 128 x 3
Learning Rate Scheduler	Piecewise
Number of Filters(layer)	8,16,32,64,128
Padding Direction	’right’ (1,1)
Validation Frequency (Vlf)	500
Stride	(1,1)
Optimiser	SGDM
Regularization Function	l2Regularization
Max Epochs	40
Verbose Frequency	50
Mini BatchSize (Bs)	32
Shuffle	’Every-epoch’
Learning Drop rate	60 iterations

1.5 Experimental Design, Data Collection, Analysis, and Discussion

1.5.1 Experimental Setting

The experiments were trained and validated using Matlab 2019b. A system with an 8GB RAM and an NVIDIA GeForce MX150 with 4GB of memory was used to do the initial network modelling. The processor was an Intel Core i5 8250U running at 1.8 GHz. Workstations powered by Intel Xeon E3 processors, equipped with NVIDIA Gforce GTX graphics cards and 32 GB of RAM, were used for training, testing, and validating the network with benchmarked datasets.

1.5.2 Collection of information

There are grayscale pictures of 48x48 facial expressions in the FER2013 Dataset on Kaggle. By using an automated registration process, the faces are now evenly spaced and virtually centred in all of the images. The objective is to sort every face into one of seven categories according to the emotion conveyed by its expression: angry, disgusted, afraid, happy, sad, surprised, or neutral. On the training set, you may find 31,759 instances. The number of photos in the test dataset is 3,813 while the number of images in the validation dataset is 3,589. With 34,512 photos of 100 Indian actors and actresses from more than 100 films, the Indian Movie Face Database (IMFDB) is a huge, free face database. There is a great deal of variation in size, expression, and posture across the photos because they were all physically extracted from video frames. One hundred and thirteen images of ten Japanese female models posing with seven different expressions on their faces are included in the JAFFE collection. There are 123 individuals represented in the 717 photos in the CK+ collection. The 2018 Deep Facial Expression Survey by Li details the most up-to-date developments and commonly utilised databases. Despite the abundance of datasets that can be utilised for emotion analysis, there is a dearth of classroom-ready datasets that feature faces of Indian descent. In addition to being limited to the fundamental emotions, learning environments can be extended to include several classifications, each of which can affect the exact evaluation of a student's involvement in a certain class. The major goal of this research is to use the current dataset for individuals of Indian descent in order to elaborate the sentiments. The integration of foundational lessons, engagement recognition, and emotions centred around learning allows for this to be achieved. The DAISEE , iSAFE, and ISED databases are utilised for the analysis of students in an Indian classroom. Affective computing dataset details are provided in Table 1.2.

Table 1.2: Emotional Analysis Dataset

S. No.	Name of the author	Name of the dataset	Database details	Affective states	Emotions enlisted
1	Setty et al.	IMFBD dataset	100 movies videos	Posed	Fundamental emotions
2	Dhall et al.	AFEW database	957 videos	Temporal data	Fundamental emotions
3	Happyet al.	ISED database	428 video data from 50 participants	Collected from the wild	Fundamental emotions
4	Sapinski et al.	Multimodal database	560 images with 16 subjects	Learning centred	Learning based emotions
5	Bian et al.	Spontaneous expression database	30184 images from 82 students	Online learning	Learning based emotions
6	Daisee et al.	DAISEE database	9068 videos from 112 users	Engagement recognition	Learning based emotions
7	Lyons et al.	JAFFE	7 different expressions consisting of	Acted expressions	Fundamental emotions

			213 images		
8	Goodfellow I J et al.	FER-2013	35685 images	Collected from the wild	Fundamental emotions
9	Shivendra Singh et al.	iSAFE	395 videos from 44 volunteers	Acted expressions	Fundamental emotions
10	Kaur et al.	Student Engagement Database	78volunteers with 5mins video	Head pose and eye Gaze	Behavioural Cues

1.5.3 Recognising Facial Expressions using EmoNet

Results for True Positive, False Positive, True Negative, and false Negative are calculated by randomly selecting images from the testing data and uploading them to the network. The numbers displayed below were generated using the confusion matrices that were computed for each dataset. We estimate the F1-score, sensitivity, specificity, accuracy, and precision for each category individually after we run the analysis. The results of evaluating the network's performance using various datasets are summarised in Table 1.3. Incorrectly classified instances of the Neutral and Fear classes are rare in the dataset. It is important that the Sad, Neutral, and Fear faces be visually distinguishable when used for evaluation and training purposes. Over time, EmoNet's small-sized filters and kernels learn to distinguish between basic, intermediate, and advanced features. The network's capacity to learn and comprehend critical aspects has been improved by these and the other layers mentioned in table 1.1. The model is now better equipped to categorise the test data thanks to these adaptive acquisitions. Not only that, but the weight-sharing technique has also increased the network's accuracy by 2%. The significant size of the retrieved trained feature-map is one factor that enhances the classification accuracy. We stop training the network when we see no change in the variation of the loss function. The maximum number of epochs for this model is 40 because of the repeated early halting method employed during training. In order to end training at the optimal epoch—that is, one in which the L_r has not changed—early termination was chosen. Hyperparameter optimisation and fine tuning were employed when the parameters were being selected and integrated into the network. The original model had 68 layers, but in order to obtain the best possible performance, it was decreased to forty layers in the final version.

Table 1.3: Metric Table based on EmoNet

Data	EMOTIONS	Acc%	Pre%	Sen%	Spe%	F1%
FER2013	ANGER	93.3	75.4	80.7	95.4	77.9
	DISGUST	99.5	98.4	80.3	100	88.4
	FEAR	91.8	70.9	76.0	94.5	73.4
	HAPPY	94.7	92.9	89.3	97.1	91.1
	NEUTRAL	91.9	82.6	77.3	95.7	79.8
	SAD	91.2	74.5	77.6	94.2	76.1
	SURPRISE	96.7	87.0	87.4	98.1	87.2
CK+	ANGER	98.1	92.2	95.0	98.6	93.6
	DISGUST	97.8	91.3	92.3	98.6	91.8
	FEAR	97.2	88.2	91.1	98.1	89.6
	HAPPY	99.0	97.6	94.2	99.7	95.9
	NEUTRAL	97.2	88.2	91.1	98.1	89.6
	SAD	97.5	89.5	92.4	98.3	90.9
	SURPRISE	98.2	94.8	92.9	99.1	93.8
	ANGER	100.0	100.0	100.0	100.0	100.0
	DISGUST	99.4	96.0	100.0	99.4	98.0

JAFFE	FEAR	98.6	93.1	96.4	98.7	94.7
	HAPPY	99.4	100	96.2	100.0	98.0
	NEUTRAL	100.0	100.0	100.0	100.0	100.0
	SAD	99.4	100.0	96.0	100.0	98.0
	SURPRISE	99.4	96.0	100.0	99.4	98.0
IMFDB	ANGER	94.6	69.5	79.4	96.2	74.1
	DISGUST	92.0	71.5	80.6	94.0	75.8
	FEAR	99.0	99.2	92.3	99.9	95.6
	HAPPY	89.5	83.0	76.8	94.2	79.8
	NEUTRAL	86.0	81.3	75.0	91.4	78.0
	SAD	92.3	67.2	75.6	94.7	71.2
	SURPRISE	96.9	71.3	79.7	98.0	75.3

*Acc: Accuracy, Pre: Precision, Sen: Sensitivity,
Spe: Specificity, F1: F1-Score.*

Reducing layers using methods that included trial and error allowed us to achieve our target of creating an emotion detection framework with a network that was more compact than existing models. There are a total of 709 layers in the DenseNet201 network, 177 in the ResNet50 network, 72 in the ResNet18 network, 372 in the ResNet101 network, and 41 in the VGG19 network. Given that there is a wide variety in the number of layers in models and the dimensions of the input network images: Even though it uses fewer layers and smaller face photographs than previous models, EmoNet still manages to produce remarkable results. Thanks to cross-database learning, the network is now more adaptable to fresh data, which enhances its prediction capabilities in hidden environments. Emonet is just as productive in both expected and unexpected settings. Writers and the still images that were used to assess their accuracy are listed in Table 1.4. The results and procedures used on these data sets are detailed in the table. In spite of obstacles like face occlusion, location displacement, and data proximity, EmoNet achieves high percentages of accuracy when classifying new test data. The ambiguity of facial expressions, bias reduction, and picture generalisation were the primary goals of building EmoNet. On two-dimensional data, EmoNet has performed satisfactorily in terms of accuracy and F 1 score; yet, there is need for improvement. The vanishing gradient problem becomes more noticeable as the network size increases, even if CNNs work effectively on shallow networks.

- Even with dropout layers and learning rate monitors, the error rate increased due to disappearing gradients.
- CNNs provide a challenging operation when it comes to optimising the inputs and parameters of the network.
- The need to transform feelings into active participation in the classroom prompted the gathering of up-to-date data and the regular tuning of a network specifically for them. Performance is lower with more recent training.
- A buildup of features and discrepancies in the output was the consequence of adding additional layers.
- The efficiency of the network was reduced as the size of the image increased.
- Higher dimensional processing is necessary for colour image processing.

Table 1.4: Comparative Analysis of methods in FER

Dataset	**Author**	**Accuracy%**
FER2013	Tang et.al [1]	71.2
	Devries et al.[2]	67.2
	Zhang et al.[3]	75.1
	Guo et al. [4]	71.3
	Kim et al.[5]	73.7
	Pramerdorfer et al. [6]	75.2
	Sang et al.[7]	75.2

	EmoNet	83.6
JAFFE	Liu et al. [8]	91.8
	Hamester et al. [9]	95.8
	Xie [10]	94.75
	Fathallah et al.[11]	94.75
	Deepak et al. [12]	93.24
	EmoNet	99.5

1.5.4 ResNet with a Shallow Domain for Intent Detection

Datasets from ISED, iSAFE, and Daisee were analysed with the use of the Shallow Residual Network. All of the faces that make up the network are of Indian origin. The authors made these images for an online learning environment, thus it was only natural to use the same network to analyse the online courses offered during the pandemic. We train the network to identify True Positives, False Positives, True Negatives, and True Negatives by randomly selecting images from the testing set. Photos that have dimensions of 128 × 128 ± 3 are sent to the Residual network. The network is used for image classification after randomly selecting images from the test data. Based on which the confusion matrix is derived. There are a total of seven classes trained using the 508 photos found in the ISED and iSAFE datasets. Likewise, 5295 photos are used as training data from the Daisee dataset. The network can learn features from the beginning when training on datasets with a learning rate set at 1.00E-04. The network optimises for weight vector dispersion and employs a piecewise learning rate scheduler to enhance the learning rate by reducing it. The size of the mini-batch is set to 32 after each epoch. Figure 1.3 displays two networks that skip. Additional 74 layers were added by residual layers and skip networks. Yet, it lacks the layers of ResNet101, ResNet50, and ResNet 32. Activation of the network occurs in ReLu layers, whereas pooling layers reduce the dimensionality of recovered features. After many rounds of tweaking the entire network according to OHP performance, the layers were meticulously placed. Finding an optimal solution was a lengthy process and the training accuracy for Deep Networks with multiple layers and residual networks was below 70%. The issue with gradient descent and feature pooling caused this.

This optimal network for emotional comprehension is developed using multiple classes. The F1 score is used to quantify the algorithm's efficacy because the presented data is uneven. After 500 iterations, which give suggestions about the training progress, and simultaneously with the measurement of the loss function, the training data was validated. The training samples are chosen at random after each epoch to avoid over- or under-fitting. Figure 1.5 displays the model's efficiency index as a function of training accuracy (T) and validation accuracy (V) as iteratively applied to the Residual Network model. Also included in this picture are the results of the training and validation processes. Because it checks the model's training efficacy periodically, validation frequency (V_f) is an important parameter. Vf is initialised for this model after 500 iterations. The validation loss is reduced and learning rate is increased by averaging the scores from each layer using the posterior class probabilities. Table 1.5 and Table 1.6 show the details of the confusion matrix (CM) that was obtained across the two datasets because the classifier used images that weren't used during training the network. Every class also gets its own F1 score, recall, accuracy, precision, and sensitivity along with the CM. Results show that emotion recognition and detection have been greatly affected by the residual connections introduced to CNNs. The matrix provides a thorough understanding of each class's performance in the data since it includes the proportion of False Positives and False Negatives for each class.

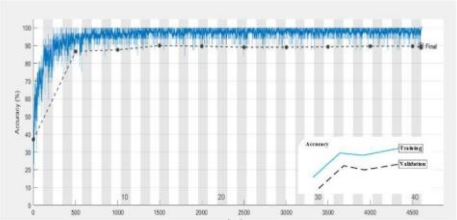
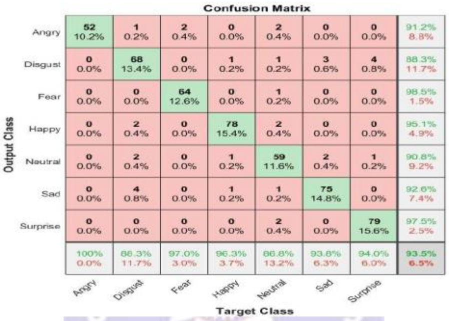


Fig. 1.5: Training and Validation Accuracy for Daisee dataset

In order to decrease the classification error rate, training weights are inputted into subsequent network parts, allowing the network to adapt and categorise new data. Table 1.4 displays EmoNet's performance metrics on the dataset, including Accuracy, Precision, Recall, Specificity, and F1-score. In terms of accuracy, Emonet performs better than the other methods listed in Table 1.4. An individualised and holistic approach improved the model's performance, even though a number of the dataset's individual classifications misclassified emotions. In order to train the network, a total of forty epochs were used. There have been multiple rounds of trial and error with the network's hyperparameters. Also, the amount of iterations may be determined conclusively using the Loss function. In the cross-validation process, 70% of the data was used for training and 30% for validation purposes inside the dataset. Network performance enhancement when assessing overfitting. At each V_f, the network checks the data for validity. The network's consistency is demonstrated by the significant correlation between validation accuracy and training accuracy. The model's validation accuracy on Daisee was 90.8% and on the combined ISED and iSafe datasets it was 88.76%, all thanks to this training. Emotions in these datasets rely heavily on residual connections as a learning structure.

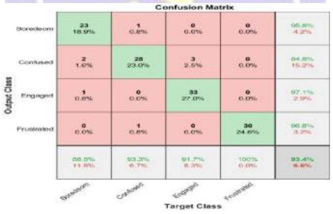
Table 1.5: Confusion Matrix for ISE and iSAFE Dataset



Confusion Matrix

Output Class \ Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	
Angry	52 10.2%	1 0.2%	2 0.4%	0 0.0%	2 0.4%	0 0.0%	0 0.0%	91.2% 8.8%
Disgust	0 0.0%	68 13.4%	0 0.0%	1 0.2%	1 0.2%	3 0.6%	4 0.8%	88.3% 11.7%
Fear	0 0.0%	0 0.0%	64 12.6%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	98.5% 1.5%
Happy	0 0.0%	2 0.4%	0 0.0%	78 15.4%	2 0.4%	0 0.0%	0 0.0%	95.1% 4.9%
Neutral	0 0.0%	2 0.4%	0 0.0%	1 0.2%	59 11.6%	2 0.4%	1 0.2%	90.8% 9.2%
Sad	0 0.0%	4 0.8%	0 0.0%	1 0.2%	1 0.2%	75 14.8%	0 0.0%	92.6% 7.4%
Surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.4%	0 0.0%	79 15.6%	97.5% 2.5%
	100% 0.0%	88.3% 11.7%	97.0% 3.0%	96.3% 3.7%	86.8% 13.2%	93.8% 6.3%	94.0% 6.0%	93.5% 6.5%

Table 1.6: Confusion Matrix for Daisee Dataset



Confusion Matrix

Output Class \ Target Class	Boredeom	Confused	Engaged	Frustrated	
Boredeom	23 18.9%	1 0.8%	0 0.0%	0 0.0%	95.8% 4.2%
Confused	2 1.6%	28 23.0%	3 2.5%	0 0.0%	84.8% 15.2%
Engaged	1 0.8%	0 0.0%	33 27.0%	0 0.0%	97.1% 2.9%
Frustrated	0 0.0%	1 0.8%	0 0.0%	30 24.6%	96.8% 3.2%
	88.5% 11.5%	93.3% 6.7%	91.7% 8.3%	100% 0.0%	93.4% 6.6%

Table 1.5 shows that the network generated large outcomes using the remaining connections, as indicated by many parametric findings. Following are the typical outcomes for the ISED and iSafe datasets: The EmoNet, a notable classifier for the database, achieved an accuracy of 93.5%, an error rate of 6.5%, a sensitivity of 92.4%, a specificity of 95.6%, a precision of 93.25%, a false positive rate of 1.44%, and a Kappa's value of 0.65. The inter-observer ability is measured using Cohen's κ. The dependability of a data classification system. As indicated in equation 1.9, the inter rater is computed using the relative observed score and the likelihood. It is easy to compute the agreement using the provided observer accuracies and the network-acquired categorization accuracy

$$K\,appa(\kappa) = \frac{p0 - pe}{1 - p_e} \tag{1.9}$$

Using the Daisee dataset, we evaluated the Lightweight ResNet and found: An outstanding classifier for the database, with a sensitivity of 93.6%, specificity of 97.8%, precision of 93.37%, false positive rate of 2.18%, and an error rate of 6.6%. Kappa's coefficient is 0.825. The proposed method used large training images and optimised layers to enhance the shallow network's ability to learn, understand, and gather data. To achieve accurate classification, network-learned features are utilised. The network is improved by minimising disappearing gradient and overfitting by using activation functions to produce feature maps after each layer. The results of using emotional identification models in the ISED, iSafe, and Daisee databases are displayed in Table 1.5. Compared to CNN and earlier definition approaches, the new methodology performs better. By averaging recall and accuracy, F1-measures the classifier's performance. Both the accuracy of the data and the performance of the classifier with various datasets are examined by these measures. When tested against the Emonet, this model achieved an accuracy of 90.53% on ISED data and 91.78% on the iSAFE database. This was in contrast to the Emonet, which achieved 83% and 85%, respectively. Table 1.5 provides the results of the comparative examination of the parametric evidences. The table displays the results of the class's performance in relation to different emotions. The network's performance on the four behavioural classes and seven basic emotional classes is one of its most notable features. We compare the network to existing models that have used databases. The ROC is a measure of how well a classifier predicts future outcomes. Sensitivity and specificity are compared using this metric. The ROC plot calculates the true positive rate and the false positive rate, which are displayed on the Y and X axes, respectively. ROC indicates the true positive rate. Figure 1.8's top left corner shows one true positive and zero false positives. An AUC calculation takes into account both true and false positives. A larger AUC indicates a more effective classifier. The "upward gradient" of the curve illustrates the best way to increase true positives and decrease false negatives.

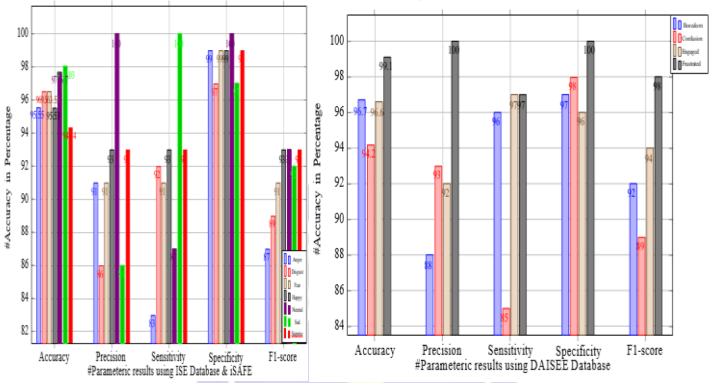


Table 1.7: Metric Table based on Shallow ResNet

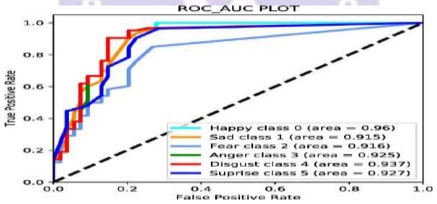


Fig. 1.6: ROC_AUC Plot on iSafe and Ised Dataset

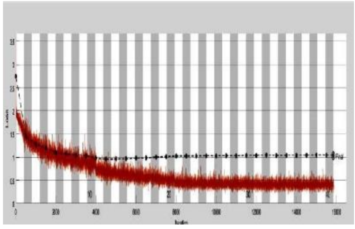


Fig. 1.7: EmoNet Loss Estimate

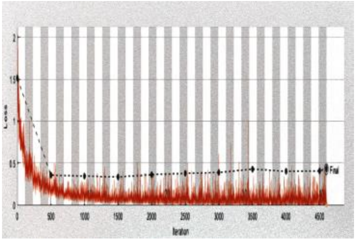


Fig. 1.8: Shallow ResNet Loss Estimate

Figure 1.8 shows the area under the curve (AUC) for each class; the Happy class ranks high for correct categorization, while the Sad class ranks low for accurate diagnosis. But with an average accuracy of 92% in a video scenario, the model truly comes into its own. Classes linked to feelings of fear and sadness had the highest number of false positives, as shown in Table 1.7. Differentiating between classes, such melancholy and fear, disgustand surprise, isn't always easy, even though several frames can be used to predict a face's expression of a certain emotion. One optimisation approach that was used to create deep learning neural systems is Stochastic Gradient Descent (SGD). A key component of any optimisation strategy is iteratively estimating the error for the present state of the model. Choosing an error function, often called a loss function, to predict the model's loss is essential for updating the weights and reducing the loss on the next assessment. Whether you're doing regression or classification, neural network models learn the mapping from inputs to outputs by looking at examples. The loss function you employ should be suitable for the predictive modelling task at hand. Also, make sure the loss function you choose is compatible with the output layer's setup.Shallow ResNets and EMONet are used to construct the Multi-Class cross entropy loss. Over-fitting issues, as discussed earlier, are caused by the Validation loss rate in fig 1.9, which keeps going up over 1. The EmoNets Architecture use of ResNets to get past these problems. Figure 1.10 clearly shows the decrease in loss rate. The networks achieve a 40-epoch training completion rate with a loss rate of 0.5 in less than 5,000 iterations. By utilising cross validation and multi-Class cross entropy loss, the network has enhanced learning features while maintaining the necessary momentum for under-fitting and over-fitting.

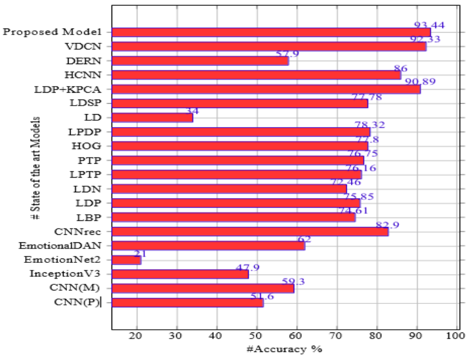
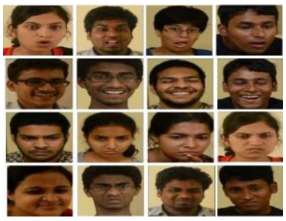


Table 1.8: Comparison of the proposed model with the state-of-the-art methods

Convolutional Neural Network(Plain)[189], Convolutional Neural Network(Modified) [189],Inception V3 [189], EmotionNet 2 [189], EmotionalDAN[189] ,CNNrec [190] , LBP (Local Binary Pattern) [191], LDP (Local Directional Pattern) [191], LDN (Local Directional Number) [191], LPTP (Local Directional Ternary Pattern) [191], PTP (Positional Ternary Pattern) [191], HOG (Histogram Of Gradients) [191], LPDP (Local Prominent Directional Pattern) [191], Landmark Detection [192], Local directional-structural pattern [29] LDP+KPCA [193], Hybrid CNN[30], Deep Engagement Recognition Network [31], Very Deep Convolutional Network [32].

The results of the network are contrasted with those of the most popular and extensively trained networks. You can see how the suggested model stacks up against the other models out there in Table 1.8. Models trained on the ISED, iSAFE, and Daisee datasets were subject to this evaluation. These static photos are taken from the Ised Database (Fig. 1.9). As shown in Figure 1.10, the seven network-derived classification classes are given as bounding boxes along with the emotions that correspond to them.



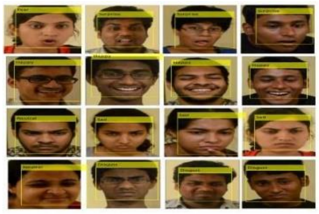


Fig. 1.9: Classroom Image **Fig. 1.10: Classified Emotion**

The research into the network's functioning yielded three observations. The first thing to notice is that the reported 0.265 training error rate has significantly decreased degradation concerns. Learning efficiency is enhanced by a decrease in training error, which is caused by the appropriate depth of the network. The second improvement is a 30% reduction in the training and validation time complexity, which can be seen in Figure 1.3's identification connections. As a third point, the network may identify optimal solutions since it uses the SGD processor. Despite the network's shallowness, the gradient descent technique may train it on smaller batches, enabling it to create numerous layers of features. At its heart, the classification unit is these traits, which are in charge of producing accurate results by firing the right neurons to generate probability on the Weighted layers. Insight into the network's validation output can be gained during training through Ten-Fold Cross validations. Additional data regarding the efficiency of the network is provided by this. The network has been optimised for greater performance by including the Optimised Hyper Parametric Settings described in Section 1.5.

1.6 Conclusion

The main goal of this project is to create a model of an emotion identification system for faces that can correctly detect different types of emotions. This paradigm can be utilised by systems that recognise facial emotions. The model's dependence on Deep Neural Networks is a major contributor to the classification system's accuracy. Exposing the model to data from many databases improves its performance in training and classification. The model can handle faces with different sizes, illumination, and registration angles. Progressive scaling makes it easier to assimilate data of different sizes by doing away with the need to resize data at the conclusion of each epoch. Despite the availability of cutting-edge networks that could facilitate transfer learning, issues including image orientation, size, cross-dataset compatibility, and resource utilisation must be addressed. EmoNet can function effectively with 40 layers, in contrast to other models that have 100 or more layers. An 8% improvement in classification accuracy on FER2013 and a 0.2% improvement on JAFFE were achieved through training and validation on numerous datasets using tiny filters (3x3,5x5). In less than 2.77 seconds, this model successfully classified 3589 unknown photos into seven separate groups. The original intent of EmoNet was to classify different types of facial expressions. This is the reason why the network has an average uptime of 90%. To solve the vanishing gradients problem, optimise the data and parameters, convert emotions into classroom engagement detection, and work with higher spatial and volumetric dimensions of images, this chapter intends to evaluate the efficacy of adding residual layers to the current CNN model. To monitor how actively students engage with a learning platform, the residual connections of the shallow network were constructed from the ground up. This study delves into both behavioural and emotional categories. This study uses students' facial traits to predict and classify real-world photos, in order to find out how accurate the proposed method is. To improve the system's learning and classification capabilities, this method employs residual networks to strengthen links between layers. To test how well the model works, cross-validation is employed. This network, which was trained using three datasets from India, has the potential to identify students' emotional and behavioural intentions. All told, there are 74 layers in the network. A learning model can efficiently check photographs during training with an average rate of 86.87% using the shallow network. We compare the model to other newly developed models and use real-world data to evaluate its detection performance. The effectiveness of classification accuracy on the datasets has been significantly enhanced by using OHP to integrate the network.

References

[1] Yichuan Tang. Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239, 2013.

[2] Terrance Devries, Kumar Biswaranjan, and Graham W Taylor. Multi-task learning of facial landmarks and expression. In 2014 Canadian conference on computer and robot vision, pages 98–103. IEEE, 2014.

[3] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In Proceedings of the IEEE International Conference on Computer Vision, pages 3631–3639, 2015.

[4] Yanan Guo, Dapeng Tao, Jun Yu, Hao Xiong, Yaotang Li, and Dacheng Tao. Deep neural networks with relativity learning for facial expression recognition. In 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pages 1–6. IEEE, 2016.

[5] Bo-Kyeong Kim, Suh-Yeon Dong, Jihyeon Roh, Geonmin Kim, and Soo-Young Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 48– 57, 2016.

[6] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903, 2016.

[7] Dinh Viet Sang, Nguyen Van Dat, et al. Facial expression recognition using deep convolutional neural networks. In 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pages 130–135. IEEE, 2017.

[8] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1805–1812, 2014.

[9] Dennis Hamester, Pablo Barros, and Stefan Wermter. Face expression recognition with a 2-channel convolutional neural network. In 2015 international joint conference on neural networks (IJCNN), pages 1–8. IEEE, 2015.

[10] Siyue Xie and Haifeng Hu. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Transactions on Multimedia, 21(1):211–220, 2018.

[11] Abir Fathallah, Lotfi Abdi, and Ali Douik. Facial expression recognition via deep learning. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pages 745–750. IEEE, 2017.

[12] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. Pattern Recognition Letters, 120:69–74, 2019.