Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

MoodScope: Navigating Emotions through Convolutional Neural Networks

Vansh Sharma[1] , Srishti Sharma[1] , Waseem Qureshi[1] , Nadeem Anwar[2]
[1]Computer science and Information Technology, Meerut Institute Of Technology ,Meerut, India ,
[2]Assistant Professor , Information Technology , Meerut Institute Of Technology ,Meerut, India ,

Abstract:- — Facial expressions play a crucial role in human communication, presenting a longstanding challenge for emotion recognition. Leveraging advancements in technology, this paper explores the feasibility of detecting emotions in facial images. We introduce a Deep Learning model based on Convolutional Neural Networks (CNNs) to recognize facial emotions. The proposed method involves three key steps: Face Detection, Feature Extraction, and Emotion Classification, aiming to identify seven cardinal human emotions—Anger, Fear, Disgust, Joy, Neutrality, Sadness, and Surprise.

The use of Convolutional Neural Networks is pivotal in achieving high accuracy and performance in Facial Emotion Recognition. This technology holds wide applicability in various domains, including student predictive learning, forensics, and social media platforms. The anticipated impact of Facial Emotion Recognition extends to enhancing communication and understanding emotional nuances in diverse applications.

Keywords*: CNN, Multiview, Gaussian, Laplacian, FER 2013, CK+ dataset.*

1. Introduction

We humans have a set of emotions, which we convey through basic facial expressions. Hence they can be considered as a non-verbal means of communication. Emotions are expressed through a number of means like hand and body movements, words and facial expressions[1]. We humans can easily determine the emotion of a person just by looking at them. But the task here is to train our machines to do so[2].

Hence in a way, we are providing sight to our machines. We are providing them the ability to perform human-like tasks. Due to the advancements in computer vision, this task is no longer so difficult[3].

[4] Talk about the idea of multi-modal FER and describe the advantages of combining data from several sources. Give an explanation for the inclusion of many modalities, including speech, body language, and physiological signs.

The review should cover the experimental results presented in the paper, demonstrating the effectiveness of the proposed gradient-based learning approach. This could include improvements in recognition accuracy and processing speed[5].

Human facial emotion recognition has been a vital topic of discussion over decades. To identify the emotional state of a person can be a critical task. In the past few years, vast research has been made on facial emotion recognition[2]. The major objective of facial emotion recognition is to identify human emotion through facial expressions and images. However, the task is to provide emotion detection with high accuracy. Different people have different ways of expressing the same emotion which makes the task more challenging. To achieve better results, CNN model of Deep Learning is used[6]. A Convolutional Neural Network (CNN) specializes in image processing, excelling at recognizing intricate patterns within images for tasks like classification and facial recognition[7]. This model basically teaches the computer what humans can do by training it with large datasets[8]. Hence it minimizes human effort. Through emotion recognition, we can determine the emotional state of anyone in different situations. Facial emotion recognition involves three steps: face detection, feature extraction

and emotion classification. When we are detecting emotions in real time two aspects that we have to take care of are accuracy and efficiency.

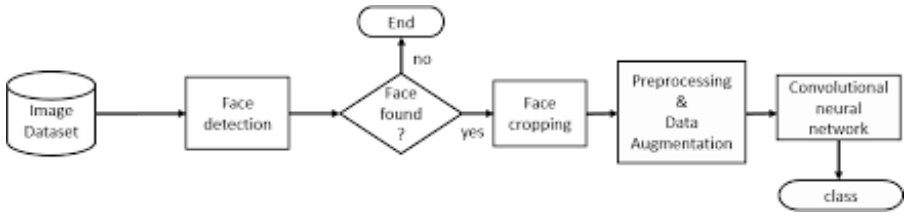


Fig1: Overview of the Workflow

2. Literature Review

The seminal paper by LeCun, Bengio, and Hinton [9] examines deep learning and provides a detailed overview of its concepts, methods, and applications. The authors elucidate hierarchical representations with less deep roots, and highlight the effectiveness of these models in areas as diverse as computer vision, natural language processing, and speech recognition They to eliminate Furthermore, it clarifies the forces and highlights the important role of back propagation and stochastic gradient descent in training effectiveness in deep correlations. By combining theoretical approaches and empirical evidence, this landmark work establishes deep learning as a cornerstone of modern artificial intelligence, paving the way for unprecedented advances in machine learning in.

Viola and Jones with their paper "Robust real-time face detection" [10] published in the International Journal of Computer Vision presented a landmark work on computer vision. The authors presented a new method for face recognition in time actually emerged, becoming one of the most widely used methods in the industry. Their method combined neck-like features with cascading classifiers trained by AdaBoost, and enabled better identification even under harsh conditions such as lighting conditions and complex backgrounds achieved by the Viola-Jones algorithm incredible speed and accuracy, and introduced facial recognition , surveillance systems, human- computer interaction and paved the way for many applications and subsequent research focused on their work, refining and extending the algorithm in its power became a cornerstone in the development of modern facial recognition technology.

Sutskever, Martens, Dahl, and Hinton [1] examine the critical elements of originality and speed in deep learning programs and examine their importance and impact. They emphasize the critical role these features play in the efficiency and effectiveness of deep learning systems. Through their analysis, they shed light on the profound effects of onset and speed on the properties of matching neurons, and highlighted their ability to significantly improve training efficiency and model performance on the snow. By providing empirical evidence and theoretical perspectives, this review contributes valuable knowledge to the field, providing guidelines for successfully introducing neural network theory and implementing applications for curriculum development ideal in deep learning programs.

Krizhevsky, Sutskever, and Hinton's [11] seminal work on ImageNet classification with deep convolutional neural networks represents a pivotal moment in the development of deep learning. Their pioneering research demonstrated the effectiveness of the deep transformation process for processing large sets of images. The use of multivariate deep neural networks achieved remarkable improvements in their image classification accuracy, establishing new dimensions in the field .

Simonyan and Zisserman [12] further advanced the field of image recognition with their introduction of very deep convolutional networks. By increasing the depth of convolutional neural networks, they aimed to improve the discriminative power and hierarchical feature representation of the models. Through extensive experimentation and analysis, they demonstrated the efficacy of their approach in achieving state-of-the-art performance on large-scale image recognition tasks. This work highlighted the importance of network depth in learning complex representations and inspired subsequent research in designing deeper architectures for various computer vision tasks.

Russakovsky et al. (2015) provided a comprehensive overview of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [13], offering valuable insights into the progress and achievements in the field of computer vision. By organizing a large-scale competition focused on image classification and object detection tasks, they facilitated collaboration and benchmarking among researchers and practitioners. The ILSVRC dataset and competition served as a catalyst for advancing the state-of-the-art in visual recognition, driving innovation in deep learning algorithms and methodologies. This paper remains a foundational reference for understanding the landscape of large-scale image recognition challenges and the evolution of benchmark datasets in computer vision.

He, Zhang, Ren, and Sun (2016) introduced deep residual learning as a novel architecture for image recognition [14], addressing the challenge of training very deep neural networks. By introducing residual connections that enable the network to learn residual mappings, they proposed a solution to the degradation problem encountered when increasing network depth. Through extensive experimentation, they demonstrated that deep residual networks could achieve superior performance compared to traditional convolutional neural networks, effectively tackling image classification tasks with unprecedented accuracy.

Szegedy et al. (2015) proposed a pioneering approach in "Going Deeper with Convolutions" [15] introducing inception modules and advancing the design of deep convolutional neural networks (CNNs). By employing a combination of convolutional filters with different receptive fields within a single layer, they aimed to capture complex patterns at multiple scales, effectively enhancing the expressive power of the network. Through their inception architecture, they achieved significant improvements in both efficiency and accuracy, demonstrating the effectiveness of their approach on various image classification tasks. This work laid the groundwork for the development of more sophisticated CNN architectures and inspired further research into multi-scale feature extraction and hierarchical representation learning.

Abadi et al. (2016) introduced TensorFlow [8], a powerful and flexible open-source platform for large-scale machine learning tasks. By providing a scalable framework that supports distributed computing across heterogeneous systems, TensorFlow enables researchers and practitioners to efficiently train and deploy deep learning models.

Zeiler and Fergus (2014) presented a seminal work on visualizing and understanding convolutional networks [16], shedding light on the inner workings of these complex models. By employing visualization techniques such as deconvolutional networks and activation maximization, they provided insights into the hierarchical feature representations learned by convolutional neural networks (CNNs). Their study elucidated the role of different layers in capturing both low-level and high-level features, facilitating a better understanding of how CNNs process and interpret visual information.

Cireşan et al. (2010) proposed deep, big, simple neural networks for handwritten digit recognition[17], introducing a straightforward yet effective approach to leveraging deep learning techniques for pattern recognition tasks. By designing deep neural networks with a simple architecture and large amounts of trainable parameters, they achieved state-of-the-art performance on handwritten digit recognition benchmarks. Their work demonstrated the potential of deep learning in tackling real-world classification problems, paving the way for the widespread adoption of deep neural networks in various domains beyond computer vision.

The function of multimedia components such as pictures, sounds, and maybe video in expressing emotions during teleconferences is probably covered in the study. Examine the ways in which integrating various multimedia components can improve the communication's emotional context[18].

Ekman addresses criticisms of the idea of universal facial expressions of emotion, especially those from James Russell, in his paper "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique," which was published in Psychological Bulletin (Vol. 115, No. 2)[19]. The strong evidence that Ekman's research provides for the cross-cultural constancy of specific facial expressions linked to fundamental emotions makes it noteworthy.

3. Proposed Methodology

For facial emotion recognition (FER), the first step is the preprocessing of images. Feature extraction and feature classification on the pre-processed are the next steps that are achieved through Convolution Neural Network (CNN)[6][11]. The activation function employed in this study is the Rectified Linear Unit (ReLU)[1].

3.1 Preprocessing of Data

The input images may contain noise and vary in illumination, size, and colour. So, Preprocessing operations will be performed on the images to obtain more accurate and faster results. Preprocessing is basically used to convert the raw dataset into suitable form. This step mainly includes refining of the dataset. This includes resizing, grayscale conversion and normalizing the data images so that the emotions can be classified in a better way.[6][10]

Resizing: The images are resized to smaller representation so that it requires less storage and less transmission time.

Gray scaling: It is a technique of creating an image whose pixel value relies on the amount of light present in the input- coloured image. Gray scaling is done because coloured images are challenging for algorithms to process.

Normalization: It is done to remove the differences in illumination and to obtain the improved image.

Preprocessing will be done using digital filters, which will remove the unwanted information like noise and preserve the desired signal. It will be done using smoothing filters like Gaussian Filters and Laplacian Filter.

i. Gaussian Filter

A 2-D Convolution operator utilized for image blurring and noise elimination. Characterized by a Gaussian Hump-shaped Kernel, as depicted in Fig.2, the Gaussian Filter assigns higher weights to pixels near the center, gradually decreasing towards the edges.

This design allows the filter to perform a convolution operation, leveraging its separability property in the Cartesian plane for computational efficiency. The primary purpose of Gaussian Filters lies in their ability to smooth images effectively.

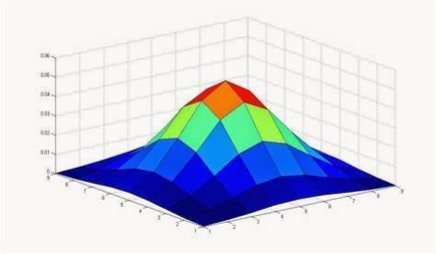


Fig2: Gaussian Hump shaped Kernel

ii. Laplacian Filter

The Laplacian filter, rooted in the second spatial derivative of an image, serves the purpose of enhancing image sharpness. The formulation of the Laplacian filter is expressed by the following equation:

$$f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \quad (1)$$

$$\frac{\partial^2 f}{\partial x^2} = f(x+1,y) + f(x-1,y) - 2f(x,y) \quad (2)$$

3.2 Face Detection

The primary step of any Facial Emotion Recognition system is face detection[10]. Face Detection is done using Haar cascades which is also known as Viola Jones Detectors. Haar Cascades are classifiers that detect an object in a trained image or video. They are trained over a set of positive and negative sets of facial images.

Haar cascades have proven to be an efficient and accurate method of object detection in images.

Three dark areas of the face such as eyebrows are detected by Haar characteristics. And, two dark areas of the face are identified by a trained computer and their location is determined by using fast pixel calculation. Haar cascades successfully extract the unnecessary background information from the images and identify the facial region.

The process of face detection using Haar cascades is implemented in OpenCV. Papageorgiouetal. first developed this technique by utilizing rectangular features.

3.3 Emotion Classification

During this stage, the system assigns the image to one of the seven universal expressions as outlined in the FER2013 dataset: Happiness, Sadness, Anger, Surprise, Disgust, Fear, and Neutral.

The training process involves the utilization of CNN, a neural network type renowned for its effectiveness in image processing. The initial phase consists of partitioning the data set into training and testing sets. This was followed by the implementation of a training program with identified training sessions. Notably, no prior Feature Extraction was applied to the dataset before merging with the Convolutional Neural Network (CNN). The strategy used was to experiment with different CNN architectures to minimize overfitting and improve accuracy with the validation set. Feature Extraction and Feature classification are done by CNN architecture.

4. CNN Architecture

In our method, the CNN structure consists of three Convolution Layers (32, 64, and 128 filters for each layer) with increasing mesh depth. Using Rectified Linear Unit (ReLU) as the activation function, each Convolutional Layer maintains the shape of the input and output data using the same padding Batch Normalization is used to increase the performance of each convolutional layer. Then, Max Pooling of size 2x2 is applied to reduce the resolution of the convolved image.

After this convolutional processing, image classification is smoothed and then fed into a 2-layer classifier called Softmax. This classifier includes a dense hidden layer with 128 nodes. In addition, a 20% Dropout layer is added for regularization, followed by an output layer with 7 nodes representing 7 classes of facial expression The complete CNN architecture is depicted in Fig 3, showing steps consecutively in the proposed manner.

In summary, our model provides a comprehensive framework for facial emotion detection by combining Convolutional Layers with ReLU operation, Batch Normalization, Max Pooling, and a 2-level Softmax classifier.

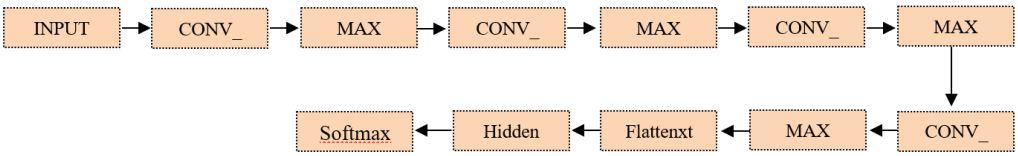


Fig3: Flowchart of CNN Architecture.

4.1 Convolution Layer

The pre-processed images are fed to the CNN architecture. This is the first layer of the CNN architecture. In this layer, convolution is performed between the input layer and a filter. It gives the output as a feature map which gives information about the image. 32 filters are used in the first layer, while 64 and 128 filters are used in the second and third layer.

4.2 *Rectified Linear Unit*

The Rectified Linear Unit (ReLU) acts as an activation function in the network, which introduces significant nonlinearity. Its primary function includes deciding whether a muscle should be activated. The use of ReLU provides a distinct advantage by converting negative inputs to zero, as the project explains

The Function f(x) = max (0, x)

4.3 Pooling Layer

The pooling layer plays an important role in dimension reduction of feature maps. In our model, we used a max pooling layer using a 2x2 mask. This layer specifically identifies the maximum number of objects in the region defined by the mask, thus contributing to the overall efficiency of the network.

4.4 Fully Connected Layers

The fully connected Layers connect the neurons between different layers. They are the layers where input from one layer is connected to every activation unit of the other layer. It also consists Biases and weights. This layer is responsible for classification of images and it is placed before the output layer.

4.5 Dropout

This layer prevents overfitting on the training data. The dropout layer eliminates the contribution of some neurons in the subsequent layer. We have used 20% dropout which means that 20% of the nodes will be discarded.

5. Multiview Facial Emotion Recognition

Multiview facial expression recognition (MFER) involves recognizing human emotions by capturing facial images from different angles and head positions This method aims to provide accurate facial expression recognition by cameras, and to obtain facial expressions a prominent.

The process includes face recognition, feature extraction, and emotion classification. For each camera scan, Face Detection Algorithms identify faces, while Feature Extraction Algorithms extract relevant information from each face image. This affects things like the position and movement of facial landmarks, the intensity of muscle movements, the overall shape and appearance of the face.

In contrast to well-established developments in facial expression recognition (FER), multi-attention facial emotion recognition is a relatively recent and challenging research area We can trace the concept to research on facial expression invariance seen around 2002, e.g. The different focus on capturing emotions from different perspectives, which is reflected in studies with titles such as, adds complexity to the validation process, building research an emphasizing its progress in this area.

6. FER2013 Repository

In this project, we use publicly available data sets, i.e. FER 2013 and CK+, both available on Kaggle were used. These data sets contain face images of 48x48 pixels, each labelled with a corresponding emotion. The FER 2013 specifically includes seven specific emotions: anger, happiness, fear, disgust, neutrality, sadness, and surprise. Notably, the FER 2013 data set presents images with variations in focus, illumination conditions, and scale. Example images from the FER 2013 dataset are shown in Figure 4 for reference.

Fig4: FER2013 Expression Samples.

7. CK+ Repository

Using the same CNN model developed for this project, we trained and tested it on the CK+ dataset. Available on Kaggle, this publicly accessible data has a 48x48 pixel image. The CK+ dataset contains a total of 981 images, each showing seven different emotions: happy, sad, scared, disgusted, contempt, shock, and anger. Figure 5 shows example images from the CK+ dataset, and provides an overview of the different facial expressions captured in this dataset.

Fig5: CK+ Expression Samples.

7.1 Data Enhancement

A sufficient amount of data is necessary to ensure excellent accuracy in our network. Using data enhancement techniques will be necessary to properly enhance our training dataset. Image enhancement appears as an important process, whereby the existing dataset is extended by applying transformation techniques such as scaling, cropping, flipping, rotation, and brightness adjustment to the original images in.

The Tensorflow library simplifies this process with the use of the ImageDataGenerator function. This function optimizes the image without modifying the original dataset. By using transformation techniques, image enhancement actively combats the overfitting problem, increasing the generalizability of the model.

The flow from directory method, in combination with ImageDataGenerator, proves to be an effective way to optimize images. This approach allows images to be read directly from the directory, simplifies the development process, and contributes to the overall robustness of the training data set.

8. Results

The convolutional neural network (CNN) model is built using a sequence of observations and trained on the FER2013 and CK+ datasets, including images representing seven specific emotions: anger, fear, disgust, surprise, neutrality, sadness, of interest are the main training parameters, such as total images that can be trained, times, Function functions, learning rates, optimizers, and loss functions are carefully organized.

The Adam optimizer is used to adjust the number of classes during training, and it also updates the network load based on the training data. The model includes four convolution layers with 32, 64, 128, and 256 filters, using ReLU as the activation function in each layer. Maximum pooling layers follow the convolutional layers, and the resulting feature maps are embedded in a Dense Layer with 20% Dropout. The details of the CNN model is shown in Table 1.

The hierarchical cross-entropy acts as a chosen loss function, which is suitable for situations with more than two lines. A low loss score indicates a good performance of the model. The training spans 100 epochs, cycling through the data for iterative refinement. Then, the OpenCV library is used to test the existing FER2013 test dataset on sample and real-time data.

After testing, the model exhibits an accuracy of 78% on the FER2013 test dataset. Notably, when applied to the CK+ data set, the same CNN model achieves an accuracy of 97%. Analytical criteria, such as the confusion matrix,

are used to compare the predicted emotions with the original emotion scores, facilitating the accuracy of the overall model.

The confusion matrix is a tool for evaluating the performance of a model by comparing the predicted results with the actual results. It has four important values: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Figure 6 shows the confusion matrix obtained from the FER 2013 dataset, while Figure 7 shows the estimates obtained from the CK+ dataset.

INPUT	LAYER	ACTIVATION	OUTPUT
48X48	Conv.(3x3, 32 filters	ReLU	46x46x32
46x46	Conv.(3x3, 64 filters	ReLU	44x44x64
	Max Pooling (2x2)	-	22x22x64
22x22	Conv.(3x3, 128 filters)	ReLU	20x20x128
	Max Pooling (2x2)	-	10x10x12
10x10	Conv.(3x3,256 filters)	ReLU	8x8x256
	Max Pooling (2x2)	-	4x4x256
	Max pooling (2x2)	-	4x4x256
4x4	Flatten	None	4096
4096	Dense (Hidden)	ReLU	512
	Dropout (20%)	-	
	Dense (Output)	Softmax	7

Table 1: Details of CNN Architecture

The illusion matrix, through the predicted sentiment scores generated by the CNN and the resulting truths, takes the form of a class matrix. Each row is an instance in the predicted class, while each column represents an instance in the actual class. including rows and columns.

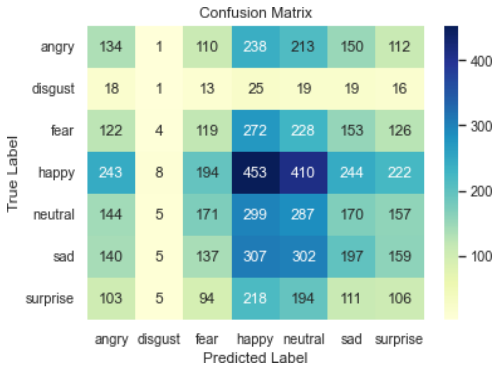


Fig6: Confusion Matrix from FER 2013 Dataset Evaluation.

To evaluate the performance of a CNN model, several metrics such as sensitivity, specificity, and recognition accuracy are derived from the confusion matrix. These metrics provide a comprehensive assessment of a model's power without resorting to platitudes.

Sensitivity, also known as True Positive Rate (TPR) or Recall, indicates how well the model is able to identify positive cases among all truly positive cases It is an important metric in binary classification problems. High sensitivity indicates that the model detects positive emotions correctly, whereas low sensitivity indicates that positive emotions may be incorrectly classified as negative.

Statistically, sensitivity is calculated as the sum of true positives (TP) and true positives (TP) and false positives (FN):

Sensitivity = TP / TP + FN = 1 - FNR

Specificity, refers to the percentage of real negatives that the model properly predicts. When a model has high specificity, it is good at predicting negative emotions. And, when the model has low specificity, then the model will predict positive emotions as negative emotions. It is commonly used in binary classification problems. It can be computed mathematically as:

Specificity = TN/ TN + FP = 1 – FPR

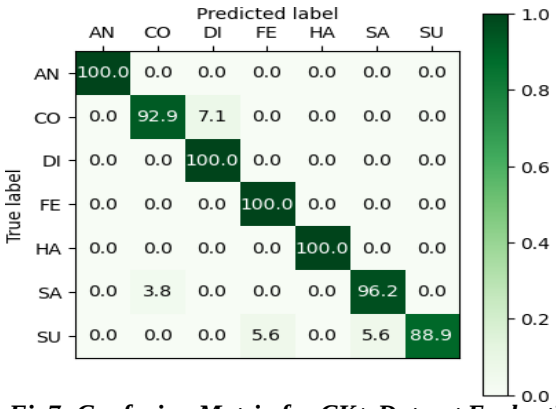


Fig7: Confusion Matrix for CK+ Dataset Evaluation.

Accuracy is a measure of the ability of a model to correctly predict classes, usually expressed as a percentage. When models are trained and tested with different kernel sizes, the accuracy varies accordingly.

The performance of the model was evaluated using kernel sizes of 3x3, 5x5, 7x7, and 9x9, and the accuracy results on the FER 2013 and CK+ data sets are summarized in Table 2.

Kernel Size	Accuracy on FER2013 Dataset	Accuracy on CK+ Dataset
3 x 3	78.25	97.81
5 x 5	76.31	96.66
7 x 7	73.54	95.31
9 x 9	71.61	94.12

Table 2: Accuracy Comparison of Different Kernels from FER 2013 and CK+ Datasets.

Figures 8 and 9 depict the accuracy of various kernel sizes on the two datasets presented graphically.

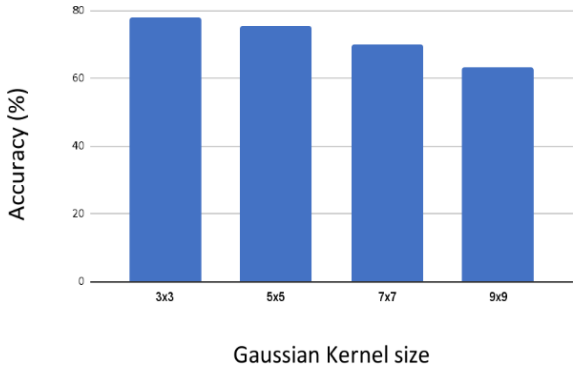


Fig8: Performance of Various Gaussian Kernel Sizes under Maximum Pooling on FER2013 Dataset.

Two types of accuracy are generally considered in training a model: training accuracy and validation accuracy. While Validation accuracy measures the performance of the model on new unseen data, Training accuracy measures how well the model performs on the training dataset itself. Generally, the training accuracy improves as the model learns from the data. However, if the model is too complex and starts memorizing the training set instead of learning the underlying model, the accuracy of the validation starts to decrease and this process is referred to as overfitting.

The training and validation accuracies of the model on the FER2013 and CK+ datasets are shown in Figure 10 and Figure 11 in terms of epochs, respectively. These diagrams provide insight into how the accuracy of the model evolves during the training process, thereby identifying potential overfitting or underfitting problems.

Fig9: Training and Validation Accuracy on FER 2013 Dataset.

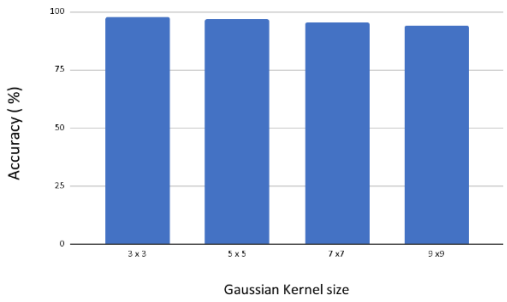
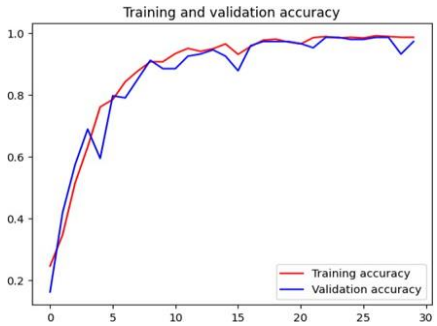


Fig10: Performance of Various Gaussian Kernel Sizes with Maximum Pooling on CK+ Dataset.



Based on the training and validation accuracy charts, the model appears to perform well without overfitting and exhibits good performance in all senses.

Data Augmentation techniques have been used to further increase the accuracy of the model. These methods are necessary when available training data are limited.

Data enhancement expands the dataset by introducing various transformations such as adding noise, cropping, flipping, and rotation of the original image, thus improving the accuracy of the emotion detector Besides, it introduces variety includes the dataset and increases model robustness by connecting to a wider input images. This increased diversity contributes to the overall performance of the model.

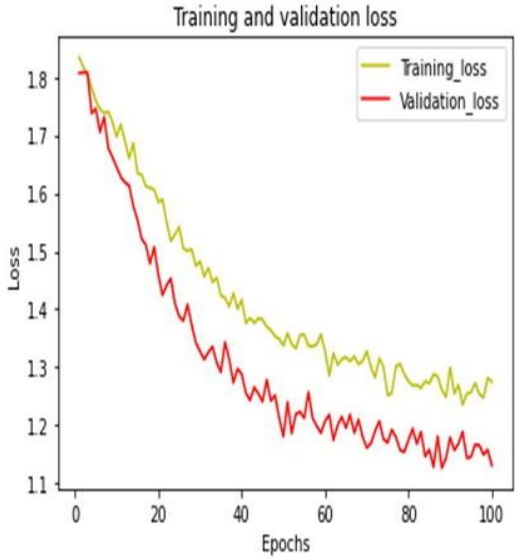


Fig11: Training and Validation accuracy obtained on CK+ dataset.

In order to detect overfitting, it is essential to monitor both training and validation loss during training. The training and validation loss are plotted against the number of training epochs in the graph.

Figures 12 and 13 shows the graph obtained for training and validation loss in case of FER 2013 and CK+ dataset.

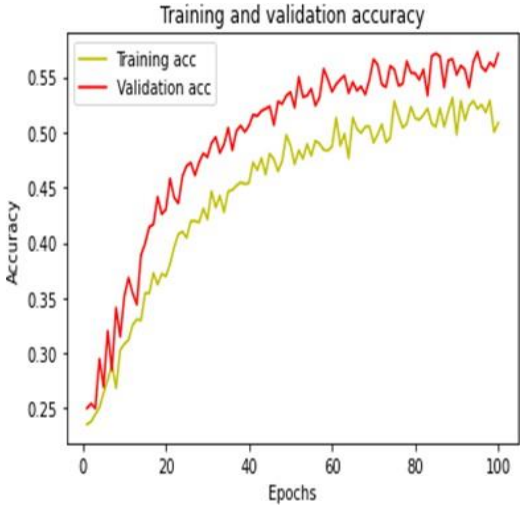


Fig12: Loss During Training and Validation on FER 2013 Dataset.

The end result should be the classification of the emotion in the image. For example, if we take an image with the original Surprise label, the model predicts the emotion of that image was Surprise, which is similar to the original image's emotion as shown in Figure 14. Since the original label and predicted label are the same. The correct prediction of the emotion shows that our model is working accurately.

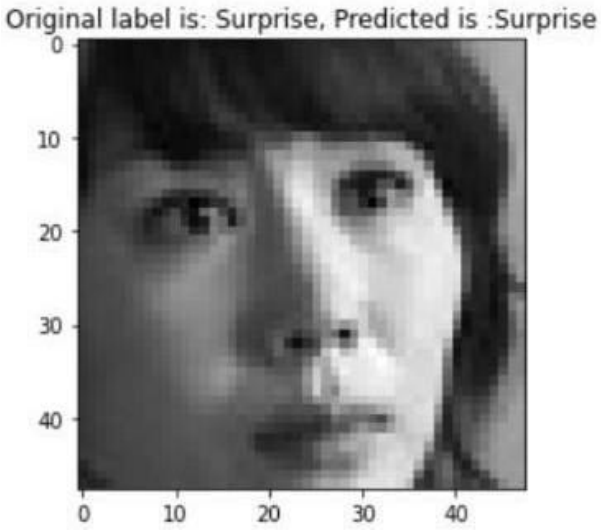


Fig13: Emotion Prediction by the CNN Model.

On obtaining the results by training the same CNN model on FER2013 and CK+ dataset, we found that the loss obtained in case of CK+ dataset was much less than that in case of FER2013 dataset. Hence, more accuracy was obtained in case of CK+ dataset which 97%.

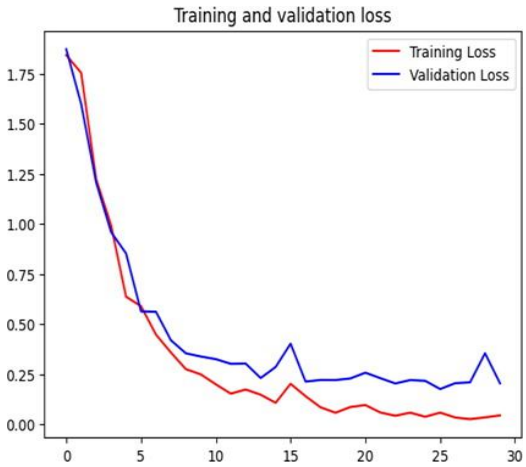


Fig14: Training and Validation Loss Curve for CK+ Dataset.

When the CNN model is trained and tested on FER 2013 and CK+ dataset, each emotion showed different accuracy for both the datasets depending upon the type of images present in the two datasets. The accuracy of each emotion is showed in the table 3.

Emotions	FER 2013	CK+
Happy	61.2	100
Sad	56.31	96.2
Fear	54.3	100
Surprise	60.1	88.9
Angry	51.8	100
Disgust	44.8	100
Neutral	58.3	92.9

Table 3: Emotion Accuracy Comparison between FER 2013 and CK+ Datasets.

9. Conclusion

The research paper presents a Realtime Emotional Reflective User Interface using Convolutional Neural Networks (CNN). The CNN model is trained on FER2013 dataset, which provides rich labelled face images for training and testing. Through FER2013 training, CNN achieves a high degree of accuracy, enabling accurate real-time detection of emotions from video and image feeds.

Furthermore, the paper examines the effectiveness of methods such as data enrichment to improve model performance. Potential applications of Real-Time Sensation Recognition using CNN span areas such as psychiatric research, marketing research, and human-computer interaction as technology advances and larger datasets, the model holds promise for greater accuracy and reliability, opening the way for future applications and research efforts.

Refrences

[1] Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 1139-1147).

[2] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[3] Ranzato, M., Huang, F. J., Boureau, Y. L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-8).

[4] De Silva, Liyanage C., Tsutomu Miyasato, and Ryohei Nakatsu. "Facial emotion recognition using multi-modal information." Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications *(Cat.*. Vol. 1. IEEE, 1997).

[5] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient- based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[6] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep Learning (Vol. 1). MIT press Cambridge.

[7] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated

recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

[8] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[10] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International Journal of Computer Vision, 57(2), 137-154.

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252.

[14] 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).

[15] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[16] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[17] Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. Neural computation, 22(12), 3207-3220.

[18] De Silva, Liyanage C., Tsutomu Miyasato, and Fumio Kishino. "Emotion enhanced multimedia meetings using the concept of virtual space teleconferencing." In Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems, pp. 28-33. IEEE, 1996.

[19] Khorrami, P., & LeCun, Y. (2015). Deep convolutional networks with noisy labels. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15) (pp. 549-558).