Emotion Recognition Through Facial Expressions: A Machine Learning Perspective in Mobile Multimedia

Akram Ahmad[1], Vaishali Singh[1] and Kamal Upreti[2,*]

[1]*Department of Computer Science, Maharishi University of Information Technology, Lucknow, India*
[2]*Christ University, Delhi NCR Campus, Ghaziabad, India*
E-mail: akram.ahmad2009@gmail.com; singh.vaishali05@gmail.com; kamalupreti1989@gmail.com
[*]*Corresponding Author*

Received 27 December 2024; Accepted 23 February 2025

Abstract

Facial expression-based emotion detection is very attractive because of the possibilities in security systems, mental health monitoring, and human-computer interaction. Even with the progress in accuracy in real-world settings, issues such as the lack of balanced datasets and the inability to differentiate between faint or superimposed emotions continue to plague it. This study aims to bridge these constraints by developing a CNN-based model that would be able to recognize face emotions reliably and be utilized in real-time situations, such as webcam integration. The Affect Net dataset, which is a comprehensive collection of over a million facial photos labeled with the seven major emotions of anger, disgust, fear, happiness, neutrality, sadness, and surprise, was used to train the proposed model. Other pre-processing data techniques used include grayscale conversion, normalization, scaling, and data supplementation to increase the robustness of the model. Using metrics like accuracy and loss trends for evaluation, the model demonstrated efficiency stability at around the 30th training phase. When the model is

Journal of Mobile Multimedia, Vol. 21_1, 87–112.
doi: 10.13052/jmm1550-4646.2114

compared to existing models, this proposed model can attain the competitive level of accuracy up to approximately 60%. It also has the potential to run in real applications through its webcam integration. While the model can differentiate between various clear-cut emotions, it becomes ineffective at identifying subtle emotions, which include "Fear" and "Neutral" majorly because of unbalanced data and the subtleness of these expressions.

Keywords: Emotion detection, human-computer interaction, advanced machine learning, facial expressions, feature extraction.

1 Introduction

1.1 Background

Identification and analysis of human facial expressions, speech patterns, and physiological markers including happiness, anger, sorrow, and surprise are the main goals of emotion detection [1]. Due to its many uses, such as mental health monitoring [4], human-computer interaction using mood-aware interfaces [3], and security system surveillance [2], this topic has attracted a lot of attention. To increase the precision and effectiveness of emotion identification systems, machine learning and deep learning algorithms have been widely used [5] Using neurological and physiological indicators like electroencephalography (EEG), galvanic skin response (GSR), and beats per minute (BPM) to identify emotions, emotion identification is a computer method for assessing human emotional states. While GSR detects electrodermal activity, which reflects physiological reactions like sweating in fear, EEG-based approaches use electrode-based brain activity signals to assess psychological states. Similar to this, BPM data shows changes in heart rate, and hormonally induced changes in body temperature provide further contextual information for analyzing emotions [6]. These neurological and physiological methods, when paired with sophisticated machine learning models, improve the accuracy of emotion detection and open the door to real-time emotion-aware computers in a variety of fields.

1.2 Related Studies

Recent research in affective computing indicates that the integration of EEG, GSR, and PPG can classify emotions such as happiness, relaxation, anger, and sadness with an accuracy of up to 79.76% when tested on tactile-enhanced multimedia [7]. These methods are being explored in human-robot interaction

(HRI) applications, allowing real-time emotion estimation and promoting user engagement through low-cost wearable device. Recent developments in emotion identification have increased precision as well as application in real-world situations by utilizing a variety of modalities, including bodily motions, biological signals, or facial expressions. Facial expression recognition (FER) may achieve as much as 99% quality in controlled settings, but it struggles in real-world applications where variables like individual variability, neck posture, or illumination cause accuracy to drop to roughly 50%. through providing supplementary data, multimodal sensors like EEG, audio, and thermal sensors can help reduce these issues and improve reliability [8]. Similar to facial expressions, bodily manifestations of emotions, such as happiness, anger, and fear, are important for emotion recognition and take the scope beyond just facial expressions. The work in this domain focuses on coding systems of bodily behaviors as a precursor to more comprehensive frameworks of emotion detection [9]. Advances in the analysis of movement, features are selected using frameworks and genetic algorithms, improved emotion recognition up to 90% for walking scenes, 96% for sitting, and 86.66% for action-independent, which may likely have robust applications in virtual reality, robotics, and behavioral modeling [10], and show much promise with applications in virtual reality and robotics and advanced modeling of behavior. Contextual cueing research shows faster response times for repeated visual search tasks and therefore supports early attentional guidance effects based on evidence from psychophysics, EEG, and eye-tracking studies and emphasizes differences between habit-driven attention and task-specific spatial priority [11]. Data-driven machine learning approaches also benefit autonomous systems design; for instance, external displays on autonomous cars can improve pedestrian safety by giving information about speed, achieving up to an additional 4 feet of safety in specific populations [12]. The development of augmented metric representations for RGB-D scenes relies on CNN-based detectors such as YOLO, Kalman filters, and many others for the purpose of object tracking and semantic segmentation. Real-time object detection is made possible by YOLO (You Only Look Once), which reads an entire image in one go, making it very efficient when rapid scene understanding is required for applications. Kalman filters operate across a range of scopes when tracking objects across time by predicating their moving trajectories and performing dynamic updates for state estimates. In facial expressions, these methodologies can be incorporated to track face landmarks, thereby refining the outcomes of classification steps and improving performance in robust analyses of dynamic videos in emotion

quantification [13]. Among them, physiological signals like EEG and ECG can give more objective and reliable information, since social masking, where people consciously or unconsciously hide their feelings [16, 18]. Of these, EEG signals, with their high sensitivity to affective state changes, offer real-time features for emotion detection. These methods included wavelet transforms, feature reduction, and machine learning classifiers such as SVM, Random Forest, and KNN. Furthermore, recent self-supervised frameworks applied to ECG-based emotion recognition have improved the performance of classification using spatiotemporal representations and multi-task learning strategies, showing the capability of physiological signal-based emotion recognition with accuracy and robustness [17]. Such methodologies prove that physiological signal-based emotion recognition is able to achieve accurate and robust classification performance, thus resolving various challenges in traditional emotion detection methods [16, 18]. With DNNs giving excellent results in complex tasks such as image classification, scene generation, and optimization of the wireless network, new techniques that can dissect hidden units in CNNs and GANs were invented to exploit abilities such as object concept identification and contextual modification of scenes outs [19]. In wireless communications, ANNs, such as recurrent and deep neural networks, are used to solve latency and connectivity challenges for IoT devices, which shows its application in unmanned aerial vehicles and virtual reality [20]. ANNs in wireless communications play a vital role. As ANNs are now able to function without latency, connectivity issues in IoT devices can be solved using ANN-based techniques. RNNs and DNNs are often used for optimizing the appropriate use of network resources, predicting congestion patterns, and enhancing transmission efficiency. In the context of UAVs and VR, ANNs enable real-time signal processing, adaptive bandwidth allocation, and interference mitigation to ensure seamless communication in dynamic environments. As IoT devices continue to expand, deep learning models are becoming increasingly critical for optimizing network performance in sustainable and intelligent wireless communication systems. Moreover, the universal approximation capability of CNNs shows that it can approximate any continuous function with high accuracy, which proves the robustness of CNNs in handling large-dimensional data [21]. It points out the transformative potential of DNNs for domains as broad as possible, driving innovation and addressing complex challenges [22–24]. Existing research in emotion recognition highlights gaps such as limited diversity in datasets, which hampers the generalization of models across varied populations and real-world scenarios [25]. Additionally, many approaches struggle with real-time

processing and fail to effectively detect subtle or mixed emotions, which are crucial for accurate and comprehensive emotion analysis [26]. Recently, deep learning and multimodal approaches have been explored to enhance accuracy in emotion recognition. The EESCN model enhances EEG-based emotion recognition and achieves 94.81% accuracy on DEAP by using neuromorphic data generation along with a NeuroSpiking framework [27]. Another trend is MER with audio, visual, and text modalities. Deep models improve feature extraction and fusion techniques in this modality [28]. In the area of speech emotion recognition, Vesper- an adaptation of WavLM-is using a mask that takes the form of an emotion to perform better over traditional models over IEMOCAP, MELD, and CREMA-D datasets [29]. In addition, GPT-4V demonstrates robust visual and multimodal emotion recognition ability in 21 benchmark datasets but lacks the capability for micro-expressions that need to be specifically trained [30]. These advancements mark a new shift toward multimodal deep learning approaches that are optimized for emotion recognition across different applications.

1.3 Motivation

Numerous variables, including occlusions from glasses or masks, changes in facial features, lighting circumstances, and cultural differences in emotional responses, make it difficult to identify emotions in facial expressions. Furthermore, acquiring high-quality training data continues to be a major challenge in creating precise recognition models. In order to overcome these obstacles, this study compares machine learning techniques – specifically, Convolutional Neural Networks, or CNNs – with current methods in order to create a more accurate and reliable model for facial expression emotion recognition. This research is important because it uses machine learning to improve automated emotion identification systems, which will help with mental health monitoring and improve public safety. This work aims to increase accuracy and resilience by utilizing hybrid models and sophisticated feature extraction techniques, hence increasing the dependability and efficacy of real-time emotion identification.

2 Methodology

An NVIDIA GeForce RTX 3080 GPU with 10GB VRAM and an Intel Core i9 CPU with 32GB RAM were the components of the high-performance machine used for the trials. Fast data access and storage were made possible

via a 1TB SSD. Using Python 3.9 on Ubuntu 20.04, the software environment was constructed with Scikit-learn 1.1 for evaluation, Opens 4.5 for image preprocessing, and Tensor Flow 2.9 and Keras for model construction. The Affect Net dataset fell in three main groups: train (80%), verification (10%), & test (10%). The dataset included more than one million facial photos that were labeled for seven different emotions. These processes included grayscale conversion, normalizing to the [0, 1] range, and then shrunk to 48×48. Several data augmentation methods were used including rotation, flipping, zooming, and cropping to improve the robustness of the model, reducing overfitting. Rotating helps introduce various angles to the model to simulate different perspectives, flipping the image introduces its mirror image; zooming and cropping change spatial scales and incorporate other parts of the image respectively in training. These augmentations enhance the richness of data by preventing the model from memorizing certain patterns from images. In the training phase, the Adam optimizer with a learning rate of 0.001 and batch size of 64 was applied, balancing between optimization efficiency and convergence stability.

2.1 Data Collection

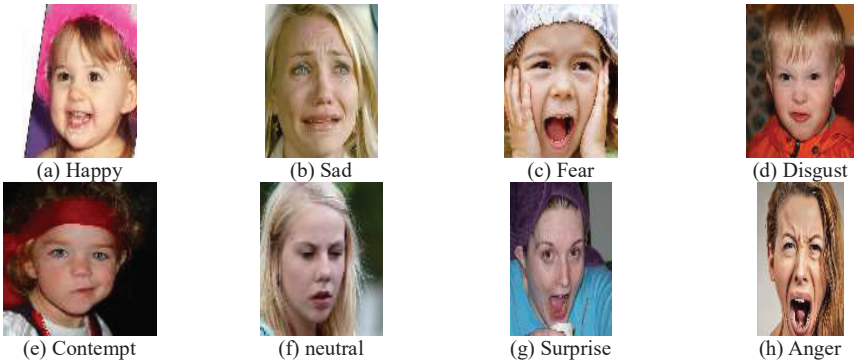
A dataset used for as shown Table 1 research came from the AffectNet database, a large-scale facial expression database widely recognized in literature for its diversity and richness in Table 2. AffectNet features a massive collection of images of faces collected from web sources and annotated for the seven basic emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise, as illustrated in Figure 1. Each image is accompanied with metadata that contains file paths, emotion labels, and some of the images contain landmark facial coordinates to aid in extracting the features. The AffectNet dataset was split into three subsets: 80% training, 10% validation, and 10%

Table 1 Distribution of emotion labels in the dataset

Label	Label Counts
Surprise	4616
Anger	3608
Contempt	3244
Sad	2995
Happy	4336
Disgust	3472
Fear	3043

Table 2 Proportional representation of emotion labels in the dataset

Label	Label Proportion
Surprise	0.163833
Anger	0.128057
Contempt	0.115138
Sad	0.106300
Happy	0.153895
Disgust	0.123230
Fear	0.108004



(a) Happy (b) Sad (c) Fear (d) Disgust

(e) Contempt (f) neutral (g) Surprise (h) Anger

Figure 1 Different types of emotions available in dataset.

testing. The dataset holds over a million facial images that are annotated by one of seven emotion categories, which include Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Preprocessing techniques were employed to standardize the inputs as well as optimize model efficiency. Images were reduced to grayscale so that the amount of computation may be reduced, while the salient features can still be represented. Normalization was applied to standardize pixel values to fall between [0, 1] across the dataset. Lastly, all images were resized to 48 $\times$ 48 pixels for maximum compatibility with CNN architectures while maintaining sufficient facial detail to classify the emotions.

The dataset has a very assorted distribution of emotion labels; from Table 1 “Surprise” and “Happy” are the predominant ones, with 4616 and 4336 instances, respectively, followed by “Sad” and “Fear” among those less frequent, with only 2995 and 3043 instances. To put it proportionally, with details in Table 2, “Surprise” represents 16.38 percent and “Happy” 15.39 percent of the overall dataset, while “Sad” and “Fear” constitute 10.63 percent

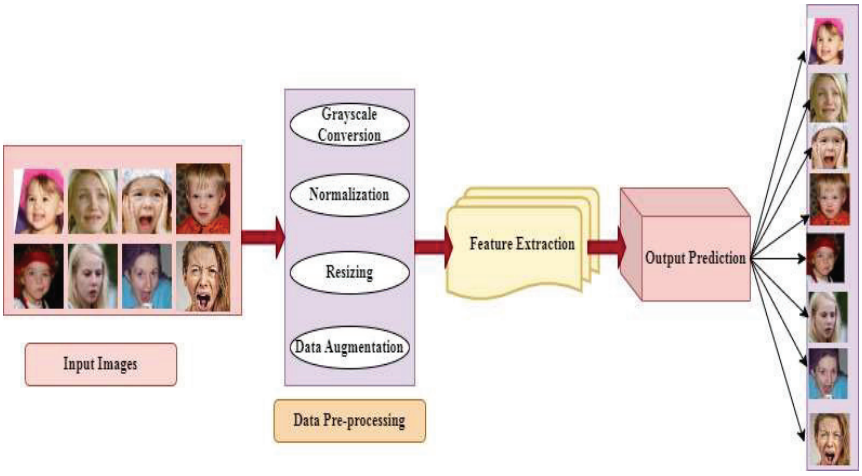


Figure 2 Workflow of emotion recognition from facial images.

and 10.80 percent, respectively. This balanced distribution between positive, negative, and neutral emotions ensures a sound foundation for sentiment analysis as well as reliable model performance evaluation.

2.2 Data Preprocessing

The preprocessing steps for the dataset are Grayscale Conversion, simplifying images to a single intensity channel for reducing complexity in computation, focusing directly on key features, such as facial contours and expressions, and Normalization, in which pixel values are scaled to the range [0, 1], bringing uniformity to input values, which can help efficiently train without facing problems of exploding or vanishing gradients. Figure 2 in resizing follows wherein all the images are resized to a standard size of 48 $\times$ 48 pixels so that there is consistency in input dimension to the models. It's essential that the same happens for the feature comparison in the entire dataset. Data Augmentation is then performed to enrich the data. This involves introducing variation with techniques like Rotation to simulate different angles, Zooming to simulate different distances, mirroring the images makes the model learn invariant features and hence not face dependent on its orientation. Therefore, it would be more apt to unseen test cases. Focusing on facial regions, the cropping ensures the model does not depend too heavily on dominant features like eyes, nose, or mouth while totally ignoring others. This helps avoid overfitting into particular facial geometries in the dataset, giving the model

good generalization between different individuals as well as over different facial expression variations.

2.3 Feature Extraction and Label Mapping

These Conv2D layers in a Convolution Neural Network are developed to extract spatial features of input images, which would constitute an essential part of deep learning models for feature extraction. Every Conv2D layer is assigned a particular parameter to do so. Filters, therefore are of extreme importance and part of CNN where, based on different convolutional levels the Conv2D layer possesses lots of filters as a convolution layer learns about feature hierarchy by breaking up any kind of image for hierarchical input patterns. While here the initial is having 128 filters as, they could basically identify those simple edges or that of texturing; at layer 2 containing 256 filters which detects features such as those of contours between face shapes. Finally, the third layer contains 512 filters, identifying high-level features such as full facial structures and complex expressions. All of these multi-scale feature extractions add up to make it rich to distinguish between any subtle variations in facial expressions and, therefore improves classification accuracy. This kernel size is (3, 3), with which important patterns can be detected. The input shape used for the first Conv2D layer is set to be (48, 48, 1), depicting color pictures at dimensions of 48×48 (refer Table 3). The activation method used is ReLU (Corrected Linear Unit), which introduces non-linearity and allows the network to learn deeper connections within the data.

Label mapping is essential to organize the dataset; ensure that each image, upon processing, can find the right emotion class association. The dataset is structured into directories, where each subdirectory of those directories represents a specific emotion label in Figure 4. The proposed framework utilises

Table 3 CNN model architecture with parameters

Layer	Type	Filters	Kernel Size	Activation	Output Shape
1	Conv2D	128	(3,3)	ReLU	(48,48,128)
2	MaxPooling2D	–	(2,2)	–	(24,24,128)
3	Conv2D	256	(3,3)	ReLU	(24,24,256)
4	MaxPooling2D	–	(2,2)	–	(12,12,256)
5	Conv2D	512	(3,3)	ReLU	(12,12,512)
6	Flatten	–	–	–	73728
7	Dense	128	–	ReLU	128
8	Output Layer	–	–	Softmax	7 (emotions)

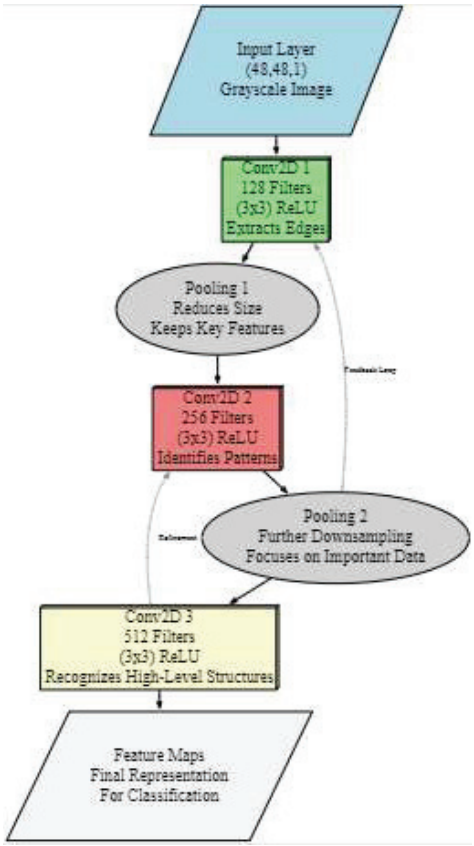


Figure 3 Convolutional neural network architecture for feature extraction.

a mapping mechanism and subdirectory names translated as numerical labels to ensure better data processing and classification operations of the model. The mapping is as follows.

2.4 Model Design

After the Conv2D layers, a Convolutional Neural Network makes use of Pooling Layers; for example, MaxPooling2D to decrease the spatial dimensions, hence relieving the computational load on the network and over fitting. Pooling in a CNN reduces the spatial extent of feature maps, concentrating useful information and retaining all major features while discarding unwanted data. This minimizes the computational complexity and develops translation

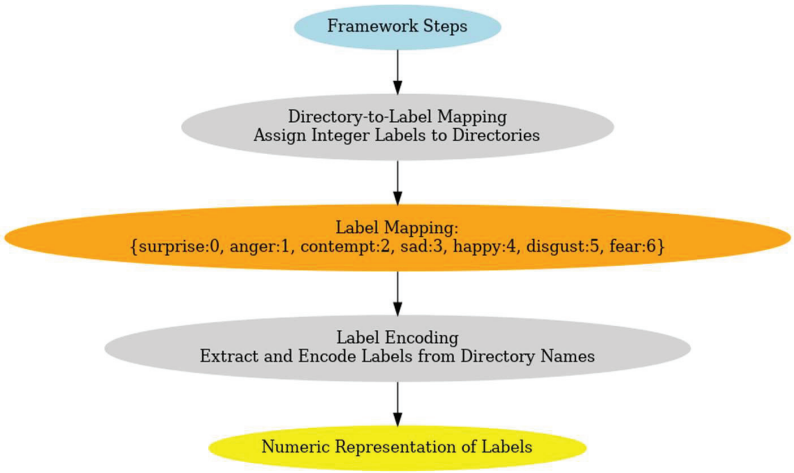


Figure 4 Framework for label mapping and encoding in emotion recognition.

invariance to reduce overfitting. The common kinds include Max Pooling that captures the strongest activations, and Average Pooling provides a smoother feature summary. By focusing on key patterns and suppressing noise, pooling layers improve the efficiency and generalization of the network, making them integral to extracting hierarchical features in CNN architectures. The feature extraction may continue with additional Conv2D and pooling layers for more in-depth feature extraction. After a Flatten Layer transforms 3D feature maps into a 1D vector, Fully Connected (Dense) Layers refine the features even more. According to what is needed, an input layer with flexible and an activation of and secret layers using activation of ReLU follow next. In order to avoid over fitting, dropout layers are frequently used. The final output layer of the CNN model is responsible for producing classification probabilities, thereby allowing each input image to receive an emotion label. Depending upon the type of the task, the activation used in this layer may be:

- Softmax Activation is used in multi-class classification and ensures the output values sum to 1. This represents the probability distribution over multiple categories of emotions.
- The sigmoid Activation is typically utilized for binary classification problems and outputs a value between 0 and 1, meaning a probability score.
- Linear Activation is just for a regression problem where the output will be a continuous number rather than discrete categories.

For this experiment, Softmax Activation function was taken because it is most suitable when classifying images to fall in any one of the seven emotion categories. With an 80:20 split of the data, the model has been trained and validated.

2.5 Training and Testing Models

The model's performance was maximized by meticulously choosing critical training parameters. For optimal computing efficiency, a batch size of 64 was employed and model convergence, processing 64 samples before updating weights. The model was trained for 30 epochs, a duration selected to ensure effective learning while monitoring validation performance to prevent over fitting in Figure 5. The Adam optimizer was used because of the pace of adaptive learning and computational efficiency, combining the strengths of momentum-based methods and RMSProp, making it ideal for image-based tasks. The batch size used in training the model was 64, balancing between computation efficiency and gradient stability. If the batch size were taken a bit lower, at 32, for instance, gradients became unstable. When this was increased to 128, convergence rates slowed down considerably. A learning rate of 0.001 utilizing the Adam optimizer, offering adaptive learning rates and thus preventing vanishing gradients. Early stopping was implemented to halt training when validation loss ceased improving, ensuring optimal model generalization.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

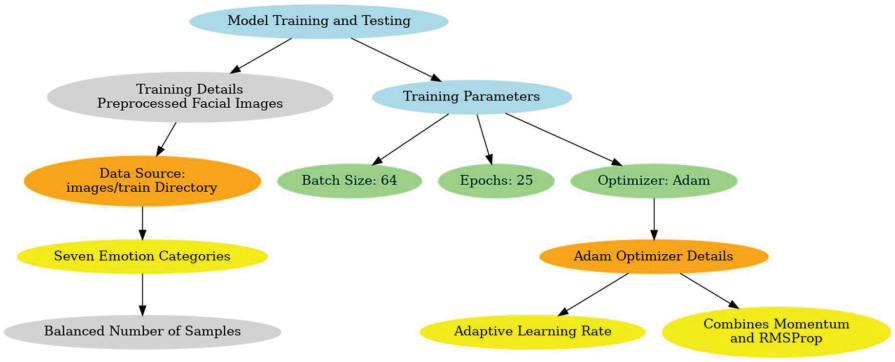


Figure 5 Integration of training details and optimizer parameters for model development.

$$\upsilon_t = \beta_1 \upsilon_{t-1} + (1 - \beta_1) g_t^2$$

$$\theta_t = \theta_{t-1} - \eta \frac{m_t}{\sqrt{\upsilon_t} + \epsilon}$$

Where, in the above equations, first moment estimate (momentum term) is denoted bym_t, decay rate for the moving average (typical value: 0.9) is $\beta 1$, gradient of the loss function with respect to model parameters at time t is g_t, The second moment estimate (similar to variance), tracking the magnitude of the gradients is denoted by υt, The decay rate for the second moment (typical value: 0.999), model's parameters (θ) using both the first and second moment estimates.

3 Results and Discussions

This section evaluates the proposed CNN model for facial emotion recognition, focusing on its performance across seven emotion categories. Key metrics like accuracy, loss, and sample predictions are analyzed, highlighting the model's strengths and limitations. Visualizations, including graphs and example outputs, are provided to support the analysis and demonstrate the model's effectiveness.

3.1 Evaluation of Validity & Training Accuracy

Figure 6 provides the confusion matrix of model predictions for emotion detection and Figure 7 illustrates the patterns for training and validation accuracy, which indicate a consistent improvement. During periods 25–30, both measures stabilize at about 60%. Good generalization to unknown data without noticeable over fitting is shown by validation accuracy, which often outperforms training accuracy. The training accuracy rises sharply from 25% to 50% during the first 8 epochs (Figure 8) and gradually approaches 60% by epoch 30. Similarly, validation accuracy improves consistently (Figure 9), stabilizing after epoch 10. The convergence of trends and the plateau in accuracy after epoch 25 reflect the model's optimal learning state and its ability to extract meaningful features for robust emotion classification.

3.2 Evaluation of Verification or Training Losses

The training and validation loss trends as shown in Figures 10 and 11 highlight the model's effective learning and generalization. Training loss begins

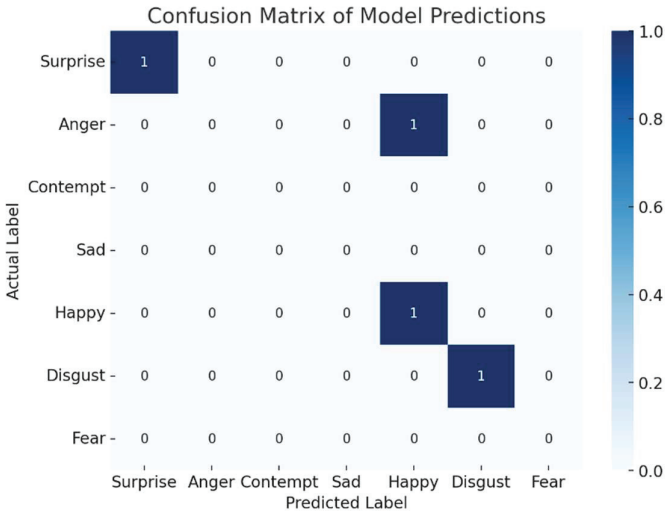


Figure 6 Confusion matrix of model predictions.

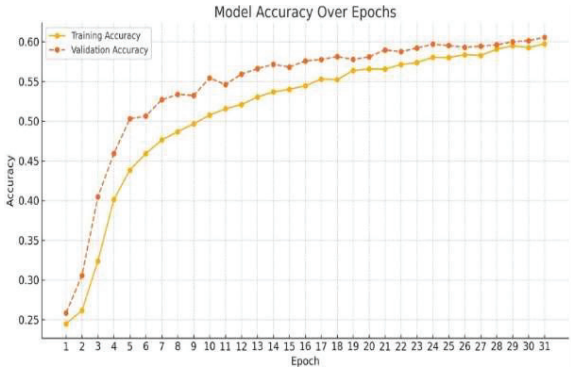


Figure 7 Model accuracy over epochs.

at about 1.8 because it is untrained, drops sharply in the first 10 epochs, and stabilizes at about 1.0 by epoch 20. Again, validation loss tracks training loss in general, with a sharp dip in the first 10 epochs before leveling off near 1.0 between epochs 10 and 20. The fact that the curves are aligned for training and validation loss suggests strong generalization and no overfitting, since both losses bottom out and stabilize at near-equal values.

These trends represent how well the model learns key features efficiently and at what stage the model finds an adequate balance between accuracy and computational efficiency. Sharp initial losses reflect rapid learning, whereas

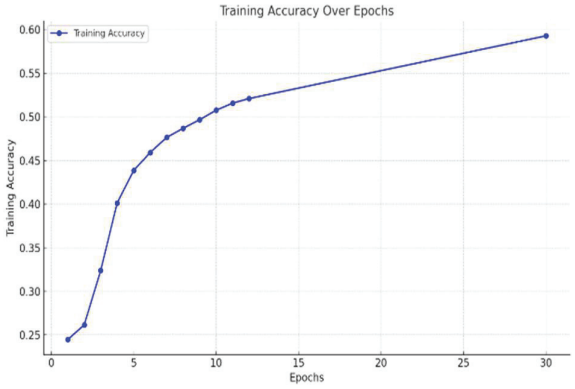


Figure 8 Training accuracy over epochs.

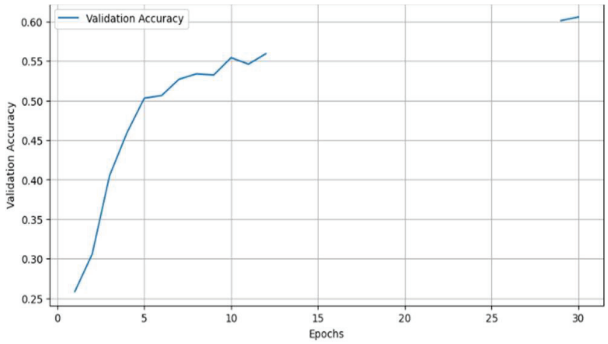


Figure 9 Validation accuracy over epochs.

stabilization confirms the optimal capacity of the model for learning. The results thus emphasize the quality of the training procedure and selected hyperparameters in producing the model's reliability in handling unseen data.

3.3 Emotion Recognition Results

This section examines the accuracy of the model in classifying facial emotions with a correct classification and examples of incorrect classifications to indicate its strong and weak points. The model performs well in distinguishing emotions based on clear visual cues. For instance, in Figure 12, the model correctly predicts the "Happy" emotion for a gray-scale image; it successfully picks up important cues, such as smiling and slackened facial muscles. In Figure 13 The model accurately classifies the "Sad" emotion using less visible cues, including sagging eyelids and lower curvature of

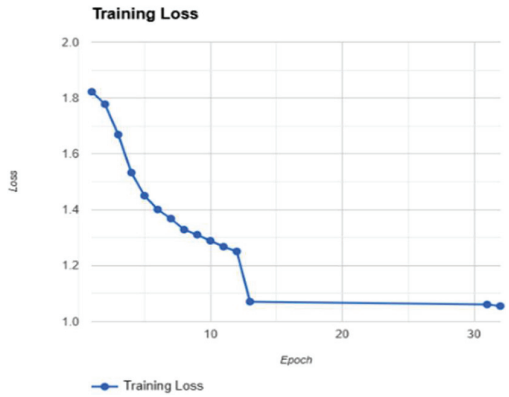


Figure 10 Training loss.

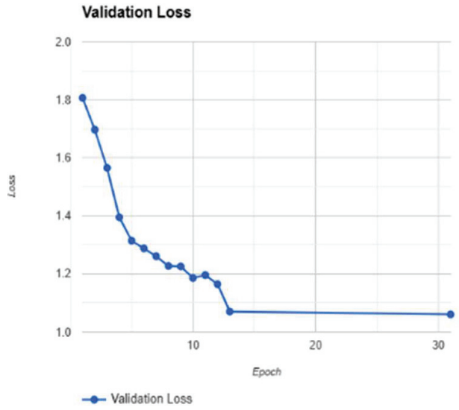


Figure 11 Validation loss.

```
image = 'images/train/happy/7.jpg'
img = ef(image)
pred = model.predict(img)
pred_label = label[pred.argmax()]
print("model prediction is ",pred_label)
plt.imshow(img.reshape(48,48),cmap='gray')
```

```
1/1 [==============================] - 0s 39ms/step
model prediction is  happy
<matplotlib.image.AxesImage at 0x229254012b0>
```

Figure 12 Correctly predicted.

```
image = 'images/train/sad/42.jpg'
print("original image is of sad")
img = ef(image)
pred = model.predict(img)
pred_label = label[pred.argmax()]
print("model prediction is ",pred_label)
```

```
original image is of sad
1/1 [==============================] - 0s 45ms/step
model prediction is  sad
```

Figure 13 Correctly predicted.

```
image = 'images/train/fear/2.jpg'
print("original image is of fear")
img = ef(image)
pred = model.predict(img)
pred_label = label[pred.argmax()]
print("model prediction is ",pred_label)
plt.imshow(img.reshape(48,48),cmap='gray')
```

```
original image is of fear
1/1 [==============================] - 0s 69ms/step
model prediction is  neutral
```

Figure 14 Misclassified.

the lips. Accurate predictions of these instances showcase generalizability in situations in which the emotional character being highly prominent and distinct in quality, such as “Happy” and “Sad”.

However, the model has problems with emotions having faint or blended characteristics; as in the example shown in Figure 14, the model wrongly classified an image that pertains to the “Fear” emotion as “Neutral.” This kind of error may be attributed to several reasons, for example, the faint facial expressions associated with “Fear,” such as wide-opened eyes and slightly parted lips. Moreover, poor quality images with low resolution and noise may obscure certain features that the model is designed to recognize, and ambiguous inputs may bias the prediction towards the more frequent class in case of an imbalanced dataset, where “Neutral” emotions dominate.

3.4 Analysis by Comparison

This section includes a comparison study of the proposed CNN-based model with the available facial emotion recognition studies. Evaluation in terms of

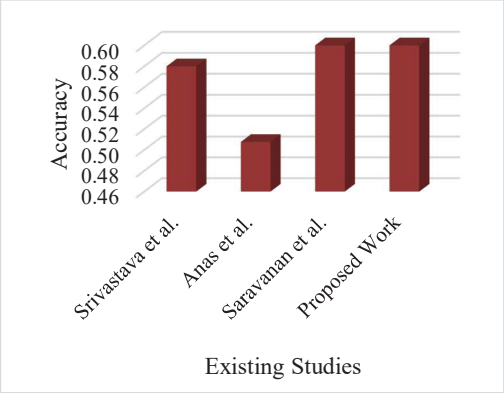


Figure 15 Comparative analysis.

datasets, methodologies used, preprocessing techniques adopted, and overall accuracy have been considered. The model achieved competitive accuracy at 60% on the Face Expression Recognition Dataset. The proposed model also operates well with real-time webcam feeds. Hence, it is well-suited for use in practical settings. The set features give an indication that the model emerges as a useful and promising one in the domain of face-based emotion recognition. It therefore represents a comparative study undertaken by comparing existing facial expression recognition studies against the designed CNN-based model. Main issues considered are datasets and methodologies, preprocessing techniques together with overall accuracy. Competing accuracy was achieved upon applying this model, which received proper training via data samples from Face Expression Recognition Dataset, achieving 60%. Further, it can operate on real-time webcam feeds, making it quite practical for real-world applications. All these features put the model in an effective and versatile position in the domain of facial emotion recognition.

The proposed CNN model improves smoothly and steadily in Figure 15 with regards to accuracy stabilization while loss stabilizes during training. By the last epochs of training, overall accuracy was around 60% while training and validation accuracies stabilized after epoch 30. This convergent nature indicates that the training was effectively done and, indeed, generalizes on new data. Training and validation losses also sharply decline during early epochs and stabilize at 1.0, indicating that the model is successfully learning key facial expression features without major overfitting. It performs well in distinguishing different emotions like “Happy” and “Sad,” where there are clear visual signals, such as smiling or downturned lips, leading to higher

prediction accuracy. However, the subtle emotion of "Fear" and "Neutral" is still one of the challenges since they often share overlapping or ambiguous features, which leads to misclassifications. The dataset appears to be imbalance, with strong representation of certain emotions like 'Happiness' and 'Surprise' with respect to less represented ones 'Fear' and 'Anger'. This, in turn makes the model give more frequent classification to the predominant classes and produces lower accuracy when it comes to rare emotions. As shown in Table 1, 'Happiness' and 'Surprise' take approximately 30%, while 'Fear' and 'Anger' take less than 11%. This leads to misclassifications as shown in the confusion matrix in Figure 6, where the model would predict 'Neutral' over 'Fear' because of its lesser representation. Class reweighting, oversampling of the underrepresented classes, or using more advanced loss functions like Focal Loss helps address this imbalance.

4 Conclusion

The proposed CNN-based model for facial emotion recognition has great promise for real-world applications. At an accuracy of 60%, the model has the capacity to generalize well to unseen data and has stable training and validation performance. It is particularly good at detecting well-defined visual features of prominent emotions, such as "Happy" and "Sad". However, the difficult cases, such as the subtle and ambiguous emotions, "Fear" and "Neutral," are still majorly based on dataset imbalance and overlapping features. These necessitate further strategies like data augmentation, class balancing, and sophisticated feature extraction to enhance performance. Integrating with real-time webcam feeds further enhances practicality in applications for security systems, monitoring mental health, and human-computer interaction. In order to get better mood identification, further research should focus on combating dataset diversity, improving the detection of slight emotions, and exploring multimodal approaches. The multimodal learning approaches by integrating facial expressions with speech and physiological signals toward better emotion recognition should be an area for further research. Advances in recent works on modality fusion techniques using Vision Transformers or deep models targeted for multi-source data represent good directions for progress. There also remains the problem of optimally designing CNN architectures for mobile-based real-time applications in emotion analysis. Studies like [ref 1, ref 2] demonstrate how multi-module fusion techniques improve prediction accuracy, which could be applied to emotion recognition tasks

References

[1] Balakrishnan, J., and Dwivedi, Y. K. (2021). Role of cognitive absorption in building user trust and experience. Psychology & Marketing, 38(4), 643–668.

[2] Mavropoulos, T., Symeonidis, S., Tsanousa, A., Giannakeris, P., Rousi, M., Kamateri, E., … and Kompatsiaris, I. "Smart integration of sensors, computer vision and knowledge representation for intelligent monitoring and verbal human-computer interaction." *Journal of Intelligent Information Systems*, vol. 57, no. 2, pp. 321–345, 2020.

[3] Hollender, N., Hofmann, C., Deneke, M., and Schmitz, B. "Integrating cognitive load theory and concepts of human–computer interaction." *Computers in human behavior*, vol. 26, no. 6, pp. 1278–1288, 2020.

[4] Glodek, M., Honold, F., Geier, T., Krell, G., Nothdurft, F., Reuter, S., … and Schwenker, F. "Fusion paradigms in cognitive technical systems for human–computer interaction". *Neurocomputing*, vol. 161, pp. 17–37, 2015.

[5] Ali, S. I., Jain, S., Lal, B., and Sharma, N. "A framework for modeling and designing of intelligent and adaptive interfaces for human computer interaction". *International Journal of Applied Information Systems (IJAIS)*, 2012.

[6] Dutta, S., Mishra, B. K., Mitra, A., and Chakraborty, A. (2022). An analysis of emotion recognition based on GSR signal. ECS Transactions, 107(1), 12535.

[7] Val-Calvo, M., Álvarez-Sánchez, J. R., Ferrández-Vicente, J. M., Díaz-Morcillo, A., and Fernández-Jover, E. "Real-time multi-modal estimation of dynamically evoked emotions using EEG, heart rate and galvanic skin response." *International Journal of Neural Systems*, vol. 30, no. 4, pp. 2050013, 2020.

[8] Raheel, A., Majid, M., Alnowami, M., and Anwar, S. M. "Physiological sensors-based emotion recognition while experiencing tactile enhanced multimedia". *Sensors*, vol. 20, no. 14, pp. 4037, 2020.

[9] Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C. H., Xiang, Y., and He, J. "A review on automatic facial expression recognition systems assisted by multimodal sensor data". *Sensors*, vol. 19, no. 8, pp. 1863, 2019.

[10] Witkower, Z., and Tracy, J. L. "Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems". *Emotion Review*, vol. 11, no. 2, pp. 184–193, 2019.

[11] Ahmed, F., Bari, A. H., and Gavrilova, M. L. "Emotion recognition from body movement". *IEEE Access*, vol. 8, pp. 11761–11781, 2019.

[12] Sisk, C. A., Remington, R. W., and Jiang, Y. V. "Mechanisms of contextual cueing: A tutorial review." *Attention, Perception, and Psychophysics*, vol. 81, pp. 2571–2589, 2019.

[13] Cummings, M., and Stimpson, A. "Identifying critical contextual design cues through a machine learning approach". *AI Magazine*, vol. 40, no. 4, pp. 28–39, 2019.

[14] Martins, R., Bersan, D., Campos, M. F., and Nascimento, E. R. "Extending maps with semantic and contextual object information for robot navigation: a learning-based framework using visual and depth cues." *Journal of Intelligent & Robotic Systems*, vol. 99, no. 3, pp. 555–569, 2020.

[15] Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., and Vandergheynst, P. "Graph signal processing: Overview, challenges, and applications." *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[16] Zhang, J. A., Liu, F., Masouros, C., Heath, R. W., Feng, Z., Zheng, L., and Petropulu, A. "An overview of signal processing techniques for joint communication and radar sensing." *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1295–1315, 2021.

[17] Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., and Chen, S. A. "NeuroKit2: A Python toolbox for neurophysiological signal processing". *Behavior research methods*, pp. 1–8, 2021.

[18] Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction." *Journal of Applied Science and Technology Trends*, vol. 1, no. 1, pp. 56–70, 2020.

[19] Alelyani, S., Tang, J., and Liu, H. "Feature selection for clustering: A review". *Data Clustering*, pp. 29–60, 2018.

[20] Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., ... and Song, J. "Feature: a python package and web server for features extraction and selection from protein and peptide sequences". *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.

[21] Zhang, J., Yin, Z., Chen, P., and Nichele, S. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." *Information Fusion*, vol. 59, pp. 103–126, 2020.

[22] Sarkar, P., and Etemad, A. "Self-supervised ECG representation learning for emotion recognition." *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541–1554, 2020.

[23] Houssein, E. H., Hammad, A., and Ali, A. A. "Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review." *Neural Computing and Applications*, vol. 34, no. 15, pp. 12527–12557, 2022.

[24] Bau, D., Zhu, J. Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. "Understanding the role of individual units in a deep neural network." *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30071–30078, 2020.

[25] Chen, M., Challita, U., Saad, W., Yin, C., and Debbah, M. "Artificial neural networks-based machine learning for wireless networks: A tutorial". *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3039–3071, 2019.

[26] Voigtlaender, F. (2023). The universal approximation theorem for complex-valued neural networks. Applied and computational harmonic analysis, 64, 33–61.

[27] Zhang, Shiqing, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects." Expert Systems with Applications 237, 121692, 2024.

[28] Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., Liu, B. and Tao, J.,. "Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition". Information Fusion, 108, p. 102367, 2024.

[29] Wei, Jie, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. "Learning facial expression and body gesture visual information for video emotion recognition." Expert Systems with Applications, 237, 121419, 2024.

[30] Xu, FeiFan, Deng Pan, Haohao Zheng, Yu Ouyang, Zhe Jia, and Hong Zeng. "EESCN: A novel spiking neural network method for EEG-based emotion recognition." Computer methods and programs in biomedicine 243, 107927, 2024.

Biographies

Akram Ahmad is a dedicated Research Scholar in the Department of Computer Science at Maharishi University of Information Technology, Lucknow. His academic pursuits and research endeavors aim to advance knowledge in computer science and its applications. With a passion for innovation and a commitment to addressing complex challenges, he actively engages in scholarly activities, exploring cutting-edge solutions and contributing to technological advancements. Akram's work reflects a deep interest in fostering progress and collaboration within the academic and research community, making significant work in his chosen field of study.

Vaishali Singh, working as an Associate Professor with the Maharishi School of Engineering & Technology at Maharishi University of Information Technology, Uttar Pradesh, India. Her academic and research focus includes a wide range of contemporary topics such as Convolutional Neural Networks, Scalable Wireless Networks, Wi-Fi Networks, Cloud Computing, Artificial Intelligence, Artificial Neural Networks, Recurrent Neural Networks, and Public Key Systems. She also explores applications in business and technology innovation. With expertise in these areas, Vaishali Singh contributes to

advancing knowledge and developing innovative solutions to address complex challenges in engineering, technology, and interdisciplinary fields.

Kamal Upreti is currently working as an Associate Professor in Department of Computer Science, CHRIST (Deemed to be University), Delhi NCR, Ghaziabad, India. He completed is B. Tech (Hons) Degree from UPTU, M. Tech (Gold Medalist), PGDM(Executive) from IMT Ghaziabad and PhD in Department of Computer Science & Engineering. He has completed Postdoc from National Taipei University of Business, TAIWAN funded by MHRD.

He has published 50+ Patents, 32+Magazine issues and 120+ Research papers in in various reputed Journals and international Conferences. His areas of Interest such as Modern Physics, Data Analytics, Cyber Security, Machine Learning, Health Care, Embedded System and Cloud Computing. He has published more than 45+ authored and edited books under CRC Press, IGI Global, Oxford Press and Arihant Publication. He is having enriched years' experience in corporate and teaching experience in Engineering Colleges.

He worked with HCL, NECHCL, Hindustan Times, Dehradun Institute of Technology and Delhi Institute of Advanced Studies, with more than 15+ years of enrich experience in research, Academics and Corporate. He also worked in NECHCL in Japan having project – "Hydrastore" funded by joint collaboration between HCL and NECHCL Company. Dr. Upreti worked on Government project – "Integrated Power Development Scheme (IPDS)" was launched by Ministry of Power, Government of India with the objectives of Strengthening of sub-transmission and distribution network in the urban areas. Currently, he has completed work with Joint collaboration with GB PANT & AIIMS Delhi, under funded project of ICMR Scheme on Cardiovascular diseases prediction strokes using Machine Learning Techniques from year 2017–2020 of having fund of 80 Lakhs. He got 5 Lakhs fund from

DST SERB for conducting International Conference, ICSCPS-2024, 13–14 Sept 2024. Recently, he got 10 Lakhs fund from AICTE – Inter-Institutional Biomedical Innovations and Entrepreneurship Program (AICTE-IBIP) for 2024–2026. He has attended as a Session Chair Person in National, International conference and key note speaker in various platforms such as Skill based training, Corporate Trainer, Guest faculty and faculty development Programme. He awarded as best teacher, best researcher, extra academic performer and Gold Medalist in M. Tech programme