Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

Emotify: Real-Time Emotion-Based Music Player

Ms. Purva Kalambate, Seema Hanchate, Ms. Poonam More
Department of Electronics and communication
Usha Mittal Institute of Technology, SNDT Women's University, Mumbai, India.

***Abstract*—Emotify is an intelligent music player system that leverages real-time facial emotion recognition to detect the user's mood and automatically play songs that align with their emotional state. The system is built using a Convolutional Neural Network (CNN) with three convolutional layers and is trained on a combination of two well-known datasets—FER2013 and CK+48—for robust emotion classification. It accurately identifies seven emotions: anger, sadness, happiness, neutral, disgust, surprise and fear. Compared to existing machine learning models such as VGG16, Xception and InceptionV3—which often involve complex architectures and higher computational demands—Emotify stands out for its simplicity, speed and efficiency. While models like VGG16 and InceptionV3 achieved accuracies in the range of 67–76%, Emotify attains a notably higher accuracy of 90% with fewer parameters and faster inference time. Its lightweight design ensures suitability for real-time deployment, delivering an engaging and personalized music experience based on users' emotional states.**

***Keywords*—Music Player System, CNN, Real-Time Emotion Detection, Deep Learning, Mood Detection.**

I. INTRODUCTION

Facial Emotion Recognition (FER) involves examining facial cues to identify emotions such as happiness, anger, fear, or sadness. Since facial expressions effectively reflect internal feelings, FER plays an essential role in enhancing the emotional responsiveness of digital systems. Its application is growing in areas like immersive gaming, intelligent advertising, user engagement platforms and healthcare diagnostics [1]. By incorporating FER into human-computer interaction, developers can create emotionally aware systems that respond more naturally to user behaviour. Affective computing—a field focused on designing systems that perceive and react to emotions—relies extensively on FER to detect and interpret human affect. In the healthcare sector, this technology supports professionals in monitoring emotional conditions, identifying mood disorders, and providing customized treatment strategies [2][3].

In today's demanding lifestyle, individuals frequently encounter emotional strain due to work overload, economic difficulties, and the aftermath of global challenges. Music is a widely adopted method for emotional relief, though its effectiveness greatly depends on how well it resonates with a person's current emotional state [4]. When music fails to align with the listener's mood, its calming impact may be reduced. Given the strong link between emotions and music, using FER to guide music recommendations can help deliver songs that match or improve a user's emotional condition. Such systems have the potential to enhance mental well-being and support emotional balance over time [5].

The paper is arranged as follows: The introduction is followed by a second section. Section 2 presents Related Work, focusing on emotion/expression recognition and the various approaches considered by researchers. Next, Section 3 provides a Background, detailing the main components of the proposed architecture. Section 4 summarizes the implementation used in this study. In Section 5, the comparison with ml model, followed by the experiments and results in Section 6. Finally, Section 7 concludes the paper, summarizing the key findings and outlining future research directions.

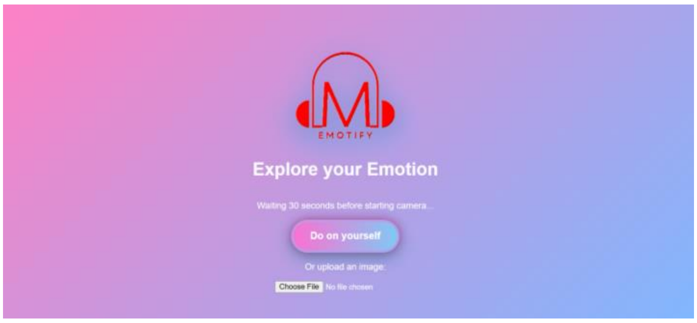
II. RELATED WORK

Recent studies in Facial Emotion Recognition (FER) have demonstrated significant progress using deep learning techniques, particularly Convolutional Neural Networks (CNNs). In the study by the author, a CNN-based model trained on FER2013 and AffectNet datasets achieved 71.61% accuracy while reducing training time [1]. In another study, the authors explored VGG16 and VGG19 architectures, where VGG19 outperformed with an accuracy of 92.5% [2]. The authors further proposed that VGG16 and DenseNet121 surpassed ensemble models, achieving 86% accuracy, highlighting the strength of deep learning models in capturing subtle facial cues [3]. The authors proposed a system using MobileNetV2 and a CNN-based emotion module that effectively detected emotions like happiness or sadness, enabling real-time music recommendations on mobile devices [4]. Another author proposed a hybrid model where PCA was used for dimensionality reduction followed by SVM with a polynomial kernel, achieving 100% accuracy in mood-based audio selection [5]. A comprehensive review analyzed methodologies, technological advancements, and challenges in FER, emphasizing the extension of recognition beyond basic emotions and advocating for multimodal approaches [6]. In the given study, deep learning models such as ResNet50 and Inception were preferred for identifying dynamic facial expressions due to their effective feature extraction and the integration of attention mechanisms improved region-specific classification [7]. The authors reported a CNN-based model that demonstrated generalization with a training accuracy of 92%, validation accuracy of 88% and a test accuracy of 85% after 20 epochs [8]. In another review, FER techniques for educational research were categorized into manual labelling and machine learning-based automatic annotation methods, providing a framework for emotional understanding in academic environments [9].

In the domain of music recommendation, several studies have integrated FER to personalize playlist generation. In another study, the model achieved 66% accuracy and users reported satisfaction with the emotion-based playlist recommendations [10]. The proposed system in a separate study processed real-time video feeds and utilized musical features like tempo, pitch, and volume for emotion-to-music mapping [11]. The authors developed a CNN-based system integrated with OpenCV, TensorFlow and Tkinter, which classified facial emotions and played music in real-time based on predictions [12]. A Streamlit-based web application was proposed, where the webcam captured facial expressions and matched detected emotions with relevant music [13]. In the given study, a CNN model trained on the FER dataset accurately recognized Happy, Fear, Sad and Surprise emotions with 96%, 97%, 93% and 94% accuracy respectively, showcasing the advantages of deeper CNN architectures [14]. The authors proposed a CNN-based model to improve accuracy and reduce time and cost in music recommendation by identifying emotions from facial images [15]. In a comparative study, the authors showed that CNN outperformed SVM, with respective accuracies of 81% and 77% [16]. Further, the authors proposed a 6-layer CNN with max pooling that achieved 83% accuracy and could detect happy, sad, and neutral emotions [17]. Another study aimed to design a real-time music recommendation system using CNN architecture composed of convolutional, pooling and fully connected layers to improve user satisfaction [18]. In the proposed methodology, several classifiers were evaluated for music emotion recognition, where Random Forest achieved the highest success rate at 75%, followed by SVM at 63%, Decision Tree at 60% and Naive Bayes at 50% [19]. A recent FER method proposed by the authors achieved an accuracy of 93.32%, outperforming many existing approaches [20]. The authors also explored advanced face detection and recognition models such as MTCNN and FaceNet, which extracted facial embeddings to enhance prediction accuracy [21]. Finally, a personalized music recommendation system was proposed using a chatbot that assessed a user's emotional state through general conversation and based on the cumulative score, generated a mood-based playlist [22]. The authors' proposed model, based on the VGG-19 architecture and optimized using CK+, JAFFE, and FER2013 datasets, outperforms existing image sentiment analysis systems by achieving 99% accuracy on CK+, 93% on JAFFE, and 65% on FER2013 [23]. A novel FERConvNet_HDM model, combining hybrid denoising methods with convolutional neural networks, outperforms VGG16 and VGG19 by achieving 85% accuracy on FER2013 and 95% on the proposed LRFE dataset for facial expression recognition on low-resolution images [24]. The author's proposed CNN model, combined with a novel hybrid filtering method, significantly outperforms traditional models by achieving 85% accuracy on FER2013, highlighting the effectiveness of hybrid filtering for facial emotion recognition on low-resolution images [25].

III. PROPOSED MODEL

The proposed system employs a Convolutional Neural Network (CNN)-based architecture for facial emotion recognition designed to identify six core emotional states: Happy, Sad, Fear, Anger, Neutral and Surprise. This emotion detection mechanism is seamlessly integrated with a music recommendation engine leveraging the YouTube API to deliver emotion-specific song playback. Users are provided with three interaction options: they can upload an image from their local device for emotion detection, manually search for a song without using the emotion recognition feature or capture a real-time image using a webcam to detect emotions and automatically receive music recommendations. For real-time webcam usage, the system offers a 7-second interval during which users can adjust their position within the camera frame to ensure optimal facial visibility. After this period, the system captures an image; if the image quality is compromised due to poor lighting, facial misalignment or motion blur, the system automatically reinitiates the capture process until a clear image is acquired.



Explore your Emotion

Waiting 30 seconds before starting camera...

Or upload an image:

Fig.1. User input options for music recommendation

Once a clear or uploaded image is obtained, it undergoes a series of preprocessing steps including resizing to meet the model’s input specifications and converting it to grayscale to enhance computational efficiency and model accuracy as shown in fig 2. The refined image is then analysed by the trained CNN model to determine the user's emotional state, which is displayed on-screen along with a notification indicating that the system is retrieving a suitable music match. Simultaneously, the system logs the detected emotion along with the current timestamp into a text file enabling the monitoring and tracking of users' mood patterns over time. During a brief waiting period, the backend engine utilizes the YouTube API to fetch a song aligned with the detected emotion which then begins playing automatically to enrich the user’s interactive experience. Additionally, a curated list of recommended songs is displayed offering users an extended range of musical choices tailored to their emotional profile.

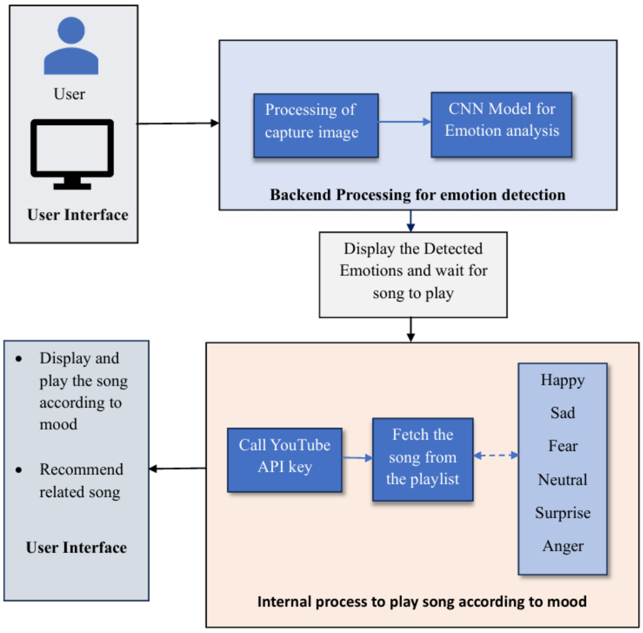

User
User Interface
Processing of capture image
CNN Model for Emotion analysis
Backend Processing for emotion detection
Display the Detected Emotions and wait for song to play
Display and play the song according to mood
Recommend related song
User Interface
Call YouTube API key
Fetch the song from the playlist
Happy
Sad
Fear
Neutral
Surprise
Anger
Internal process to play song according to mood

Fig. 2. Proposed System Architecture

IV. IMPLEMENTATION

The methodology for real-time facial expression recognition using a CNN model involves several steps:

A. Preprocessing

The captured images were pre-processed to enhance their quality and ensure compatibility with the CNN model. Image was resized to a standardized dimension of 48×48 pixels. Additionally, the image was converted to grayscale to reduce computational complexity while retaining critical features necessary for accurate facial expression analysis.

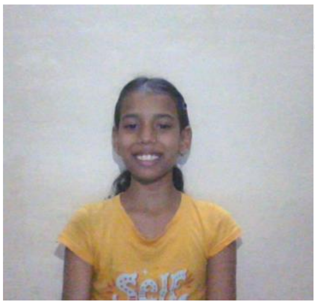


Fig. 3. Before pre-processing of capture image

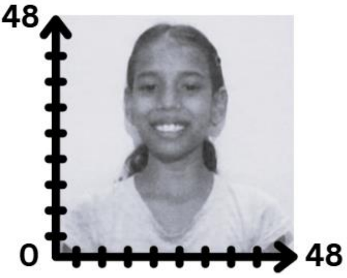


Fig. 4. After pre-processing of capture image

In fig 3 and 4, The images are pre-processed to enhance the quality and suitability for input into the CNN model.

B. CNN Model Architecture Design

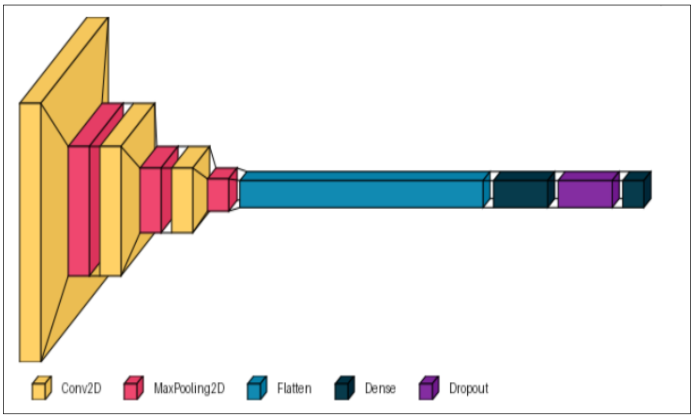


Fig.5. CNN Model Architecture

The CNN model architecture implemented in this system was deliberately kept simple and efficient to support real-time emotion recognition from facial images. The model was trained on a combination of the FER2013 and CK+48 datasets to enhance its accuracy and generalization across diverse facial expressions. The structure consisted of several essential layers as shown in fig 5 and 6:

- Convolutional Layers: These layers were responsible for detecting fundamental visual features such as edges and textures. By using only two convolutional layers, the model focused on extracting the most critical information while maintaining a lightweight structure, suitable for faster processing.
- Max-Pooling Layers: Following each convolutional layer, a max-pooling layer was applied to down sample the feature maps. This reduced the overall data size while preserving key information, thereby enhancing computational efficiency and reducing memory usage.
- Flatten Layer: After feature extraction, the multi-dimensional feature maps were flattened into a one-dimensional vector. This transformation prepared the data for input into the fully connected layers.
- Dense Layers: These layers served as the decision-making components of the model, interpreting the learned features to classify the user's emotional state.
- Dropout Layers: To prevent overfitting and improve the model's ability to generalize, dropout was introduced by randomly disabling a fraction of neurons during training.

The decision to use only two convolutional layers offered several advantages: it kept the architecture less complex, reduced training time, minimized the risk of overfitting and made the system more suitable for environments with limited computational resources. This streamlined design proved effective for delivering fast and accurate emotion predictions without significant compromises in performance.

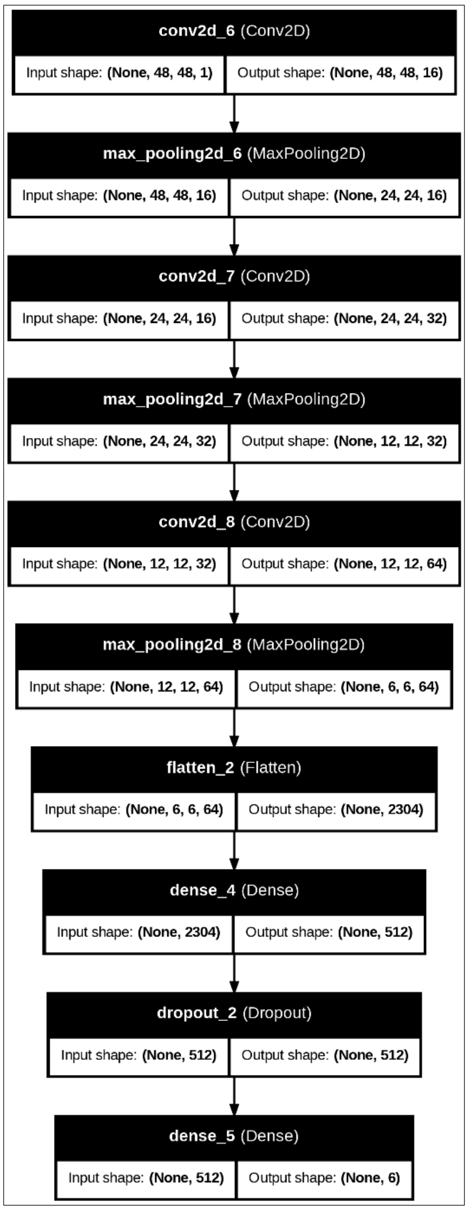

conv2d_6 (Conv2D)
Input shape: (None, 48, 48, 1)
Output shape: (None, 48, 48, 16)
max_pooling2d_6 (MaxPooling2D)
Input shape: (None, 48, 48, 16)
Output shape: (None, 24, 24, 16)
conv2d_7 (Conv2D)
Input shape: (None, 24, 24, 16)
Output shape: (None, 24, 24, 32)
max_pooling2d_7 (MaxPooling2D)
Input shape: (None, 24, 24, 32)
Output shape: (None, 12, 12, 32)
conv2d_8 (Conv2D)
Input shape: (None, 12, 12, 32)
Output shape: (None, 12, 12, 64)
max_pooling2d_8 (MaxPooling2D)
Input shape: (None, 12, 12, 64)
Output shape: (None, 6, 6, 64)
flatten_2 (Flatten)
Input shape: (None, 6, 6, 64)
Output shape: (None, 2304)
dense_4 (Dense)
Input shape: (None, 2304)
Output shape: (None, 512)
dropout_2 (Dropout)
Input shape: (None, 512)
Output shape: (None, 512)
dense_5 (Dense)
Input shape: (None, 512)
Output shape: (None, 6)

Fig 6. Detailed CNN Architecture

C. Music integration

The YouTube Data API was integrated to fetch and stream songs based on the user's detected emotion. It is an open-source and freely available service, chosen for its ease of use, cost-effectiveness and access to a vast library of music content. Once the emotion is identified by the CNN model, the system queries the API to retrieve a playlist that matches the recognized mood. One song from this playlist is automatically selected and played to align with the user's emotion, while the remaining tracks are displayed as mood-based recommendations.

V. COMPARISON WITH ML MODEL

To evaluate the performance of the proposed CNN model, a comparative analysis was conducted against several well-established deep learning architectures, including VGG16, VGG19, MobileNet, Xception and Inception. While the standard models demonstrated high accuracy, they required significantly more computational resources, resulting in slower inference times that are less suitable for real-time applications. In contrast, the proposed CNN model, despite its simpler architecture, achieved competitive accuracy with much faster processing speeds, making it more effective for real-time facial emotion recognition tasks. The model accuracies trained on the FER2013 dataset and the proposed combined dataset (FER2013 and CK+48) were compared to evaluate performance improvements.

A. VGG16

VGG16 is a widely adopted deep learning architecture recognized for its simplicity and effectiveness in image classification tasks. It comprises 16 weight layers, primarily composed of small 3×3 convolutional filters followed by max-pooling layers, and concludes with fully connected layers. The consistent and deep structure of VGG16 enables it to learn hierarchical features efficiently, making it well-suited for facial emotion recognition (FER). As shown in Figure 7, the VGG16 model achieved an accuracy of 74% on the custom facial emotion dataset, demonstrating its capability to extract meaningful facial features. According to prior research, VGG16 has achieved a test accuracy of 69.14% on the standard FER2013 dataset. [21]

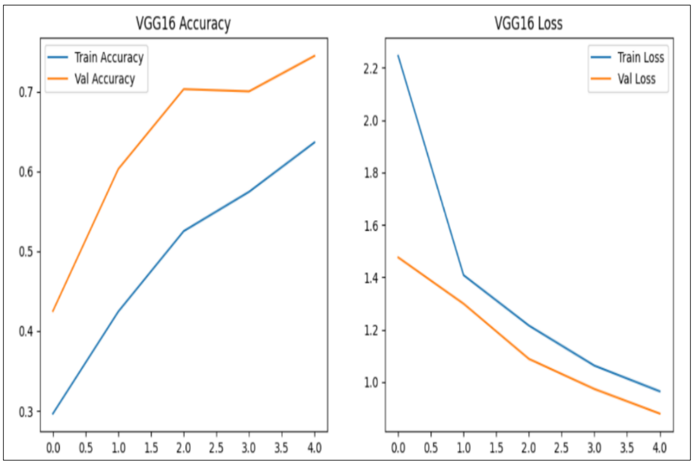


Fig 7. VGG16 accuracy and losses graph

In fig 8, the confusion matrix shows that the VGG16 model performed well for Happy (56), Neutral (57) and Surprise (53), with high correct predictions. Anger (44) was moderately accurate, while Fear (27) and Sad (31) had lower accuracy due to frequent misclassifications. Fear was often confused with surprise and sad with neutral.

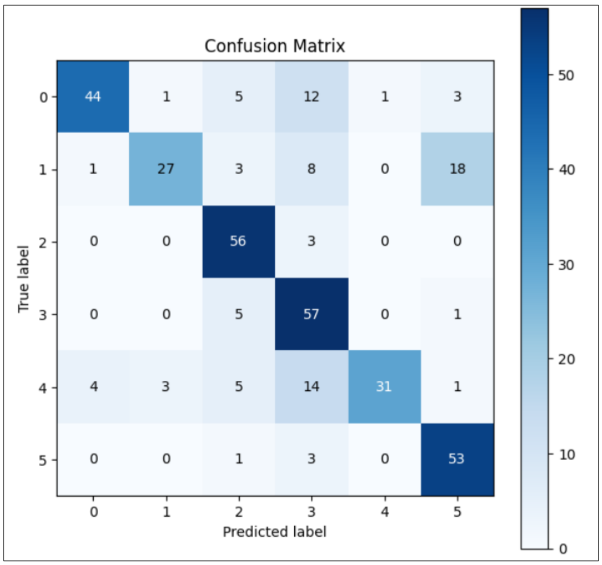
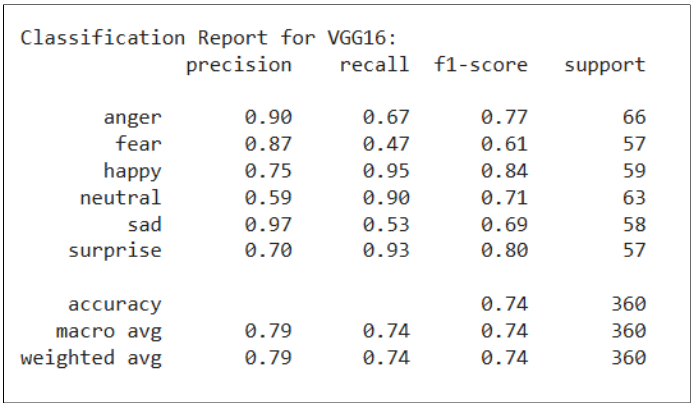


Fig 8. VGG16 Confusion matrix



```
Classification Report for VGG16:
              precision    recall  f1-score   support

       anger       0.90      0.67      0.77        66
        fear       0.87      0.47      0.61        57
       happy       0.75      0.95      0.84        59
     neutral       0.59      0.90      0.71        63
         sad       0.97      0.53      0.69        58
    surprise       0.70      0.93      0.80        57

    accuracy                           0.74       360
   macro avg       0.79      0.74      0.74       360
weighted avg       0.79      0.74      0.74       360
```

Fig 9. VGG16 performance matrix

The VGG16 model achieved an overall accuracy of 74% on the test dataset. In fig 8, it performed best in recognizing Happy (F1: 0.84) and Surprise (F1: 0.80) while struggling slightly with Fear (F1: 0.61) and Sad (F1: 0.69) due to lower recall. Neutral had high recall (0.90) but lower precision (0.59), indicating overprediction. The model showed a balanced performance with a macro and weighted F1-score of 0.74.

B. VGG19

VGG19 is a deep convolutional neural network architecture characterized by its uniform structure, with multiple stacked convolutional layers using small (3×3) filters. It has proven highly effective for feature extraction and has achieved strong performance in various image classification tasks, including facial emotion recognition. As shown in Fig. 10, the VGG19 training results indicate that validation accuracy steadily increased across epochs, reaching over 57%, while training accuracy reached around 37%. On the FER2013 dataset, VGG19 achieved an accuracy of approximately 53%, demonstrating its capability to handle challenging facial emotion recognition tasks [24].

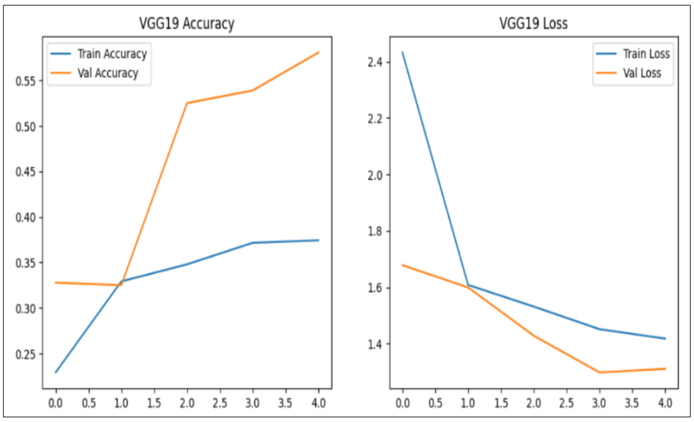


Fig 10. VGG19 accuracy and losses graph

In fig 11, The VGG19 confusion matrix shows good accuracy for emotions like classes 2, 4 and 5, but struggles with class 1 and class 3, where many samples are misclassified. This suggests VGG19 handles some emotions well but has difficulty with overlapping expressions.

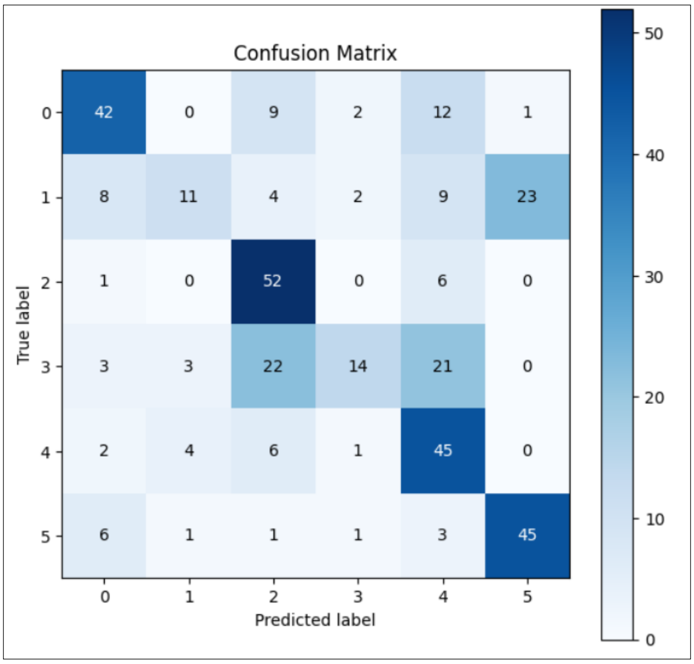
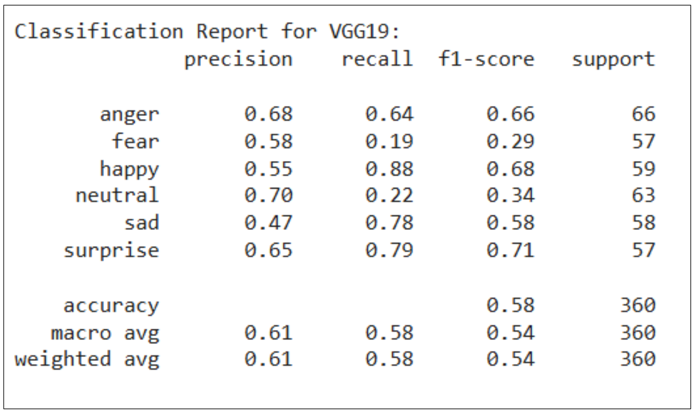


Fig 11. VGG19 Confusion matrix

In fig 12, it performed relatively well in recognizing emotions like happy (F1-score: 0.68) and surprise (F1-score: 0.71) but showed poor recall for fear (0.19) and neutral (0.22), indicating difficulty in correctly identifying those emotions. The macro average F1-score was 0.54, suggesting that the model struggled to balance precision and recall across all classes.



```
Classification Report for VGG19:
              precision    recall  f1-score   support

       anger       0.68      0.64      0.66        66
        fear       0.58      0.19      0.29        57
       happy       0.55      0.88      0.68        59
     neutral       0.70      0.22      0.34        63
         sad       0.47      0.78      0.58        58
    surprise       0.65      0.79      0.71        57

    accuracy                           0.58       360
   macro avg       0.61      0.58      0.54       360
weighted avg       0.61      0.58      0.54       360
```

Fig 12. VGG19 performance matrix

C. MobileNet

MobileNet is an efficient and lightweight convolutional neural network architecture developed specifically for mobile and embedded applications. It utilizes depthwise separable convolutions to significantly reduce computational complexity and model size, making it well-suited for real-time tasks such as facial emotion recognition. As shown in fig 13, The training and validation accuracy showed steady improvement, with validation accuracy exceeding 80% by the fourth epoch. The loss curves consistently declined, indicating effective learning and minimal overfitting. On the FER2013 dataset, MobileNet achieved an accuracy of approximately 57%, which is lower than its performance on the proposed dataset.

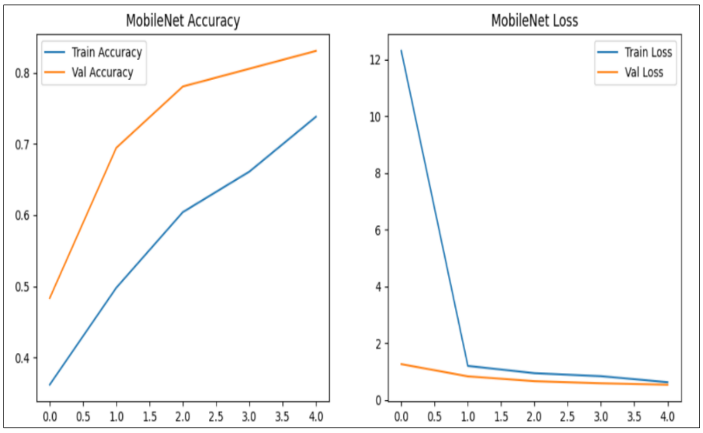
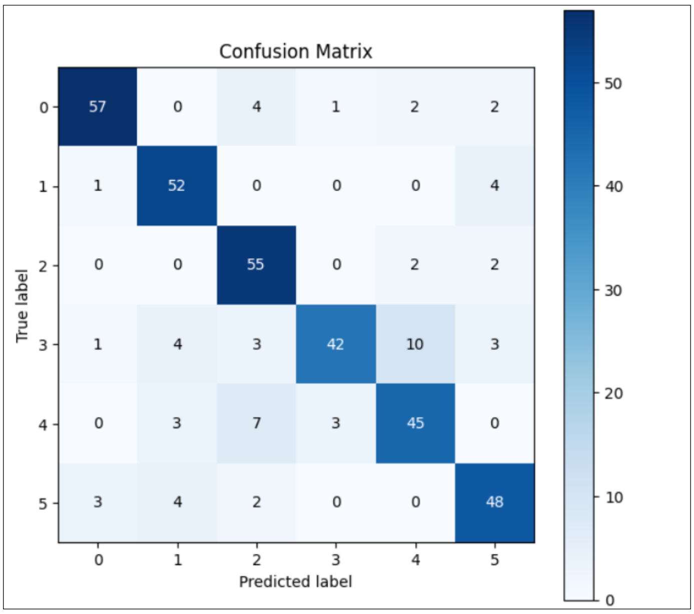


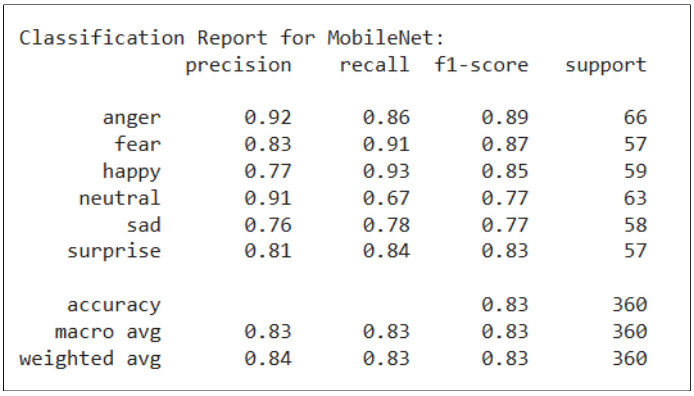
Fig 13. MobileNet accuracy and losses graph

As shown in fig 14, this confusion matrix for MobileNet shows strong performance across most emotion classes. High values along the diagonal indicate correct predictions, especially for classes like 'anger' (57), 'fear' (52) and 'happy' (55). Misclassifications were relatively few, with some confusion between 'sad' and 'neutral' and 'surprise' and 'fear'.



	0	1	2	3	4	5
0	57	0	4	1	2	2
1	1	52	0	0	0	4
2	0	0	55	0	2	2
3	1	4	3	42	10	3
4	0	3	7	3	45	0
5	3	4	2	0	0	48

Fig 14. MobileNet Confusion matrix



```
Classification Report for MobileNet:
              precision    recall  f1-score   support

       anger       0.92      0.86      0.89        66
        fear       0.83      0.91      0.87        57
       happy       0.77      0.93      0.85        59
     neutral       0.91      0.67      0.77        63
         sad       0.76      0.78      0.77        58
    surprise       0.81      0.84      0.83        57

    accuracy                           0.83       360
   macro avg       0.83      0.83      0.83       360
weighted avg       0.84      0.83      0.83       360
```

Fig 15. MobileNet performance matrix

In the above fig, the classification report for MobileNet shows an impressive test accuracy of 83.06%. Emotions like happy, fear, and anger are classified with high precision and recall, especially happy with a recall of 0.93 and anger with a precision of 0.92. The model maintains balanced performance across all classes, with both macro and weighted averages of precision, recall and F1-score at 0.83, indicating strong and consistent emotion recognition.

D. InceptionV3

InceptionV3 is a deep convolutional neural network developed by Google, known for its efficient architecture utilizing Inception modules and factorized convolutions to reduce computational costs. It is approximately 48 layers deep and processes input images of size 299×299. Designed primarily for image classification tasks, InceptionV3 effectively balances network depth and performance. When applied to facial emotion recognition, it achieved an accuracy of around 67%, demonstrating decent performance for a general-purpose model. On a different dataset, InceptionV3 achieved an accuracy of approximately 71% [1].

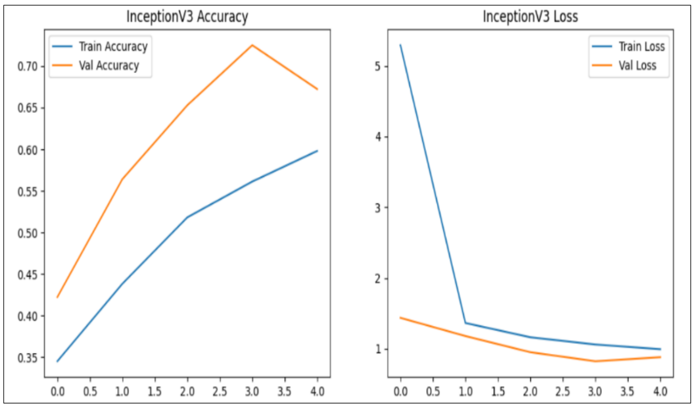
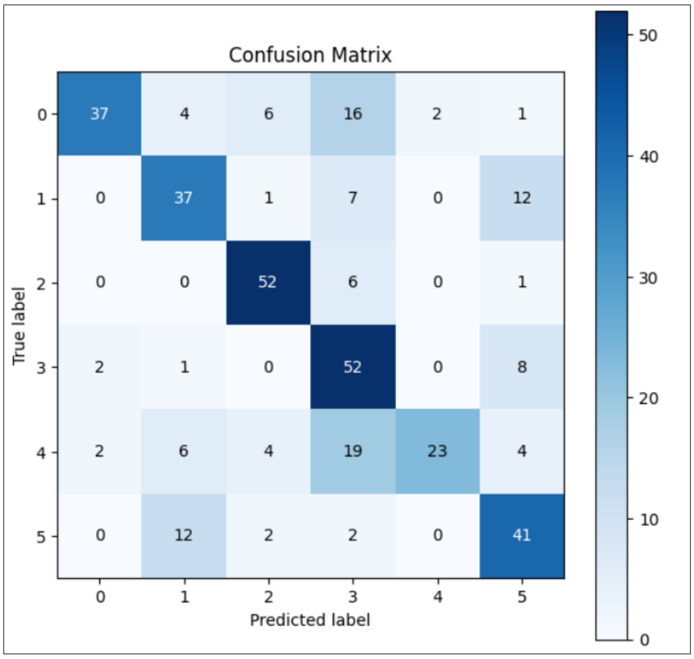


Fig 16. InceptionV3 accuracy and losses graph

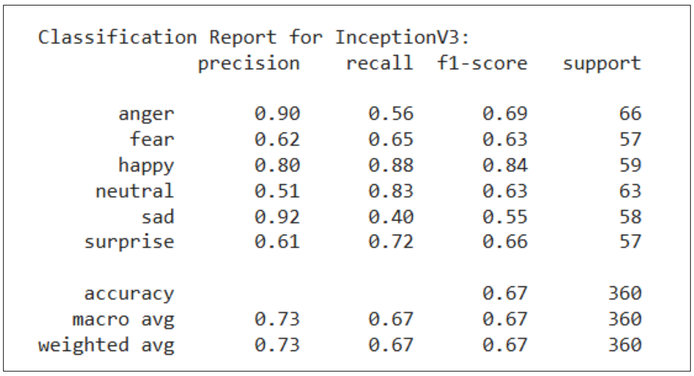
In the fig 17, the model performs well in classes 2 and 3, with 52 correct predictions each, reflecting high accuracy in those categories. Class 5 also shows strong results with 41 correct classifications. However, there is some confusion in classes like 4 and 1, where misclassifications are more frequent. This analysis helps in identifying specific classes that may need more data or fine-tuning to improve model performance.



True label \ Predicted label	0	1	2	3	4	5
0	37	4	6	16	2	1
1	0	37	1	7	0	12
2	0	0	52	6	0	1
3	2	1	0	52	0	8
4	2	6	4	19	23	4
5	0	12	2	2	0	41

Fig 17. InceptionV3 Confusion matrix

As given in fig 18, Among the six emotion classes, the model performs best on "happy" (F1-score: 0.84) and "sad" shows the lowest recall (0.40), meaning many sad instances were missed. Precision is highest for "sad" and "anger". The overall macro and weighted F1-scores stand at 0.67, reflecting balanced performance across classes.



```
Classification Report for InceptionV3:
              precision    recall  f1-score   support

       anger       0.90      0.56      0.69        66
        fear       0.62      0.65      0.63        57
       happy       0.80      0.88      0.84        59
     neutral       0.51      0.83      0.63        63
         sad       0.92      0.40      0.55        58
    surprise       0.61      0.72      0.66        57

    accuracy                           0.67       360
   macro avg       0.73      0.67      0.67       360
weighted avg       0.73      0.67      0.67       360
```

Fig 18. InceptionV3 performance matrix

E. Xception

Xception is a deep convolutional neural network developed by Google as an extension of the Inception architecture. Xception consists of 71 layers and processes input images of size 299×299. It is widely used for image classification tasks due to its strong performance. When applied to facial emotion recognition, Xception achieved an accuracy of around 76%, demonstrating its effectiveness. However, on the FER2013 dataset, the model achieved an accuracy of approximately 52%, reflecting the challenges posed by the dataset's complexity [25].

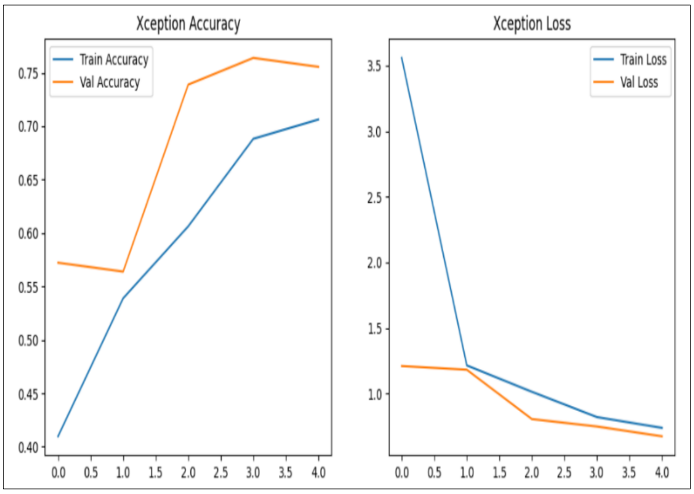


Fig 19. Xception accuracy and losses graph

In fig 20, the Xception model's confusion matrix shows high accuracy for Happy (56), Neutral (54) and Anger (52). Surprise (44) also performs well, while Sad (26) and especially Fear (23) face more misclassifications. Fear is often confused with Anger and Surprise. Overall, the model handles maximum emotions well but struggles slightly with subtle ones like Fear and Sad.

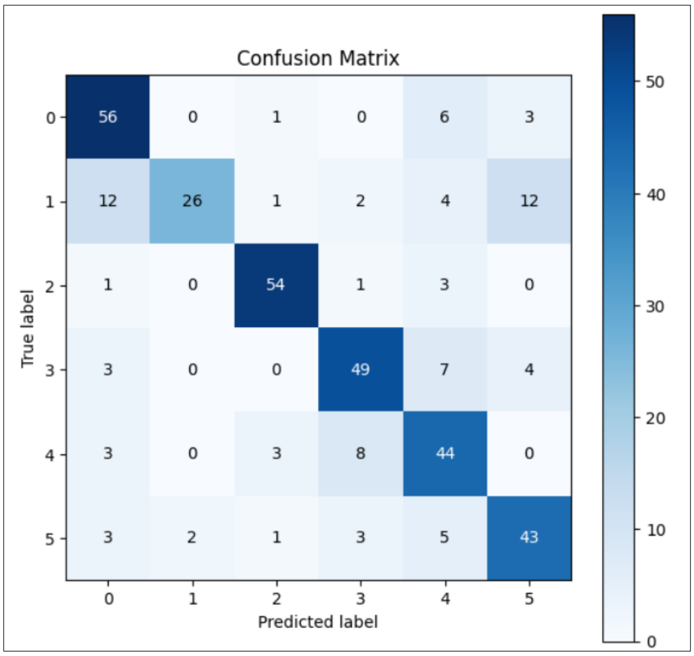
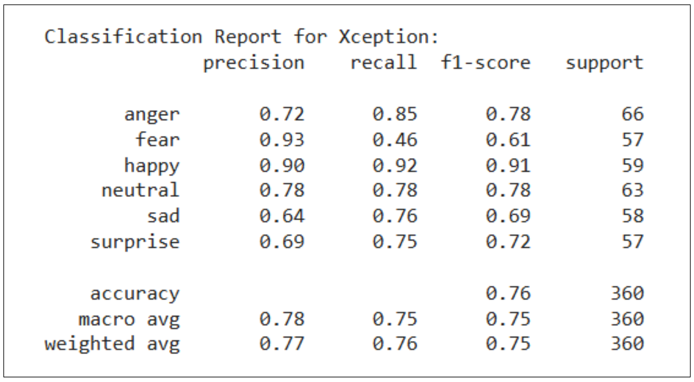


Fig 20. Xception Confusion matrix

As shown in fig 21, the model performs best on Happy with an F1-score of 0.84, followed by Anger and Neutral. Fear has the lowest recall at 0.46, indicating frequent misclassification. The macro and weighted F1-scores are 0.75, reflecting balanced performance across classes.



```
Classification Report for Xception:
              precision    recall  f1-score   support

       anger       0.72      0.85      0.78        66
        fear       0.93      0.46      0.61        57
       happy       0.90      0.92      0.91        59
     neutral       0.78      0.78      0.78        63
         sad       0.64      0.76      0.69        58
    surprise       0.69      0.75      0.72        57

    accuracy                           0.76       360
   macro avg       0.78      0.75      0.75       360
weighted avg       0.77      0.76      0.75       360
```

Fig 21. Xception performance matrix

VI. RESULTS

A. Proposed model accuracy and loss

The accuracy plot shows a significant rise in training accuracy from around 0.2 to nearly 1.0, indicating that the model has effectively learned from the training data as shown in fig 22. Validation accuracy also increases from approximately 0.3 to around 0.9, though it begins to plateau after the 15th epoch, suggesting limited improvement in generalization beyond that point. The model achieved a high validation accuracy of 90% when classifying six emotion classes, reflecting strong performance in facial emotion recognition.

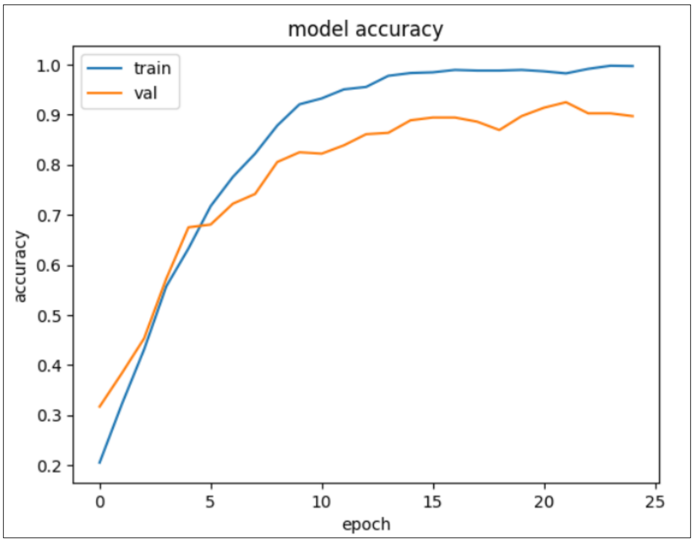


Fig 22. Proposed model accuracy

In the fig 23, the training loss decreases steadily from about 1.8 to below 0.05, confirming effective model optimization. Validation loss follows a similar downward trend until around epoch 10, after which it fluctuates slightly and stabilizes around 0.5. The gap between training and validation loss in later epochs indicates some level of overfitting, but it remains within a reasonable range.

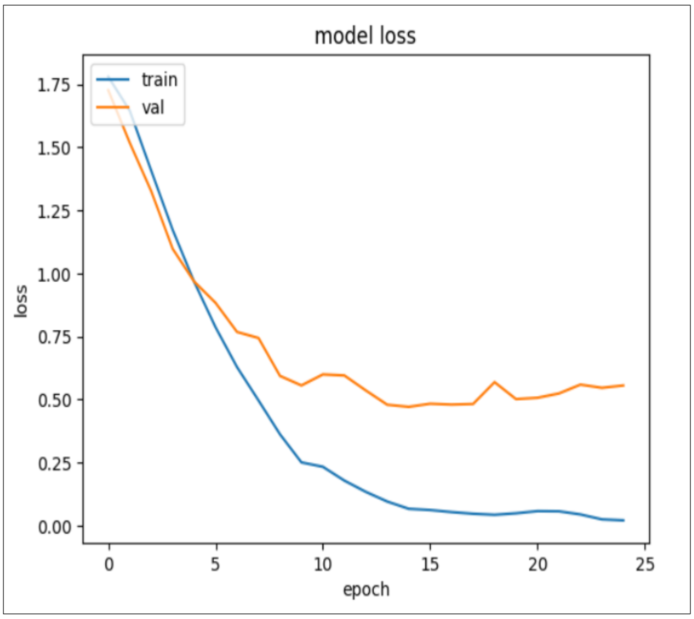


Fig 23. Proposed model losses

B. Proposed model confusion and performance matrix

The confusion matrix reflects strong model performance, with the maximum predictions aligning along the diagonal. For instance, most classes have over 85% correct predictions, indicating reliable classification. However, some misclassifications are observed—such as class 5, which shows scattered errors with a few samples predicted as class 2 and class

3. Similarly, class 0 and class 4, while mostly accurate, show minor confusion with class 1. These numerical patterns suggest that while the model generalizes well, it could still benefit from improved class separation, especially for emotions with similar visual cues.

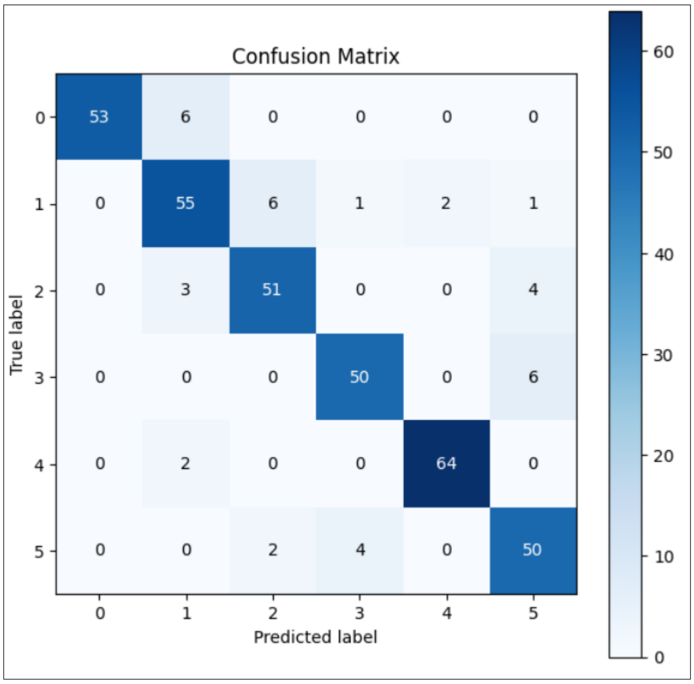
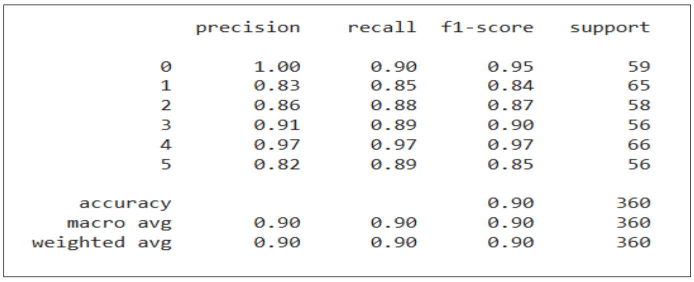


Fig 24. Proposed model Confusion matrix

The model achieved an overall accuracy of 90%. The precision, recall and F1-score are balanced across all classes, with macro and weighted averages all at 0.90. Class 0 had a perfect precision of 1.00 but a slightly lower recall of 0.90 due to a few missed detections. Class 4 stood out with precision and recall at 0.97, indicating strong and consistent classification. Classes 1, 2 and 5 had slightly lower scores, with precision values ranging from 0.82 to 0.86 and recall from 0.85 to 0.89, suggesting some confusion among neighboring classes.

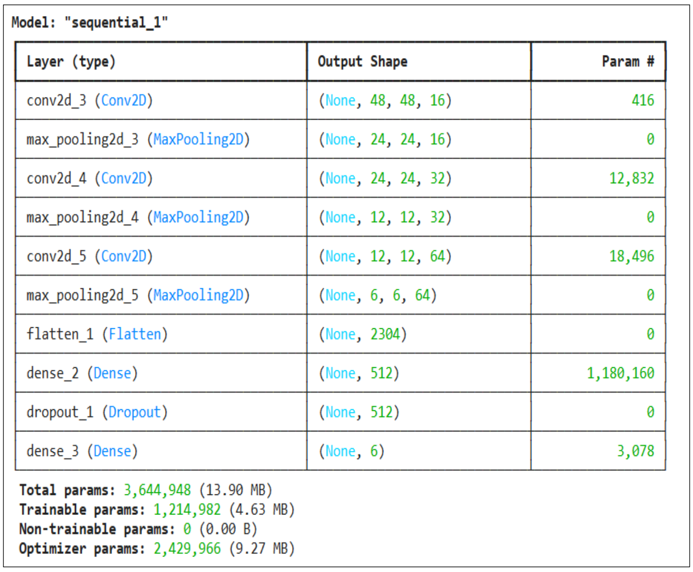


```
              precision    recall  f1-score   support

           0       1.00      0.90      0.95        59
           1       0.83      0.85      0.84        65
           2       0.86      0.88      0.87        58
           3       0.91      0.89      0.90        56
           4       0.97      0.97      0.97        66
           5       0.82      0.89      0.85        56

    accuracy                           0.90       360
   macro avg       0.90      0.90      0.90       360
weighted avg       0.90      0.90      0.90       360
```

Fig 25. Proposed model performance matrix

C. Proposed model summary

As given in fig 26, this CNN model presents a well-structured and efficient architecture suitable for facial emotion recognition tasks. It maintains a strong balance between performance and complexity, with approximately 3.6 million total parameters, of which only 1.2 million are trainable—making it lightweight and ideal for smaller datasets or deployment on resource-constrained devices. A dropout layer is used for regularization, helping to reduce overfitting and improve generalization. Although it lacks the depth and complexity of advanced models like Xception or ResNet, it offers a reliable and interpretable foundation for classifying six basic emotions effectively.



Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 48, 48, 16)	416
max_pooling2d_3 (MaxPooling2D)	(None, 24, 24, 16)	0
conv2d_4 (Conv2D)	(None, 24, 24, 32)	12,832
max_pooling2d_4 (MaxPooling2D)	(None, 12, 12, 32)	0
conv2d_5 (Conv2D)	(None, 12, 12, 64)	18,496
max_pooling2d_5 (MaxPooling2D)	(None, 6, 6, 64)	0
flatten_1 (Flatten)	(None, 2304)	0
dense_2 (Dense)	(None, 512)	1,180,160
dropout_1 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 6)	3,078

Total params: 3,644,948 (13.90 MB)
Trainable params: 1,214,982 (4.63 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2,429,966 (9.27 MB)

Fig 26. Proposed model summary

D. Proposed model comparison with existing model

In the table 1, compares the performance of several deep learning models on six emotion classes—Anger, Happy, Sad, Neutral, Surprise and Fear—along with their overall classification accuracy. Among the models, the Proposed Model outperforms all others, achieving the highest overall accuracy of 90% and showing strong results in detecting Surprise (0.97), Anger (0.95) and Neutral (0.90) emotions.

Table I. Compares the performance of deep learning model

Model	Anger	Happy	Sad	Neutral	Surprise	Fear	Overall accuracy
VGG16	0.77	0.61	0.84	0.71	0.69	0.80	0.74
VGG19	0.66	0.29	0.68	0.34	0.58	0.71	0.58
MobileNet	0.89	0.97	0.85	0.77	0.77	0.83	0.83
InceptionV3	0.69	0.63	0.84	0.63	0.55	0.66	0.67
Xception	0.78	0.61	0.91	0.78	0.69	0.72	0.76
Proposed Model	0.95	0.84	0.87	0.90	0.97	0.85	0.90

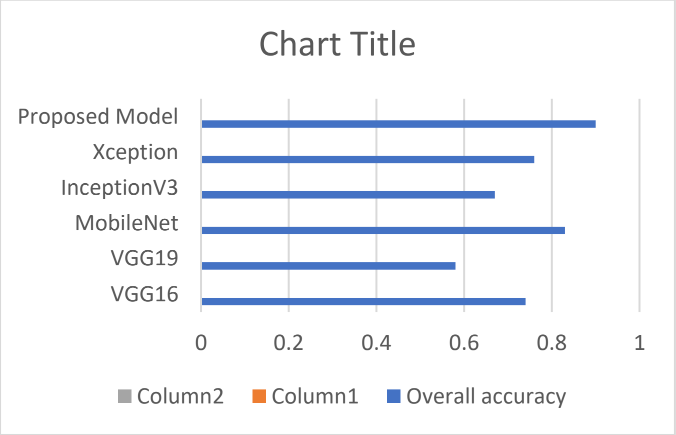


Fig 27. Graphical representation of model accuracy comparison

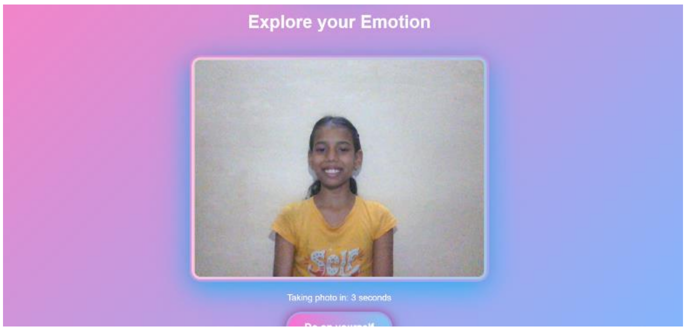
In table 2, Comparison of different facial emotion recognition approaches using the FER2013 dataset, with the proposed system also incorporating the CK+48 dataset. Reference [1] achieved an accuracy of 71.61% using a basic CNN model, while [6] and [16] show improved results with 85% and 81% accuracy respectively, likely due to enhancements in architecture or training techniques.

Table II. Comparison of existing model

References	Dataset	Methodology	Accuracy(%)
[1]	FER2013	CNN	71.61
[8]	FER2013	CNN	85
[16]	FER2013	CNN	81
Proposed system	FER2013 and CK+48	CNN	90.14

In contrast, the Proposed System, which combines FER2013 and CK+48 datasets and uses a CNN-based approach, significantly outperforms the others with an accuracy of 90.14%. This improvement suggests that leveraging multiple datasets and potentially a more optimized architecture enhances the model's ability to generalize and accurately classify emotions.

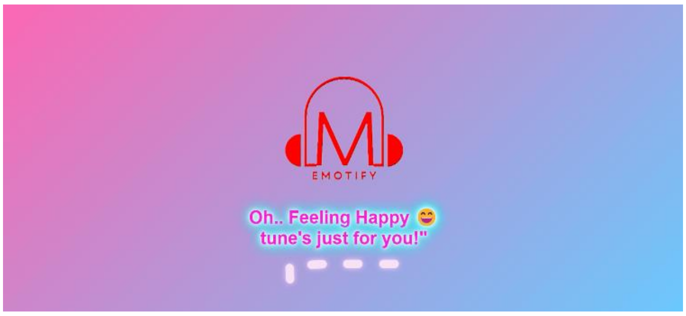
E. Output of WebApp


Explore your Emotion

Taking photo in: 3 seconds

Fig 28. Image capture by using webcam

In fig 28, real-time interface showing live webcam feed and countdown timer for automatic image capture, used for facial emotion recognition and music recommendation.



Oh.. Feeling Happy 😊
tune's just for you!"

Fig 29. Emotion detected: Happy

In fig 29, The emotion detection result screen displays the detected mood, as "Happy," prominently, allowing users to immediately see the recognized emotion.

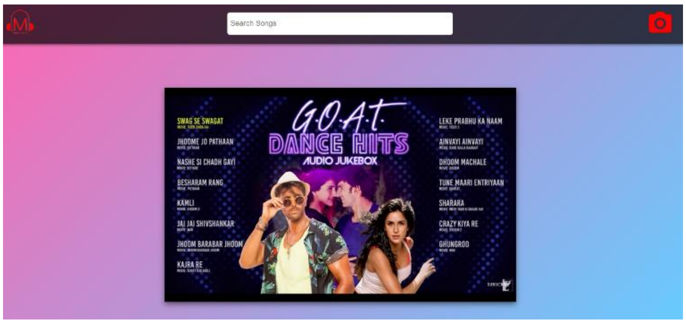
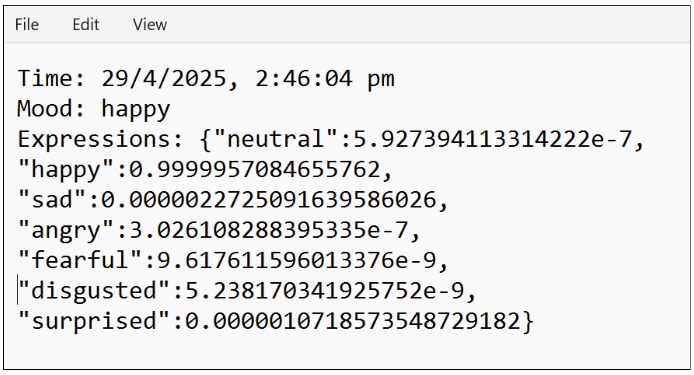


Fig 30. Happy song plays

After the emotion is detected, the system triggers the corresponding playlist or song based on the recognized mood. The detected emotion is "Happy," the system would play upbeat or energetic music that matches a joyful mood. The song is selected from a streamed using an API to provide a personalized music experience aligned with the user's emotional state.



```
Time: 29/4/2025, 2:46:04 pm
Mood: happy
Expressions: {"neutral":5.927394113314222e-7,
"happy":0.9999957084655762,
"sad":0.0000022725091639586026,
"angry":3.026108288395335e-7,
"fearful":9.617611596013376e-9,
"disgusted":5.238170341925752e-9,
"surprised":0.0000010718573548729182}
```

Fig. 31. Logged Emotion Data with Confidence Scores

This output shows the emotion detected by your device at a specific time. It records the timestamp, the predicted mood ("happy" here), and the detailed probability scores for different emotions. The mood is selected based on the highest probability. Saving this data helps track emotions over time and can be used for analysis or debugging.

VII. CONCLUSION AND FUTURE WORK

Emotify is a music-playing system based on facial expression recognition using a CNN-based model, demonstrating high accuracy and robustness across six emotion classes. By integrating the FER2013 and CK+48 datasets, the model benefits from a diverse and balanced dataset, leading to better generalization and enhanced performance, outperforming architectures like VGG16, Xception, and InceptionV3 with an overall accuracy of 90.14%. Beyond detecting emotions, Emotify recommends and plays songs based on the user's current emotional state, offering a personalized and interactive music experience. Although the system is highly effective, future improvements could include integrating multimodal data (such as voice tone or heart rate), using advanced deep learning techniques like attention mechanisms or transformers, expanding the emotion categories, enabling user-customizable music preferences, and deploying the system on mobile or IoT platforms to increase accessibility and real-world usability.

ACKNOWLEDGEMENT

We extend our sincere gratitude to SNDT Women's University for providing the platform and resources necessary to undertake this research. Our heartfelt thanks go to the Usha Mittal Institute of Technology and its dedicated faculty for their consistent support and encouragement throughout our academic journey. We sincerely thank our respected Principal, Dr. Yogesh Nerkar, for his visionary leadership and constant encouragement throughout our research journey.
We are also immensely grateful to our research guide, Dr. Seema Hanchate and co-guide, Ms. Poonam More, for their dedicated mentorship, thoughtful guidance and unwavering support, which contributed to the success of this work. Their expertise, encouragement and commitment played a vital role in shaping and enriching the quality of this work.

REFERENCES

[1] M. Mehrotra, K. P. Singh and Y. B. Singh, "Facial Emotion Recognition and Detection Using Convolutional Neural Networks with Low Computation Cost," 2024 2nd International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2024, pp. 1349-1354, doi: 10.1109/ICDT61202.2024.10489678.

[2] P. Mishra, A. S. Verma, P. Chaudhary and A. Dutta, "Emotion Recognition from Facial Expression Using Deep Learning Techniques," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543313.

[3] G. Bhoomika, V. D. Pujitha, M. Sindusha, C. S. Rekha and B. Suvarna, "Facial Emotion Recognition: A Comparative Study of Pre-trained Deep Learning Models," 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC), Gwalior, India, 2024, pp. 377-382, doi: 10.1109/AIC61668.2024.10730808.

[4] A. V. Gadagkar, S. S, S. Begum and A. S. M, "Emotion Recognition and Music Recommendation System based on Facial Expression," 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690441.

[5] K. Seshaayini, S. L. Srinithya, P. Verma, A. Visalatchi and N. Neelima, "Emotion Recognition Based Music Player," 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 2023, pp. 01-05, doi: 10.1109/ICECCT56650.2023.10179716.

[6] S. Deshmukh, S. Chaudhary, M. Gayakwad, K. Kadam, N. S. More and A. Bhosale, "Advances in Facial Emotion Recognition: Deep Learning Approaches and Future Prospects," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-3, doi: 10.1109/MITADTSoCiCon60330.2024.10574908.

[7] C. M. Prashanth, D. Sree Lakshmi, M. S. Gangadhar, K. Swathi and V. S. Rao, "Facial Emotion Detection: A Comprehensive Survey," 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), Coimbatore, India, 2024, pp. 80-86, doi: 10.1109/ICC-ROBINS60238.2024.10533999.

[8] Uneza, D. Gupta and S. Saini, "Facial Expression Analysis: Unveiling the Emotions Through Computer Vision," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522418.

[9] B. Fang, X. Li, G. Han and J. He, "Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review," in IEEE Access, vol. 11, pp. 112060-112074, 2023, doi: 10.1109/ACCESS.2023.3322454.

[10] R. Shekhawat, S. H. Tedla, J. S. P. Peter and P. E. B, "Syncing Music With Emotions Using Computer Vision," 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2024, pp. 522-527, doi: 10.1109/ICCPCT61902.2024.10672743.

[11] M. S. Guthula and M. Bordoloi, "Music Recommendation System Using Facial Detection Based Emotion Analysis," 2024 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 2024, pp. 296-301, doi: 10.1109/ICETCI62771.2024.10704201.

[12] S. Saranya, M. Varshana Devi, M. Colin Powell, D. Dhanya Bharathy and K. Devatharshini, "Emotion Based Music Recommendation System," 2024 International Conference on Smart Systems for Electrical, Electronics, Communication and Computer Engineering (ICSSEECC), Coimbatore, India, 2024, pp. 255-260, doi: 10.1109/ICSSEECC61126.2024.10649404.

[13] M. M. Joseph, D. Treessa Varghese, L. Sadath and V. P. Mishra, "Emotion Based Music Recommendation System," 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2023, pp. 505-510, doi: 10.1109/ICCIKE58312.2023.10131874.

[14] J. S. Joel, B. Ernest Thompson, S. R. Thomas, T. Revanth Kumar, S. Prince and D. Bini, "Emotion based Music Recommendation System using Deep Learning Model," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 227-232, doi: 10.1109/ICICT57646.2023.10134389.

[15] M. K. Singh, P. Singh and A. Sharma, "Facial Emotion Based Automatic Music Recommender System," 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 699-703, doi: 10.1109/ICSEIET58677.2023.10303573.

[16] P. Sudhakaran, P. K. Nair and A. Suraj, "Music Recommendation using Emotion Recognition," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-7, doi: 10.1109/MysuruCon55714.2022.9972635.

[17] R. Arya, C. Bhatt and M. Mittal, "Music Player Based on Emotion Detection Using CNN," 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), Vijaypur, India, 2022, pp. 1-5, doi: 10.1109/NKCon56289.2022.10126761.

[18] B. Zhai, B. Tang and S. Cao, "Music Recommendation System Based on Real-Time Emotion Analysis," 2022 International Conference on Culture-Oriented Science and Technology (CoST), Lanzhou, China, 2022, pp. 334-338, doi: 10.1109/CoST57098.2022.00075.

[19] D. Unni, A. M. D'Cunha and D. G, "A Technique to Detect Music Emotions Based on Machine Learning Classifiers," 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), Chennai, India, 2022, pp. 136-140, doi: 10.1109/ICPS55917.2022.00033.

[20] P. Vijayeeta and P. Pattnayak, "A Deep Learning approach for Emotion Based Music Player," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 278-282, doi: 10.1109/OCIT56763.2022.00060.

[21] O. Ghosh, R. Sonkusare, S. Kulkarni and S. Laddha, "Music Recommendation System based on Emotion Detection using Image Processing and Deep Networks," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-5, doi: 10.1109/CONIT55038.2022.9847888.

[22] A. Nair, S. Pillai, G. S. Nair and A. T, "Emotion Based Music Playlist Recommendation System using Interactive Chatbot," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1767-1772, doi: 10.1109/ICCES51350.2021.9489138.

[23] Gaurav Meena, Krishna Kumar Mohbey, Ajay Indian and Sunil Kumar,"Sentiment Analysis from Images using VGG19 based Transfer Learning Approach," International Conference on Industry Sciences and Computer Science Innovation, Ajmer, India, Procedia Computer Science 204 (2022) 411–418, doi: 10.1016/j.procs.2022.08.050

[24] Bodavarapu PNR, Srinivas PVVS. (2021) Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques. Indian Journal of Science and Technology. 14(12): 971-983. https://doi.org/10.17485/IJST/v14i12.14

[25] Pavan Nageswar Reddy Bodavarapu and P.V.V.S Srinivas, "An Optimized Neural Network Model for Facial Expression Recognition over Traditional Deep Neural Networks" International Journal of Advanced Computer Science and Applications(IJACSA),12(7), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120751