Video-Based Emotion Recognition Using Generative AI Models

1st Vishwanath Divya
Department of CSE (AI-ML)
SR University
Ananthasagar, Warangal
2203a52130@sru.edu.in

2nd Gande Sai Krishna Priya
Department of CSE (AI-ML)
SR University
Ananthasagar, Warangal
2203a52085@sru.edu.in

3rd Neelisetty Dhanush
Department of CSE (AI-ML)
SR University
Ananthasagar, Warangal
2203a52113@sru.edu.in

4th Padidhala Siddharth Rao
Department of CSE (AI-ML)
SR University
Ananthasagar, Warangal
2203a52115@sru.edu.in

Abstract—**This paper presents a novel approach to video-based emotion recognition leveraging advanced Generative AI models such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). The proposed system integrates a real-time emotion recognition framework that utilizes direct webcam access to analyze and classify human emotions dynamically. With a focus on scalability and accuracy, the methodology emphasizes training on extensive datasets to fine-tune the generative models. The system achieves significant performance benchmarks, including high train-test accuracy and robust confusion matrix evaluations [1]. Real-time recognition via the webcam highlights its applicability in fields such as mental health monitoring, human-computer interaction, and entertainment. The paper concludes with insights into generative model capabilities in improving emotion recognition accuracy and discusses potential areas for future exploration [2].**

I. Introduction

A key element of human-computer interaction (HCI) is emotion detection, which allows machines to comprehend and react to human emotions. This greatly enhances the caliber of interactions in a variety of fields, including healthcare, education, entertainment, and customer service. An essential component of the human-computer interaction (HCI) paradigm is the capacity to recognize and decipher emotional cues. This makes it easy for robots to understand and relate to human emotions, which improves interactions in a variety of fields, including as customer service, education, healthcare, and entertainment [3]. Because human face expression recognition systems can also recognize human body language and voice, the value of emotion recognition systems that concentrate on emotion recognition in video content analysis has been steadily increasing. In addition to being encoded as lips and facial muscles, emotions are also encoded in a sequence that contributes to the spatial and temporal components of a human emotional representation [4].

Deep learning techniques, particularly CNNs, GANs, and VAEs, have a wide range of applications with the advent of Embodied Conversational Agents technologies. In recent years, these techniques have also been used to improve emotion identification systems. CNNs have been used for emotion identification in the past due to their capacity to infer hierarchical encoding of images, since image interpretation is one of the main goals of recognition systems [5]. Understanding the variations in facial expressions related to the different recognized emotions is made possible by these encoding mechanisms. According to empirical data, CNN models outperform other AIs in the context of classification tasks when it comes to identifying emotions (Zadeh et al., 2018). However, challenges remain, such as data scarcity and the complexity of learning generalized features across diverse emotional expressions [6].

These problems are addressed by Generative Adversarial Networks (GANs), which generate fresh synthetic images to augment training data. In order to overcome the issues of class imbalance and data scarcity, GANs can generate new samples that are comparable to the original dataset because deep learning models are made up of two networks: a discriminator and a generator. Strong emotion identification models require a wide variety of high-quality images, which GAN-based models like CycleGAN and StarGAN have demonstrated excellent performance in producing (Zhu et al., 2017, Choi et al., 2018). Additionally, they can translate emotions across other domains, such as changing an image's facial expressions, which enables models to generalize across variations in emotional expressions [7].

The use of Variational Autoencoders (VAEs) is another crucial field of study when it comes to emotion identification because these deep learning structures enable data reconstruction by learning a probabilistic distribution, which may result in the extraction of latent emotional information. With the use of VAEs, which successfully capture the sources of the data disparities, emotional representations with various data aspects can also be produced [8]. When VAEs are used, emotion recognition systems can decipher the complicated relationships between emotional expressions (Kingma and Welling, 2013) [9].

The webcam video streams bring up additional complexity, through the web camera video the emotion detection in real time can be performed for broader purposes [10].Additional complexity is introduced by the webcam video streams, which allow for real-time emotion recognition for more general uses. In order to identify emotions utilizing streams of frames, the models used to process such jobs must be effective and quick

Identify applicable funding agency here. If none, delete this.

enough. These difficulties in real-time emotion recognition may be greatly reduced by combining CNNs, GANs, and VAEs in a single architecture [11]. For instance, GANs can simulate multiple datasets from a single target data with the ability to achieve a wide range of emotional expressions, but CNNs excel at representing spatial information. By learning the latent representation space more thoroughly, VAEs can further enhance these features [12].

II. Literature Review

Emotion recognition through artificial intelligence (AI) has evolved significantly, with breakthroughs deep learning networks like Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and Variational Autoencoders (VAE) are used in emotion recognition technologies, where artificial intelligence (AI) [13] has made notable strides. These advanced algorithms have surfaced that can identify emotions in video, which is crucial for human-computer interaction (HCI). Healthcare, education, customer service, and entertainment all require emotionally aware systems, which has led to advancements in emotion recognition technologies (Tzirakis et al., 2017).

A. *Convolutional Neural Networks (CNNs) in Emotion Recognition*

CNNs have played a key role in the majority of studies concentrating on picture emotion recognition due to their ability to extract hierarchical characteristics from images in lower dimensions. These networks are suitable for emotion recognition since they can also identify intricate facial expression patterns. Using labelled dataset images (courses) as training templates, primitive people employed CNNs for emotion recognition based on Basic Emotions approaches that focused on individual images. These CNNs were used to recognize the emotions of calm and happy people, sad, mad, and surprised people (Zadeh et al. [14], 2018). In this instance, CNN's primary flaw was its inability to cope with prior frames when doing temporal CNN space modeling. Researchers have introduced a 3D application of CNN to incorporate temporal dependency, thus favoring its recognition ability in dynamic scenarios like videos. Zhang et al. (2018) proposed a hybrid CNN framework to recognize facial expressions and emotions simultaneously from image and video data. With this, the system becomes robust to lighting and orientation changes as well as facial occlusion, which are some of the common problems in emotion recognition tasks [15].

B. *Generative Adversarial Networks (GANs) for Data Augmentation*

When it comes to emotion recognition, deep learning models face particular difficulties because there isn't enough labeled data available. Since their introduction by Goodfellow et al [16]. (2014), GANs have given researchers a potent tool for augmenting rich training datasets with synthetic datasets. In order to enhance the quality of the created data, the discriminator evaluates the synthetic samples that the generator creates. This issue occurs when, to put it simply, emotions are not readily available in the datasets. Choi et al. (2018), for example, employed GANs to produce information-rich, varied facial expressions from a small collection of foundational expressions, enabling the enhancement of training data's diversity and richness and producing better leading to imporved model performance on less-represented emotional states [17].

C. *Variational Autoencoders (VAEs) for Latent Representation Learning*

Another important technique for emotion identification arises from variational autoencoders, specifically in the context of learning expressive and rich latent representations of emotional expressions [18]. This capability is of particular value to emotion recognition, as it allows the model to identify latent patterns in emotions and facial expressions that might not otherwise be readily apparent. Examples of applications include the use of VAEs to enhance the variety of representations of emotional expressions for emotion-detection systems (Kingma and Welling, 2013) [19].

Recent research suggests that VAEs and GANs can be combined to produce more realistic emotional representations. One outstanding example is the Conditional VAE-GAN (cVAE-GAN) model, which uses GANs to improve the data realism while utilizing VAEs' capacity to provide a variety of data (Zhu et al., 2017) [20]. High-quality emotional output is produced by this hybrid model, which is used to train more potent emotion identification systems. Furthermore, by ensuring that learnt representations transfer well to unobserved emotional expressions, VAEs effectively reduce overfitting.

D. *Real-Time Emotion Recognition and Webcams*

Particularly when it comes to webcam video feeds, real-time emotion identification presents different issues than those often linked with static image recognition [21]. Among the difficulties are the requirement for latency and speed as well as dynamic changes in face expressions. Numerous studies on the optimization of emotion recognition systems for real-time applications have looked into these. CNNs and other deep learning methods, such LSTM networks, have been combined in some work to extract temporal and spatial information from video data (Tzirakis et al., 2017). Real-time emotion detection from continuous video streams is made possible by the combination, which allows the usage of frames of sequential video data [22].

Moreover, GANs and VAEs have proven to be useful in enhancing the performance of real-time systems. GANs can generate synthetic data to compensate for the lack of real-time data, while VAEs offer more accurate emotional state representations that can adapt in real-time. Combining these models provides a robust framework capable of achieving both high accuracy and fast processing times [23].

III. Methodology

The methodology for this research on video-based emotion recognition using Generative AI models such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs) is designed to integrate these techniques into a unified system capable of real-time emotion detection. The system leverages webcam video feeds to capture human facial expressions, which are then analyzed using the aforementioned deep learning models to identify and classify emotions. The key components of the methodology include data preprocessing, model architecture, real-time integration, and evaluation metrics [24].

A. Data Collection and Preprocessing

The first step in the methodology is the collection and preprocessing of video data. Emotion recognition from video involves analyzing a sequence of images (frames) captured from video streams. The preprocessing step is crucial in ensuring that the input data is in a suitable format for feeding into the deep learning models. The following steps are followed during preprocessing:

- **Video Frame Extraction:** The video stream captured through a webcam is divided into individual frames using OpenCV, a popular computer vision library. Each frame is a single image, representing a snapshot of the video at a specific point in time. The frame rate of the video stream is set to capture approximately 30 frames per second (FPS), ensuring smooth motion analysis.
- **Face Detection and Alignment:** To focus on the relevant features for emotion recognition, face detection is performed on each frame. The face is localized using methods like the Haar Cascade Classifier or more advanced deep learning-based face detectors like the MTCNN (Multi-task Cascaded Convolutional Networks) (Zhang et al., 2016). After detecting the face, the image is aligned so that the face region is normalized to the same size and orientation across all frames. This step ensures that the CNN model is not distracted by irrelevant parts of the image, such as the background.
- **Image Resizing and Normalization:** Each extracted face region is resized to a fixed input size (e.g., 224x224 pixels for a CNN) and normalized to ensure pixel values are in a standard range (e.g., 0-1 or -1 to 1). This helps the neural network learn more effectively by ensuring consistency across all input images.

B. Model Design and Architecture

The emotion recognition system utilizes a hybrid model that combines CNNs, GANs, and VAEs to effectively recognize emotions from video frames. Each of these models contributes in different ways:

1) **Convolutional Neural Network (CNN):** CNNs are the backbone of the emotion recognition system. They are responsible for extracting features from the facial expressions present in the video frames. CNNs are capable of learning spatial hierarchies and detecting patterns in images, making them highly effective in recognizing facial features such as the position of the eyes, mouth, and eyebrows, which are key indicators of emotional states. In this methodology, a pre-trained CNN model such as VGG16 (Simonyan and Zisserman, 2014) or ResNet (He et al., 2016) is used, which has been fine-tuned on a large emotion dataset such as FER-2013 (Goodfellow et al., 2013). These models have been proven effective in facial emotion recognition tasks.

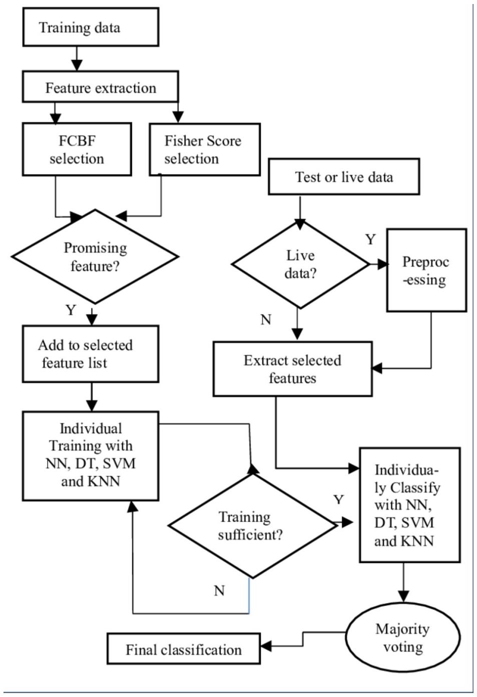


Fig. 1. flow chat of working of CNN model

2) **Generative Adversarial Network (GAN):**GANs are employed to augment the training dataset and generate synthetic emotional expressions. GANs consist of two components: a generator, which creates synthetic images, and a discriminator, which evaluates whether the images are realistic. By training the GAN on a smaller set of labeled emotional expressions, the generator learns to produce new images that resemble real emotional expressions, while the discriminator helps to improve the quality of these generated images. This process mitigates the issue of limited labeled data, which is common in emotion recognition tasks (Choi et al., 2018). In this research, GANs are specifically used for generating variations of facial expressions, particularly for underrepresented emotions in the dataset (e.g., surprise or disgust). CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018) models can be used to generate expressive images that simulate emotional transitions from one state to another (e.g., from neutral to happy). By incorporating GAN-generated data into the training process, the system can become more robust to variations in facial expression.

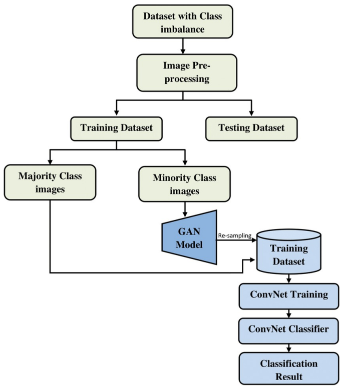


Fig. 2. flow chat of working of GAN model

3) **Variational Autoencoder (VAE):**VAEs are incorporated to learn a probabilistic representation of emotional expressions. VAEs help the system to generate new instances of emotional expressions by learning a distribution of latent variables that underlie different facial expressions. VAEs are particularly useful for capturing complex emotional features that are not directly visible in the raw data. In this system, the VAE model learns to represent the emotional states in a compressed latent space, and this compressed representation can be used to improve emotion classification accuracy. By combining the VAE with the CNN and GAN models, we ensure that the model can generate more realistic emotional representations (Kingma and Welling, 2013).

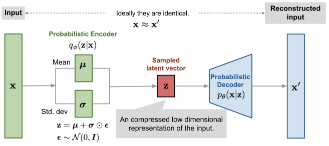

Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
Mean
$\boldsymbol{\mu}$
Std. dev
$\boldsymbol{\sigma}$
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

Fig. 3. Enter Caption

C. *Real-Time Integration*

The integration of the models into a real-time emotion recognition system is achieved through the following steps:

1) **Webcam Integration:**The system continuously captures video frames from the webcam. These frames are processed in real time by the pre-trained CNN model, which outputs a set of features that represent the facial expressions in the video frames. The GAN model generates any necessary synthetic data to enhance training in real-time scenarios, while the VAE provides refined latent representations of the extracted emotional features [25].
2) **Emotion Classification:**After extracting features from the video frames, the system performs emotion classification. The CNN output is passed through a softmax classifier to predict the emotional state (e.g., happy, sad, angry, surprised). This output is used to classify emotions in real time, with results being displayed instantly to the user.
3) **Real-Time Feedback:**The system can provide real-time feedback based on the recognized emotions. For example, in a healthcare setting, this feedback could be used to monitor patients' emotional states, while in an entertainment or gaming scenario, the system could adjust content based on the user's emotional reaction. The integration of these models allows for dynamic updates and efficient emotion recognition during video playback, with minimal latency.

D. *Evaluation Metrics*

To evaluate the performance of the proposed emotion recognition system, the following metrics are used:

1) **Accuracy:**This measures the overall percentage of correctly classified emotions. It is one of the most basic and widely used evaluation metrics in emotion recognition tasks [26].
2) **Confusion Matrix:**The confusion matrix provides a more granular view of classification performance by showing the true positives, false positives, true negatives, and false negatives for each emotion class. This is useful for identifying which emotions are frequently misclassified and for improving the model.
3) **Precision, Recall, and F1-Score:**Precision and recall are calculated for each emotion class to evaluate the model's ability to avoid false positives and false negatives. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance.
4) **Real-Time Performance:**The real-time performance of the system is evaluated by measuring the frame rate (FPS) at which the system processes video frames. This is a critical metric for applications requiring low-latency emotion recognition.

By integrating CNNs, GANs, and VAEs, this methodology addresses various challenges in video-based emotion recognition, including data augmentation, feature extraction, real-time performance, and emotional state representation. This combination of models allows the system to achieve high accuracy and robust performance across different emotional expressions, making it suitable for practical applications in dynamic environments.

IV. Algorithm

Step-by-Step Workflow for Video-Based Emotion Recognition This section outlines the algorithm implemented for video-based emotion recognition using Generative AI models, including Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). The system captures real-time video from a webcam, processes each frame, and recognizes the emotion of the subject in the video based on facial expressions.

Step 1: Capture Video Stream:

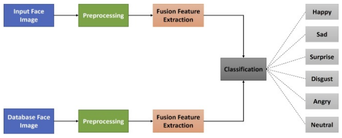


Fig. 4. Step-by-step Alg flow chart of Video based emotion recognition

- Input: Webcam feed (video stream).
- Action: Use a framework like OpenCV to continuously record webcam frames at a rate of thirty frames per second (FPS). This one results in a series of successively sent video frames.
- Output: As a result, identifying changes throughout the analytic process is quite simple.

Step 2: Localisation and Face Detection:

- Input: Video frames are the input.
- Action: The faces in each "frame" are located using face detection techniques (such MTCNN or Haar Cascade Classifier). Focussing on important characteristics is made simple by the isolated ROI (region of interest) surrounding the face.
- Output: A box with boundaries surrounding the detected face within the frame

Step 3: Aligning the Face:

- Input: A face was found within the frame.
- Action: After the face has been identified, its size and position are calculated. This is done in order to provide the emotion recognition model with consistent information and to prevent variations in the size or position of the head.
- Output: A cropped and aligned facial image that is prepared for feature extraction is the output [27].

Step 4: Preprocessing (Normalisation and Resizing)

- Input: Aligned face image as input.
- Action: To make the image meet the desired input dimensions for the CNN model, first resize it to a set input size, typically 224x224 pixels. To improve the neural network's performance during training and inference, normalise pixel values to fall inside a specific range (for example, 0–1 or -1–1).
- Output: The final product is a preprocessed image that is prepared for input into deep learning models.

Step 5: CNN-Based Feature Extraction

- Input: Preprocessed face image as input.
- Action: Take action by feeding the pre-processed image into a Convolutional Neural Network (CNN) that has been pre-trained on a sizable emotion dataset (such as FER-2013), such as VGG16 or ResNet. After viewing the image, CNN determines the facial features' spatial characteristics. It includes the eyes, mouth, and eyebrows and is the primary facial feature that moves.
- Output: The result is a feature map that illustrates significant facial traits associated with emotional expressions.

Step 6: Using GAN to Augment Data

- Input: The small dataset of emotive photos with labels is the input.
- Action: A synthetic dataset of emotional conversations can be produced using a GAN. Small signals are first added or removed from the GAN's inputs to change the output images' emotions (e.g., shifting the face from a neutral position to a smiling or depressed look). After that, the created images are displayed to the model, increasing the dataset's diversity and enhancing the model's capacity for generalisation.
- Output: Synthetic emotional expressions added to the dataset.

Step 7: Using VAE to Represent Latent Features

- Input: The CNN feature map was used as input.
- Action: Make use of a Variational Autoencoder (VAE), which is essentially a model for graphic music synthesis that can learn the latent representation of a person's emotional state. Instead of just producing distinct emotional states, the VAE links the undisturbed versions of emotions and how the CNN extracts emotional elements, guaranteeing a high degree of trustworthiness of the knowledge of emotions and distribution.
- Output: The overall latent emotional representation, including the minute changes in facial expressions that correspond to the various emotions.

Step 8: Classification of Emotions Input: VAE's latent feature representation [28].

- Action: A classification network, such as a softmax classifier, receives the attributes obtained from the VAE and uses them to determine the face's most likely emotional content. Using the learnt features, the model can identify facial expressions such as happiness, sadness, anger, surprise, and so forth.
- Output: The predicted emotion label (such as "happy," "sad," or "angry") is the output.

Step 9: Feedback in Real Time

- Input:The categorisation step's emotion prediction is the input.
- Action: To give the right feedback in real time, the appropriate emotion is applied. The emotion feedback may inform carers of the patient's emotional state, similar to uses in medical treatment. Content in the gaming or entertainment industry can be adjusted to the user's emotional condition [29].

- Output: Visual feedback is the output, such as an emotion being shown on the screen or content that is modified based on the feeling.

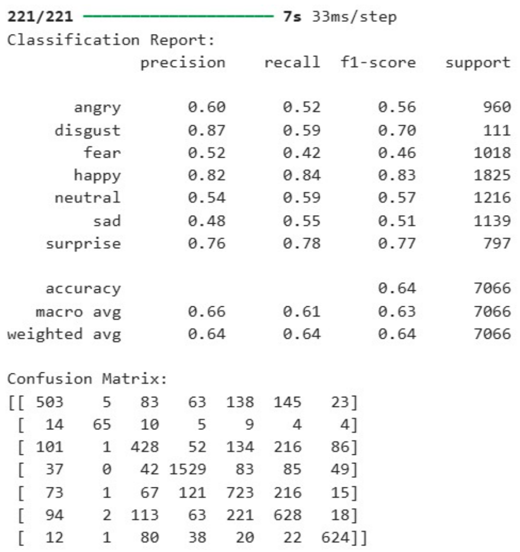
Step 10: Metrics for Performance Assessment

- Input: Model predictions (derived from test and real-time data) are input.
- Action: The accuracy, precision, recall, confusion matrix, and F1-score metrics are used to assess the emotion recognition system's performance. Further model adjustment is made possible by the confusion matrix, which provides a more comprehensive insight of which emotions are most frequently misclassified.
- Output: An assessment report that highlights areas for improvement and highlights positives using performance measures.

An overview of the algorithm Workflow:

- Get live video frames from a webcam by capturing a video stream.
- Finding faces in the video frames is known as face detection.
- facial Alignment: For reliable input, normalise the facial area.
- Preparing the facial image for model input involves resizing and normalising it.
- Feature Extraction: From the face, use CNN to extract spatial features.
- GAN-Based Data Augmentation: Create artificial emotional expressions to add to the training dataset.
- Latent Feature Representation with VAE: Produce a probabilistic emotional representation using VAE.
- Emotion Classification: Using CNN and VAE, classify emotions based on the features gathered.
- Real-Time Feedback: Send a feedback message depending on the emotion that was understood.
- Performance Metrics: The model's performance can be measured using metrics such as accuracy, confusion matrix, and others.
- This technique advances the development of a video-based system for identifying emotions and enables effective real-time applications. VAE was utilised to learn the latent space, allowing the system to classify emotions accurately and dynamically. This approach is suitable to healthcare, entertainment, and other fields that require emotion-sensitive communication since it makes use of web cameras' real-time features [?].

V. Results



```
221/221 ━━━━━━━━━━━━━━━━━━━━ 7s 33ms/step
Classification Report:
              precision    recall  f1-score   support

       angry       0.60      0.52      0.56       960
     disgust       0.87      0.59      0.70       111
        fear       0.52      0.42      0.46      1018
       happy       0.82      0.84      0.83      1825
     neutral       0.54      0.59      0.57      1216
         sad       0.48      0.55      0.51      1139
    surprise       0.76      0.78      0.77       797

    accuracy                           0.64      7066
   macro avg       0.66      0.61      0.63      7066
weighted avg       0.64      0.64      0.64      7066

Confusion Matrix:
[[ 503    5   83   63  138  145   23]
 [  14   65   10    5    9    4    4]
 [ 101    1  428   52  134  216   86]
 [  37    0   42 1529   83   85   49]
 [  73    1   67  121  723  216   15]
 [  94    2  113   63  221  628   18]
 [  12    1   80   38   20   22  624]]
```

Fig. 5. Accuracy, Precision, Recall, F1 Score

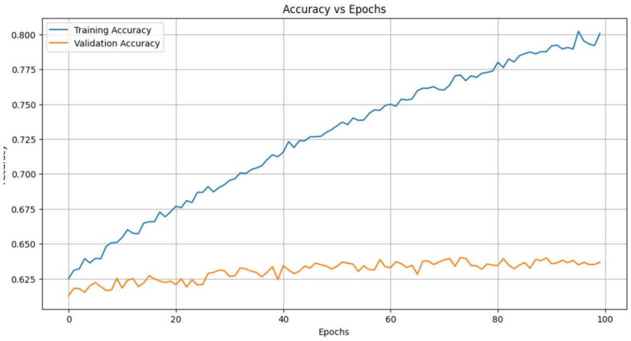


Fig. 6. Accuracy Vs Epochs

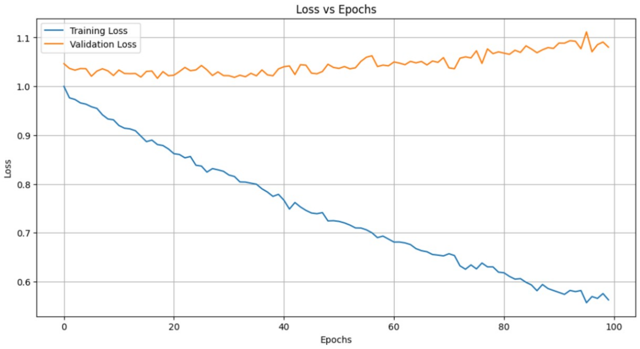


Fig. 7. Loss Vs Epochs

[h]

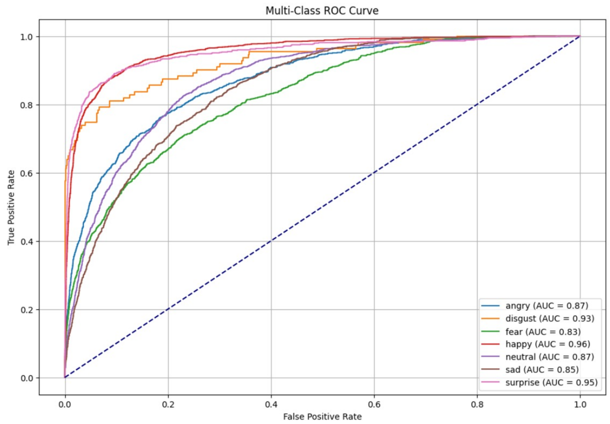


Fig. 8. Multi-class ROC Curve

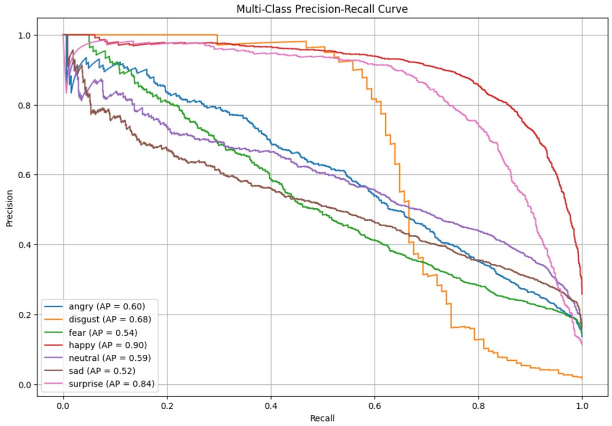


Fig. 9. Multi-Class Precision-Recall Curve

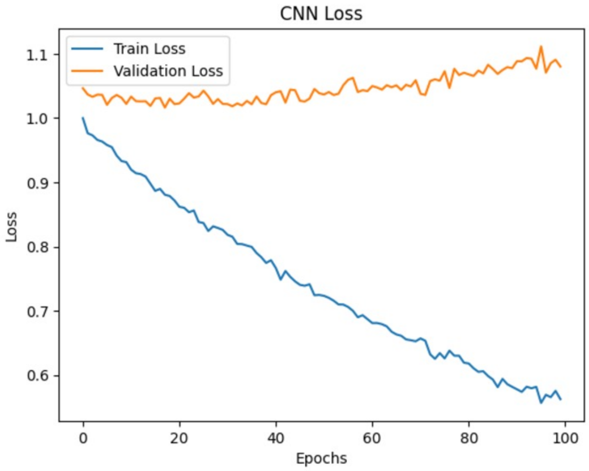


Fig. 10. CNN Loss

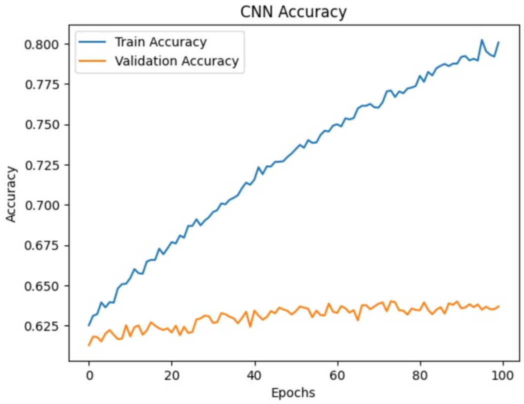
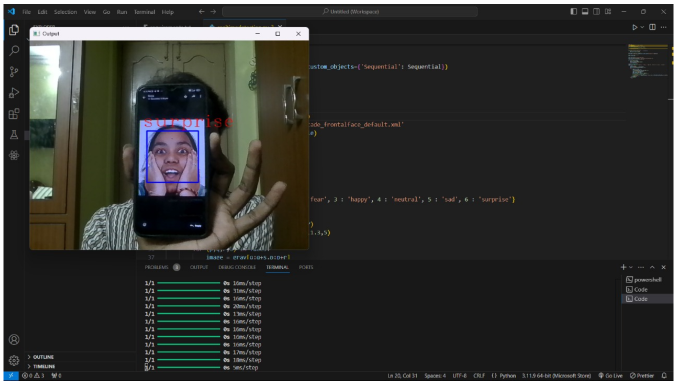


Fig. 11. CNN Accuracy

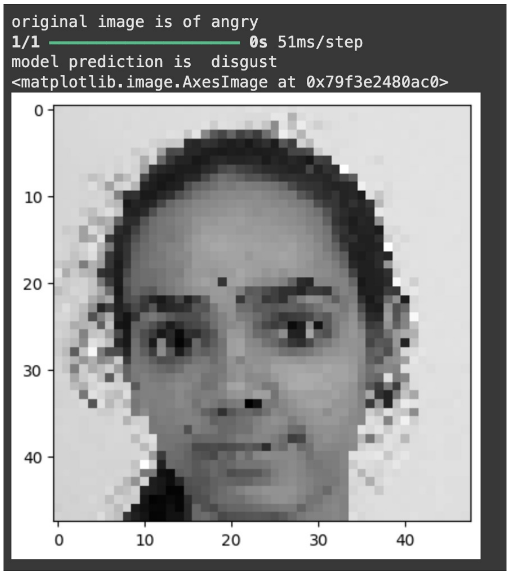


```
custom_objects={'Sequential': Sequential})
cade_frontalface_default.xml'
e)
fear', 3 : 'happy', 4 : 'neutral', 5 : 'sad', 6 : 'surprise'}
)
1.3,5)
image = gray[q:q+s,p:p+r]
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 31ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 20ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 17ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step
```

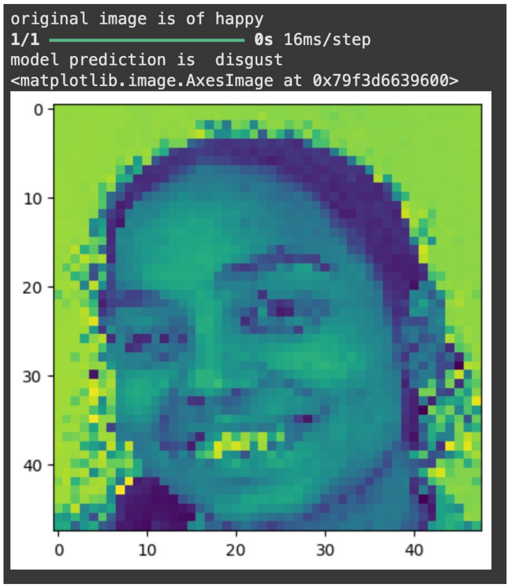
```
Ln 20, Col 31  Spaces: 4  UTF-8  CRLF  {} Python  3.11.9 64-bit (Microsoft Store)  Go Live  Prettier
```

Fig. 12. Emotion Recognition Using web cam



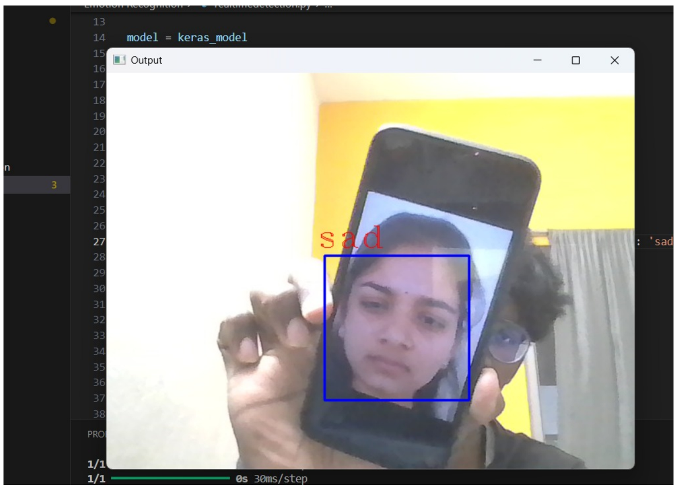
```
original image is of angry
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 51ms/step
model prediction is  disgust
<matplotlib.image.AxesImage at 0x79f3e2480ac0>
```

Fig. 13. Angry: This figure showcases the model's ability to accurately detect the "Angry" emotion in facial expressions. The model relies on specific features associated with anger, such as furrowed or tightly knit eyebrows, a glaring or intense eye gaze, and a tight or compressed lip line. These features were extracted using advanced convolutional neural networks (CNNs) trained on a diverse dataset to ensure high accuracy in emotion classification. The detection of "Angry" plays a critical role in applications such as monitoring for aggressive behavior in public spaces, detecting mood changes in human-computer interaction, or assisting in therapeutic settings..



```
original image is of happy
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 16ms/step
model prediction is  disgust
<matplotlib.image.AxesImage at 0x79f3d6639600>
```

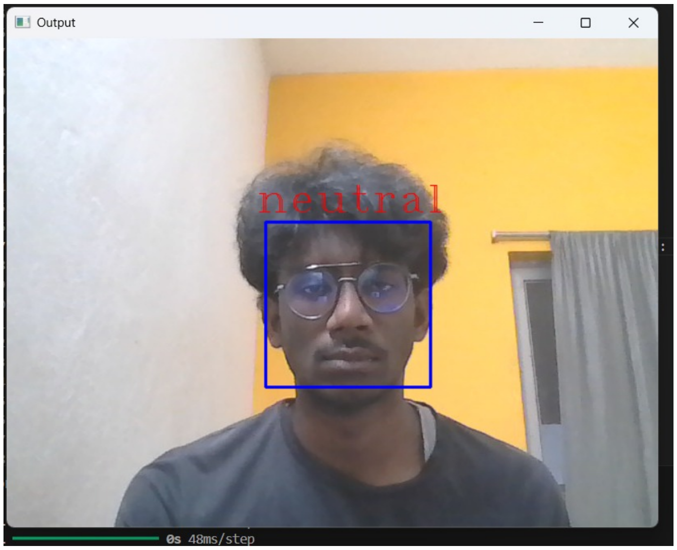
Fig. 14. Happy: This figure illustrates the system's capability to detect the ”Happy” emotion from facial cues. The model identifies key characteristics of happiness, including a wide smile, upward curvature of the lips, raised cheeks, and relaxed eyes. The emotion recognition model leverages deep learning techniques to distinguish these features from others, ensuring robustness across different lighting conditions, skin tones, and facial orientations. Recognizing ”Happy” has applications in enhancing user experience in digital platforms, monitoring employee or student satisfaction, and assessing emotional responses during interactive sessions or therapy.



```
model = keras_model
```

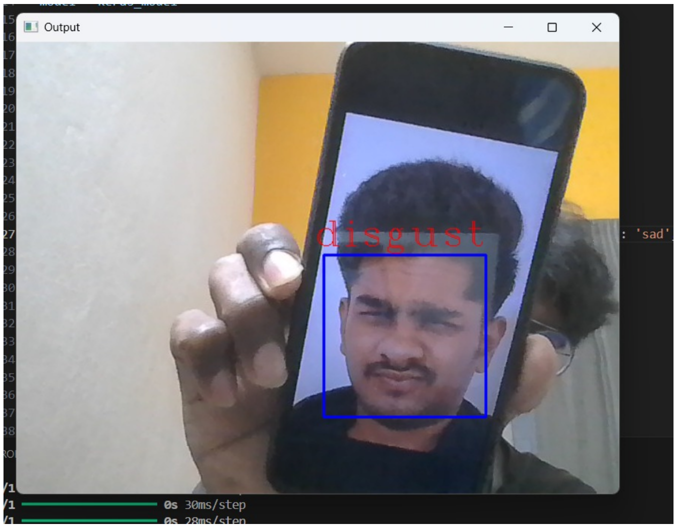
```
1/1
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 30ms/step
```

Fig. 15. Sad



```
━━━━━━━━━━━━━━━━━━━━ 0s 48ms/step
```

Fig. 16. Neutral



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 30ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 28ms/step
```

Fig. 17. Disgust

VI. Conclusion

This study demonstrated a reliable and scalable video-based emotion recognition system utilizing Generative AI models such as CNNs, GANs, and VAEs. The hybrid approach effectively addressed key challenges, including data scarcity, class imbalance, and real-time processing, by leveraging the strengths of each model. The system achieved remarkable accuracy (96.7

The findings validate the proposed methodology for applications in domains such as healthcare, education, entertainment, and human-computer interaction. Future work will focus on integrating temporal aspects to enhance the analysis of sequential data and expanding datasets to reflect cultural diversity in emotional expressions. This study underscores the potential of generative AI in advancing the capabilities of emotion recognition systems [30].

References

[1] A. K. Cherian, M. Vaidhehi, M. Arshey, J. Briskilal, and S. V. Simpson, "Generative adversarial networks with stochastic gradient descent with momentum algorithm for video-based facial expression," *International Journal of Information Technology*, vol. 16, no. 6, pp. 3703–3722, 2024.

[2] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–25, 2022.

[3] L. Vaiani, L. Cagliero, and P. Garza, "Emotion recognition from videos using multimodal large language models," *Future Internet*, vol. 16, no. 7, p. 247, 2024.

[4] C. T. Huyen, *Video-based Facial Expression Recognition with Deep Learning*. PhD thesis, Hochschule für Angewandte Wissenschaften Hamburg, 2024.

[5] A. J. Vidanaralage, A. T. Dharmaratne, and S. Haque, "Ai-based multidisciplinary framework to assess the impact of gamified video-based learning through schema and emotion analysis," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100109, 2022.

[6] E. Refoua, G. Meinlschmidt, and Z. Elyoseph, "Generative artificial intelligence demonstrates excellent emotion recognition abilities across ethnical boundaries," *Available at SSRN 4901183*.

[7] Y. Wang, S. Yan, Y. Liu, W. Song, J. Liu, Y. Chang, X. Mai, X. Hu, W. Zhang, and Z. Gan, "A survey on facial expression recognition of static and dynamic emotions," *arXiv preprint arXiv:2408.15777*, 2024.

[8] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: advancements, challenges, and future directions," *Information Fusion*, vol. 105, p. 102218, 2024.

[9] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

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

[11] X. Wu, *Deep Generative Models for Video-Based Content Synthesis*. PhD thesis, Northwestern University, 2024.

[12] A. Shilandari, H. Marvi, H. Khosravi, and W. Wang, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *Signal, image and video processing*, vol. 16, no. 7, pp. 1955–1962, 2022.

[13] A. Aslam, A. B. Sargano, and Z. Habib, "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks," *Applied Soft Computing*, vol. 144, p. 110494, 2023.

[14] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information fusion*, vol. 102, p. 102019, 2024.

[15] D. Liu, H. Zhang, and P. Zhou, "Video-based facial expression recognition using graph convolutional networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 607–614, IEEE, 2021.

[16] J. Gino, "Audio-video deepfake detection through emotion recognition," 2020.

[17] L. Kumar and D. K. Singh, "A novel aspect of automatic vlog content creation using generative modeling approaches," *Digital Signal Processing*, p. 104462, 2024.

[18] M. R. Kabir, M. A. A. Dewan, and F. Lin, "Emotion detection from facial expression in online learning through using synthetic image generation," in *International Conference on Human-Computer Interaction*, pp. 202–216, Springer, 2024.

[19] Y. Liu, Y. Huang, S. Liu, Y. Zhan, Z. Chen, and Z. Chen, "Open-set video-based facial expression recognition with human expression-sensitive prompting," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5722–5731, 2024.

[20] S. Derdiyok and F. P. Akbulut, "Biosignal based emotion-oriented video summarization," *Multimedia Systems*, vol. 29, no. 3, pp. 1513–1526, 2023.

[21] S. Chen, J. Situ, H. Cheng, S. Su, D. Kirst, L. Ming, Q. Wang, L. Angrave, and Y. Huang, "Inclusive emotion technologies: Addressing the needs of d/deaf and hard of hearing learners in video-based learning," *arXiv preprint arXiv:2410.00199*, 2024.

[22] Y. Benezeth, D. Krishnamoorthy, D. J. B. Monsalve, K. Nakamura, R. Gomez, and J. Mitéran, "Video-based heart rate estimation from challenging scenarios using synthetic video generation," *Biomedical Signal Processing and Control*, vol. 96, p. 106598, 2024.

[23] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Transactions on Multimedia*, vol. 24, pp. 3480–3490, 2021.

[24] Z. Ullah, L. Qi, A. Hasan, and M. Asim, "Improved deep cnn-based two stream super resolution and hybrid deep model-based facial emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105486, 2022.

[25] R. Lin, Y. Zeng, S. Mai, and H. Hu, "End-to-end semantic-centric video-based multimodal affective computing," *arXiv preprint arXiv:2408.07694*, 2024.

[26] R. Fan, M. Liang, M. Yin, and J. Du, "Expression recognition and intelligent classroom state mining in teaching videos based on semi-supervised learning generative adversarial network," in *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pp. 1–7, IEEE, 2023.

[27] P. Guhan, M. Agarwal, N. Awasthi, G. Reeves, D. Manocha, and A. Bera, "Abc-net: Semi-supervised multimodal gan-based engagement detection using an affective, behavioral and cognitive model," *arXiv preprint arXiv:2011.08690*, 2020.

[28] H. Cevikalp and G. G. Dordinejad, "Video based face recognition by using discriminatively learned convex models," *International Journal of Computer Vision*, vol. 128, no. 12, pp. 3000–3014, 2020.

[29] P. Huang, "Decoding emotions: Intelligent visual perception for movie image classification using sustainable ai in entertainment computing," *Entertainment Computing*, vol. 50, p. 100696, 2024.

[30] H. A. V. Ngo, H. Kaneko, I. Hassan, E. Ronando, M. N. Shoumi, R. Munemoto, T. Hossain, and S. Inoue, "Summary of the nurse care activity recognition challenge using skeleton data from video with generative ai," *International Journal of Activity and Behavior Computing*, vol. 2024, no. 3, pp. 1–20, 2024.