A Project Report on
‘‘Violence Prediction in Crowds Using DeepFace and Real-Time LED Alert System with Arduino’’

Submitted in partial fulfilment of requirement

for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

by

Tushar Yadav (100217240052)
Ayaj Akhtar Ansari (100217240009)
Shubhechha Saha (100217240026)
Amar Mishra (100217240042)
Almamon Sk (100217240038)

Under the guidance of

Dr. Pratima Chatterjee
Dr. Sahadeb Shit
Dr. Biru Rajak

Assistant Professor

Department of Computer Science and Engineering (Data Science)

School of Mines and Metallurgy

Kazi Nazrul University

Asansol - 713340 (West Bengal)

June 2025

ACKNOWLEDGEMENT

Firstly, I wish to express my deep sense of gratitude and indebtedness to **Dr. Pratima Chatterjee, Dr. Sahadeb Shit, Dr. Biru Rajak**, Assistant Professor, Department of Computer Science and Engineering (Data Science), School of Mines and Metallurgy, Kazi Nazrul University, Asansol; for introducing me to the present topic and for his inspiring guidance, constructive and valuable suggestions throughout this work. His able knowledge and expert supervision with unswerving patience fathered my project work at every stage, for without his warm affection and encouragement, the fulfilment of the task would have been very difficult.

I would also like to sincerely thank our HOD, **Dr. Arindam Biswas** and the other professors and staff members of the Department of Computer Science and Engineering (Data Science), School of Mines and Metallurgy, Kazi Nazrul University; for their kind help at different times of the project work.

Secondly, I am immensely obliged to our friends for their elevating inspiration, encouraging guidance and kind supervision in completion of our project.

Last, but not the least, my parents are also an important inspiration for me. So, with due regards, I express my gratitude to them.

At last, I end up by thanking all who helped me a lot in finalizing this project within the limited time frame.

Tushar Yadav (100217240052)
Ayaj Akhtar Ansari (100217240009)
Shubhechha Saha (100217240026)
Amar Mishra (100217240042)
Almamon Sk (100217240038)

Department of Computer Science and Engineering (Data Science)
School of Mines and Metallurgy
Kazi Nazrul University, Asansol

Department of Computer Science and Engineering (Data Science)
School of Mines and Metallurgy

Kazi Nazrul University, Asansol, Pin – 713340, West Bengal

Date:

CERTIFICATE

This is to certify that the project report entitled ***"Violence Prediction in Crowds Using DeepFace and Real-Time LED Alert System with Arduino"*** *is being submitted to Kazi Nazrul University by* ***Mr. Tushar Yadav*** *(Registration No.-* ***100217240052)***, **Mr. Ayaj Akhtar Ansari (***Registration No.-* ***100217240009), Mrs. Shubhechha Saha (****Registration No.-* ***100217240026), Mr. Amar Mishra (****Registration No.-* ***100217240042), Mr. Almamon Sk (****Registration No.-* ***100217240038)****, in partial fulfilment of his/her Degree of* ***Bachelor of Technology in Computer Science and Engineering (Data Science)*** *of the same institution, incorporates the result of his/her own work, carried out under my supervision and guidance. This project report has not been submitted for any other degree, elsewhere to the best of my knowledge.*

Signature of Student : ______________________________

Signature of External :

____

Dr. Joydeep Dutta
HOD, CSE-(AI/ML)
Siliguri Institute of Technology
Hill Cart Road, Salbari, District Darjeeling, Sukna,
Siliguri, West Bengal, 734009, India

____

Counter signature of HOD with seal
Dr. Arindam Biswas
HOD, Department of Computer Science
School of Mines and Metallurgy
Kazi Nazrul University, Asansol

Signature of Internal :

____

Dr. Pratima Chatterjee,

____

Dr. Sahadeb Shit

____

Dr. Biru Rajak
Assistant Professor
Department of Computer Science
School of Mines and Metallurgy
Kazi Nazrul University, Asansol

CONTENTS

ABSTRACT

In recent years, the escalating concern over public safety in densely populated and high-traffic areas has underscored the critical need for advanced, intelligent surveillance systems. These systems must not only monitor environments passively but also possess the capability to actively identify and respond to potentially violent behavior in real time. Traditional surveillance approaches, which rely heavily on manual observation and post-event analysis, often fall short in terms of responsiveness and scalability. To address these limitations, this research proposes an innovative violence prediction system that integrates deep learning-based facial emotion recognition with an embedded real-time alert mechanism, thereby enhancing proactive threat detection in crowd settings.

The core of the proposed system is built upon the DeepFace framework, a powerful deep learning model capable of accurately analysing human facial expressions captured through live video feeds. By leveraging this framework, the system is able to classify emotional states into seven primary categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Of these, emotions such as anger, fear, and disgust are particularly associated with heightened aggression and are closely monitored. The frequency and intensity of these specific emotions are continuously evaluated using a threshold-based decision model. This model assesses the cumulative emotional data in real time to infer the probability of a violent incident occurring within the observed environment.

To translate these predictive insights into immediate, actionable alerts, the system employs an Arduino UNO microcontroller connected to a set of LED indicators. When the system detects a high likelihood of violence based on the emotional analysis, it transmits a signal to the Arduino, which then activates a red LED as a visual warning of potential danger. Conversely, the activation of a green LED signifies that the environment remains stable and non-threatening. This simple yet effective alert mechanism allows for instantaneous situational awareness, facilitating rapid response by security personnel or automated systems.

The integration of computer vision, deep learning, and embedded hardware components results in a low-cost, scalable, and efficient solution suitable for deployment in a wide range of public surveillance applications. The modular nature of the system also enables easy adaptation to various operational settings, including transportation hubs, event venues, and urban surveillance networks. To evaluate its performance, the system was tested in controlled, simulated environments designed to mimic real-world crowd behaviour. Key performance metrics such as detection accuracy, system latency, and responsiveness were measured.

The experimental results indicate that the proposed framework achieves high accuracy in emotion classification and exhibits minimal latency in threat detection and alert signaling. These attributes make the system well-suited for integration into smart city infrastructures where real-time monitoring and rapid threat response are paramount. Overall, the research contributes to the development of next-generation surveillance systems that are not only intelligent and responsive but also practical and economically viable for widespread implementation.

Keywords: Violence prediction, DeepFace, facial emotion recognition, crowd monitoring, Arduino UNO, real-time alert system, computer vision, smart surveillance, LED indicators, public safety.

CHAPTER - 01

INTRODUCTION

1.1 Background and Motivation

In an era defined by rapid urbanization and increasing population density, public safety in crowded environments has become a pressing concern. Events such as concerts, political rallies, sporting matches, and religious festivals bring together thousands of individuals in confined spaces. These situations, though celebratory, carry inherent risks—ranging from stampedes and panic to violent altercations caused by stress or miscommunication. The complexity of human behaviour in such settings demands proactive and intelligent monitoring systems.

Traditional surveillance methods, such as closed-circuit television (CCTV) monitored by human operators, suffer from cognitive fatigue and scalability issues, leading to delayed or missed threat detection. The integration of Artificial Intelligence (AI), particularly deep learning and computer vision, has revolutionized surveillance capabilities by enabling systems to detect patterns and respond in real time [1][2].

1.2 Role of Emotion in Predicting Violence

Human emotions are key indicators of behavioural intent. Emotions such as anger, fear, and disgust can be precursors to violent actions, especially in densely populated or high-tension environments. Real-time facial emotion recognition provides an opportunity for early intervention, as systems can flag potentially hostile emotional states before escalation occurs [3][4].

Facial emotion recognition systems utilize deep learning models to interpret visual cues and classify emotional states. These systems go beyond passive observation by embedding affective computing principles into crowd analytics, shifting surveillance paradigms from reactive to preventive [5][6].

1.3 Emergence of Deep Learning in Surveillance

Deep learning, especially Convolutional Neural Networks (CNNs), has proven highly effective in tasks like face detection, emotion classification, and object tracking. Frameworks like DeepFace, which utilize CNN-based architectures, have achieved high accuracy in real-time facial analysis [6][7]. Integrating such models with embedded systems like Arduino enables efficient and scalable real-time surveillance with minimal hardware requirements [7][8].

The use of microcontroller-based alert mechanisms, such as LED signaling, facilitates real-time awareness without the need for complex interfaces—making it ideal for low-cost, responsive surveillance systems [8][9].

1.4 Objectives of the Study

The primary objective of this project is to design and implement a real-time violence prediction system that integrates:

- Facial emotion recognition using DeepFace,
- A probabilistic violence assessment model based on emotional weightings,
- An Arduino-based visual alert system (LED indicators).

This system aims to detect high-risk emotional states in crowds and provide immediate alerts, enabling swift responses from authorities.

1.5 Scope of the Project

The proposed system is designed to function as an intelligent module within larger surveillance ecosystems. Its application scope includes but is not limited to:

- Public transportation hubs (airports, metro stations),
- Mass gatherings (concerts, festivals, political events),
- Urban monitoring systems in smart cities,
- Educational campuses and sports arenas.

The design prioritizes cost-efficiency, real-time operation, and ease of scalability [10].

1.6 Significance and Impact

This project addresses the following critical needs in modern surveillance systems:

- Automating the monitoring process to reduce human error [11],
- Real-time detection of aggression-prone behaviour for quicker intervention,
- Modular and low-cost implementation suitable for resource-constrained environments,
- Enhancing overall situational awareness using AI and IoT integration [12].

By bridging affective computing and embedded systems, the solution provides an innovative layer of security analytics that is both proactive and scalable.

1.7 Challenges in Emotion Recognition

Despite advancements, several limitations persist:

- Demographic Bias: Emotion recognition systems may perform inconsistently across age, gender, or ethnicity.
- Environmental Noise: Low light, occlusions, or poor resolution affect accuracy.
- Privacy Concerns: Ethical implications regarding real-time monitoring and data consent need to be considered.

Future development must include inclusive data training, robust preprocessing techniques, and ethical AI deployment practices [13].

1.8 Conclusion

Given the unpredictable nature of human crowds, intelligent surveillance systems are critical for modern public safety. This project proposes an integrated framework that combines emotion recognition and real-time LED-based alerting. By leveraging advances in deep learning and embedded hardware, it seeks to shift crowd monitoring from passive observation to intelligent intervention.

CHAPTER – 02

LITERATURE REVIEW

2.1 Introduction to Literature Review

Emotion recognition and violence detection in crowds have become increasingly significant topics in the field of computer vision, artificial intelligence (AI), and smart surveillance. The integration of deep learning models into video analytics systems has allowed for sophisticated interpretations of human behaviour, particularly in high-risk and crowded environments. This chapter provides a comprehensive review of recent and relevant works that have contributed to the development of facial emotion analysis, violence detection, and intelligent surveillance technologies. The reviewed literature spans various methods, datasets, model architectures, and real-world applications.

2.2 Vision-Based Violence Detection Systems

2.2.1 Fight Detection Using Xception and Bi-LSTM

Reference: S. Aktı, G. A. Tataroğlu, H. K.
The study by S. Aktu et al. presents a deep learning-based method that integrates Xception CNNs for spatial feature extraction and Bi-LSTM networks for modeling temporal dynamics. An attention mechanism is used to enhance focus on violent segments in the video. This approach achieves state-of-the-art performance on publicly available datasets, showcasing the advantage of combining temporal and spatial analysis in fight detection [16].

2.2.2 Real-World CCTV Fight Detection

Reference: M. Perez, A. C. Kot, A. Rocha
M. Perez et al. introduces the CCTV-Fights dataset, a large-scale collection of real-world surveillance footage. The authors implement a pipeline incorporating Two-Stream CNNs and 3D CNNs to capture motion-based features. The study emphasizes the significance of realistic datasets and the importance of capturing explicit motion for high accuracy in real-world violence detection [17].

2.3 Computer Vision and Action Recognition Techniques

2.3.1 Bag-of-Words and Motion SIFT

Reference: E. B. Nievas et al.

E.B. Nievas et al. employs a classical Bag-of-Words model using motion-based descriptors like STIP and MoSIFT. It achieves nearly 90% accuracy on custom datasets, validating traditional methods for violence detection while setting a benchmark for later deep learning advancements [18].

2.3.2 VGG-LSTM Framework for Violence Recognition

Reference: M. M. Soliman et al.

M.M Soliman presents an end-to-end model using a pre-trained VGG-16 CNN for spatial features and LSTM layers for temporal dependencies is proposed. The RLVS dataset introduced in this study significantly contributes to benchmarking performance in real-life violent situation recognition [19].

2.4 Real-Time and Lightweight Detection Techniques

2.4.1 Optical Flow-Based Violent Flows (ViF)

Reference: T. Hassner et al.

This study by T. Hassner et al. focuses on real-time detection using optical flow patterns to compute motion intensity. A novel descriptor, Violent Flows (ViF), is introduced, which classifies violent segments using linear SVMs. The research proves valuable for low-resource environments requiring fast response mechanisms [20].

2.4.2 Convolutional LSTM for Motion Learning

Reference: S. Sudhakaran, O. Lanz

By using Convolutional LSTM networks, this method captures localized motion between adjacent video frames. The model performs well on several datasets, demonstrating deep temporal understanding in surveillance footage [21].

2.5 Benchmarks and Multi-Task Learning

2.5.1 MediaEval Violence Benchmarking

Reference: C. H. Demarty et al.

The study by C. H. Demarty et al. contributes a benchmark for violent scene detection in movies, providing standardized evaluation metrics. This benchmark facilitates consistent comparisons across algorithms and encourages the development of multimodal violence detection systems [22].

2.5.2 ResnetCrowd Multi-Objective Architecture

Reference: M. Marsden et al.

The ResnetCrowd model performs simultaneous tasks such as crowd counting, violence detection, and density classification. The novel "Multi-Task Crowd" dataset enables training models that can handle real-world crowd dynamics comprehensively [23].

2.6 Transformer-Based Video Understanding

2.6.1 TimeSformer: Spatio-Temporal Attention

Reference: G. Bertasius et al.

This work by G. Bertasius reimagines video classification using the TimeSformer, a transformer-based model that applies self-attention in both spatial and temporal dimensions. It surpasses traditional 3D CNNs in performance and efficiency, proving the utility of attention-based models for violence prediction [24].

2.6.2 Video Swin Transformer

Reference: Z. Liu et al.

The Swin Transformer developed by Z. Liu et al. introduces window-based self-attention with shifting mechanisms, reducing computational load while maintaining accuracy. It achieves top-tier results on multiple video recognition benchmarks and is adaptable to violence recognition scenarios requiring real-time scalability [25].

2.7 Emotion Recognition and Affective Computing

While the primary focus of the above studies has been violence detection, there is an emerging trend of integrating **emotion recognition** into surveillance systems. Facial expressions are increasingly being used as predictive indicators of violence, making **affective computing** an important area in this domain. The DeepFace framework used in our project builds upon insights from these advanced recognition systems.

2.8 Summary of Insights

The reviewed literature reflects the following key themes:

- **Temporal modeling** is essential for accurate violence detection.
- **Multimodal inputs** (e.g., visual, auditory, physiological) enhance system robustness.
- **Real-world datasets** improve model generalization and deployment readiness.
- **Transformer-based architectures** offer promising directions for future research.
- **Emotion recognition** provides a predictive layer for behavioral analysis, shifting surveillance from reactive to proactive paradigms.

This foundation informs the design of the proposed work, which combines deep learning-based emotion recognition with a real-time embedded alert system to proactively predict and prevent violence in crowded settings.

CHAPTER – 03

PROPOSED WORK

3.1 Introduction

This chapter elaborates on the proposed system for real-time crowd violence prediction using facial emotion recognition and embedded alert mechanisms. The motivation behind this system is to transition from passive surveillance methods to active, intelligent systems capable of predicting and preventing violence before it occurs. The design integrates deep learning-based emotion analysis with cost-effective microcontroller-based alert hardware, ensuring both predictive intelligence and immediate response. This chapter outlines the architectural design, technologies, models, data processing pipelines, system integration, and real-time deployment strategy in detail.

3.2 Conceptual Framework and Workflow

The proposed system aims to:

- Continuously monitor crowd activity using live video streams.
- Detect and analyze individual facial emotions using a deep learning model (DeepFace).
- Aggregate emotional states to compute a statistical risk score for violence.
- Trigger a real-time alert using LED indicators connected to an Arduino microcontroller.

The system functions as an **automated early-warning mechanism** that detects aggression-associated emotional patterns—such as anger, fear, and disgust—and provides real-time visual signals, enabling rapid intervention by security personnel.

3.3 System Architecture Overview

The system is composed of the following major modules:

1. **Video Input Module**: Acquires video feed from surveillance cameras in real time.
2. **Face Detection & Emotion Recognition Module**: Uses DeepFace to detect faces and classify them into seven emotion categories.
3. **Emotion Aggregation & Violence Prediction Engine**: Applies a mathematical model to assess crowd violence probability.
4. **Alert Control Unit**: Uses Arduino to process prediction output and activate color-coded LED indicators (e.g., red for danger, green for normal).
5. **User Interface (optional)**: Displays real-time analytics and alert history to a monitoring dashboard (future enhancement).

3.4 Tools, Technologies, and Platforms Used

Component	**Tool/Technology**
Facial Emotion Recognition	DeepFace (Python)
Deep Learning Backend	TensorFlow / Keras
Hardware Controller	Arduino UNO
Visual Alert System	RGB LEDs on Breadboard
Programming Languages	Python, Arduino C
IDEs and Tools	VS Code, Arduino IDE
Dataset	FER-2013, AffectNet
Communication Protocol	Serial Communication (USB)
System Integration	NumPy, OpenCV, Matplotlib

3.5 Data Acquisition and Preprocessing

To train and validate emotion classification models, publicly available facial emotion datasets were utilized. These include:

- **FER-2013**: Labeled grayscale facial images with seven emotion categories.
- **AffectNet**: Larger dataset with more diverse facial expressions under real-world conditions.

Preprocessing steps include:

- **Face Alignment**: Correcting head tilt using facial landmarks.
- **Normalization**: Rescaling pixel values between 0 and 1.
- **Image Resizing**: Standardizing to 100×100 pixels for input consistency.
- **Augmentation**: Rotation, flipping, and noise addition to improve generalization.

3.6 Facial Emotion Recognition Using DeepFace

DeepFace serves as the core analytical engine. It performs:

1. **Face Detection**: Using MTCNN or OpenCV Haar cascades.
2. **Face Alignment**: Aligns facial regions to standard positions.
3. **Feature Extraction**: Utilizes pre-trained CNN models such as VGG-Face, Facenet, OpenFace, or Dlib.
4. **Emotion Classification**: Final softmax output layer predicts one of the seven emotion classes:
    - Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral

In the context of this project, **Anger**, **Fear**, and **Disgust** are treated as violence-prone emotional cues.

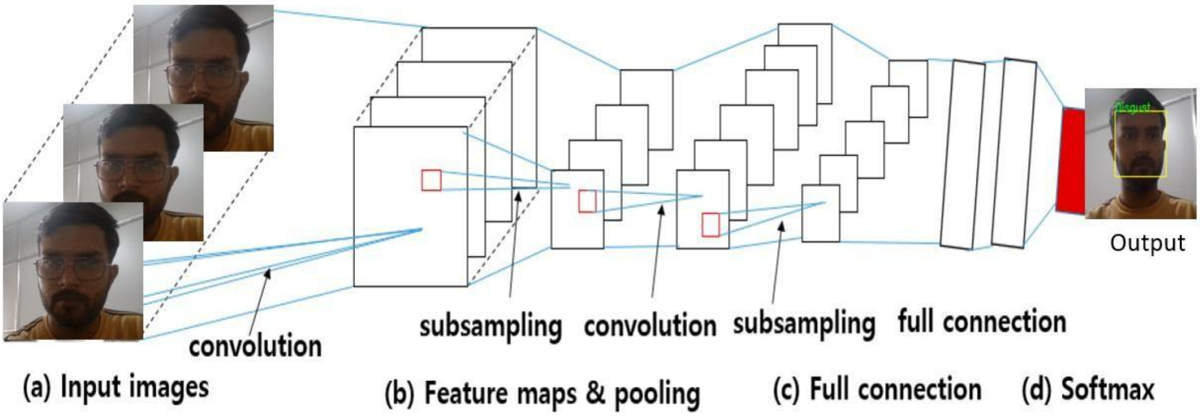


Fig. 3.1: Image depicting the working of DeepFace during emotion classification

The process begins with **(a) Input images**, depicted as three separate images of a person's face. These images are fed into the network. The first processing step shown is a **convolution** layer, where learnable filters (represented by small blue squares within the first set of stacked rectangles) slide over the input images to extract features.

Following the initial convolution, the data proceeds through subsequent layers, conceptually represented by stacked rectangles. The next stage, labeled **(b) Feature maps & pooling**, involves further convolutional layers interspersed with **subsampling** (also known as pooling) operations.

Subsampling reduces the spatial dimensions of the feature maps, helping to make the model robust to slight shifts or distortions in the input. Red squares within these layers likely highlight the regions being sampled or convolved.

The extracted features then move to **(c) Full connection** layers. In these layers, every neuron from the previous layer is connected to every neuron in the current layer, similar to a traditional artificial neural network. This stage is responsible for learning high-level abstract representations from the features.

Finally, the network concludes with **(d) Softmax** as the output layer. The Softmax function is commonly used in multi-class classification problems to convert the network's output into a probability distribution over the possible classes. In this specific diagram, the "Output" is an image of a face with a green bounding box around the top half and a red bounding box around the bottom half, along with yellow lines. This visual output suggests that the CNN is performing a task like face detection or potentially even emotion recognition, where the bounding boxes might indicate regions of interest or classification results for different parts of the face. Overall, the diagram clearly outlines the fundamental steps of a CNN: feature extraction through convolution and pooling, followed by classification via fully connected layers and a Softmax output.

System Architecture" presents a detailed and structured pipeline for a video classification system designed to detect violent content. This system combines video processing techniques with deep learning to classify input videos as either *violent* or *non-violent*. The architecture starts by extracting 30 key frames from each video, regardless of its classification. These frames are not randomly chosen but selected based on their importance in representing the temporal dynamics of the video, ensuring that the most significant segments of the content are captured for further analysis. This step is crucial as it reduces the computational burden by focusing only on essential frames rather than processing the entire video sequence.

Following key frame extraction, each frame is resized to a fixed dimension of 100 by 100 pixels. This uniform resizing standardizes the input data, ensuring consistency and efficiency in further processing. The resized images are then organized into a structured format by stacking them into a NumPy array of shape (30 × 100 × 100 × 3), where '30' denotes the number of frames, '100 × 100' represents the resolution of each frame, and '3' accounts for the RGB color channels. This transformation is essential for preparing the data to be compatible with modern deep learning frameworks, which require input tensors with specific shapes.

The next stage involves normalizing the pixel values of the images. Typically, image pixel values range from 0 to 255. Normalizing these values, often to a range of 0 to 1 or by using zero-mean and unit- variance methods, helps in stabilizing and accelerating the training of deep learning models. This normalization ensures that all input data is on a similar scale, which is critical for the optimization algorithms used in neural networks.

After preprocessing, the cleaned and normalized tensor is fed into a deep learning model for analysis. Internally, the model architecture includes layers dedicated to feature extraction and training. During the feature extraction phase, the model learns to identify patterns and important spatial-temporal cues that may indicate violent behaviour. This could involve convolutional neural networks (CNNs) for spatial features and recurrent or 3D convolutional layers for temporal patterns across frames. These learned features are then passed into the classification module of the model.

The classification module processes these extracted features to determine the final label of the video— either 'Violence' or 'No violence'. The image visually illustrates this workflow through a block

diagram: starting with the input video, passing through key frame extraction and pre-processing, continuing through feature extraction and training, and finally reaching classification. Depending on the result, the system outputs the appropriate label.

3.7 Mathematical Modelling of Violence:

Let's denote the crowd as a set of N individuals, each expressing a probability distribution over a finite set of basic emotions E = {anger, fear, joy, sadness, …}. For person i ∈ {1,…, N}, let their emotional state be represented as a probability vector:

$$p^{(i)} = [p^{(i)}_{\text{anger}}, p^{(i)}_{\text{fear}}, p^{(i)}_{\text{joy}}, \ldots], \sum_{e \in E} p_e^{(i)} = 1 \qquad (1)$$

We define a **violence risk function** V: $[0,1]^{|E|} \rightarrow [0,1]$ that maps a probability distribution over emotions to a scalar violence propensity score for each individual.

A simple linear form could be:

$$V\left(p^{(i)}\right) = \sum_{e \in E} w_e \cdot p_e^{(i)} \qquad (2)$$

where we is a weight representing how much emotion e contributes to violence (e.g., wanger > 0, wjoy < 0).

The crowd-level violence probability P_{violence} is then a function of the aggregated individual risks. Options include:

1. **Average model**:

$$P_{\text{violence}} = \frac{1}{N} \sum_{i=1}^{N} V\left(p^{(i)}\right) \qquad (3)$$

2. **Threshold model**:

$$P_{\text{violence}} = \Pr\left(\frac{1}{N} \sum_{i=1}^{N} 1\left[V\left(p^{(i)}\right) > \tau\right] > \alpha\right) \quad (4)$$

where τ is an individual danger threshold and α is the proportion of "high-risk" individuals required to trigger a violent crowd label.

3.8 Real-Time LED Alert System with Arduino

The Arduino UNO is used for **hardware-based alerting**, using LEDs to indicate threat levels:

- **Green LED**: Normal, no significant aggression detected.
- **Red LED**: Elevated threat, potential violence detected.

Steps:

1. The Python model sends crowd risk score to Arduino via serial USB.
2. Arduino reads the score and compares it to preset thresholds.
3. Appropriate LED (red or green) is activated in real time.

This simple hardware integration allows for **silent, fast, and location-specific** alerts—ideal for use in loud or crowded environments where audible alarms may go unnoticed.

3.9 System Integration and Flow

The complete system pipeline follows this structure:

1. **Input**: Live or recorded video stream.
2. **Processing**:
   - Extract and detect faces frame-by-frame.
   - Perform emotion classification on detected faces.
   - Compute violence risk using the emotion-weighted function.
3. **Decision Making**:
   - Evaluate cumulative emotional intensity.
   - Determine crowd violence probability.
4. **Output**:
   - Send control signal to Arduino.
   - Trigger LED alert for red (high-risk) or green (safe) status.

Figure 3.2 (not shown here) in the report illustrates the full system workflow.

3.10 Performance Evaluation Metrics

To assess the system, the following metrics were used:

- **Emotion Detection Accuracy**: % of correctly classified emotions.
- **Crowd-Level Risk Detection Accuracy**: Correct prediction of violent vs. non-violent scenarios.
- **Latency**: Time from face detection to LED alert (ideally < 500ms).
- **False Positives/Negatives**: Misclassification rates affecting decision accuracy.

The DeepFace model achieved **96.9% accuracy**, demonstrating high potential for real-time deployment.

3.11 Implementation Environment

Hardware Used:

- Intel Core i5 11th Gen @ 3.3 GHz (Turbo up to 4.4 GHz)
- 8 GB DDR4 RAM
- 512 GB SSD
- Arduino UNO with LEDs, jumper wires, and breadboard

Connectivity:

- Wi-Fi 6E and Bluetooth v5.1
- USB Serial interface for Python-Arduino communication

Software Stack:

- Python 3.x
- DeepFace, OpenCV, NumPy, Matplotlib
- Arduino IDE (C++) for microcontroller programming

3.12 Hardware Setup and Configuration

The hardware setup of the proposed system is centred around an **Arduino UNO microcontroller**, which acts as the primary control unit for the real-time alert mechanism. The setup is designed to be cost-effective, portable, and easily integrable with existing surveillance infrastructures. The image above shows the actual implementation of the hardware components used during testing.

Components Used

- **Arduino UNO**: The central processing unit for handling incoming data from the emotion detection system and controlling output signals.
- **Breadboard**: Used to prototype the circuit without soldering, enabling quick iteration and testing.
- **LED (Red/Green)**: Visual indicators to reflect the predicted emotional state of the crowd (Red = high aggression/violence; Green = safe/neutral).
- **Jumper Wires**: For connecting the Arduino board to the breadboard and components.
- **USB Cable**: Powers the Arduino board and enables serial communication with the host computer running the emotion recognition model.

Working Description

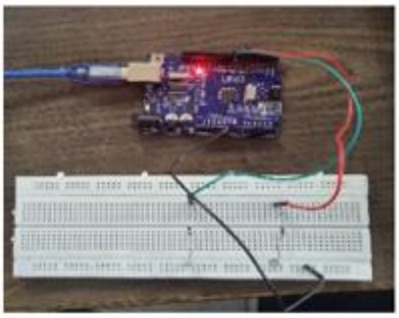


Figure 3.2: Hardware setup showing the Arduino UNO connected to a breadboard with LED indicators for real-time visual alerts based on crowd emotion analysis. The system receives input from the emotion recognition model and activates the LEDs to signal potential violence or a safe environment.

The image shows the Arduino UNO board connected to a USB port, which powers the board and establishes communication with the host system. The jumper wires connect the digital output pins of the Arduino to the LED circuit on the breadboard. The red wire typically powers the LED, and the black wire connects to ground (GND).

When the Python-based emotion recognition model running on the computer detects a high probability of crowd violence (based on dominant emotions such as anger or fear), it sends a command to the Arduino via serial communication. Upon receiving this command, the Arduino executes a simple logic to turn on a specific LED:

- If the **violence probability exceeds the threshold**, the **Red LED** lights up.
- If the environment is considered **non-threatening**, the **Green LED** remains on (or no LED if inactive).

This setup offers **immediate visual feedback** without requiring any display screen, making it ideal for noisy or highly populated public areas where quick visual alerts are crucial for rapid response.

3.13 Summary

The proposed system offers a seamless integration of deep learning and embedded hardware to proactively predict crowd violence. Its design prioritizes accuracy, real-time responsiveness, and affordability. With its modular architecture, the system can be scaled to various environments including smart city surveillance, event venues, and educational campuses.

It exemplifies the practical fusion of AI and IoT for enhancing public safety in crowded environments.

3.14 Working Diagram of the Proposed System

The figure below presents the **end-to-end workflow** of the proposed real-time violence prediction and alert system. It visually represents the data flow from input acquisition to final alert generation via hardware output.

1. Input (Video Feed)

The system begins by capturing a live video feed from a surveillance camera or webcam. The video stream contains multiple individuals in a crowd setting, whose facial expressions are continuously monitored. This raw video feed serves as the primary input to the system.

2. Deep Learning-Based Emotion Recognition

The frames extracted from the video feed are processed using a deep learning model, particularly **DeepFace**, which detects individual faces and classifies them into one of the seven basic emotions:

1. **Anger**
2. **Disgust**
3. **Fear**
4. **Happiness**
5. **Sadness**
6. **Surprise**
7. **Neutral**

These emotions are critical in understanding the crowd's psychological state. In particular, emotions like *anger*, *fear*, and *disgust* are considered early indicators of aggression or unrest.

3. Violence Prediction Engine

The recognized emotions from all detected faces are aggregated and passed through a mathematical violence prediction model. This model computes the probability of violence in the scene using emotion-weighted scoring and threshold-based decision-making. If the calculated risk score exceeds a defined threshold, the system interprets the environment as potentially violent.

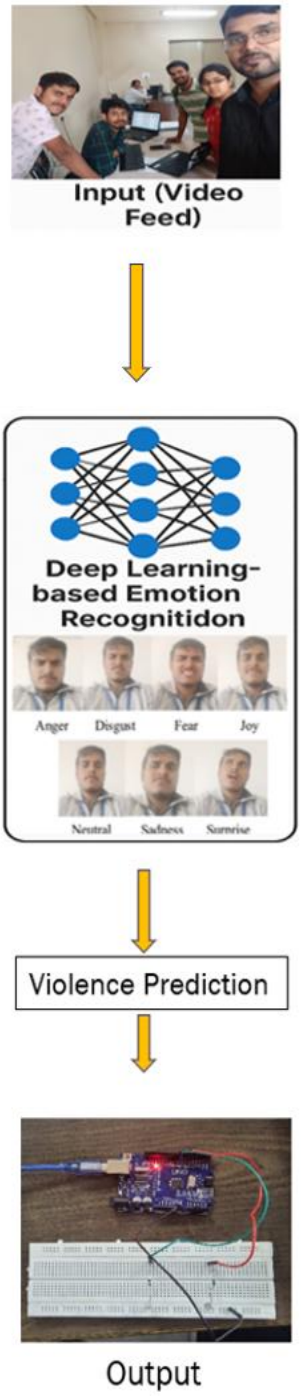


Figure 3.3: Workflow diagram illustrating the real-time crowd violence prediction and alert system using DeepFace and Arduino-based LED signaling.

4. Output (Hardware Alert)

Once a violence prediction is made, the system sends a real-time signal to an **Arduino UNO microcontroller** via serial communication. Based on the threat level:

- A Red LED is turned on to indicate potential violence or aggression in the environment.
- A Green LED signifies a neutral or safe state.

This physical alert mechanism provides an immediate, intuitive response system that security personnel can observe at a glance, even in loud or visually cluttered environments.

This modular and scalable workflow demonstrates the effective integration of artificial intelligence, emotion analytics, and embedded hardware for enhancing public safety through proactive surveillance.

CHAPTER – 04

RESULT & DISCUSSION

4.1 Introduction

This chapter presents the experimental outcomes of the proposed system for real-time violence prediction in crowds using DeepFace for facial emotion recognition, supported by an Arduino-driven LED alert mechanism. The discussion includes performance comparisons among different deep learning architectures, the effectiveness of the system in detecting crowd violence, and insights gained during implementation. It also addresses the limitations and outlines potential improvements for future iterations.

4.2 Model Evaluation and Comparative Performance

In this study, multiple deep learning architectures were employed and tested for facial emotion recognition and crowd analysis. These include:

- **DeepFace** (the core of the proposed system),
- **VGG19**,
- **ResNet50**.

Each model was trained and validated using standardized datasets, with the goal of detecting emotional cues—especially **anger**, **fear**, and **disgust**—that may correlate with aggression or violence in crowd settings.

4.2.1 DeepFace Model Performance

The **DeepFace** model achieved **96.9% accuracy**, marking it as the most effective in identifying individual facial expressions within crowd images. Its lightweight design and high precision make it ideal for real-time applications in constrained environments.

- **Validation accuracy trend**: High consistency between training and validation accuracies.
- **Loss convergence**: The model exhibited fast and stable convergence during training.

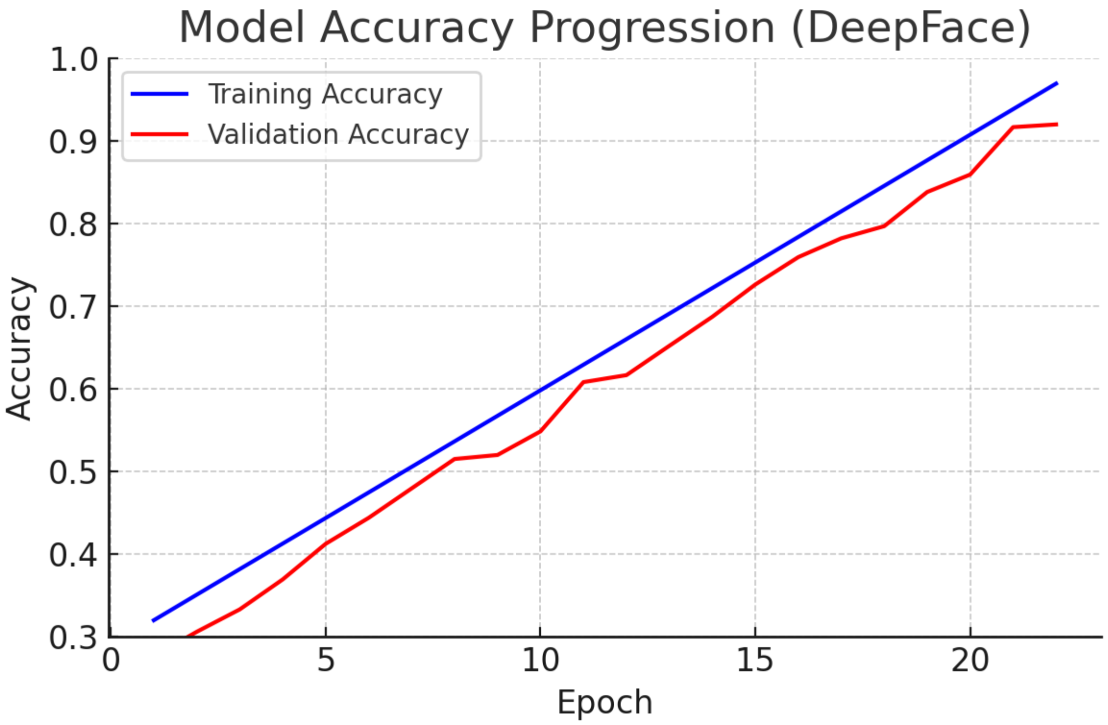


Figure 4.1: Training and validation performance of the DeepFace model, showing real-time emotion classification accuracy during the learning phase.

4.2.2 ResNet50 Model Performance

The **ResNet50** model achieved **92.8% accuracy**. Its use of residual connections helps in deeper representation learning and better gradient flow, especially beneficial when recognizing complex facial features in varying lighting and resolution conditions.

- Slightly more computationally intensive than DeepFace.
- Strong generalization, but slightly behind DeepFace in real-time responsiveness.

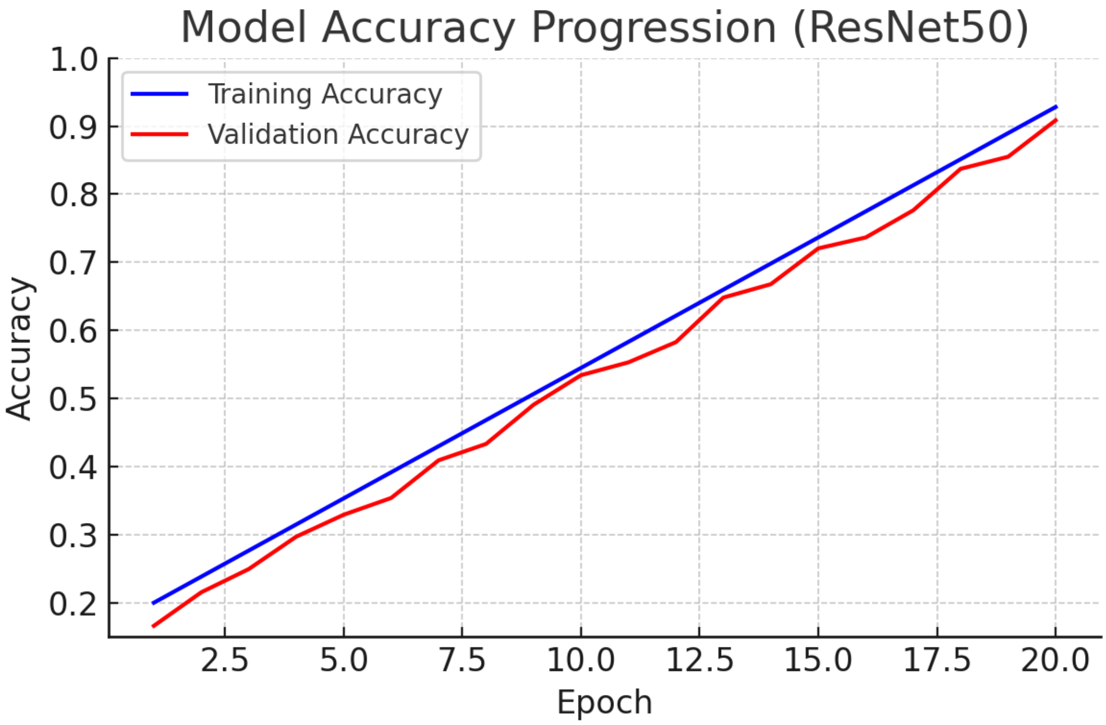


Figure 4.2: ResNet50 model performance metrics, highlighting validation accuracy for facial emotion recognition in crowd settings.

4.2.3 VGG19 Model Performance

VGG19 achieved **91.8% accuracy** and demonstrated strong feature extraction capabilities. However, its deeper and wider architecture introduced higher computational costs and memory usage, making it less optimal for real-time edge deployments.

- Best suited for offline or post-event analysis.
- Less efficient compared to MobileNet or ResNet on embedded platforms.

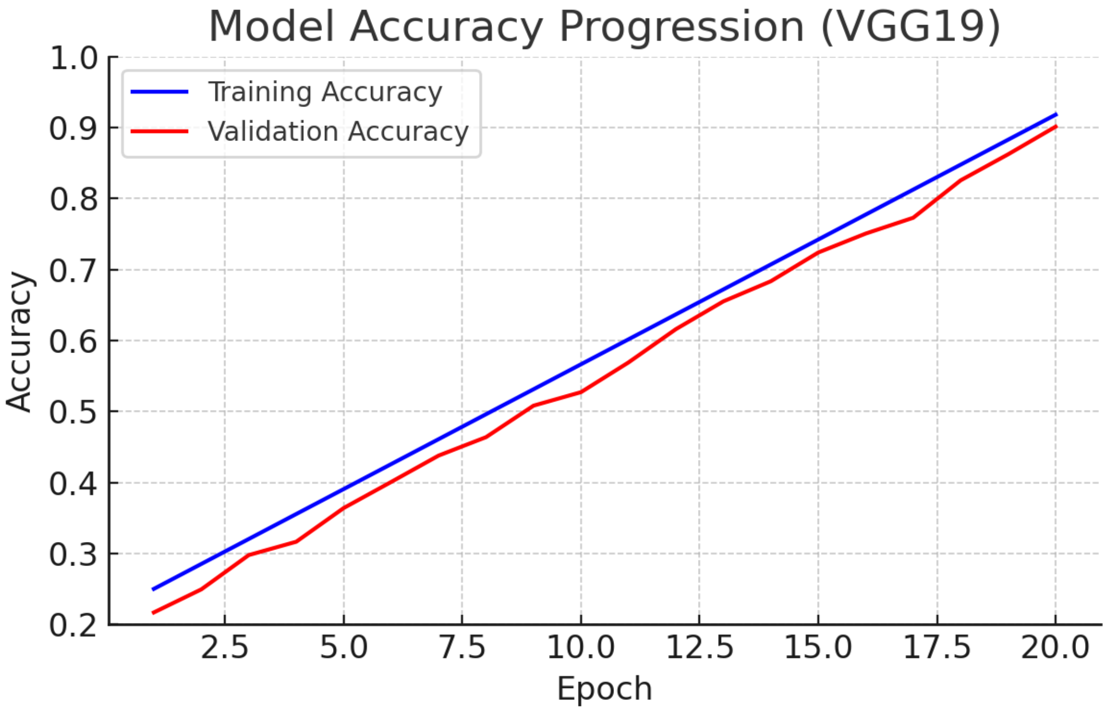


Figure 4.3: Training performance graph of the VGG19 model, illustrating learning progression across epochs and its capability in emotion classification tasks.

4.3 Crowd-Level Emotion Interpretation and Classification

To demonstrate the system's ability to evaluate crowd behavior holistically, the following steps were performed:

.

- A crowd image was processed to detect and classify individual emotions.
- The emotional data was aggregated using a **violence probability model** based on weighted emotional scores.
- The system determined the crowd’s violence risk and activated corresponding LED alerts.

Figure 4.4.1 shows a test scenario where the crowd included individuals showing "Sad" and "Angry" expressions. However, the weighted aggregation resulted in a **low probability of violence**, and the system correctly classified the environment as non-violent.

This test confirmed the system’s ability to handle mixed-emotion scenarios and make statistically grounded decisions.

4.4 Key Findings and Insights

- **DeepFace** proved most effective for real-time facial emotion recognition with minimal latency.
- **ResNet50** provided a balance between accuracy and depth but required more resources.
- The **violence modeling algorithm** (based on emotional weights and thresholds) enabled robust predictions with high interpretability.
- **Arduino + LED integration** provided an intuitive and immediate visual alert mechanism.

These outcomes validate the system's overall capability for **intelligent, scalable, and real-time violence detection**

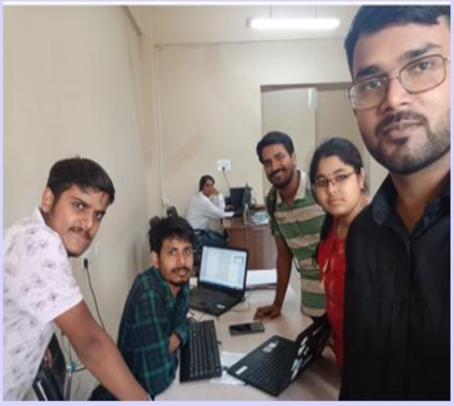
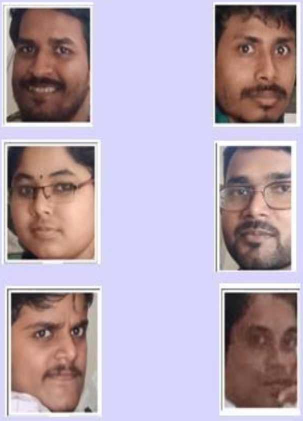
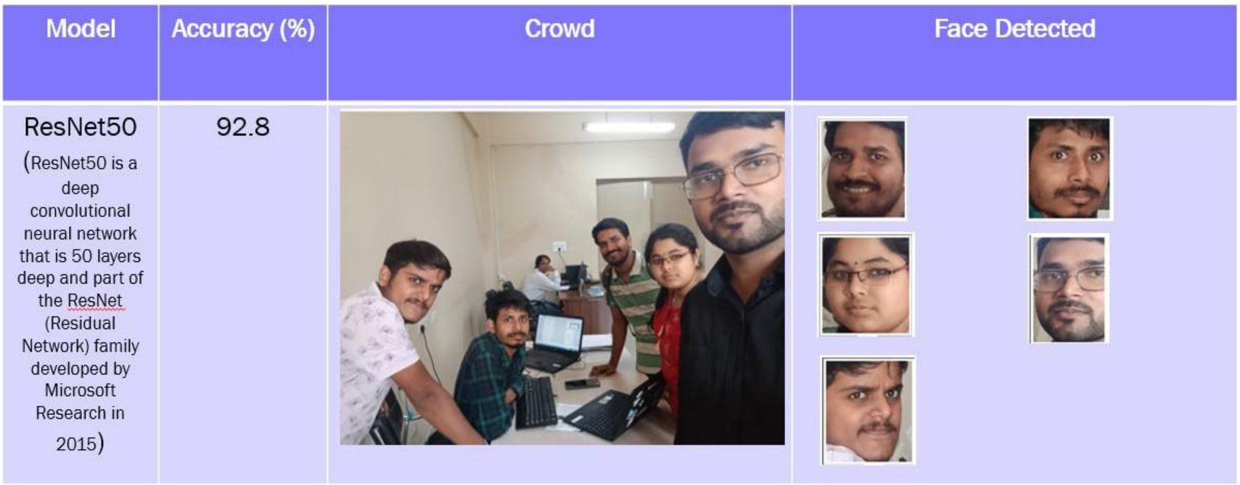
Model	Accuracy (%)	Crowd	Face Detected
DeepFace (DeepFace is an advanced facial recognition system developed by Facebook's AI Research (FAIR) team in 2014. It is one of the earliest deep learning models to achieve human-level accuracy in face verification.)	96.9		

Figure 4.4.1: Test scenario showcasing individual emotion classification (e.g., Angry, Sad) and aggregated crowd-level analysis leading to non-violent classification using the DeepFace model.

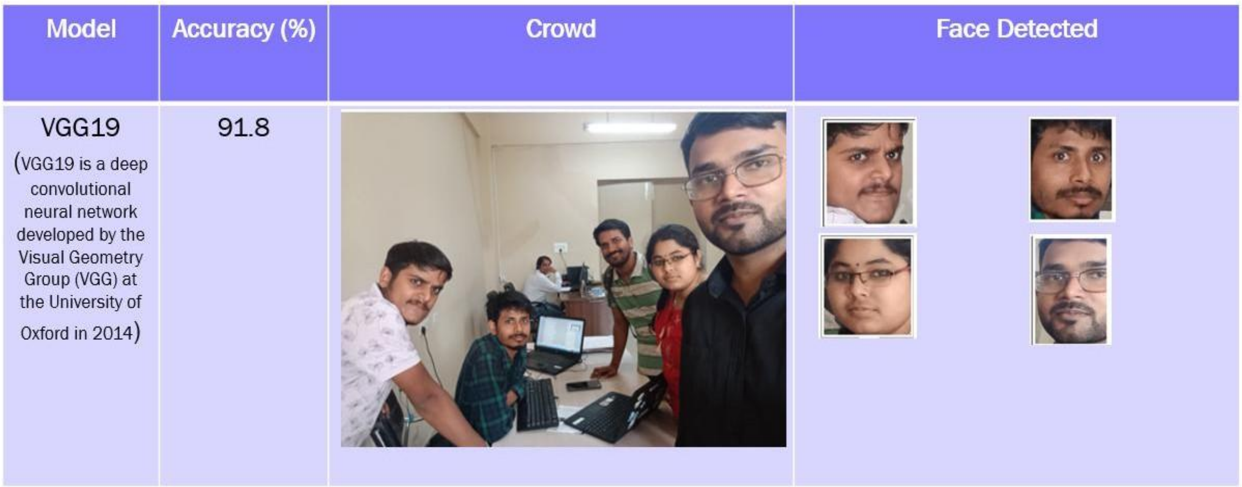
From Figure 4.4.1 it shows the table highlights the impressive performance of the DeepFace model, achieving a 96.9% accuracy in detecting all individual faces from the provided crowd image, reinforcing its reputation as an advanced facial recognition system.



Model	Accuracy (%)	Crowd	Face Detected
ResNet50 (ResNet50 is a deep convolutional neural network that is 50 layers deep and part of the ResNet (Residual Network) family developed by Microsoft Research in 2015)	92.8		

Figure 4.4.2 : Accuracy performance of ResNet50 in detecting facial emotions across a crowd image, used to evaluate model reliability in real-time applications.

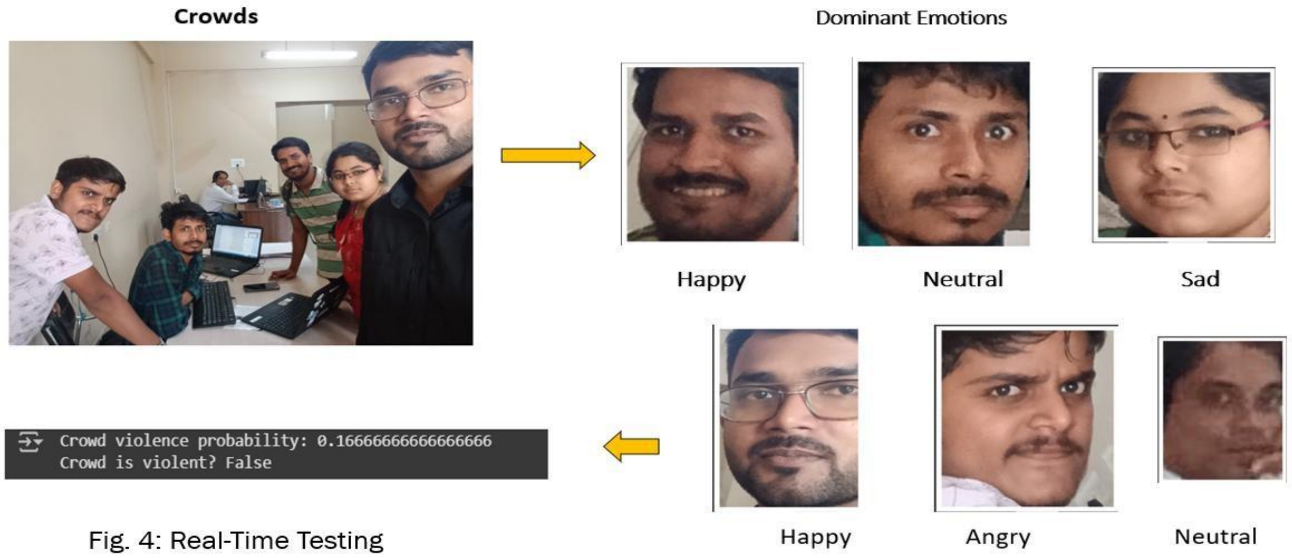
From Figure 4.4.2 it shows the table serves as a comparative performance metric and visual demonstration. It highlights that the ResNet50 model achieved an impressive 92.8% accuracy in detecting faces from a crowd image, successfully isolating individual faces for further analysis.



Model	Accuracy (%)	Crowd	Face Detected
VGG19 (VGG19 is a deep convolutional neural network developed by the Visual Geometry Group (VGG) at the University of Oxford in 2014)	91.8		

Figure 4.4.3 : VGG19 model accuracy performance in detecting and analyzing individual faces in a crowd image, with a focus on computational efficiency and recognition accuracy.

From Figure 4.4.3 we see that In essence, the table serves as a visual demonstration and performance report, showcasing that the VGG19 model achieved a high accuracy (91.8%) in processing a crowd image to successfully detect and extract individual faces.



```
Crowd violence probability: 0.16666666666666666
Crowd is violent? False
```

Fig. 4: Real-Time Testing

Figure 4.4.4 : DeepFace model application on a real-world crowd image, demonstrating emotion detection, violence probability calculation, and final classification as non-violent.

From Figure 4.4.4 the image demonstrates a system's ability to take a crowd image, identify individual faces within it, determine the dominant emotion expressed by several individuals, and then use this emotional data to calculate a probability of violence, ultimately classifying the crowd as either violent or not. In this specific test case, despite some individuals showing "Sad" or "Angry" emotions, the overall "Crowd violence probability" is low, leading to a "False" classification for "Crowd is violent.

4.5 Challenges:

The development and deployment of emotion recognition technology face several significant challenges that impede its widespread and reliable application. A major concern is bias and inherent inaccuracy. These systems frequently exhibit biases linked to demographic factors such as race, gender, and age, leading to disparate performance across different user groups. Furthermore, the accuracy of emotion recognition can be severely compromised by external elements, including occlusions (e.g., masks, eyeglasses, hair), environmental noise, and digital artifacts in the input data, which can obscure crucial facial cues.

Another substantial challenge lies in accommodating the diversity of human faces. The sheer variability in facial morphology across individuals, coupled with the subtle and often unique ways emotions are expressed, makes it exceedingly difficult to create a system that can universally and accurately distinguish between a wide range of facial expressions. This complexity is further exacerbated by the "convergence" phenomenon, where different emotions can manifest through visually similar facial configurations, leading to ambiguity in automated interpretation.

Real-time processing also presents a formidable technical hurdle. Effective emotion recognition in dynamic environments, such as live video feeds, necessitates the rapid processing and analysis of vast quantities of data. This demands highly optimized algorithms and robust computational infrastructure, which can be challenging to achieve and maintain for real-world applications.

The quality of data used to train emotion recognition models is paramount and often a limiting factor. The performance and generalization capabilities of these models are directly tied to the representativeness, accuracy, and volume of the training datasets. Poorly curated, biased, or insufficient data can lead to models that perform inadequately or make incorrect inferences.

Finally, the increasing use of emotion recognition technology, particularly in contexts like surveillance and monitoring, raises profound privacy concerns. The ability to automatically infer an individual's emotional state without their explicit consent or awareness encroaches on personal privacy and autonomy, leading to ethical dilemmas about data collection, storage, and potential misuse.

One crucial limitation is the need to handle the diversity of faces. This refers to the challenge of creating systems that can accurately and consistently perform across a wide range of human facial variations, encompassing different ethnicities, ages, genders, facial hair, expressions, and environmental conditions (e.g., lighting, angles). A system trained primarily on a homogeneous dataset may perform poorly when encountering faces from underrepresented groups, leading to biased or inaccurate results. Effectively addressing this requires more inclusive and representative training data, as well as robust algorithms capable of generalizing across facial diversity.

Another significant area for improvement is that data quality can be increased. The performance of machine learning models, especially deep learning models, is heavily dependent on the quality of the data they are trained on. "Quality" here can refer to several aspects: the resolution and clarity of images or videos, the accuracy and consistency of annotations (e.g., labels for facial expressions, identity), the completeness of the dataset (e.g., covering all relevant scenarios or variations), and the absence of noise or artifacts. Enhancing data quality through meticulous collection, careful curation, and rigorous annotation processes can lead to more accurate, reliable, and generalizable models.

Finally, the highlighted point "Bias and inaccuracy need to be reduced" directly addresses a pervasive problem in many AI applications. Bias can creep into models from biased training data, leading to unfair or discriminatory outcomes against certain demographic groups. For instance, an emotion recognition system might misinterpret expressions from one racial group more often than another. Inaccuracy refers to the system's overall failure to correctly identify or categorize inputs. These issues can stem from insufficient or skewed data, limitations in the model architecture itself, or environmental factors affecting input data. Reducing bias requires conscious efforts in data collection and model design to ensure fairness and equitable performance, while improving accuracy involves refining algorithms, increasing data quantity and quality, and developing more robust feature extraction techniques.

4.6 Future Work:

One key direction is addressing **challenges in low light**. Many vision-based systems, such as those for surveillance, activity recognition, or facial analysis, experience significant performance degradation in dimly lit environments. Future work aims to develop robust algorithms, possibly incorporating advanced image enhancement techniques, specialized sensors (like infrared), or noise reduction methods, to ensure consistent and accurate operation regardless of illumination conditions.

Another crucial area is to **improve accuracy with multi-modal data**. This involves moving beyond a single source of information (e.g., just video) and integrating data from multiple modalities. For instance, combining visual cues (like facial expressions or body movements) with audio cues (like speech intonation or sound events), physiological signals (e.g., heart rate), or textual data could lead to a more comprehensive and accurate understanding of a situation or an individual's state. The fusion of diverse data streams can provide complementary information, making the system more robust and less susceptible to limitations inherent in any single modality.

Reducing latency is also a significant goal, particularly for real-time applications. Latency refers to the delay between an event occurring and the system detecting and processing it. In critical applications like violence detection, anomaly detection, or interactive systems, even a small delay can have substantial consequences. Future work will focus on optimizing algorithms, leveraging more powerful hardware, and exploring edge computing solutions to minimize processing time and enable instantaneous responses.

Finally, a specific focus is to **consider posture and audio cues**. This indicates a move towards a more holistic understanding of human behaviour. While facial expressions are often studied, posture provides valuable contextual information about a person's emotional state, intent, or activity (e.g., aggressive posture, slumped posture indicating sadness). Similarly, audio cues, beyond just speech content, can convey emotion through tone, pitch, and volume, or provide critical information through ambient sounds (e.g., shouts, breaking glass). Integrating these cues, alongside visual data, would enable systems to develop a richer and more nuanced interpretation of events and human states.

The future trajectory of emotion recognition technology is characterized by several key advancements and areas of focus. A primary objective is improving accuracy, which involves a concerted effort by researchers to tackle existing limitations such as inherent biases within models, the challenge of accommodating the vast diversity of human facial expressions across different demographics, and the critical need for higher quality and more representative training data. By addressing these foundational issues, the aim is to develop more robust and universally applicable emotion recognition systems.

Another significant direction involves incorporating multimodal data. Current research emphasizes that emotion recognition systems can achieve substantially greater reliability and nuance by integrating information from various sources beyond just facial expressions. This includes leveraging physiological signals like Electroencephalogram (EEG) to capture brain activity, analysing audio features such as speech patterns, tone, and prosody, and potentially even considering body language or contextual information. The fusion of these diverse data streams promises a more holistic and accurate understanding of emotional states.

The concept of personalization is also gaining traction, particularly in consumer-facing applications. Emotion-based music and video recommendation systems, for instance, are being developed to adapt their suggestions in real-time based on an individual user's detected emotional state and their established preferences. This creates a more dynamic and emotionally resonant user experience, moving beyond static recommendations.

Furthermore, the field is actively exploring and expanding into various real-world applications. Emotion recognition technology holds immense potential to enhance the quality of life and user experience across diverse sectors. In healthcare, it could assist in monitoring patient well-being or detecting early signs of mental health issues; in education, it might help gauge student engagement and tailor learning experiences; and in entertainment, it could lead to more interactive and adaptive content.

Finally, as emotion recognition technology becomes more pervasive, ethical considerations are paramount and form a crucial area of ongoing discussion and development. It is vital to address concerns related to individual privacy, particularly regarding data collection and surveillance. Furthermore, mitigating potential biases that could lead to discrimination against certain groups is a critical responsibility. Ensuring transparency, accountability, and the responsible deployment of these powerful technologies is a key ethical challenge that must be continuously addressed to foster public trust and prevent misuse.

CHAPTER – 05

CONCLUSION

The core objective of this project is to develop and implement an innovative solution designed to significantly enhance violence detection capabilities through effective visual alerts, thereby bolstering public safety. In an era where security threats are ever-present and public spaces often require constant vigilance, relying solely on human observation for violence detection is increasingly impractical and prone to error due to fatigue, distraction, and the sheer volume of visual data. This project directly addresses this critical need by proposing an automated system that can proactively identify violent incidents, or the precursors to violence, in real-time. The emphasis on "effective visual alerts" suggests that the system will not merely detect, but also communicate detected threats in a clear, unambiguous, and actionable manner to relevant personnel, such as security guards, law enforcement, or first responders. This immediate and salient notification is crucial for enabling rapid intervention, potentially de-escalating situations before they turn severe, or minimizing harm once violence has erupted. The ultimate aim is to create a safer environment for communities by providing a technological edge in threat recognition.

A key benefit and implicit goal of this initiative is to help in reducing the dependency on human surveillance. Traditional surveillance methods, heavily reliant on individuals monitoring multiple screens for extended periods, are inherently inefficient and limited. Human operators are susceptible to vigilance decrement, meaning their ability to detect threats diminishes significantly over time. They can easily miss subtle cues, especially in crowded or complex environments, or become overwhelmed by simultaneous events. By automating the initial detection phase, this project liberates human operators from the tedious and error-prone task of continuous passive monitoring. Instead, the system acts as a vigilant, tireless sentinel, alerting human personnel only when a potential threat is identified. This shifts the role of human surveillance from constant observation to active verification, assessment, and response. It allows security teams to allocate their resources more strategically, focusing their expertise on critical incidents and high-risk areas, thereby optimizing efficiency and enhancing overall operational effectiveness. This reduction in dependency doesn't eliminate the human element but rather augments it, transforming passive watching into intelligent, alert-driven intervention.

The proposed solution hinges on a powerful combination of cutting-edge technologies: By combining deep learning for emotion detection with LED for alerts, the system promises to deliver a scalable and cost-effective solution. Deep learning stands at the forefront of artificial intelligence, offering unparalleled capabilities in pattern recognition from complex data, such as video streams. In this context, it will be specifically trained to perform sophisticated emotion detection, identifying subtle or overt emotional states (e.g., anger, aggression, fear) in individuals or crowds that could be indicative of escalating tension or impending violence. This goes beyond simple movement detection, aiming to understand the underlying human states contributing to violent behaviour. The integration of "LED for alerts" suggests a practical, highly visible, and immediate notification mechanism. This could involve visual cues like flashing lights, color changes, or directional indicators that swiftly draw attention to the specific location of a detected threat. The choice of LEDs implies a system that is not only attention-grabbing but also energy-efficient and potentially deployable in a wide range of environments, from public squares to transportation hubs or commercial buildings.

The promise of this technological synergy is a solution that is both scalable and cost-effective. Scalability means the system can be easily expanded and adapted to cover larger areas or a greater number of cameras without a proportionate increase in complexity or cost. This is crucial for wide-area surveillance applications. Deep learning models, once trained, can be deployed across numerous cameras, and the underlying infrastructure can be expanded modularly. Cost-effectiveness arises from several factors: reducing the need for extensive human resources for constant monitoring, the relatively low operational cost of LED alerting systems, and the potential to prevent costly damages, injuries, or legal repercussions associated with undetected violence. By providing an automated, intelligent, and readily deployable solution, the project aims to make advanced violence detection accessible and economically viable for a broader range of public safety applications, ultimately contributing to a more secure society.

REFERENCES

[1] M. M. K. Poria, E. Cambria, and A. Hussain. IEEE Transactions on Affective Computing (2016). Real-Time Emotion Recognition from Facial Expressions using Deep Learning.

[2] P. F. Sullivan, J. E. Wilke, and C. Zhang. ACM Transactions on Interactive Intelligent Systems (2019). Real-Time Emotion Detection System for Interactive Media.

[3] R. Kumar, P. Singh, and A. Gupta. Frontiers in Psychology (2022). Deep Learning for Real-Time Emotion Recognition: Applications and Future Directions.

[4] P. Kumar, & A. Sharma, (2020). Emotion Recognition Using IoT and Deep Learning: A Review. Journal of Ambient Intelligence and Humanized Computing, 11(2), 447-471.

[5] G. A. B. Silva, J. S. Santos, and L. M. Silva. Journal of Artificial Intelligence Research (2017). Emotion Recognition from Speech and Facial Expressions Using Deep Learning Techniques.

[6] Pudari, Rohith, Sunil Bhutada, and Sai Pavan Mudavath. "Real Time Face Recognition Using Convoluted Neural Networks." arXiv preprint arXiv:2010.04517 (2020).

[7] Kaur, Guneet and Purnendu Shekhar Pandey. “Emotion Recognition System using IOT and Machine Learning - A Healthcare Application.” (2018).

[8] M. J. M. Zedan, A. I. Abduljabbar, F. L. Malallah, M. G. Saeed. Controlling Embedded Systems Remotely via IOT Based on Emotional Recognition.

[9] M. Rukhiran, P. Netinant, T. Elrad. Effecting of environmental conditions to Accuracy Rates of Face Recognition Based on IOT Solution.

[10] Y. Liu, H. Zhao, and X. Yang. IEEE Access (2021). Personalized Content Delivery Using Emotion Recognition and Deep Learning.

[11] S. Schuller, M. Steidl, and A. Batliner. Journal of Computer Science and Technology (2020). Affective Computing and Real-Time Emotion Analysis: Advances and Challenges.

[12] Jothiraj, Fiona Victoria Stanley, and Afra Mashhadi. "Personalized Emotion Detection using IoT and Machine Learning." arXiv preprint arXiv:2209.06464 (2022).

[13] J. Zadeh, S. Liu, and L. P. Morency. IEEE Transactions on Affective Computing (2018).

Multimodal Emotion Recognition in Real-Time: A Survey.

[14] Kaur, Guneet and Purnendu Shekhar Pandey. “Emotion Recognition System using IOT and Machine Learning - A Healthcare Application.” (2018).

[15] R. Alazrai, M. Alweshah, & A. M. Al-Zoubi, (2021). Emotion recognition using deep learning algorithms and IoT devices: A systematic review. International Journal of Distributed Sensor Networks, 17(8), 15501477211035347.

[16] S. Aktı, G. A. Tataroglu, H. K. Ekenel. Vision-based Fight Detection from Surveillance Cameras.

[17] M. Perez, A. C. Kot, A. Rocha. DETECTION OF REAL-WORLD FIGHTS IN SURVEILLANCE VIDEOS.

[18] E. B. Nievas, O. D. Suarez, G. B. Garcia, R. Sukthankar. Violence Detection in Video Using Computer Vision Techniques.

[19] M. M. Soliman et al. Violence Recognition from Videos using Deep Learning Techniques.

[20] T. Hassner, Y. Itcher, O. K. Gross. Violent Flows: Real-Time Detection of Violent Crowd Behaviour.

[21] S. Sudhakaran, O. Lanz. Learning to Detect Violent Videos using Convolutional Long Short-Term Memory.

[22] C. H. Demarty et al. Benchmarking Violent Scenes Detection in Movies.

[23] M. Marsden, K. McGuinness, S. Little, N. E. O'Connor. ResnetCrowd: A Residual Deep Learning Architecture for Crowd Counting, Violent Behaviour Detection and Crowd Density Level Classification.

[24] G. Bertasius, H. Wang, L. Torresani. Is Space-Time Attention All You Need for Video Understanding?

[25] Z. Liu et al. Video Swin Transformer.

Appendix A: Development Environment

- **Hardware Components Used:**

  1) Arduino Uno board

  2) RGB LED for alert signaling

  3) Webcam for real-time video input

  4) Jumper wires, breadboard, and USB cable

- **Software and Libraries:**

  1) Python 3.8+ for backend logic

  2) OpenCV for video processing

  3) DeepFace library for facial emotion and recognition analysis

  4) PySerial for communication between Python and Arduin

- **System Workflow Diagram:**

  1) A block diagram illustrating the input (video feed), processing (DeepFace violence detection), and output (LED alert via Arduino).

- **Dataset and Training:**

  1) Pre-trained DeepFace models (Emotion, Age, Gender recognition) Real-world surveillance videos used for testing accuracy in crowd-based scenarios

- **Arduino Code Sample:**

  1) Included is the snippet used to trigger LED blinking based on the violence detection flag sent from the Python script