RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Report on

SHE ALERT: VIGILANCE REDEFINED

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Divya Binu (U2103077)

Diya Baiju (U2103078)

Diya Thankachan (U2103079)

E H Hrithika (U2103080)

Under the guidance of

Ms. Dincy Paul

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
April 2025

CERTIFICATE

This is to certify that the project report entitled ***"She alert: Vigilance redefined"*** *is a bonafide record of the work done by* ***Divya binu(U2103077)***, ***Diya baiju(U2103078)***,***Diya Thankachan(U2103079)***, ***E H Hrithika (U2103080)*** *submitted to Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in "Computer Science and Engineering" during the academic year 2024-2025.*

Ms. Dincy Paul
Project Guide
Assistant Professor
Dept. of CSE
RSET

Ms. Anu Maria Joykutty
Project Co-ordinator
Assistant Professor
Dept. of CSE
RSET

Dr. Preetha K. G.
Head of Department
Computer Science and Engineering
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Rev. Dr. Jaison Paul Mulerikkal CMI**, Principal of RSET, and **Dr Preetha K G**, Head of the Department of Computer Science and Engineering for providing us with the opportunity to undertake our project, "SheAlert: Vigilance Redefined".

We are highly indebted to our project coordinators, **Ms. Anu Maria Joykutty**, Assistant Professor, and **Dr. Sminu Izudheen**, Professor of Department of Computer Science and Engineering, for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Ms. Dincy Paul** for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Divya Binu

Diya Baiju

Diya Thankachan

E H Hrithika

Abstract

The Women Safety Project is an AI-powered surveillance system that is designed to improve public safety by automatically detecting suspicious behavior in real-time, focusing specifically on protecting women in public spaces. It uses advanced computer vision techniques such as gender detection, anomaly detection, facial expression recognition, and camera obstruction monitoring to provide continuous and proactive monitoring. It reduces the reliance on manual human oversight, which is prone to errors, fatigue, and delays. The system can detect aggressive gestures, distress signals, and unusual behavior patterns that may indicate harassment or violence, sending real-time alerts to authorities or security personnel for swift intervention. The camera obstruction detection feature ensures that the system is still functional even if objects or environmental factors block the view of the camera. The project aims to provide a comprehensive solution to women's safety concerns, which are designed to operate in various public environments, such as crowded public areas and workplaces. The system improves the speed and accuracy of threat detection while offering a proactive approach to preventing incidents of gender-based violence by using AI to continuously monitor and assess potential threats. This project aims to contribute to safer public spaces and reduce the risk of harm to women through timely, data-driven interventions.

Contents

List of Abbreviations

- **AI** - Artificial Intelligence
- **CNN** - Convolutional Neural Network
- **YOLO** - You Only Look Once
- **SSD** - Single Shot MultiBox Detector
- **R-CNN** - Region-based Convolutional Neural Network
- **RNN** - Recurrent Neural Network
- **LSTM** - Long Short-Term Memory
- **ResNet** - Residual Network
- **VGG** - Visual Geometry Group
- **InfiNet** - A deep learning model optimized for real-time gender detection (specific expansion not provided; presumed contextual)

List of Figures

List of Tables

Chapter 1

Introduction

The She Alert is an AI based surveillance system to boost public safety, especially in contexts where women are likely to be exposed. It constantly scans public areas in real-time to identify possible threats, minimizing human error and the need for manual monitoring. By processing movement patterns and interaction, the system is able to detect unusual activity and raise alerts for timely intervention. This proactive approach ensures a faster response, preventing incidents before they escalate.Among She Alert's strengths is the feature of gender and facial expression recognition to identify individuals' distress or fear. The system also recognizes suspicious activities based on anomalies from typical behavior patterns, thus highlighting potential threats. The system is also capable of detecting blockages in CCTV cameras to ensure continuity of monitoring. An integrated threat assessment feature also gauges the severity of detected anomalies, enhancing decision-making among security officers.She Alert sends automated alerts to law enforcement or security personnel, allowing them to respond quickly to urgent situations. Scalable and adaptable, the system can be implemented in different environments, such as public areas, transportation systems, and workplaces. Its all-encompassing safety measures not only augment security efforts but also build a safer community by using technology to deter crime and build public trust.

1.1 Background

Serious issues regarding women's safety have arisen in a public asset over the last few years. Many incidents related to harassment, assault, and gender-based abuses have occurred against individuals in various other parts of the world, urging the establishment of effective and continuous surveillance systems to address those concerns. The traditional design of security systems relied on the human-based way of monitoring security in which time

delay is generally an influence factor in an event incident in the facility due to fatigue or human error in judgment. This issue becomes all the more serious in those settings where public safety alone cannot be ensured through human observation, such as very crowded places, transport terminals, and remote areas. Therefore, this is the next step forward in which AI can be used to make improvement mechanisms automating the detection and prevention of unsafe situations. AI-based surveillance systems are beneficial in this regard as they provide automated detection of an actual threat in an event to monitoring user. The sophisticated advanced algorithm attached with video-streaming capability analyses different, specific videos and describes them to tell about dubious activities with aggressive forms of gesture or movement that could be interpreted as an attack in action or occurrence. There is a contrast between AI and human surveillance in terms of the fact that they are amenable to continuous work and do not require active human attention to initiate actions in approaches to safety proactively. Thus, the system would differentiate the various men and women through gender-based behavioral specific patterns which could be very vital in cases where woman protection from harassment and assault comes into play. Therefore, building up gender detection capabilities, along with other such behavioral analyses such as recognition of facial expression, can add a great deal as far as automatic detection of distress or aggression signs goes, thus among the most critical signs of putting women into danger. Another modern feature of surveillance is unobstructed cameras installed for monitoring, as cameras in the real world get blocked either through objects or people or caused by outdoor conditions such as fogging or dirt; these cause voids in the charge. The obstruction detection feature, therefore, incorporated in an AI-based system would facilitate prompt detection and correction of these problems to ensure monitoring systems remain effective. The emerging factor of public safety demand vis-a-vis the growing adoption of surveillance technologies in urban planning and development has once again focused attention on surveillance systems. Much depends on how artificial intelligence or non-AI will prove to be in the efficiency and efficacy of intervention-type decisions quickly made when data indicate the possibility of movement in the situation toward danger; improving response thus becomes essential.

1.2 Problem Definition

The high rates of harassment and unsafe environments experienced by women and other marginalized groups in public areas create a significant need for increased safety. Traditional monitoring systems are generally based on human observation and in attendance, and, therefore, have factions, based on that dependence; for example, if monitoring an area, they cannot act on a threat until it is perceived (often between a few seconds to several minutes). Some of the limitations of using people to monitor for potentially suspicious behavior are a delay in recognizing that the behavior is suspicious, not witnessing the behavior at all, or needing too long to respond, which can escalate something minor into a significant safety threat. Public areas like transportation hubs, shopping malls, sidewalks, streets, and office buildings have a specific surveillance challenge due to short bends and compression. Being able to differentiate between routine human activity and potentially suspicious or problematic human behavior, in these spaces takes some accuracy and attention that human monitoring alone cannot always provide. Another significant challenge is the need for real-time identification and response to threats. Current systems could not observe multiple feeds of live video data quickly enough to become aware of a suspected incident and alert responding officers or security personnel. Without having a system with an ability to operate as intelligent monitoring systems, which can observe behavior and identify threats, risks of incidents being unnoticed and responded to are increased going forward. The listed limitations are complemented by the increasing need for cost-effective and scalable solutions that are capable of conforming to various public environments and addressing the intricacy of contemporary surveillance needs. The provision of privacy and security of data while deploying such systems further complicates matters, necessitating the discovery of creative solutions to meet these complex challenges. This issue highlights the imperative to develop an AI-driven surveillance system that can promote security in public places by automating behavior monitoring, providing reliability under various conditions, and facilitating timely responses to potential dangers.

1.3 Scope

The project aims to build an AI surveillance system to make the environment safer for women and vulnerable groups. The system provides real-time monitoring while adding

advanced features, such as gender recognition, behavioral assessment, and obstacle awareness. As this transition will happen in environments such as the public transportation hub or while using public transportation, in a shopping mall, navigating urban streets, or in workplace workflow where there are many environmental constraints (congested areas, different lighting situations, different weather) we know the system will function well as we have experience with other transitions in varying environments. The new technology will be implemented to scale in a cost-effective way from an already present CCTV safety installation with efficient alerting for quick queueing to action. This product will adhere to legislation applicable in these jurisdictions and proper ethical usage will have a focus on data security and privacy. This product aims to reduce human monitoring, improve reliability in present surveillance systems, while being a valuable addition in creating safe smart public spaces.

1.4 Motivation

This idea was formulated due to the increasing amount of crimes that are faced by women in public spaces. They are exposed to harassment, assault, and unsafe conditions, which has become a significant international problem, revealing the inadequacies of an available method of protection. Most conventional surveillance technologies have been globally based on human observation, which is prone to mistakes, exhaustion, and delays in noticing and reacting to a threat. Additionally, most of the existing systems have been incapable of real-time monitoring, resulting in large coverage gaps during very crucial moments. This requires a more sophisticated system, enabling automated monitoring with a view to making women's safety and security accessible in busy public areas.

This particular research proposes the induction of artificial intelligence to fill the chasm left by conventional methods in the proactive identification and elimination of threats. This will also provide for the minimization of issue escalation through better detection times and higher intervention levels. The system will continuously monitor the preservation of women's feeling of safety and empowerment in public areas through disturbance in a hostile environment. In the long run, the intention is to purposely put this technology to work towards a smarter and safer world for women, thus instilling public confidence in safety features for the benefit of society.

1.5 Objectives

- To develop an AI based surveillance system that identifies and alerts authorities about potential threats to women in real-time.
- To apply an obstruction detection mechanism ensuring continued and unobstructed camera monitoring.
- To set up an alert system to notify the authorities immediately of an incident happening so that they can quickly respond.
- To test the performance of systems in real-world scenarios in order to demonstrate its reliability and effectiveness in improving the safety of women.

1.6 Challenges

One of the biggest challenges of this project is making sure that AI is accurate and reliable in identifying threats. In populated areas, individuals engage in countless different ways, so it's challenging to separate normal behavior from a possible threat. The system should be that intelligent enough to identify true danger without producing false alarms. It must also function under various light conditions—either harsh daylight, soft evening light, or total darkness—without compromising precision. Physical barriers such as poles, cars, or even individuals in front of the camera can also make it very difficult to detect threat.

Another significant problem is the ability to process significant video data in real time. The system should identify the threats in a timely manner and alert without any delays even when surveilancing several locations simultaneously. Adverse weather conditions such as rain, fog, or rough winds may degrade the video quality, and camera shakes caused by vibrations add noise in the video. Finally, the system needs to be affordable as well as scalable to enable it to be rolled out extensively without becoming very costly to maintain.

1.7 Assumptions

The system installations shall be carried out in places where normal CCTV camera setups provide an unobstructed view of the area under investigation. Most of these cameras are installed in high-traffic locations, such as great malls, airports, and so on. Cameras placed in public transport hubs would maximize views without leaving any blind spots. The cameras will do their jobs well under all light conditions, ensuring clear inputs by day and night. Thus, the system will ensure continuous capturing of real-time video feeds, providing an overall view of monitored spaces for pattern detection and safety supervision.

The desired AI models in Anomaly Detection and Gender Recognition would need a training dataset that is a representative of various demographics, environments, and real-life situations. This would, therefore, mean that any such applications in the ambit would, ideally, be very good at recognizing human behavior without developing any biases toward age, gender, or even ethnicity, excluding environmental lighting conditions. So all such biases will be minimized and will even allow the proper functioning of the system on all built dialects. An almost real-time interface with a negligible amount of latency is bound to guarantee lightning-fast threat identification. Alerts will almost instantly reach the concerned security personnel or establishments according to the nature of the threat and geographical location where it occurred for rapid and targeted intervention, be it prevention or cure-for example. Ordinary setups of CCTV cameras would be the arenas where the system would be installed and where cameras would provide unobstructed views of monitored areas.

1.8 Societal / Industrial Relevance

This project is significant, not only in societal terms but also for industries, in particular addressing the safety of mostly women in public and private spaces. The project relies upon intelligent devices designed to transform environments with real-time monitoring and anomaly detection in proactive interventions against harassment, assault, and even other criminal actions in public safety. The system also ensures timely emergency responses through immediate alerting of the concerned authorities, which reduces the time gap between the incidence detection and action against it, thereby minimizing injuries or loss of lives. This technology is applicable to a broad spectrum of high-risk sites, like public

stations, shopping malls, school premises, airports, parks, and workplaces, by providing continuous monitoring where traditional security measures failed.

In the industrial context, smart city projects will incorporate this project into the urban surveillance system, which increases city security in Leaps and Bounds. This abnormal activity's automatic detection will enable the law enforcement agencies to become more effective in allowing personnel to devote their efforts to critical actions in the presence of reliable documentation for their investigations. One major advantage of the system is its ability to survive the adverse conditions in terms of crowd, lighting, and vision obstructions.

This would also serve to enhance workplace safety. Surveillance is ongoing in potentially hazardous places where harassment or dangerous conduct might occur, thereby providing a workspace that might be a morale booster and instill confidence in the employees. Cost- and space-efficient; well meets the need for national highways toward a massive urban set-up and smaller individual competitors. Scalable and cost effective, it means it's a very affordable alternative to traditional security measures: versatile and flexible.

1.9 Organization of the Report

1. **Introduction** Summarizes the Women Safety Project and briefly outlines its impact, objectives and actions to make women public safer.

2. **Background** : Current landscape of monitoring systems underpinnings with some challenges in the area of public safety with the problems associated to traditional systems. Defines the cause mop needed to seek out an AI solution, with emphasis placed on woman safety.

3. **Problem Definition** : This project seeks to address a particular problem such as real-time detection of suspicious occurrences and limitations of public safety measures at present.

4. **Scope and Motivation** : Outlines the environments in which the system is meant to operate, e.g. crowded public areas, workplaces. Motivations for creating the system include arguments for the need to design an AI-powered surveillance system

relating to womens' safety.

5. **Objectives** : States the chief objectives of the Women Safety Project. For example, these may concern goals such as real-time detection, automation, minimization of human supervision, and timeliness of response.

6. **System Architecture and Methodology** : Gives the overall architecture of AI-powered surveillance system, incorporating gender and anomaly detection, emotion recognition, and camera obstruction detection. Describes the methodology and techniques used, such as computer vision algorithms, machine learning models, and real-time alert systems.

7. **Implementation** : Documentation on how the system was implemented step-wise covering technical details of the model, data processing and deployment in surveillance environments.

8. **Challenges** : Specifically this talks about the general challenges that could be faced during development, such as accuracy in a crowded environment; false alarms minimal; adaptability to different environmental conditions.

9. **Assumptions** : Discusses the ideas that have been considered into account during the development of the system regarding the environmental factors such as positioning of the camera and the requirements for real-time operation monitoring.

10. **Evaluation and Results** : Evaluates the criteria used to test the system regarding performance, accuracy, and efficiency. Quantitative and qualitative results in favor of the system being capable of detecting suspicious activities and securing public spaces.

11. **Societal / Industrial Relevance** : Making a dive into wider context and importance of Women Safety Project coupled with what possible impacts it could have on society and industry as far as making the public domain safe is concerned.

12. **Conclusion and Future Work** : The study summarizes important findings and outlines the potential that the AI based surveillance system has towards maximizing the detection of possible threats.

This project pertains to Women Safety and will develop artificial intelligence based surveillance that will carry out strong real-time identification and detection of suspicious activities related to women and will strongly contribute to making public spaces safer for women. Technologies such as face identification, blockage detection at camera-level, and behavior analysis will be a substitute for human monitoring and notify authorities as soon as such activities are noticed so that they take action. These steps will, therefore, lead aggressively to solutions for harassment and assault-related problems and give rise to a society-friendly solution.

While generally making a point that the continuous flow of data is being surveilled as opposed to making intermittent captures, "surveillance" was made suggestively because it would be compared with the fact that traditional systems would often fail, have errors, and have delays. It would include detection of suspicious behavior in real time, gender classification, obstruction monitoring, and alert-raising, all in the name of public safety-and from whatever angle-the indoor advanced AI tools within the system can recognize aggression, facial expressions, and also unobstructed-camera views. The most challenging aspects of the problem lie in complex environments and lighting conditions. It is a contribution both to society and industry because it involves not only smart city initiatives but also engenders safety at public and private spheres.

Chapter 2

Literature Survey

In this section, we shall review literature pertaining to our proposed project in an already published context, more specifically on AI-enhanced surveillance systems, gender classification, camera obstruction, and real-time threat detection. The application of these methodologies serves towards the enhancement of security, particularly for women in public spaces.

2.1 Gender Classification

As far as the identification of gender from facial pictures is concerned, it is central to human-based applications such as advertising, health, and law enforcement. Unfortunately, issues such as the unbalanced datasets with certain gender categories, variations in facial appearance, and possible age-related problems are as many factors affecting the performance of different models.

The recognition of gender has its importance in a variety of real-life applications that include human-computer interaction, security solutions, and consumer analysis. Tracing gender from facial images is easy whereas tracing age from the same image becomes a lot more complicated to achieve. Various deep learning models have been used in gender detection but ResNet[1] is preferably the most commonly used due to its excellent ability to extract deep facial features with high accuracy.

ResNet (Residual Network) is a model for deep learning that features state-of-the-art performance for image classification purposes, including gender detection. Created by Microsoft, ResNet brought forward the idea of residual learning, which ensures training very deep neural networks without experiencing the usual issue of vanishing gradients. With this architecture, information flows through shortcut connections, making it possible for deeper networks to learn more intricate features without losing important details. This

makes ResNet extremely effective for precise facial feature extraction tasks, like gender classification.

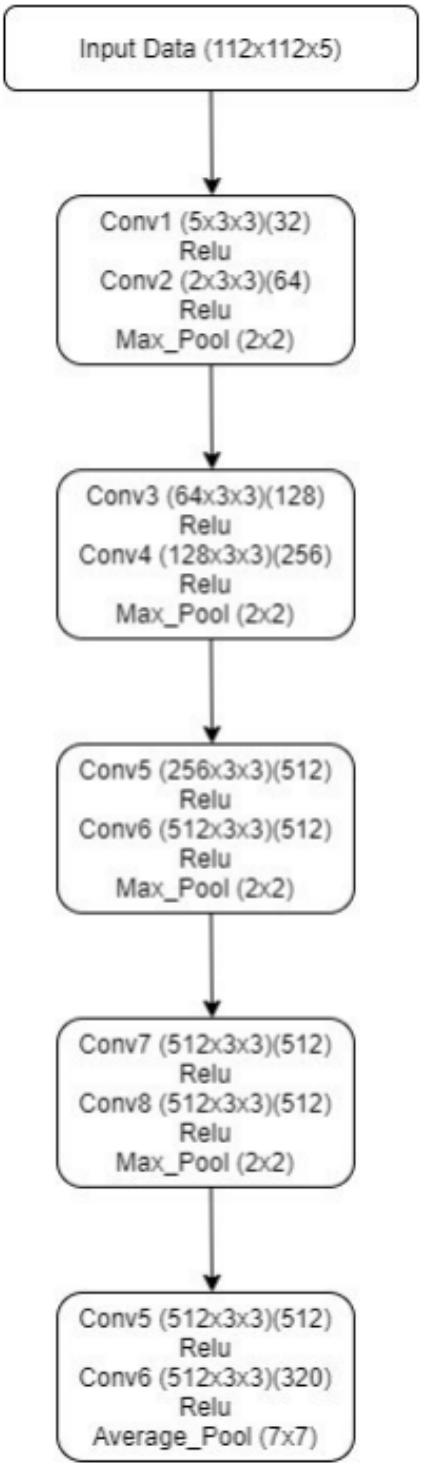


Figure 2.1: Detailed Layers of CNN

ResNet has stood out among gender detection models by consistently achieving higher

accuracy than other deep learning techniques. Research has shown that such models exceed 90% gender classification accuracy for benchmarking datasets like UTKFace and Adience. High accuracy, specifically, ensures reliability in applications that involve safety concerns. By accurately identifying the gender of a user, a more nuanced response can be generated to make sure that the alarms are triggered appropriately in contexts involving women safety.

The methodology begins with capturing video streams from surveillance cameras and preprocessing the streams into model-analyzable frames. The frames are standardized to uniform dimensions and normalized to reduce inconsistencies because of variations in lighting or resolution. We used open-source pre-trained models, particularly ResNet (Residual Network), for real-time face recognition and gender classification. It effectively combated the problems associated with deep architectures and had lots of residual connections, whereas other things have been said for preserving high efficiency when it came to handling considerable data set without losing a lot of accuracy. Specifically how the architecture helps by mitigating the vanishing gradient issues allows it to train deeperly would actually lend to being a good choice for the mission. The model hence, lays more emphasis on leveraging of ResNet that leads to optimal performance without having to necessarily retrain because, as stated, predominantly, model-learned traits generalize well for the male-female classification domain.

It uses pre-trained[2] optimized neural network weights for the identification of male and female patterns in facial features. Initially, it applies classifiers to locate the facial region from the image, and then advanced neural network architecture is used to predict gender. All of a person's video frames are scrutinized for ongoing monitoring, and prediction modifications are made instantaneously to ensure any alerts or notifications triggered vary in context. This method incorporates correction mechanisms against errors due to pose and environmental changes to make it more robust.

Validation through real-time gender detection allows for activating interventions pertaining to each gender where the alerts go to relevant authorities about the perceived threat or are directing focused assistance where needed. The convergence of this module with various behavioral analysis modules like anomaly detection would overall enhance the operations of this system due to greater flexibility in providing assistance to complicated

requirements for safety monitoring.

The analyses conducted for gender identification contribute to a wide application solution for video analysis, creating strong grounds for a gender-sensitive development of our platform towards safety by promoting the development of safer public space.

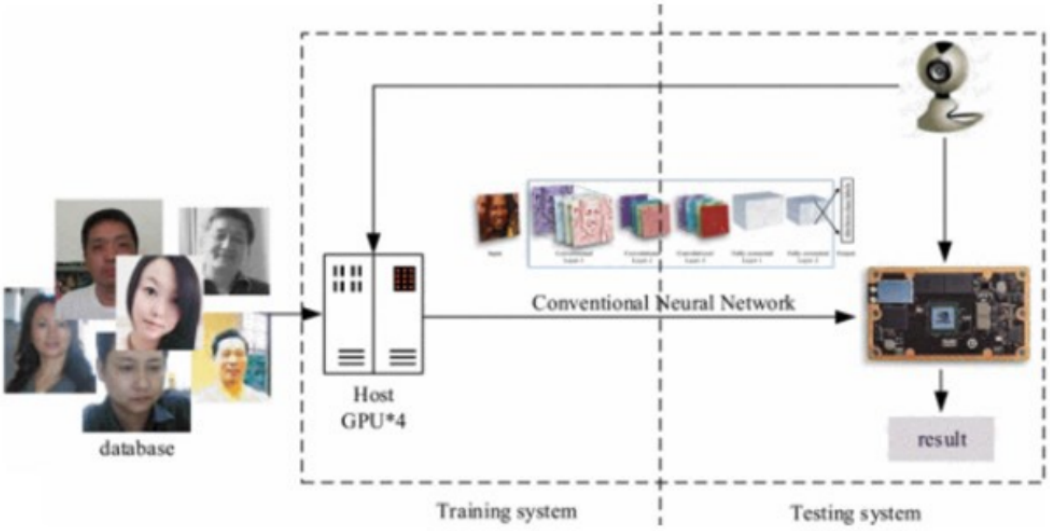


Figure 2.2: Architecture Framework of CNN

[2]

2.1.1 Results and Evaluation

The model's ability to predict gender and age in real time has been greatly improved. This will classify the gender as male or female and the expected age between 2 to approximately 80. This was trained by transfer learning with a deep learning model over about 10,000 human images. The decision was on account of time and resource constraints. The convent of the models during testing was into a wider framework belonging to video capture, which lent to an almost consistent quality across all outputs.

The model was tested on two fronts: Manual Testing with Real-Time Faces: The first set of tests was done live with a webcam stream of a 15-video sequence in real time, whereby the model went frame by frame estimating ages frame during these frames and also recognized as gender differentiation. Gender detection performance was almost perfect. Model accuracy for age was, however, slightly varied by factors between trials-her facial expressions and varied light conditions. Example: On the premise of age of 21, a subject was identified with an error range of just a few years. Celebrities also needed to

be exposed to testing. The system looked at some 100 celebrity images publicly available for general audience testing. The gender classification had an accuracy to be 90%, but the model projected the ages inconsistently because of the slight variations created by factors such as the introduction of makeup and lighting.[3]

$$\text{Accuracy} = \frac{\text{Correct Number of Predictions}}{\text{Total Number of Predictions}}$$

The gender classification model's overall accuracy is approximately estimated to be around 85%, mainly owing to the design optimization made in Keras and in the ResNet architecture, thus achieving much in operational stability. However, with all this robustness, an additional overlay of obstacles can be enumerated as: Pose variations: that is, their age and gender cannot be predicted from side-view poses. This definitely is a different ball game, since the classification barometer, in this case, lies firmly against one point of view, that of frontal face detection. Many things come into play, like changes in environmental lighting, makeup, and facial expression-these inconsistencies are slight but truly become significant to a partial degree in age prediction from one scene to another or in different weather conditions. Perhaps when providing a solution for the above-mentioned predictions, some extensions would take in other variables.

2.2 Anomaly Detection

The public's safety is greatly augmented through the real time detection of violence, theft, or any other incidents, which is accomplished by the monitors passing data to surveillance cameras. However, manual monitoring presents more issues including it being more time consuming, more error prone, and more cumbersome for the user. Therefore, it is imperative to have frameworks that assist monitoring and automatically detect unusual activity. In an effort to consolidate many of these processes, we developed a system that uses fuzzy logic combined with convolutional neural networks (CNN) and recurrent neural networks (RNN). In this system first we use fuzzy logic to extract keyframes from video streams. Then CNN-based transfer learning is utilized to extract spatial features, then LSTM based RNN is utilized for sequence learning. Together these methods allow our system to effectively detect abnormal events. The framework is tested on the UCF50 and UCF-Crime datasets and outperformed many other traditional methods in recognizing

unusual activity in complex video data streams.[4]

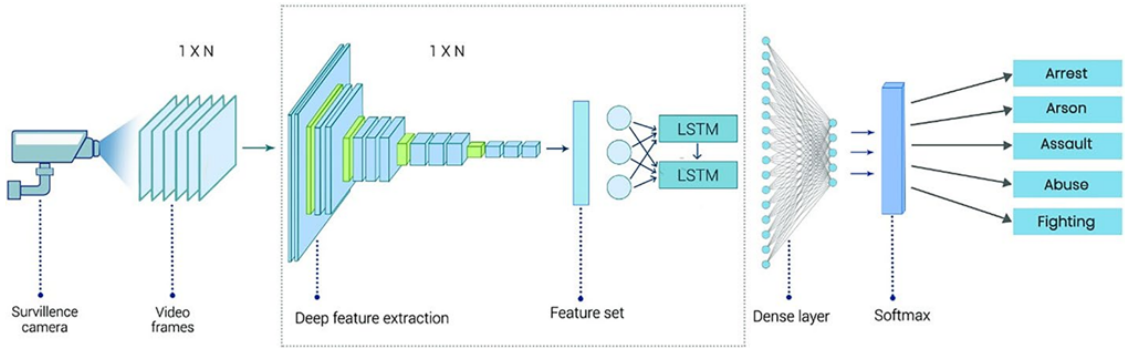


Figure 2.3: Framework for abnormal activity recognition

2.2.1 Frame Processing

At this stage, the video has been Figure 3 disassembled to discrete frames. For each frame, keyframes are defined by finding the histogram distance between two frames in the CIELab color space, which closely mirrors the way humans perceive color. This process involves computing histogram distances Dh for L, a, and b components of the constructs of the two video frames in fuzzy logic with trapezoidal membership functions to remain with frames that contain key information during that frame selection. This process retains frames during the keyframe selection that has the highest dh and is therefore more likely to include key points than frames that did not pass the first fuzzy filtration process. The histogram distances DH are computed as Revised 3 and 4 show the fuzzy membership functions and distances for each of the components. In a similar fashion to Subsection 3.5, the fuzzy distance summary of each frame for each of the CIELab components L, a, and b determine the keyframe definition based on the results of each frame distance measure then summarized to develop a single expression for selection of keyframes. This process hastens the selection of frames and that, with the required frames, will leave meaning for most of the other parts of the video. The histogram distances Dh are computed as:

$$D_h = \sum_{i=0}^{M} \sum_{j=0}^{N} \left(S_{ij}^{n} - S_{ij}^{n+1} \right)$$

Let's suppose S and n are the stacks of the frames and the number of frames, respectively. M denotes the number of rows and N denotes the number of columns in a frame. The Fuzzy logic model uses the histogram distance calculated for consecutive frames.Next,

take the mean for the subsequent frame differences ((d)):

$$\mu(q) = \frac{n(b) + n(a) + n(L)}{3}$$

With the help of calculated µ(q) create the trapezoidal membership function dynamically, where small, medium, and large are the parameters. Then calculate the value of the parameter using Eqs:

$$A = \mu(q) - \mu(q) \cdot 0.4 \tag{2.1}$$

$$B = \mu(q) - \mu(q) \cdot 0.3 \tag{2.2}$$

$$C = \mu(q) - \mu(q) \cdot 0.2 \tag{2.3}$$

$$D = \mu(q) - \mu(q) \cdot 0.4 \tag{2.4}$$

$$E = \mu(q) - \mu(q) \cdot 0.5 \tag{2.5}$$

$$F = \mu(q) - \mu(q) \cdot 0.8 \tag{2.6}$$

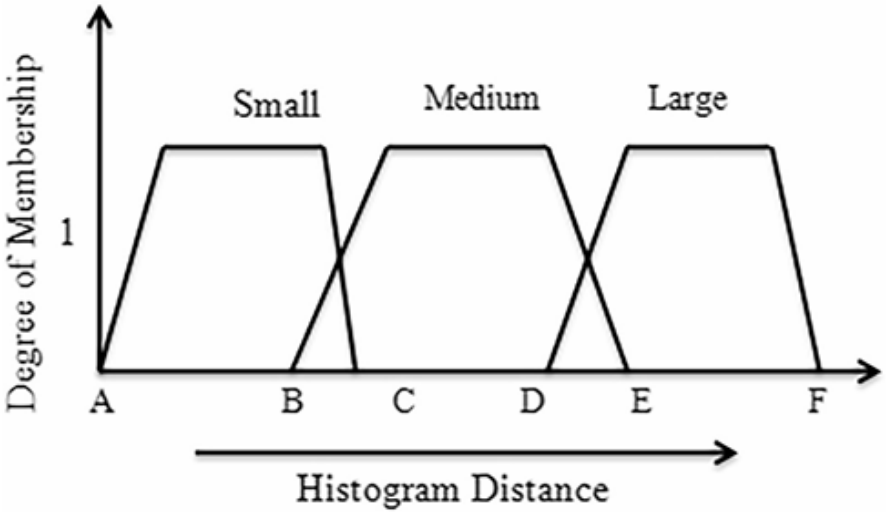


Figure 2.4: Histogram distance for selection of keyframes

A keyframe is one in which the histogram distance between succeeding frames is minimal or big. Then we extract the deep features from selected frames.[5]

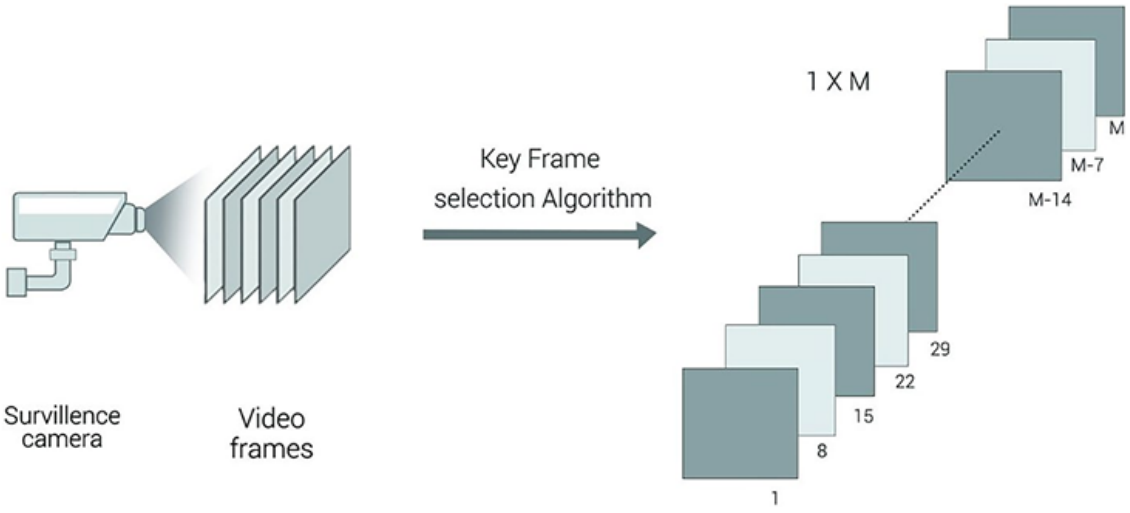


Figure 2.5: Keyframes selection in video

2.2.2 Feature Extraction

When selecting keyframes, the previously trained model, InceptionV3, will be used to extract deep spatial features. To reiterate, the mentioned model has yielded valid results throughout challenging visual tasks. The system performs better from the use of transfer learning, as it is able to apply pre-trained CNN weights to identify relevant features without additional significant retraining. This method produces a vector that comprises a series of relevant spatial features relevant to activity recognition.[6]

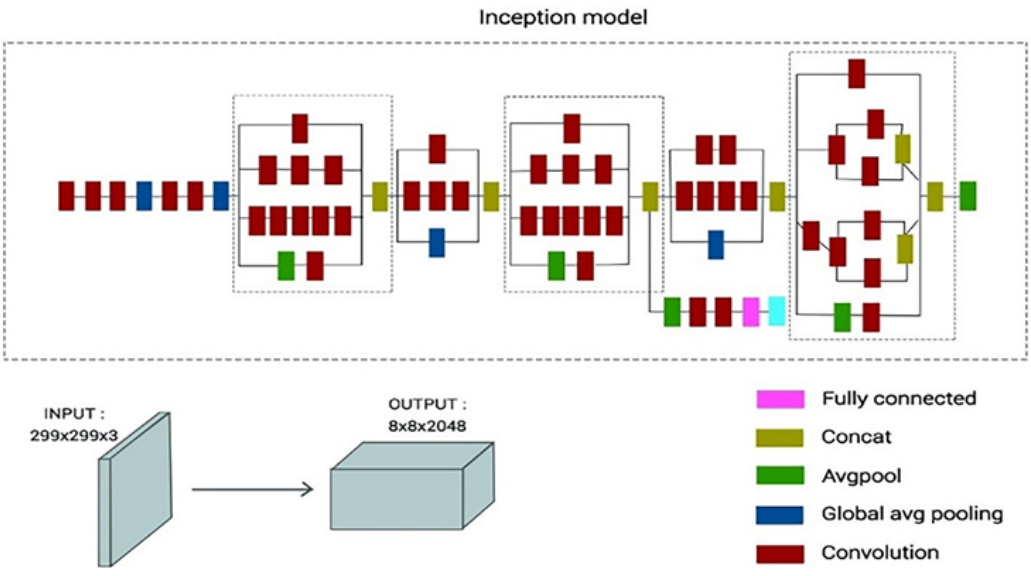


Figure 2.6: InceptionV3 as a feature extractor

2.2.3 Anomaly Detection

The model was also trained on temporal sequences with the extracted features, using an RNN that incorporated the LSTM layers. The architecture includes an extra stacked layer of LSTMs, dense layers, and dropout regularization as a means to mitigate overfitting. The UCF-50 dataset contains well curated videos of human actions, in which the model achieved a high accuracy of 95.04, showing just how robust deep learning can be on structured data. However, on the more challenging UCF-Crime dataset of untrimmed real-world videos, it dropped to 49.04, still below the Siamese and CNN-LSTM models. It does a great job to learn temporal dependencies and detect anomalies based on the activities, where an anomaly is detected by softMax in the output layer.

2.2.4 Datasets and Preprocessing

For the purpose of testing the proposed model, we shall investigate the UCF-50 and UCF-Crime datasets. The UCF-50 dataset consists of 50 class categories of human motion including playing musical instruments and horseback riding, all drawn from videos on YouTube. The videos themselves were cut short in duration and grouped so that solutions in controlled, well-lit situations were highly easier to pursue. By contrast, UCF-Crime contains a set of untrimmed surveillance videos capturing real-world anomalies such as assault, arson, abuse, etc. These videos include dense backgrounds, random camera motion, and varying object sizes, thereby rendering the detection system a much more challenging dataset to use. During the preprocessing stage, the videos were first trimmed into frames and then a fuzzy logic-based keyframe selection algorithm was applied. In essence, the fuzzy logic algorithm operated with histogram distances in the CIELab color space to filter frames containing high informational content. In this way, the procedure preselected the most informative frames while suppressing all the redundant and less informative data for the input to the model. The keyframes were then fed into a pre-trained InceptionV3 model that effectively extracted highly informative spatial features. By eliminating useless frames and emphasizing meaningful content, the overall quality of input data for the continuation of the anomaly detection is improved.[7]

2.2.5 Results

The results obtained from the experiments prove the effectiveness of the model proposed on structured and real-world datasets. The accuracy on the UCF-50 dataset, which consists reasonably quantified human actions, was found to be as high as 95.04. Such a performance emphasizes the deep learning capabilities to handle structure data efficiently. Further challenge was posed by untrimmed videos within chaotic real-life scenes from the UCF-Crime dataset. When faced with these data types, our model plummeting at 49.04 was lower than the results measured for Siamese and CNN-LSTM models. Differently from such investigation, it would be possible to combine fuzzy logic keyframe selection for spatial feature extraction by InceptionV3 with sequential learning through LSTM. Its ability to generalize rightly was quite evident in its ability to define temporal patterns and was shown in its much better results concerning all varying datasets. Such results not only proved the efficacy of the proposed mechanism but also will be a very solid ground for using that mechanism even more explicitly in real-world contexts-it is especially pertinent in an area like public safety, where anomaly detection is of tremendous value.[8]

2.3 Real-Time Threat Detection and Alert Systems

ADAG stands for Activity Detector and Alert Generator which is an integral system for retail safety and security. It is completely automatic human-independent system.This has a facility of CCTV cameras for detection of suspicious activity. Modern security systems like Electronic Article Surveillance (EAS) use physical tagged objects which can easily be thieved from retail; yet these can always be subduable when acquaintance with the tags is formed and deactivation or removal is done. ADAG says that live video feeds from the CCTV cameras will show signs of behaviors like stealing in typical shoplifts, robbery actions, as well as beating up actions against a woman. Object detection is done through a Convolution Neural Network. Transfer learning is done through pre-trained models like those under ImageNet and classifiers that classify suspicious actions, when detected, alert the store keeper according to anything that matched highlighted impulses.

2.3.1 Methodology

Surveillance areas like retail sites, public places, or private premises may capture closed-circuit television (CCTV) videos. The package raw video footage for analysis will be sliced into frames for further analysis at some future date. These frames are to be preprocessed probably since there is one other common problem with surveillance footage, namely resolution noise, and another dependent on lighting conditions. Resize frames fixing to a universal dimension, usually 224 x 224 pixels; normalize pixel values to a zero-mean; some data augmentation, such as antecedent random rotation, flipping, brightness adjustment: went ahead unrestrained processing as the model was being restrained by bringing in variability with the current data that would cause it to generalize well in reality. Transfer learning in this system makes bold with the use of pre-trained models, as are those pre-trained on ImageNet datasets, consisting mostly of millions of labeled images. These are fantastic for feature extraction, thereby freeing one from having to develop a CCTV model from scratch, as the pre-trained weights can be efficiently fine-tuned to detect activities such as theft, unauthorized access, or rampancy. Hence, the computation power and data are lessened with this, but it comes at the cost of accuracy and efficiency. Fig. 2.7: Fine Tuning ResNet 50 In the heart rests the Convolutional Neural Network. This is very appropriate, as all the networks mentioned use, in general, a hierarchical approach to generalize and extract features or properties from data presented in the form of images or videos. There are layers that will perceive a mixture of modalities and shapes, such as convolutional layers, which will usually contain edge and texture detectors; pooling layers, which actually downsample the feature perspective; and finally, the fully connected last layers that present the class of the features associated with what has been learned. The fine-tuned CNN becomes the very identification model of those objects, actions, and behaviors deemed to be suspicious. What mainly supports the procedure is amply processing in real-time video streaming probing, where each frame is processed separately. It makes use of a trained CNN in the detection of abnormalities or suspicious activities in real-time, and therefore security personnel reaches conclusions in a timely manner, concerning the incidents occurring around them. In case of any threat, the system activates alarm updates which are delivered to stakeholders via emails, SMS, or the whole property under integrated monitoring. Thus, this system generates automatic alerts ensuring that

whatever incident takes place is dealt with in time, therefore minimizing the response time, and maximizing the efficiency of security environments for everybody. Testing is rather elaborate and combines publicly accessible datasets with real-world surveillance footage to show performance vol

It is mandatory to collect video data from installation sites such as retail sites, public areas, or private premises. In preparation for analysis, the raw 19 video footage is sliced into frames. The main goal of their preprocessing comes in because in most situations, the alteration is due to resolution-noise conditions and lighting conditions. Frames with a resolution of 224x224 pixel normalize the pixel values uniformly, while data augmentation is applied through techniques like random rotation, flipping, and brightness adjustment. The whole process thereby endowed the model with some measure of robustness, allowing it to generalize well in reality by injecting diversity into the data with the aid of randomization. The core importance of transfer learning within this system is that the input takes pre-existing models, such as models that were pretrained on the ImageNet dataset. Such models are unto themselves the training of millions of labeled images. On top of that, the application of feature extraction is a very good opportunity for saving time and money in constructing his/her surveillance model and fine-tuning pre-trained weights simply for detecting activities like theft, unauthorized access, or aggression. This drastically reduces computational power and data usage but compromises accuracy and hence efficiency. Figure 2.7: Fine Tuning ResNet 50 It is the Conv Net which keeps things at the centre of the system. Of such interesting, all such networks operate on opponent hierarchical generalizations to extract features from data presented in the form of images and videos. Out of these, the competing layers ones put into play numerous forms and modes-for instance, convolutional layers usually come in to play edge and texture detectors, pooling layers before play downsizing-the-space subsampling of the reduced feature perspectives, and then the fully connected last layers that ultimately deliver all that has been learnt into a class as features. In this instance, the fine-tuned CNN becomes the conventional identifying model of those objects, actions, and behaviors that they consider to be suspicious. One of the main features of the approach is real-time processing of streaming video feeds, processed independently frame by frame. It employs the trained CNN for detecting anomalies or suspicious activities in real time; thus, security personnel will be able to make timely informed findings about the happenings. And when there

is a possible threat, an alarm is triggered and sent to the relevant stakeholders either via e-mail, SMS, or integrated monitoring to the whole site. Thus far, the automatically alerting system ensures that incidents shall always be attended to on time, therefore minimizing time in responding and maximizing efficiency with security environments for all. The performance v-testing includes extensive testing with a mix of publicly available datasets and real-world surveillance footage.

Such video data should be collected from CCTV cameras installed at places of surveillance, such as retail locations, public spaces, or private premises. The raw video footage from 19 cameras are sliced into individual frames, intended for future analytical assessment. The important preprocessing of these frames makes them habitable by the general problems of surveillance footage resolution noise and those relating to lighting conditions. There is resizing of frames to an anchor size-mostly 224x224 pixels; normalizing pixel values to that degree unattractiveness; additional data augmentation random flipping, brightness modification, random rotation- went above in underprocessing as restriction model was attained in creating variability among current data that effects generalizations well in reality. This system has some meaning attached to transfer learning whose pre-trained models are employed as in models already trained pre on the ImageNet dataset. Such training models train with millions of labeled images. Besides, feature extracts are quite well and cost/time effective building own surveillance model and only fine-tuning of pre-trained weights for detection activities like thieving, unauthorized access, or aggression. The computational power and data are reduced, but the accuracy and efficiency are compromised. Figure 2.7: Fine Tuning ResNet 50 At the very heart of the system lies a Conv Net. All these networks function on rival hierarchical generalizations-among such interest-in order to extract features from data, which is presented in the form of images and videos. These include layers that decode various forms and means, including convolutional layers, which will include edge and texture detectors, pooling layers that would perform subsampling-the-space downsizing of reduced feature perspectives, and then last layers which are fully connected that finally present what has been learnt as features into a class. The fine-tuned CNN then becomes a typical identifying model for those objects, actions, and behaviors they identify as suspicious. Among the many focuses of this approach is real-time processing for streaming video feeds, where each frame is treated separately. It uses a trained CNN in detecting an anomaly or suspicious behavior in real time, thus

the timely informed conclusions by security personnel trained. Moreover, whenever a risk is sensed, the system activates the alarm and sends it to the respective stakeholders by email, SMS, or a full facility connected to the integrated monitoring. Therefore, this system alerts automatically to almost all incidents in time, which means shortened time to response and maximization of the effectiveness for security environments for all. Testing thus becomes very thorough with combinations of public datasets and real-world surveillance footage for performance vol

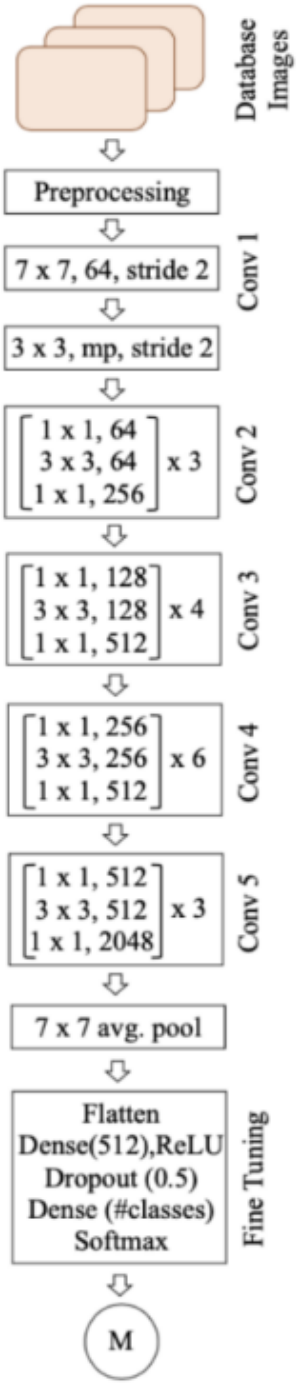


Figure 2.7: Fine Tuning of ResNet50

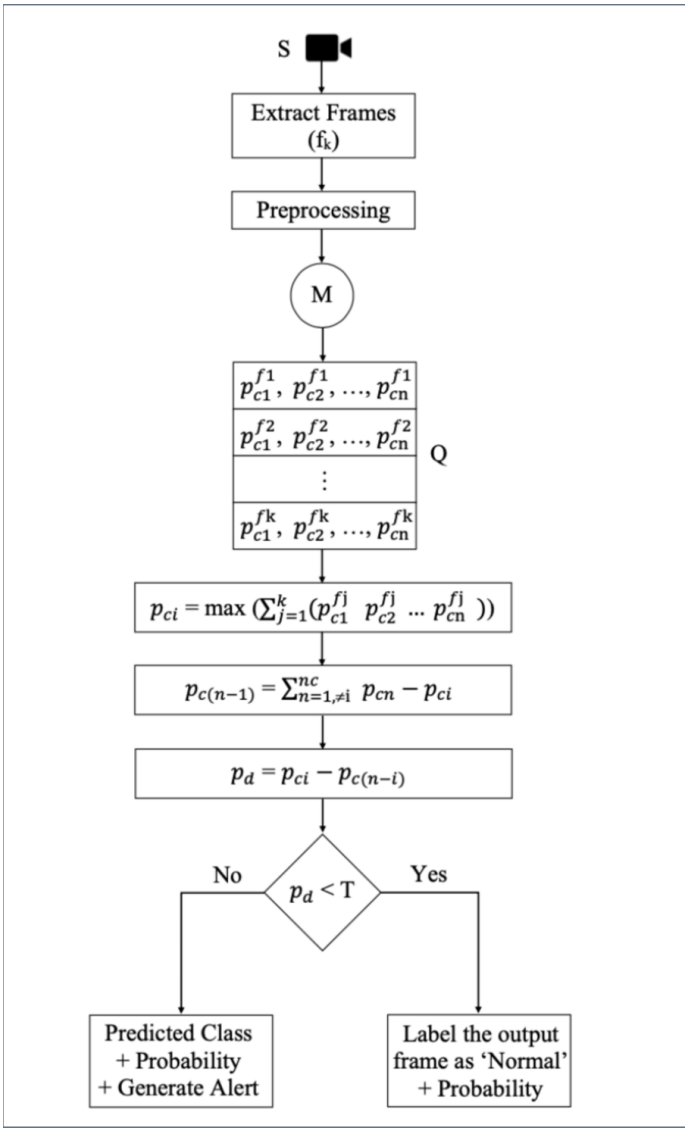


Figure 2.8: flowchart of resnet model

The method starts out by collecting video data from CCTV installed in all surveillance locations like retail sites, public places, or even private premises. The raw 19 video footage has been sliced as particular frames for analytical assessment in the future. Important preprocessing of these frames makes them much more habitable by the generic problems of surveillance footage resolution noise and those related to lighting. There is resizing of frames to an anchoring size, usually 224x224 pixels; normalizing pixel values to that degree unattractiveness; some data augmentation, using, for instance, random rotation, flipping, brightness adjustment; has gone beyond that scale of underprocessing because optimizing model-learned variability was induced by the current data that affected gener-

alization well in real life. Transfer learning carries a greater meaning in this system since pre-trained models will be used, similar to those pre-trained in the ImageNet dataset. Training such models requires millions of labeled images for training. Besides, feature extraction is well, thus saving time and money to develop his/her surveillance model and fine-tune pre-trained weights for activity detection of theft, unauthorized access, or aggressiveness. This reduces greatly computational power and data but compromises with accuracy and efficiency. ResNet 50 At the very heart of this system is a Conv Net. Of such interest, all such networks operate on opponent hierarchical generalizations to an extent. Features from the data are presented in the form of images and videos. These comprise layers that decipher numerous and ways, such as convolutional layers, which will usually include edge and texture detectors, pooling layers that would perform downsizing-the-space subsampling of reduced feature 20 perspectives, and then the fully connected last layers which finally present what has been learnt into a class as features. The fine-tuned CNN then becomes regarded as a typical identifying model of those objects, actions, and behaviors that they consider suspicious. Among this approach's emphases is real-time processing for streaming video feeds, where each frame is processed independently. It uses a trained CNN to detect anomalies or suspicious activity in real-time, so security professionals can make timely informed conclusions about the events. And whenever there is a possible threat, the system activates the alarm and sends it to relevant stakeholders either through email, SMS, or an entire facility with integrated monitoring. Thus, any almost incidents are notified automatically in real time, thus shortening time taken before response and maximizing the effectiveness for security environments for all. Testing thus is very rigorous with combinations of publicly available datasets and real-world footage surveillance,.

2.3.2 Results

Any suspicious activity would then produce positive results. The CNN based model tested high accuracy while classifying an act into Normal, Robbery, Shoplifting, Arson, Assault etc. Alerts will be generated for any of these classifications and the system has achieved threshold reduction of unnecessary notification as alerts are generated when behaviour changes only. Real-time analysis would produce nearly instantaneous feedback, which is crucial in the context of retail security. This is, therefore, fast real-time processing

that guarantees counter actions being quickly mobilized-either by security personnel or store owners-on awareness of the perceived threats. Regarding practicality, it is easy to incorporate real-time processing into an already-existing surveillance system, which would further enhance its efficiency and proactive approach to handling security matters.[9]

2.4 Expression detection

Emotion detection, one of the most consequential elements in human-computer interactions, is defined as a computerized mechanism' ability to deduce human emotional states through interpretation of facial expressions. It finds applications in various fields such as security systems, e-learning, customer experience management, and mental health. Traditional emotion detection methods relied mostly on handcrafted features and separate phases for feature extraction and classification, which may not be optimal for accuracy and generalization. Emotion detection has been revolutionized by the deep learning approach, mainly through CNNs, which enable the direct learning of features from data. MobileNetV3, probably the most lightweight but very efficient architecture in CNN, effectively performs in real-time emotion detection, thanks to its well-crafted balance between performance and efficiency in computation. The enhancement of MobileNetV3 incorporates various superior modules such as squeeze-and-excitation blocks, depthwise separable convolutions, and hard-swish activation functions. The former is extremely efficient on computations while giving stern feature extraction capabilities.Inmobile and embedded applications where computational resources are limited yet high accuracy is indispensable.[10]

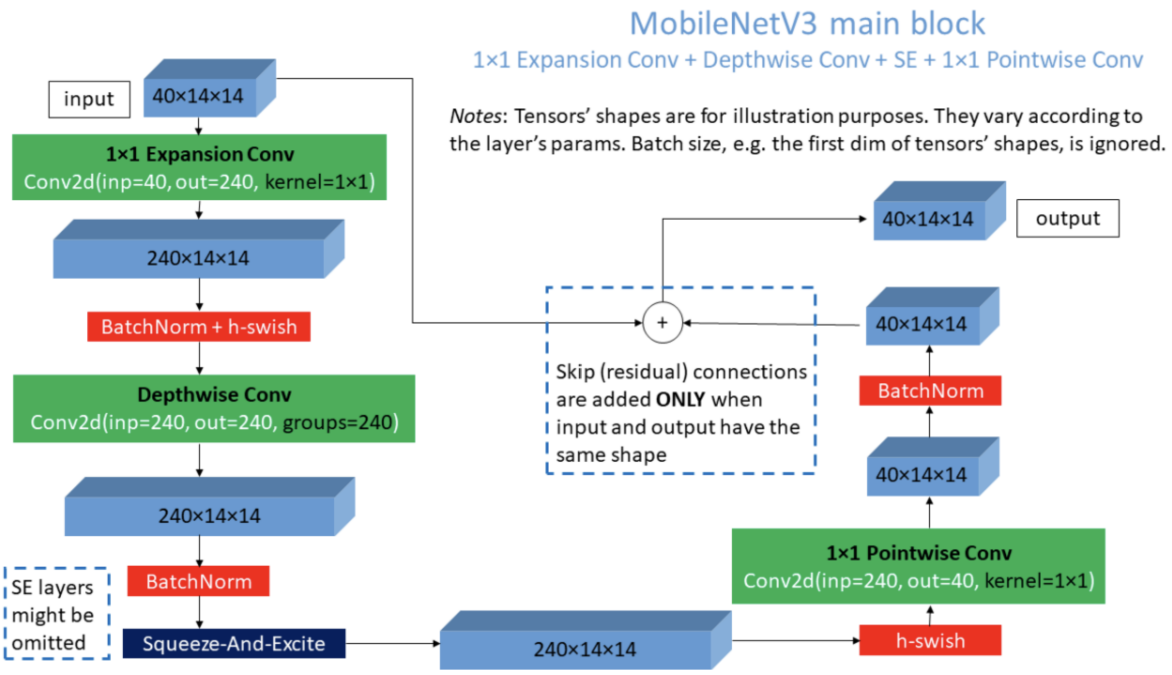


Figure 2.9: Structure of MobileNetV3

The above architecture is an extension of MobileNetV1 and V2 with critical optimization of depthwise separable convolutions, activations, and channel-wise recalibrations. In the center concept of MobileNetV3, it devotes itself to switching normal convolutions for depthwise separable convolutions, which decouple the feature extraction process into a depthwise convolution for spatial information and a pointwise convolution to gather channel-wise information. This maximizes the reduced computation without sacrificing accuracy. The other indispensable ingredient is the Squeeze-and-Excitation (SE) module, which establishes channel dependency by Scaling features dynamically. In this setting, features that carry more information are amplified, while less relevant information is suppressed, thereby helping the optimization of information processing within the network.[11]

The MobileNetV3 Hard-Swish is an activation function that is an efficient approximation of Swish in terms of its fast computation and good memory efficiency to apply within resource constraint environments. With this, smooth gradient flow is guaranteed, at low computational cost and therefore ideal for mobile implementation. The main building blocks of MobileNetV2 consist of inverted residuals with linear bottlenecks. Input channels are expanded via pointwise convolution, followed by depthwise convolution for spatial feature extraction, and finally, a pointwise convolution that compresses down to the orig-

inal size. These blocks adopt skip connections to carry information and improve gradient flow, predominantly when input and output dimensions match. MobileNetV3 has two flavors, namely MobileNetV3-Small for resource-constrained devices and MobileNetV3-Large for applications that require higher accuracy. MobileNetV3-Small has fewer parameters and uses SE modules and Hard-Swish activations selectively, hence being more suitable for battery-powered devices such as IoT ones. MobileNetV3-Large has more layers and places SE modules and Hard-Swish activations at multiple levels to enhance its representational power so that it does well at inputs with complexity without serious loss of efficiency. The last layers utilize global average pooling to reduce the spatial dimension of the feature maps, followed by a fully connected layer with dropout for regularization. Finally, a softmax layer gives probabilistic outputs such that the model can be used on a classification task.[12] Other than that, partial NAS optimization is another innovation put in place. With NAS, evaluation of architectural choices is automated, thus finding an optimal balance of parameters for each layer. The search process therefore assists in configuring, say, number of channels, placement of SE modules, and activation functions for the best compromise in efficiency and accuracy. Hence, MobileNetV3 yields a compact architecture, strong enough for embedded and real-time mobile applications with efficient yet high-performing processing.[13] It can also be fine-tuned or transferred to related tasks with ease. It recognizes emotions, thus making it ready for emotion-aware systems implementation.

2.4.1 Dilated convolution

Dilated convolution (which is also called atrous convolution) is a wonderful technique implemented in CNNs in order to expand the receptive field, which is the region of the input image that each filter can "see," without adding more parameters or increasing computations. In standard convolution, a filter moves over the image, pixel by pixel, gathering local features within a set window size. If, however, we would like to capture larger-scale features, we would normally have to stack several layers of these standard convolutions, getting heavy computationally and with more parameters to deal with. This is where dilated convolution steps in. It accomplishes this by introducing spaces, or "holes," between the filter elements, thus enabling greater spatial coverage over the image without increasing the number of parameters.

Here we now consider the dilation rate, a crucial aspect in informing us about the spacings of filter elements. At1, we say that each filter element is sitting next to each other, just like in classical convolution. When we speak about a dilation rate of 2, that just means that there is now a one-pixel gap between each of them. As the dilation rate increases more and more, the receptive field exponentially increases. Thus, through this property, those convolution layers can now obtain a more global context instead of having to work with a larger filter.

This is especially helpful for applications that have a significant need for understanding the context over a larger area, such as semantic segmentation where the identification of object boundaries at high spatial resolution is extremely important. Rather, dilated convolution captures the information from a wider context without adding more layers- it increases the receptive field but retains the fine details without downsampling feature maps.[14]

In particular, this is crucial for applications demanding a more in-depth understanding of the context over a larger area such as in semantic segmentation, where boundary identification while keeping spatial resolution is critical. Dilated convolution draws upon much wider contexts without adding more layers by simply expanding the receptive field but keeps details intact by downsampling feature maps.

It is especially effective in applications requiring greater understanding of the context over more extended areas, as in semantic segmentation, where identifying boundaries of objects while maintaining critical spatial resolution is of utmost importance.[15] Unlike other operations that increase the number of layers, dilated convolution brings in information from a wider context, increases receptive field size without down sampling the feature maps, and keeps details intact.

This is especially good for applications where more profound comprehension of the context is needed over a larger area, as with semantic segmentation, where identifying object boundary at very large spatial resolution is extremely important. Dilated convolution, rather, encroaches into much wider contexts without attaching greater number of layers simply by expanding the area of receptive fields but maintains whole details intact by not downsampling feature maps.

This is very much applicable for applications that need much deeper understanding of the context over a wider area, such as in semantic segmentation for which the object

boundaries need to be detected while keeping the higher spatial resolution. Whereas, dilated convolution draws from a wider context without using more layers by just enlarging the receptive field, but it keeps intact the high detail by not downsampling feature maps.

At the heart of it, this method is computationally efficient and versatile-allowing CNNs to capture a wide variety of features across different scales while keeping memory and computational costs to a minimum. The spacing of those zeros in the filter is governed by the dilation rate, usually denoted by d. Dilated convolution, or atrous convolution, is a very clever strategy that is used inside CNNs to increase the receptive field-the area that each filter can "see" of the input image-without adding any additional parameters or increasing computation overhead.

The pixel-by-pixel scanning of the image to search in a small surrounding area for features is entirely performed by the convolutional filters. More than one convolutional layer may be used to acquire more complex features, but this really ramps up the computational load and the number of parameters to manage. Here comes the dilated convolution. It "dilates" the filter by inducing spaces or "holes" between the filter elements. Thus, the same number of parameters covers a larger area of the image now. The dilation rate is important here; it states how filter elements are spaced. Thus, a dilation rate of 1 leaves filter elements right next to each other, just like in classical convolution. A dilation rate of 2 yields a one-pixel gap between all filter elements. As the dilation increases, the receptive field also gets bigger, which allows each convolutional layer to gather more information from the outside while not needing larger filters. This is the quality that makes dilated convolutions particularly convenient for context-sensitive tasks-perhaps the most striking example being semantic segmentation, where object boundaries must be accurately defined but the spatial resolution should at least be kept high. The widening of the receptive field by dilated convolution draws information from a broad context with fewer layers while preserving the high definition.[16]

Human beings are brought into conformity to a proposal that applies nothing more than convolutional filtering methods to an image-pixel observation, as much as the surrounding area looks for features. One could even include more than one convolution layer to encapsulate more complex features, but the same truly turns on extra computation load and the need to manage more parameters thrown into the game. This is where dilated convolution comes in: It stretches a filter by adding spaces or "holes" between

its elements, so that about the same number of parameters cover a much larger area of the image now. Dilation rate could be the utmost important here in deciding how widely spaced filter elements are. Thus, a dilation rate of 1 would like filter elements right next to one another, just as in normal convolution. A dilation rate of 2 would give every element a one-pixel gap. As dilation increases, so will the receptive field, thus allowing each convolution-layer to capture more data from the outside without the need for larger filters. This is precisely why dilated convolutions are so effective for tasks that need to understand a context over larger areas-a perfect example would be semantic segmentation, where accurate identification of object boundaries while maintaining high spatial resolution is imperative. The dilated convolution thus widens the receptive field over which information can be pulled with fewer layers but remains intact in detail because it does not down-sample the feature maps. They have been trained on data until October 2023.

Dilated convolution, therefore, can be summarized as one of the efficient and flexible convolution techniques. It allows CNNs to grab larger ranges of features across different scales whilst saving memory and computational costs. The spacing of the zeros in the filter is determined by a parameter known as the dilation rate, formally denoted as d.

Consider the diagram given below:

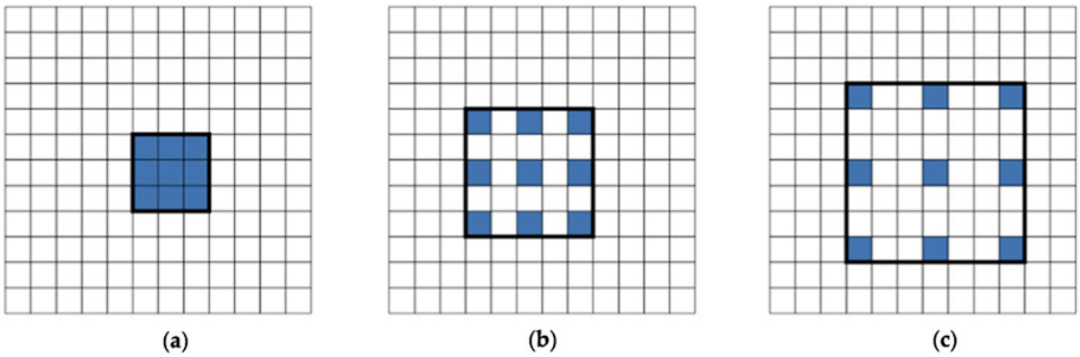


Figure 2.10: Illustration of Dilated Convolution with Different Dilation Rates

In fig (a), A dilation rate of $d = 1$ corresponds to a standard convolution.IN figure (b), A dilation rate of $d = 2$ expands the receptive field to an effective filter size of 5×5.In figure (c), A dilation rate of $d = 3$ expands the receptive field to an effective filter size of 7×7. The addition of dilation in convolution is an attempt by the model to broaden its

receptive fields: thus there can be more general capture of context for applications where global context and fine-grained details are key components, such as emotion detection in facial expressions. For emotion detection tasks, it is interesting to note that the subtle changes in features (such as slight movements around the eyes or mouth) may mean the difference between the correct and incorrect classification of an emotion. The traditional ways of increasing the receptive field have remained constrained: that is either to use larger filters or the pooling layers.[17] Larger filters therefore cover a larger area and increase the number of parameters incurred by training on the model, which leads to larger memory and increased computation costs. The pooling layers do increase the receptive field through downsampling of the input, but sometimes they destroy the required spatial information, rendering this method ineffective when requiring accuracy, as this would blur or lose even minor insignificant details. Dilated convolution overcomes all these disadvantages by providing a large receptive field while maintaining image resolution and computational efficiency. This means that the dilation convolution spreads out the filter's effective region by inserting "holes" between adjacent components to capture both local and global patterns while preserving the spatial resolution of the entire feature map. The fine-grained details are not lost through this process, allowing the network to identify the slightest variations in expressions while comprehending the larger context. This essentially makes dilated convolution an enabler toward deeper contextual understanding of facial expressions, hence improving the emotion detection capability without the associated cost of training extra parameters and loss of detail along the process. Therefore, it forms a strong candidate application requiring performance on fine details while also providing information from the broader context.

The equation of output feature map is shown below:

$$z(\mathrm{p}, q) = \sum_{h,j} f\big(p + d * h, q + d * j\big) * g\big(h, j\big)$$

The parameters p, q, h, j, f, g, and d in the equations indicate the following: The parameter p stands for horizontal coordinate in the feature map; the parameter q stands for vertical coordinate in the feature map; the parameter h stands for horizontal coordinate in the convolution kernel; the parameter j stands for vertical coordinate in the convolution

kernel; the parameter f stands for the value at a feature map; the parameter g represents the value at a convolution kernel; and d is the dilation rate. Within the framework of the advanced MobileNetV3 model, the inclusion of dilated convolution facilitates multiscale feature extraction such that the model works on facial expressions revealing emotions, such as distress or anger. Because dilated convolution derives both global and local features, it can also detect subtle emotions on facial dynamics with varying lighting, thus justifying its application for real-time jobs in public safety and surveillance.

2.4.2 SimAM

In linguistics, by definition, attention methods are devices that help a model being "focused" or "concerned" towards certain portions of the input wherein such portions may be really important for the task. An example of such model functioning in a convolutional neural network is the squeeze-and-excitation network (senet), which helps to enable and improve the feature representation power. SENet effects the squeeze-and-excitation mechanism by converting the spatial information of each channel into a channel descriptor by applying global average pooling globally across the feature map. This operation is termed as the "squeeze". Fully connected layers are then employed to determine the importance of each channel based on a given task; this is referred to as "excitation". More important channels are given higher weights while less important ones are downscaled. With this kind of selective emphasis, one can increase the chances of finding something interesting about the features, thus improving performance in classification tasks for example. The flip side of senet is that these fully connected layers introduce more parameters and make the layers deeper, thus increasing resource consumption. This in fact increases its resource utilization, which has imposed a limitation on the use of senet across a variety of applications such as mobile or embedded systems, where efficiency is the primary consideration. The Simple Attention Mechanism (SimAM) exploits channel and spatial attention, thus overcoming SENet limitation while remaining lightweight. The SENet function concentrated exclusively on channels, whereas SimAM factors in not only channels but also those areas within a channel, which is now able to attend to selected parts of the feature maps. Therefore, SimAM sets weights for all the elements of feature maps according to their importance in the computation without adding extra parameters and an overhead layer to the model. While attention, like many techniques, is known to redefine by learning more

parameters per model, SimAM, by energy-based definition, considers pixel contributions solely according to their values or pixel importance to the whole image. This also helps to enhance the ability of SimAM to understand and ascertain which among the channels shall be more useful for the task at hand and which channel is actually useful.

So attention mechanisms have some use for incorporating fine-grain facial cues into evoked emotions like apprehension, discomfort, or delight when undertakings such as emotion detection are concerned. The model is therefore improved by SimAM, focusing on a very particular area of the face, namely around the mouth and eyes-these are important attributes to be looked into for emotion detection. SimAM is also efficient in real-field applications that actually require an attention mechanism less calibrated by memory and computation, since SimAM, unlike SENet, does not contribute anything to the computationally heavy schemes of real-time systems. Thus it becomes pretty handy for low-power or mobile devices to be used in real-time scenarios, as it balances the demands for attention and the constraints placed by limited memory and processing capacity.

A central aspect of the functioning of SimAM is the determination of neuron energy, which symbolizes the significance of each neuron within the feature map. Neuron energy is evaluated concerning the output of a neuron augmented or attenuated by some weight factoring in the output from the neighboring neurons. In this, the energy can be expressed as:

$$\mathrm{e}_t(\omega_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2$$

- e_t is the energy of neuron at position t.
- y_t is the output of neuron at t.
- $\hat{y}_t$ is the predicted value of neuron at t.
- M is the total number of nearby neurons taken into account.
- y_i and $\hat{y}_i$ are the actual and predicted values of other neurons nearby.

binary labels are employed for the sake of simplicity and a regularization term is added. The ultimate energy function is defined using the following formula:

$$e_t(\omega_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(\omega_t x_i + b_t))^2 + (1-(\omega_t t + b_t))^2 + \lambda\omega_t^2$$

Taking the derivative with respect to w_t and b_t respectively, we get:

$$\omega_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda}$$

$$b_t = -\frac{1}{2}(t+\mu_t)\omega_t$$

where $\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i$ and $\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)\omega_t$ are mean and variance calculated over all neurons except t in that channel. It can significantly reduce the computation costs to avoid iteratively calculating μ and σ for each position. Therefore, the minimum energy can be obtained by the following formula:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$. Finally, according to the definition of the attention mechanism, the feature enhancement can be performed using the learned attention coefficients. The entire process is shown in the following formula:

$$\tilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X$$

Comparably enough, SimAM's calculation using three-dimensional weights is simply not too complicated; it is even possible to say that the weight of an attention module becomes very lightweight because—unlike the two-step method of CBAM that requires so much computation-to-the-two-step method of-compare with- SimAM.

This table provides comparative study of the different attention modules.

Attention Modules	Type	Operators	Parameter
SE	Channel attention	GAP, FC, ReLU	$2C^2/r$
CBAM	Channel & spatial attention	GAP, GMP, FC, ReLU, CAP, CMP, BN, C2D	$2C^2/r + 2k^2$
SimAM	Channel & spatial attention	CAP,/,⊙,+	0

In the table the abbreviation CAP represents channel average pooling, GMP refers to spatial max pooling, CMP stands for channel max pooling, GSP indicates spatial dimension standard deviation, C2D is for 2D convolution, FC represents fully connected network, and BN refers to batch normalization. k and r indicate the number of convolutional filter and reduction ratio respectively. C refers to the current feature channels.

From the table above, it can be seen that the parameter count for SimAM attention module clearly outshines others in the list. Moreover, in comparison with CBAM attention mechanism, SimAM is a simple parameter-free attention mechanism, and so it does not add extra parameters to the network.

2.4.3 Group Normalization

Deep neural networks are supposed to manage within the neurons activation patterns with care in order to provide an adequate training. This technique, which comes under the nomenclature of BN, has been specifically designed both in the direction of stabilizing the training process and speeding it up by removing the so-called "internal covariate shift," where the inputs distribution of each layer are changed in the course of training. In Batching Normailzation, the input to each layer normalized according to batch statistics (mean and standard deviation) for a batch of samples. The dataflow is normalized used input so that input values are maintained from one pass to another to speed up learning. In fact, BN further accelerates learning by even allowing larger learning rates and making the process less reliant on initializations with large weights. This much benefit for BN brings with it a major constraint: its dependency on larger batch sizes for the accurate estimation of mean and variance. Sometimes limited memory size as that in mobile phones or low power devices asks for a smaller batch size. In best cases of efficacy, this can even reduce effectiveness on the backpropagation. Model accuracy and stability could, thus, be affected, as estimates would degrade based on statistical precision.

The Group Normalization is a possible solution for the Batch Normalization when it happens that the batch size is small. It deviates from BN by partitioning the channels of each layer into small groups for computing mean and variance for each of these groups independently instead of across the entire batch. Normalization here does not depend on the batch size, as it normalizes within groups of channels rather than across the batch.

That is why GN could be said to be more stable and reliable in applications that have no memory for the use of large batches. The channel grouping of GN adds further consistency in performance and accuracy in that it largely diminishes dependency on batch statistics.

Hence, group normalization is a better alternative for batch normalization.

2.5 Decision tree

These are tree-based supervised models for the tasks of classification and regression. In the above, an internal node of the tree corresponds to a specific feature or attribute of the sample; each branch, a decision rule; and each leaf, an outcome at that node. Basically, it splits the input dataset into many smaller subsets based on the testing of attribute values, and finally, builds a tree where the decision process occurs.

Thus, it is being widely used due to easy and interpretable methods in coding categorical and continue data. The tree structures can be obtained by having every attribute node maximized for classification using the various measures of percentage, like Information Gain, Gini Index, etc. But unfortunately, it overfits the data: pruning techniques help to avoid the issue.

2.5.1 Methodology

Preprocessing of data has been considered as one of the most important steps in building a Decision Tree model. It provides clean, structured input data ready for analysis. This step would involve inputting missing values either by imputing them with mean and median for numerical data and mode in case of categorical data or perhaps extracting the records due to its insignificance. Normalization or standardization step follows, especially when performing on continuous attributes just to avoid the bias from scales. The next one involves splitting the dataset into training and test (80-20 or 70-30) for modeling and validating the model's performance in revealing patterns at the test end. This step is flexible enough to handle categorical and numerical attributes but essential for formatting data in the same way. At this stage, the feature selection may be done as some features that are not related to the analysis can be removed to save computation and avoid intro-

ducing noise into the model. Ensabling the Efficient, easy analysis of data is dependent on Preprocessing, when it comes to Decision Trees.

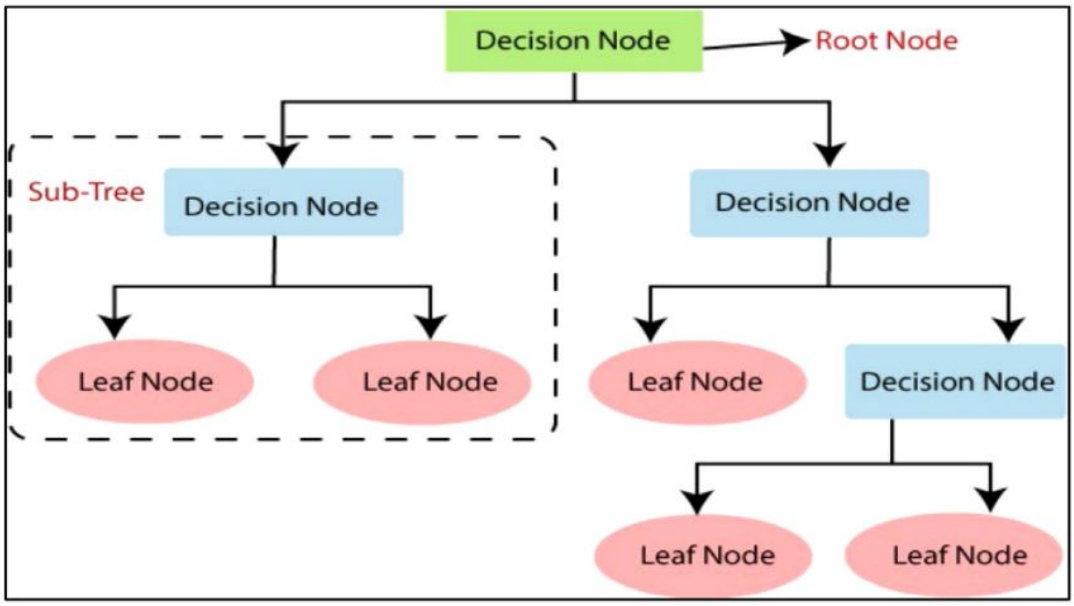


Figure 2.11: Decision Tree

[18]

Data preprocessing forms the most important step in the development of Decision Tree models since it guarantees that input data is clean, structured, and ready for analysis. This step involves the coding of missing values, whether imputing by mean and median as dictated (if numerical), or mode (if categorical) or outright deleting the record when the number of such records is less than significant. Thereafter comes normalizing or standardizing in particular with respect to continuous attributes to prevent any bias that may arise due to differences in scale. The next step comprises the splits between training and testing of the datasets into different halves-usually 80-20 or 70-30 respectively-to trap patterns into the model and validate its performance. This step, which accommodates both categorical and numerical attributes, provides that much flexibility; however, it is pertinent to ensure that entries on the attribute have a common ground. One can also have the feature selection done here by eliminating poorly associated attributes thus minimizing computation and model noise. The preprocessing phase shall ensure that upon consumption, the Decision Tree is able to analyze the data much easier and faster. The very basis of decision tree construction is to select the most informative attribute at

each node using attribute selection methods, one of them being Information Gain, which relies on the concept of entropy, which measures the level of uncertainty reduced by the splitting of a dataset using an attribute. Should a high discrimination power for the variable "presence of disease" be attributed to the attribute "age group," it would now be a potential candidate for the root or decision node. The second most widely used measure is the Gini Index, which, like Information Gain, emphasizes impurity or inequality in the dataset and assesses it before and after splitting. The two measures are in some sort of trade-off between completely special interest in the subsets and not forming trees that are too abstract in structure. This will continue to be split recursively such that at any given time an attribute with a higher score according to these measures will be chosen. This systematic way of attribute selection renders the encoded decision tree to be both accurate and interpretable since every split represents a meaningful decision rule.

The procedure for constructing a decision tree is as follows: The root node is chosen according to the attribute with the highest Information Gain or the lowest Gini Index. The data set is partitioned according to the selected attribute into smaller subsets. The process is then repeated recursively for each subset creating child nodes until the leaf node condition is satisfied or a stopping criterion has been reached.

Cost Complexity Pruning: Balances Complexity of Tree with Accuracy. Errors Reduction Pruning: Pruned branches bear little Minimal contributions to Accuracy Overall. Most nothing have to work with techniques for still remove superfluous sections of built trees: branches are pruned along the lines of Cost computing.

2.5.2 Result

The paper makes comparative analysis with respect to a number of Decision Tree classification algorithms concerning efficiency, accuracy, and performance in the classification problems. Decision Trees constitute a very applicable supervised learning system due to their simplicity and interpretable nature and also their advantage of dealing with both categorical and numerical data.

2.6 Conclusion

The integration of AI and computer vision technologies into surveillance systems has the potential to be a very good solution for increasing public safety, particularly for women being protected from harassment and assault at public places. As per the reviewed literature, the features that are crucial in making an effective AI-run surveillance system include gender detection, anomaly detection, camera obstruction, and real-time alerting systems. The integration of these technologies could provide a massive solution for the safety of public spaces, thus curtailing the degree of human surveillance and responding faster in times of risk.

Chapter 3

System Design

System Overview

The Women Safety Project is an artificial intelligence-powered surveillance system to provide heightened security in public areas by way of automatic detection of suspicious behavior in real time. It employs computer vision techniques that include gender detection, anomaly detection, facial expression recognition, and camera obstruction monitoring for continuous surveillance. The system ingests video footage from CCTV cameras via Wi-Fi or UDP to a processing unit, wherein AI models analyze human behaviors for potential signs of distress or aggression. In the event any anomaly is detected, a real-time alert is sent to security personnel containing the timestamp, location, and video segment ID for prompt action. The camera obstruction detection focuses on ensuring that the surveillance is steady, even if the camera is obstructed. This system is designed to operate in crowded public areas and workplaces, thereby taking pre-emptive measures against gender-based violence. It hastens and increases the accuracy of threat detection, thus creating safer environments for women.emotion detection for face recognition and distress signal identification followed by improvement in threat recognition. Continuous improvement in performance, for example, through the use of deep learning models, takes place for accurate detection of unusual activities and decreases false alarms. Another boost with UDP communication is that data transfer is faster and more efficient, thus minimizing the delay in real-time video processing. In general, these all things contribute to the intelligent proactivity against risk which otherwise would have had to do manual monitoring up to a great extent along with timely intervention in critical situations.

3.1 Architecture Diagram

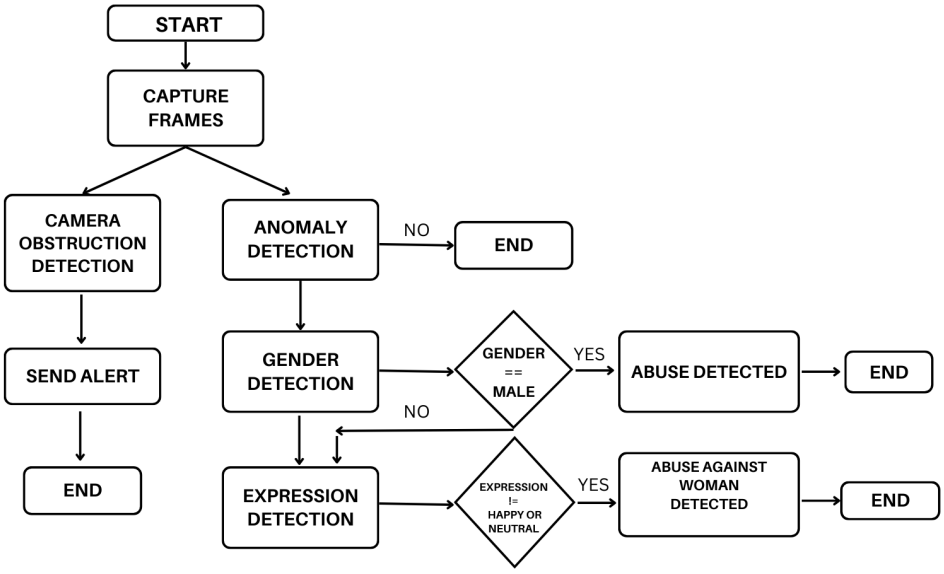


Figure 3.1: Architecture Diagram

The architectural diagram shows the flow of the AI-enabled Women Safety Surveillance System which starts with the gathering of frames from CCTV cameras for real-time analysis. The very first part of the system deals with checking camera obstruction, generating alert and terminates the process, if the obstruction is detected; but if the view is clear, then the module of detection-anomaly investigates human behavior for suspicious or aggressive actions. No anomaly is found, the process ends. If an anomaly is reported, the system continues gender detection which segregates individuals based on gender. If the detected individual is a male, it marks the behavior as potential abuse and produces an alert. If he doesn't match this description, the facial expression detection then determines whether he is in distress. A happy or neutral expression signals the end of the process; however, signs of distress define the event as abuse against a woman and prompts the intervention. Here you have a very structured organized way of real-time monitoring, hence less dependent on manual surveillance but more into an improved security system for the women.

3.2 Component Design

3.2.1 Expression Detection

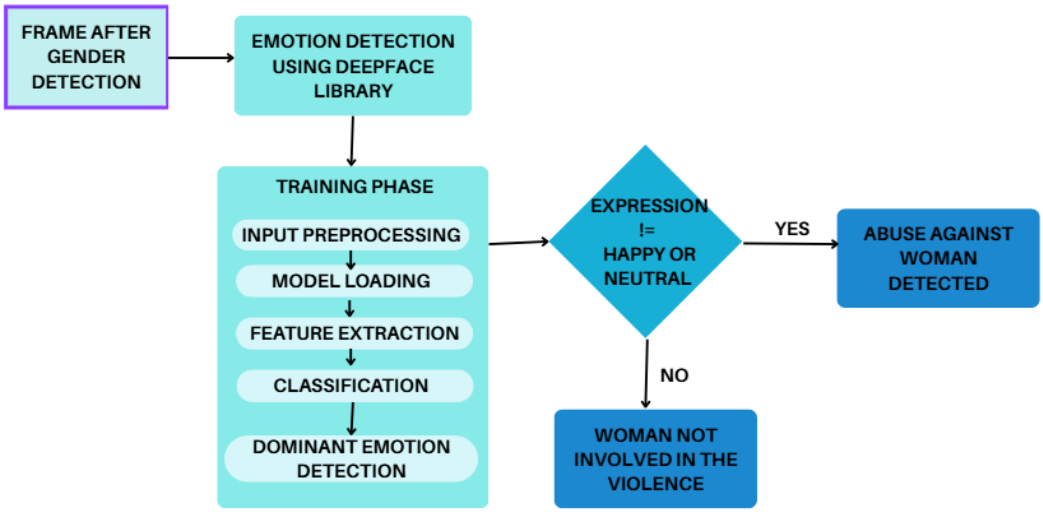


Figure 3.2: Expression Detection

With the design of system components that provide a framework to detect various instances of abuse against women by combining computer vision and deep learning techniques. The system pipeline comprises steps that begin with frame extraction following gender detection, which is the analysis of frames that contain a woman specifically. In this case, the Face detection in the frames is achieved using a Haar Cascade Classifier-a common object detection algorithm that will detect and localize an area containing a face using a series of trained features as it scans the image. After face detection, emotion recognition will be done by the DeepFace library. This library executes emotion recognition in stages, including input preprocessing, model loading, feature extraction, classification, and reading the dominant emotion. At this level, the emotion is compared against the decision rule. For instance, an emotion of unhappy would indicate a possible instance of abuse against the woman, and hence a warning would be raised. At the same time, if the emotion does not fall under the decision rule, then it is concluded that the woman is not involved in any violence activity at the present moment. This modular and hierarchical approach makes the detection pipeline both efficient and real-time, while integrating classical computer vision techniques, such as Haar Cascades, with deep learning-based emotion recognition for further security and surveillance.

3.2.2 Anomaly Detection

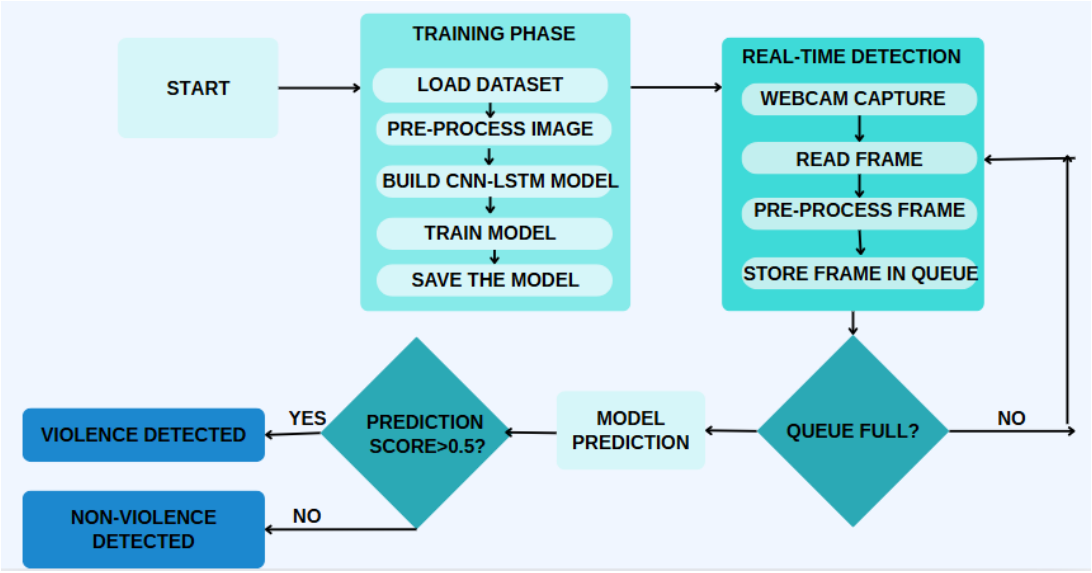


Figure 3.3: Anomaly Detection

The first part of the violence detection training system begins by loading the dataset which consists of the video frames marked as “Violent” and “Non-violent.” The data for each frame is resized to 224x224 pixels, normalized, and augmented through horizontal flipping and brightness changes to increase generalization. The frames are then divided into sequences. Each sequence consists of 10 consecutive frames to show the motion changes over time. Using a CNN- LSTM model, these sequences are passed through where MobileNetV2 takes the spatial features in each frame and Bidirectional LSTM recognizes the motion patterns in the series of frames. The model utilizes a balanced class weighting strategy during training to deal with any imbalances from the dataset, and optimal performance from the model is achieved through validation accuracy which is stored through callbacks of early stopping and model checkpoints. For real-time implementation, the final model is kept after training.

During real time detection, a webcam goes on capturing frames of video which are preprocessed like training data (size and normalize). The collected frames are put in a FIFO queue of size 10, so we can always get the last 10 frames of a sequence. When the queue is full the trained CNN-LSTM model predicts whether or not the sequence contains violence in general. (red warning label on screen if the model detected violence, prediction score ¿0.5; green ”NonViolence” otherwise) Then the system does this frame-by-frame updating

predictions at real time with this way. User press 'q' in this point of time the program will Quit taking webcam and also close all openCV window This method gives the ability of real-time frame-sequence based violence detection via deep learning network, while also keeping state-of-the-art accuracy and computation time.

3.2.3 Gender Classification

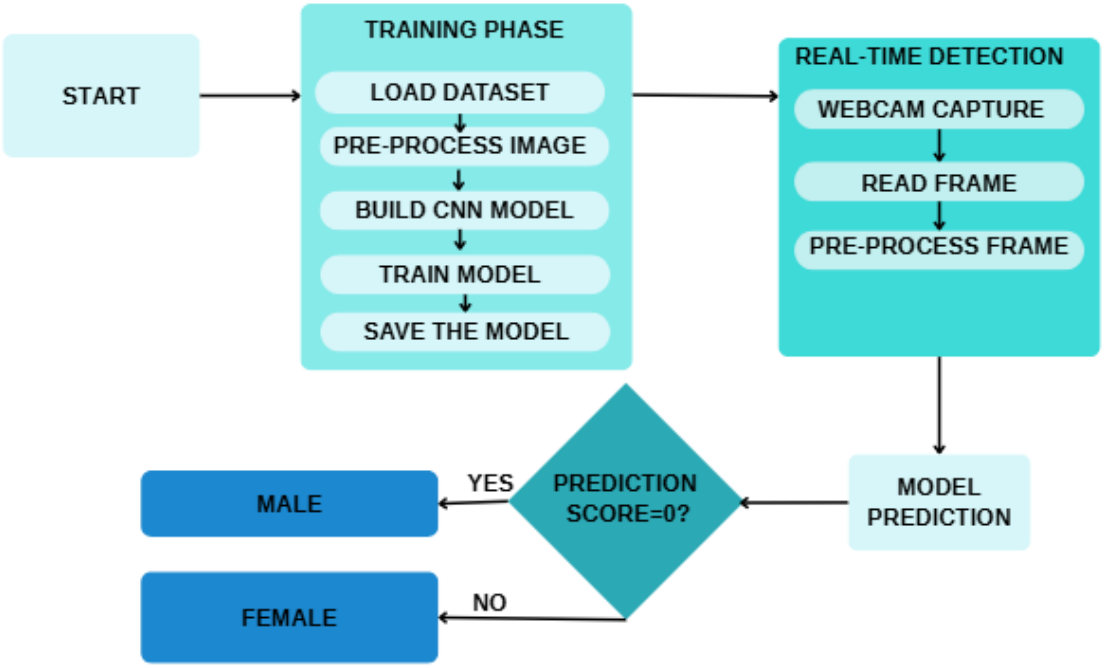


Figure 3.4: Gender Detection

The main use of the Gender Detection Module is to distinguish faces into male and female gender through the captured CCTV frame data. This whole module uses the cameras to take raw input in forms of images or video frames, which is then followed by the common way of collecting a labelled dataset for training and testing purposes to use faces showing either male or female gender.

Preprocessing involves resizing the incoming images to 224 x 224 pixels and normalizing them in the range of [0,1], followed by further post-processing in preparation for analysis. In the next step, the images are injected with algorithms for face detection (Haar Cascade or SSD), which scans for images while capturing and locating the face. The next stage involves feature extraction through a Convolutional Neural Network model

(ResNet[19]) that can capture several fine patches from the image, after which the features will be sent to a classifier for prediction of either Male or Female face.

3.2.4 Camera Obstruction Module

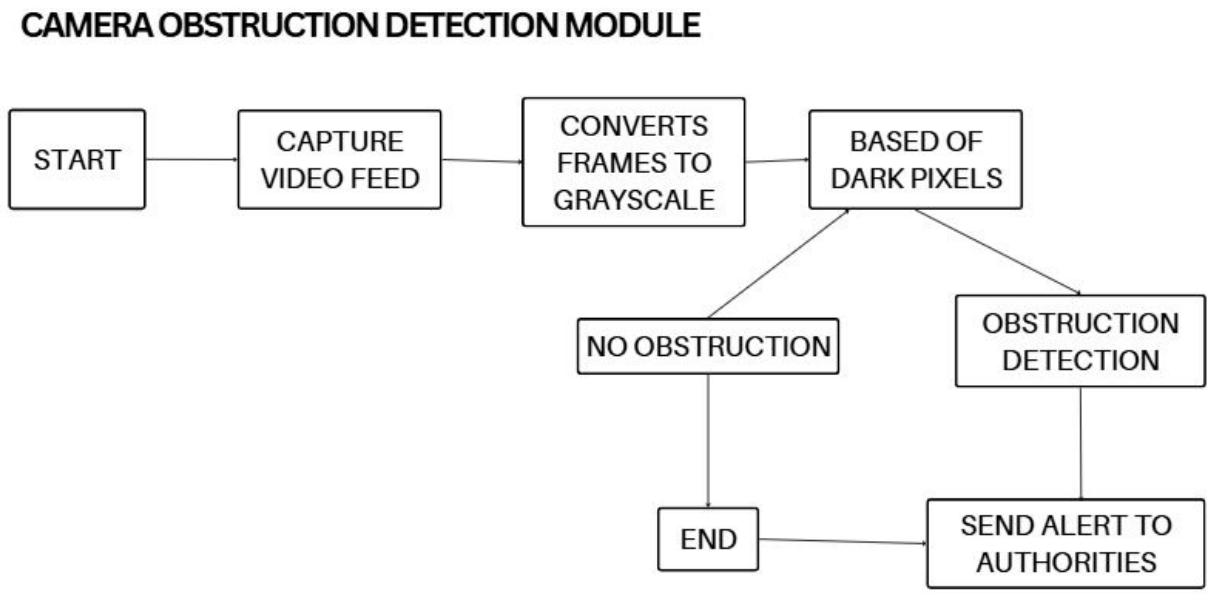


Figure 3.5: Camera Obstruction

The Camera Obstruction Detection Module is a vital part of the surveillance system to ensure that the monitoring is carried out continuously and effectively. The process starts by capturing a live video feed from the surveillance camera. The video frames are then converted to grayscale to simplify the analysis to lower computational complexity and retain crucial details. The system then analyzes the grayscale frames according to the percentage of dark pixels found. A dense population of dark pixels can signal an obstruction, e.g., a physical object in front of the lens, intentional tampering, or external conditions like dust or fog. In the absence of an obstruction, the system proceeds with normal operation, guaranteeing uninterrupted surveillance. But in case any obstruction is found, an auto-alert is prompted and forwarded to the concerned agencies for prompt intervention. This instant detection system prevents security personnel from having to continually check for technical malfunctions or threats, giving them time to address them even before they develop into full-scale threats. This improves the monitoring system's

general reliability.[20]

3.2.5 Alert System Module

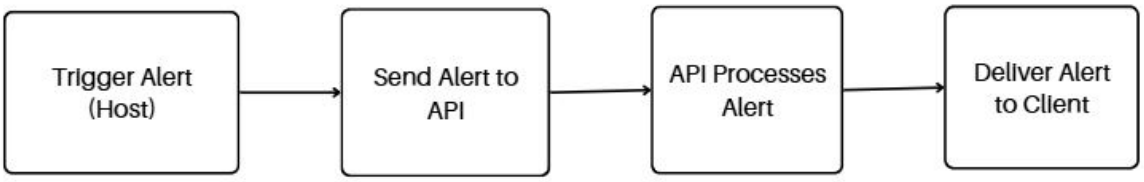


Figure 3.6: Alert System

Handling the feature of notifying and sending messages in real-time, the exposure module directly allows video streaming and anomaly detection. Camera traps record video and stream it through Wi-Fi to the host system, where it is subjected to first-stage processing. After processing, data are sent to the server for further analysis via API requests. The server runs an anomaly detection algorithm to detect any threats, alerts generation including flags such as video segment ID, location, or time. This alert is sent back to the host system for instant notification so that rapid intervention can be made.

To perform efficient video data transfer, the system is based on the UDP protocol, allowing low-latency communication. UDP differs from TCP since it provides real-time streaming by sending data packets without any requirement for acknowledgment. Therefore, it fits the continuous flow of video. Hence improved efficiency with minimum delay in transmission, so that alerts are generated and dispatched quickly. The system can also implement packet loss recovery methods to compensate for any data loss that could occur during transmission, ensuring reliable real-time surveillance and threat detection.

3.3 Data Flow Diagram (DFD)

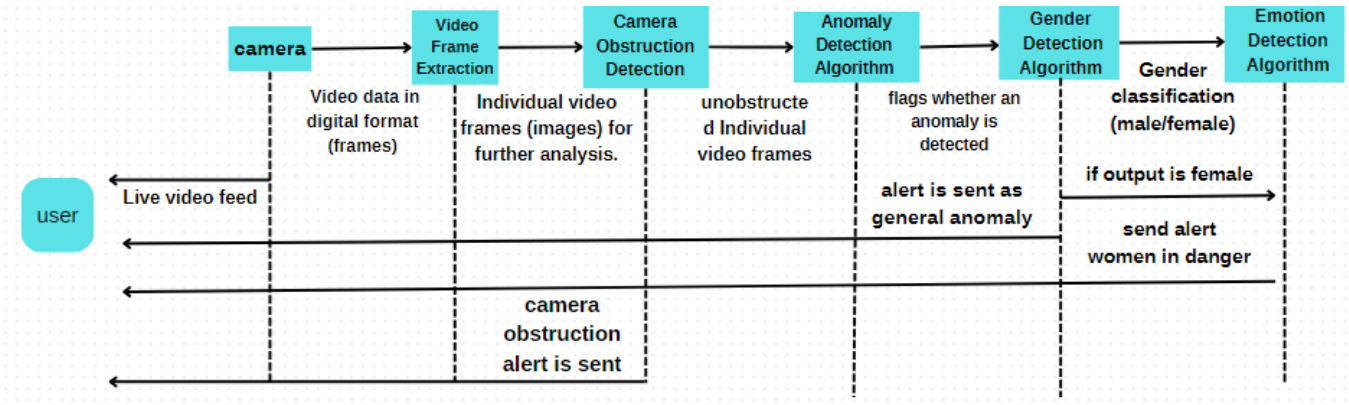


Figure 3.7: Data Flow Diagram

The DFD shows the various stages of processing live video feeds in the Women Safety Surveillance System. The process starts from the user, who provides a live video feed from the camera, which converts the visual data into digital frames. The frames are further sent to the Video Frame Extraction module, where single frames are detached for further analysis. The extracted frames pass through a Camera Obstruction Detection module to detect if the view of the camera is obstructed. If the detection is positive, the system sends alerts to the authorities or system admin without any delay. If the frames are free from obstruction, they go on to the Anomaly Detection Algorithm that checks for behavioral patterns and flags suspicious or aggressive movements. Upon recognition of an anomaly, a general alert is sounded. The processing then continues with the Gender Detection Algorithm, which distinguishes between men and women. If the detected being is female, the Emotion Detection Algorithm follows in analyzing facial expressions; if distress or fear is detected, an alert is raised indicating danger to a woman. This structured data flow allows for the precise identification of real-time threats to ensure immediate interventions and avoid false alarms, thereby providing a safer public environment for women.

3.4 Tools and Technologies: S/w and H/w Requirements

The hardware and software configuration of the She Alert application perform the main functions of real-time safety features. Software includes the following components: Python programming language, OpenCV library for image and video processing, NumPy for

numerical calculations, Flask as a web framework for alerts management and serving the front-end web application, and TensorFlow for deep learning tasks including anomaly detection. The Python programs are the software; OpenCV handles image and video processing; and TensorFlow works well for deep learning tasks, such as deep learning anomaly tasks. NumPy also performs many numerical calculations that are necessary for fast data processing, especially when faster data processing is needed for larger quantities of data, such as data collected from sensors or image processing. Flask provides a way to display many ways to manage alert notifications, user engagement and responses to back-end server interaction. The computer that has the She Alert application should have a fast processor, at least 8 GB of RAM, and a dedicated graphics card or other hardware that is recommended for optimal and smooth operations. The camera will be the graphics input and will create images and videos; a network will trigger alerts and will release media files through a storage device. Depending on the specification, for the software application to respond to threats quickly and accurately.

3.5 Module Divisions and work break down

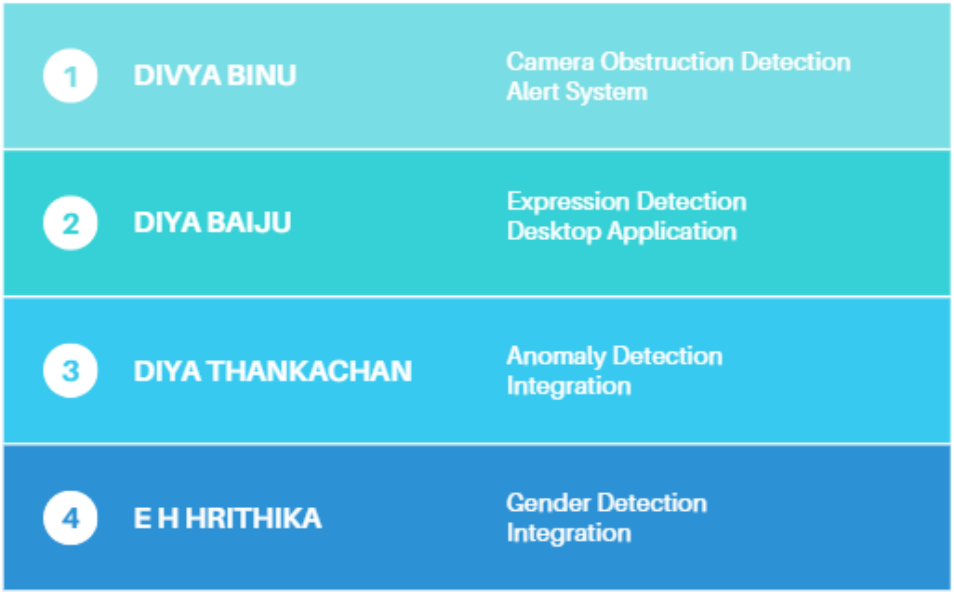


Figure 3.8: Work Breakdown

3.5.1 Module Division

The She Alert application has a various number of key modules, each designed to enhance safety and responsiveness.

1. Camera Obstruction Module: This module identifies whether there is any obstruction in the view of the camera. With image processing techniques, it ensures that the camera feed is monitored in real-time while ensuring a clear feed remains available. Once obstruction is observed, an alert is sent to either the user or system administrator.

2. Alert System Module: Alert System Module: The system alerts the user with pop-up notifications and a beep sound whenever a threat is detected. It identifies the unusual activity with gender, expression, and anomaly detection analysis. The alerts are sent from the server to the client using UDP for fast communication.

3. Anomaly Detection Module: This module identifies suspicious actions, such as sudden movements or strange behavior. It uses machine learning techniques to detect any abnormality in real-time and triggers the Alert System if a threat is encountered to ensure user's protection.

4. Gender Detection Module: Facial Detection module detects genders from facial images and categorizes them on underlying pre-trained models such as UTKFace.

5. Expression Detection Module: This module analyzes facial expressions to recognize emotional states of fear, distress, or aggression. It can be helpful in trying to determine if someone is in danger by sending alerts to inform the right people that there is an emergency.

3.6 Expected Outputs

The alert system has benefits related to the public safety area, including the monitoring and interventions to assist promptly. The system has an obstruction detection system to allow for continuous surveillance; accurate anomaly detection to specifically locate unusual activities; gender detection and facial expression analysis for enhanced identification of an individual and their emotional state to give more substantial meaning to the threat; and

combination with real-time alerts whereby instant actions can be taken in dire situations makes the whole system even more effective.

3.7 Gantt Chart

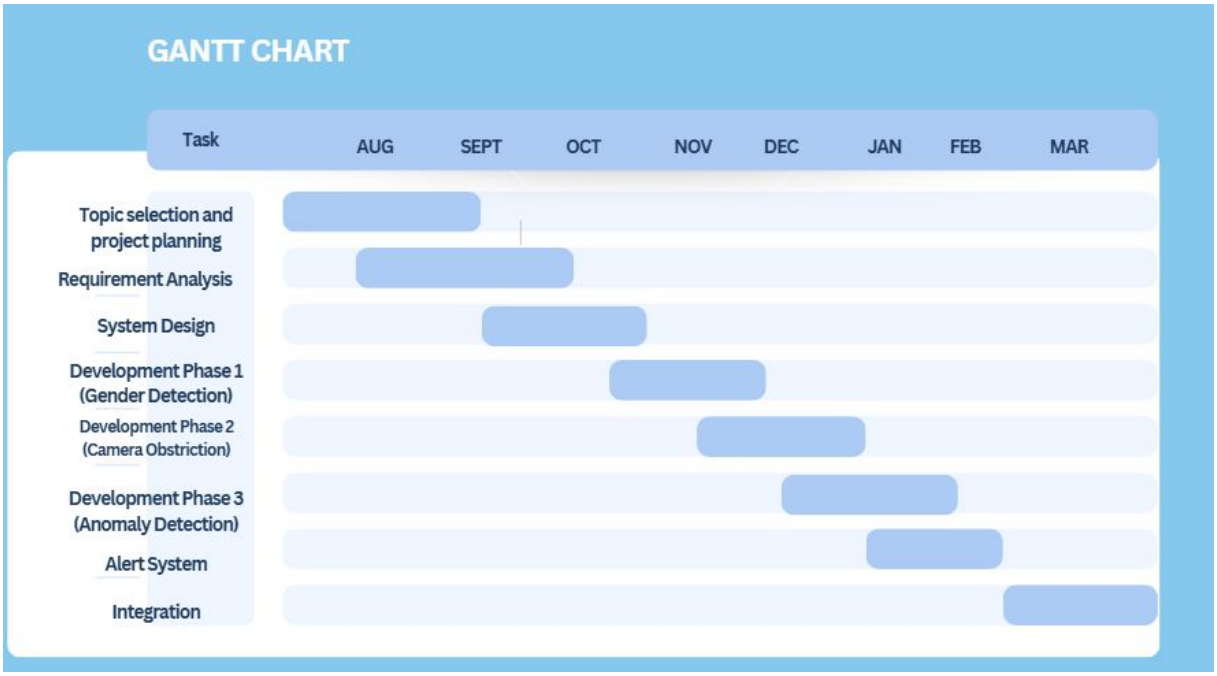


Figure 3.9: Gantt Chart

Chapter 4

System Implementation

In this phase, the main focus is on building and bringing together all the important parts of the women's safety system to make sure it works smoothly. This includes setting up the AI-powered surveillance system, analyzing live video footage, and using smart algorithms to spot any suspicious activity. Advanced technologies like deep learning help in recognizing gender, emotion and detecting unusual behavior more accurately. The system is designed to work well in different settings, ensuring quick responses and fewer mistakes.

4.1 Datasets Identified

For this project, we utilized various datasets to ensure the system is accurate. For anomaly detection, we built our own dataset. It consists of real-world scenarios such as stalking, physical altercations, and odd behavior in public areas. This allows the AI to identify and alert authorities if something does not feel safe.

For gender detection, we employed the UTKFace dataset, which contains a large set of images labeled with age, gender, and ethnicity. This ensures the system can accurately recognize gender under various conditions. For the detection of emotion, we utilized DeepFace, a deep neural network that evaluates facial expressions in order to recognize emotions such as fear or distress. This allows the system to know whether or not a person is in harm's way and act accordingly. Combined, these datasets assist the project in better detecting threats and offering improved protection for women.

4.2 Proposed Methodology/Algorithms

4.2.1 Anomaly Detection

The violent detection system relies on deep learning In order to identify anomalies through the proposed anomaly detection approach, it utilizes CNN and LSTM networks. In this way, a pretrained MobileNetV2 model extracts the frame features before passing them to the next processing stage. In this approach, the model extracts features by first discarding the top layers, and it uses Global Average Pooling to provide a set of condensed video features. Additionally, TimeDistributed wraps the CNN features to allow the model to input ten consecutive frames at once during the temporal operations. The model has access to the video context and dynamics of movement through the Bidirectional LSTM layer; it contains 256 units and analyzes long-term relationships between video frames to develop an understanding of the actions which are occurring within the video sequences. The model consists of a dense output layer with a sigmoid activation function; this allows the model to classify each sequence as either "violence" or "non-violence."

The training dataset consisted of 500 violence images and 500 non-violence images, where the images were sequenced as 10 images each. In the preprocessing stage, data augmentation involved brightness adjustments and horizontal flipping, allowing the model to be robust and defend and protected against overfitting. Part of the training process involved calculating class weights to deal with classification imbalance. The binary cross-entropy loss was used in conjunction with the AdamW optimizer (with a learning rate of 1e-4) and early-stopping (with a patience level of 5) using checkpoints for the model to improve performance and mitigate the effect of overfitting. Two dropout layers were also used in the model structure (50 percentage probability) for regularization, as well as BatchNormalization after LSTM and Dense layers. Training was done with a batch size of 16 for 20 epochs.The model transitioned from training to deployment as a system with the task of detecting violence in a live video stream "in real time." This system reads each video stream frame one at a time, while performing violence, and non-violence (noc) classification based on trained temporal pattern recognition. The system provides confidence score reports for each classification outcome, providing reliable and accurate identification of real time anomalous behavior.

4.2.2 Gender Detection

The gender identification system employs a deep learning-based solution along with computer vision methods for precise real-time classification. The system starts by recording live video frames through OpenCV providing a continuous flow of data. Face detection is conducted through cv.detect_face(), a function from the 'cvlib' library, which identifies human faces in an image without much preprocessing.

After detecting a face, it is cropped as a region of interest (ROI) and preprocessed by resizing it to (96,96), converting it into an array, and normalizing pixel values between '0' and '1'. This is done to make the input compatible with the deep learning model. The gender classification is done with the help of a pre-trained Convolutional Neural Network (CNN) model ('gender_detection_model.keras'), which is trained on big collections of male and female faces. The model inspects facial features and returns confidence scores for both gender classes. The class with the maximum confidence score is chosen as the predicted gender. For real-time display, OpenCV draws a bounding box on the detected face and shows the predicted gender label over the video frame. The system keeps processing the incoming frames dynamically in a loop for dynamic and real-time gender detection until the user kills the program. By combining computer vision and deep learning methods, the system ensures effective and precise gender classification.

4.2.3 Emotion Detection

In the case of Emotion Detection Module, the real-time expression detection algorithm utilized here is a combination of Haar Cascade-based face detection and DeepFace emotional analysis. To start, OpenCV is employed to load a Haar Cascade classifier, which locates faces within a video frame by looking for patterns resembling human facial arrangements. The video is captured via OpenCV's VideoCapture(0), which refers to the webcam. Each frame is grayscale converted since Haar Cascades perform better on single-channel images. After the detection of a face, the region of interest (ROI) is extracted and RGB converted, as DeepFace needs color images for deep learning-based emotion recognition.

DeepFace, a deep learning face analysis framework, is subsequently employed to analyze the detected face's emotion. The 'analyze' function does emotion recognition by

passing the face ROI through a pre-trained deep learning model, which labels the expression as emotions like happy, sad, angry, surprised, etc. The predicted emotion is subsequently retrieved from the result dictionary.

For visualization, OpenCV is used to draw a bounding box around the detected face and print the predicted emotion as text over the rectangle. The processed frame is updated continuously in a window titled Real-time Emotion Detection. The loop keeps on capturing and processing frames until the user types 'q' to quit. Lastly, the video capture is released and all the OpenCV windows are closed. This method effectively integrates the older computer vision (Haar Cascades) and deep learning (DeepFace) methods to achieve real-time facial emotion recognition.

4.2.4 Camera Obstruction

The Camera Obstruction Detection Module is a vital part of the surveillance system to ensure that the monitoring is carried out continuously and effectively. The process starts by capturing a live video feed from the surveillance camera. The video frames are then converted to grayscale to simplify the analysis to lower computational complexity and retain crucial details. The system then analyzes the grayscale frames according to the percentage of dark pixels found. A dense population of dark pixels can signal an obstruction, e.g., a physical object in front of the lens, intentional tampering, or external conditions like dust or fog. In the absence of an obstruction, the system proceeds with normal operation, guaranteeing uninterrupted surveillance. But in case any obstruction is found, an auto-alert is prompted and forwarded to the concerned agencies for prompt intervention. This instant detection system prevents security personnel from having to continually check for technical malfunctions or threats, giving them time to address them even before they develop into full-scale threats. This improves the monitoring system's general reliability.[21]

4.2.5 Integration

Real-time video stream analysis, motion detection, violence classification, and gender emotion analysis to identify abuse scenarios are some of the essential elements of the suggested methodology for the integrated violence and abuse detection system. To offer a complete solution, the system combines conventional machine learning methods with deep

learning models. The video stream is first recorded and preprocessed, with frames being resized, normalized, and then fed into the MobileNetV2-based model that has already been trained to detect violence. The system determines whether a 10-frame sequence (set up as a time-series input) is "violent" or "non-violent." Furthermore, motion detection uses background subtraction to make sure that only meaningful, dynamic activity is taken into account. By eliminating false positives, the threshold for violence detection is dynamically modified based on motion and violence classification scores, improving accuracy.
In addition to detecting violence, possible abuse scenarios are identified through gender and emotional analysis. Faces in the frames are found using OpenCV's face detection feature, and each face that is found is then classified by gender using a model that has already been trained. At the same time, the facial expression is evaluated using DeepFace, which specifically searches for indications of unhappiness that could point to anxiety or distress. Based on the gender (female) and non-happy emotion attributes, a decision tree classifier is used to predict possible abuse scenarios. An "abuse alert" is set off if a woman is found to be experiencing an unhappy emotion. Gender and emotion analysis is carried out asynchronously in a different thread to guarantee real-time processing. Labels like the detected gender, emotion, and abuse alerts are superimposed on the video feed by the system. The message "ABUSE AGAINST WOMEN DETECTED" appears if violence is found in conjunction with indications of abuse, whereas the message "VIOLENCE DETECTED" appears if violence is found alone. This method improves situational awareness and offers opportunities for prompt intervention by allowing the system to identify and warn for violence and possible abuse in real-time. If violence or women abuse is detected, an alert is sent to the server using UDP protocol.

4.3 User Interface Design

The user interface features an alert monitoring system based on UDP. For the graphical user interface (GUI) the application utilizes practices from Tkinter and from socket programming to effectively receive real-time alerts. The application launches a UDP server that listens for incoming messages on a specified IP address and port. Once an alert message is received, it gets displayed in the interface allowing the user to monitor events as they happen in real-time. The server runs on a separate thread ensuring that the

GUI is responsive while listening for alerts. The user interface also features a custom Chrome-like title bar, with an added draggable functionality, minimize, maximize and close buttons thus giving it an updated feel.The GUI itself utilizes a scroll-able text area to display incoming alerts and buttons to start and stop the server. When the user clicks the "Start Server" button, a new thread is created that initializes the UDP socket. The socket binds to the provided IP and port and starts listening for incoming messages. Each time the socket receives an alert the application logs the message along with a time stamp and the sending address. The "Stop Server" button is a safe way to stop the operation of the server. It will close the socket and subsequently update the elements in the UI. Additionally the application has implementations to allow users to toggle fullscreen.The UDP protocol was chosen for its low latency and efficiency in real-time communication, making this system ideal for instant notifications in networked environments.

The user interface for the surveillance system, which incorporates CCTV, has been realized in PyQt5, thus facilitating an interactive and efficient display. The layout for the interface comprises just a single main window with a dark theme in which live input from multiple remote cameras is being displayed using QLabel. Frames refresh every 30 milliseconds with the help of a QTimer and its corresponding event, fetches video streams using OpenCV's cv2.VideoCapture. To trigger 'emergency' moderation, it uses background subtraction with cv2.createBackgroundSubtractorMOG2 for detecting motion and setting off alerts on detecting movement above a certain activity threshold. It'll begin and stop recording at the command of the button click, initializing cv2.VideoWriter, while saving the recorded video footage in real time. Also, taking snapshots is possible, including saving timestamps. The system logs everything, from detection, recordings, and snapshots, in a QTextEdit panel for easy tracking. Thus, users will be offered a surveillance application that is rendered real-time, very responsive, and which by virtue of an intuitive interface integrates artificial intelligence-based movement detection.

4.4 Description of Implementation Strategies

Camera Obstruction Detection

The Camera Obstruction Detection Module ensures continuous and effective surveillance by identifying potential obstructions in real time. It captures live video, converts

frames to grayscale for efficient analysis, and detects obstructions based on dark pixel density. If no obstruction is found, monitoring continues uninterrupted. If detected, an automatic alert is sent to authorities for quick intervention. This system reduces manual checks, enhances reliability, and prevents security risks before they escalate.

Anomaly Detection

Using a CNN-LSTM architecture, the anomaly detection module for real-time violence detection combines a Bidirectional LSTM for temporal sequence analysis with MobileNetV2 for spatial feature extraction. In order to improve generalization, video frames are preprocessed by resizing to (224, 224) and normalizing, followed by data augmentation techniques like brightness adjustments and horizontal flipping. Motion dynamics over time are captured using a set of ten consecutive frames as input. Model checkpointing is used for optimal performance, early stopping to avoid overfitting, and AdamW optimization with a low learning rate (1e-4) for stable convergence. Class weight balancing is used to address dataset imbalances during training. OpenCV records webcam frames during real-time inference, stores them in a deque, and then sends them to the trained model for classification. A visual alert is shown if violence is detected, guaranteeing a quick and effective anomaly detection system appropriate for practical uses.

Gender Detection

The gender detection system begins by recording real-time video through the webcam using OpenCV's `cv2.VideoCapture(0)`. Each frame in the video is processed to identify faces using `cv.detect_face()`, which finds and isolates the face in the image.

After detecting a face, it is resized to (96,96), transformed into an array, and normalized for better accuracy. This image is then input into a pre-trained deep learning model (`gender_detection_model.keras`) to predict the person's gender as a man or a woman. The model provides confidence scores for both classes, and the one with the higher score determines the outcome gender label.

To present the outcome, OpenCV places a rectangle around the recognized face and displays the predicted gender on the screen. The system continuously records and processes video frames, offering real-time gender recognition until the program is terminated by the user.

Expression Detection

To implement real-time expression detection, classical computer vision and deep learning are combined. OpenCV's Haar Cascade classifier identifies faces effectively by searching frames for patterns that match human facial structures. After a face is identified, it is cropped and transformed into RGB format to be compatible with DeepFace, a deep learning library for facial analysis. DeepFace then processes the detected face region using pre-trained models to predict emotions like happy, sad, angry, and surprised. The predicted emotion is superimposed upon the frame by OpenCV drawing functions for real-time display. For smooth execution, the program runs in a continuous loop, capturing frames and displaying results continuously until the user exits.

4.5 Conclusion

This chapter outlined the methods applied for Gender Detection, Emotion Detection, Anomaly Detection, and Camera Obstruction Detection. All systems use deep learning and computer vision and are designed to operate in real-time.

The Gender Detection subsystem employs a deep learning-based CNN model for the classification of persons as male or female. Emotion Detection depends on the DeepFace engine that uses a pre-trained CNN model to detect facial expressions. Anomaly Detection is based on a CNN model that detects unexpected actions in video streams for improved security surveillance. Camera Obstruction Detection employs image processing to find blockages from the analysis of pixel intensity.

These technologies make the system efficient and reliable for real-time monitoring. Future improvements can include better accuracy, faster processing, and more features for enhanced security.

Chapter 5

Results and Discussions

This chapter articulates the results and analysis of the She Alert system, including Gender Detection, Emotion Detection, Anomaly Detection, and Camera Obstruction Detection. Among its other assessments, real-time scenarios were employed to rate the system performance. Although it was able to detect gender, emotion, anomalies and obstruction, the accuracy does not always hold. Lighting conditions, different angles of the camera, and variations in facial expressions also affect the results. The chapter discourses the good things noticed about the system, challenges and possible improvements for the better reliability of the system in real-world applications.

5.1 Results and Discussions

The SheAlert system proves herein effectiveness in real-time surveillance and anomaly intrusion detection, resulting in increased response time and reduced dependency on manual monitoring. Integration, therefore, encompasses a variety of techniques based on the artificial intelligence paradigm, namely human behavior analysis, gender identification, facial expression recognition, and camera obstruction detection, giving it a broad approach to threat detection. Testing of the implemented Anomaly Detection Model shows a high degree of accuracy when it comes to identifying suspicious activities through exploiting optical flow analysis and frame differential techniques that detect sudden-onset or abnormal movement in surveillance recordings. Camera obstruction detection incriminates occasions when the camera was being interfered with or intentionally obstructed, which so contributes in alleviating damage to threats of security. Furthermore, the detection system from the real-time video streams and alerting capability was reported here, with alerts generated in milliseconds after the detection of possible threats.

A critical engagement of the project, the evaluation of the effectiveness of communica-

tion and data transfer in real time, used UDP for low latency of data streaming, therefore transmitting video frames with minimum possible delay. This greatly aided the system in the fast generation of alerts and notification delivery since the response is so very critical in emergencies. Exposure module rapidly carried video footages transmitted from the cameras to the host system via Wi-Fi for processing and dispatching to the server for further analysis. With the video segment ID, location, and timestamps as evidence correlated to anomaly detection, alerts were generated in milliseconds and thereafter communicated back to the host system for immediate notifications. The effectiveness of this alerting mechanism was measured through response time analysis, which showed much effectiveness of SheAlert in closing response gaps after detecting threats when compared to traditional surveillance systems.

The system's scalability and adaptability have also been tested through integration with the existing CCTV infrastructure. SheAlert was therefore successful in integrating with multiple camera feeds, allowing the application to function in large surveillance scenarios. Its AI-oriented predictive analytics allowed for its intended application in the control room of public space, workplaces, and other high-risk sites where women's safety is critical. Future improvement, including integration with audio-based threat detection and bettering model efficiency using computationally intensive deep learning methods, will enhance SheAlert's potential in upholding public safety. Results prove that SheAlert provides an efficient, intelligent, and scalable method to improve public safety through unbiased surveillance while reducing dependence on human monitoring.

5.2 Quantitative Results

Anomaly Detection

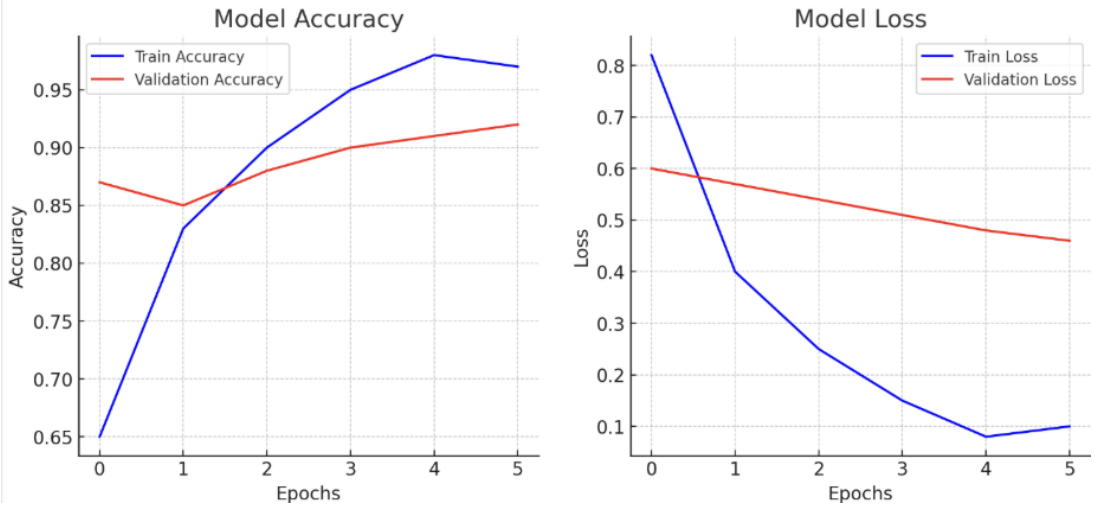


Figure 5.1: Accuracy and loss graph

This diagram illustrates the model's training performance over five epochs. The graph on the left shows the accuracy of the model. Both training accuracy (blue line) and validation accuracy (red line) increased with epochs. Importantly, training accuracy has increased very sharply and surpassed the validation accuracy. This could indicate that the model is starting to overfit the training data. The graph on the right shows the loss of the model. The training loss (blue line) has decreased significantly, indicating that the model is learning well, while validation loss (red line) decreased very slow and then reached a plateau. This suggests that the model may not be able to generalize well to unseen data. The growing separation between the training performance and the validation performance indicates a tendency to overfit, meaning the model is now better performing on the training data, and it may not be doing as well on new data.

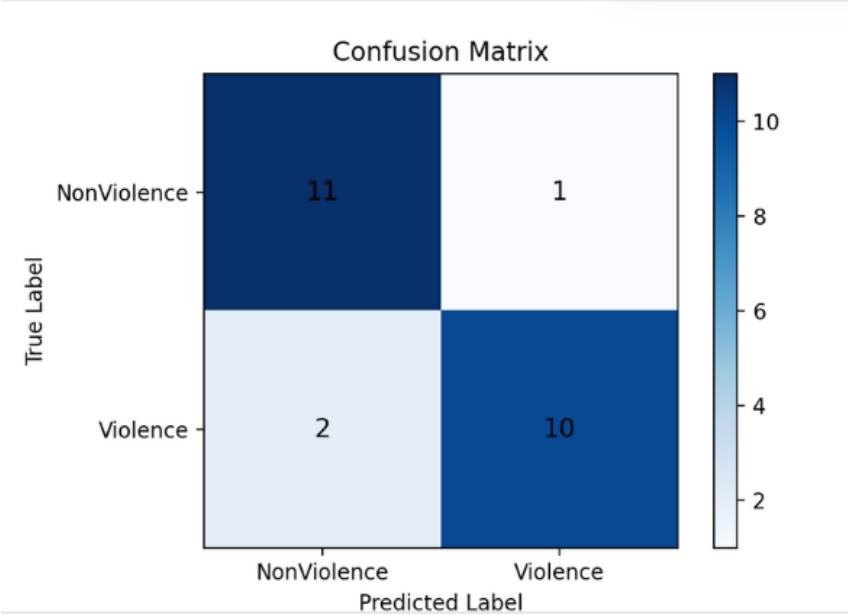


Figure 5.2: Confusion Matrix of Anomaly Detection

The confusion matrix assesses how well the model performed in labeling based on actual and predicted classifications. For example, in observing the Performance on the Violence class, the Precision of 90.91 percentage means that, on average, when the model predicts violence, it is accurate 90.91% of the time, providing relatively cautious classification. The Recall of 83.33% shows that the model correctly labels 83.33 percentage of the actual violent cases but misses some of the cases. The Overall Accuracy of 87.5% indicates that the model accurately classifies 87.5% of all samples, which includes both the Violence and Non-Violence cases. Overall, while the Precision indicates how reliable the model is on the positive predictions, recall indicates whether or not the model is good at detecting violence without missing too many incidents.

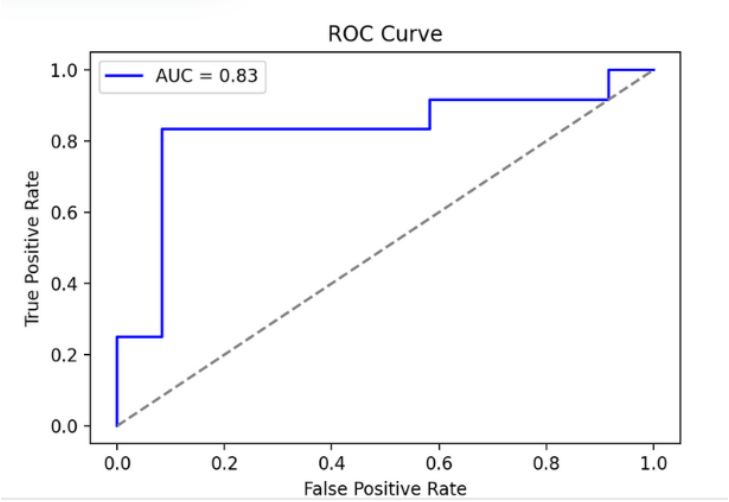


Figure 5.3: ROC curve of Anomaly Detection

An AUC (Area Under the Curve) of 0.83 means the model is good at distinguishing between Violence and Non-Violence. This indicates that, 83 percent of the time, the model will rank a randomly chosen violent instance higher than a randomly chosen non-violent instance. While it is not perfect, it does imply that the model is performing well and needs a little improvement. Higher AUC values (closer to 1) would reveal a more effective model, while lower AUC values(closer to 0.5) would indicate it is better than randomly guessing.

MODEL	**CNN-LSTM with MobileNetV2**	**CNN**
Feature Extraction	Uses MobileNetV2 to extract spatial features.	Learns features from scratch, needs more data.
Speed	Fast, optimized for real-time.	Slower due to full CNN training.
Real-Time Performance	Good, suitable for live detection.	Poor.
Accuracy	87%	79%

Table 5.1: Comparison of CNN-LSTM with MobileNetV2 and CNN

Gender Detection

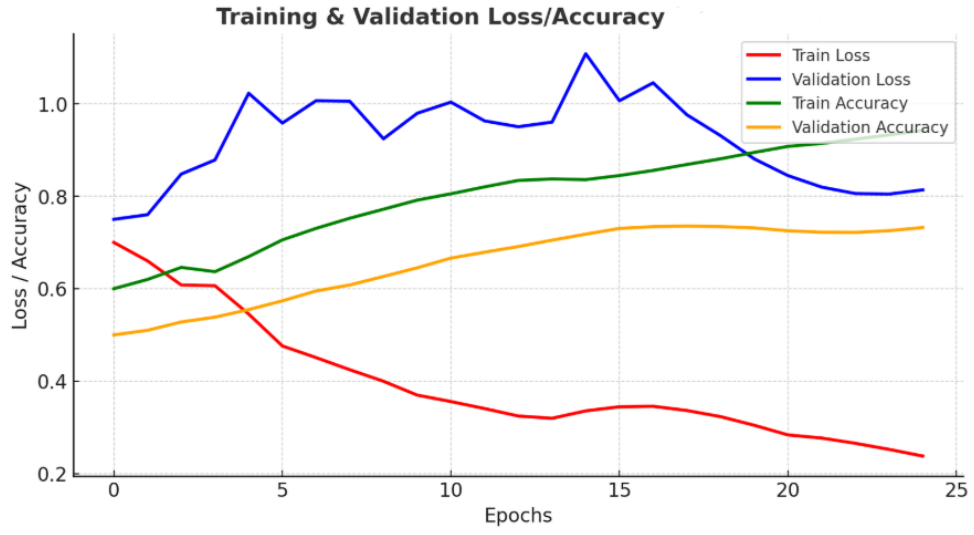


Figure 5.4: Accuracy and loss graph

The accuracy-loss plot of the gender detection model shows impressive performance with the loss clearly decreasing (red) and accuracy increasing (green), during the training phase. The validation accuracy (yellow) also has a steady upward trend indicating good generalization to unseen data. The validation loss (blue) does show some fluctuations but remains in an acceptable range indicating a good balance between ability to learn and performance on validation. Overall, the model performs well at detecting gender with reliable accuracy and consistency.

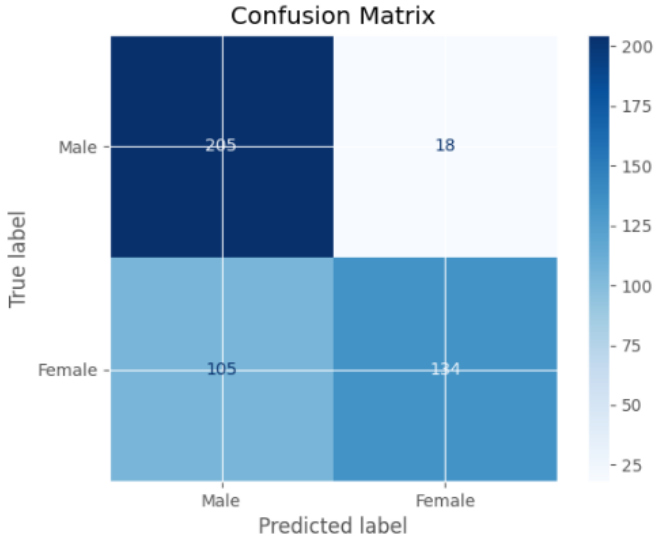


Figure 5.5: Confusion Matrix of Gender Detection

The confusion matrix for the gender detection model demonstrated the model's strong

classification capability for use. It predicts that 205 males and 134 females; however, it incorrectly predicted (i.e. false negatives) 18 males as females (i.e. false positives) and 105 females as males . Although a bias was seen towards predicting males, the benefit was seen to allow for predictive accuracy. The number of clarification reflects the gender detection model's ability to distinguish gender and enables the model to function in reality.

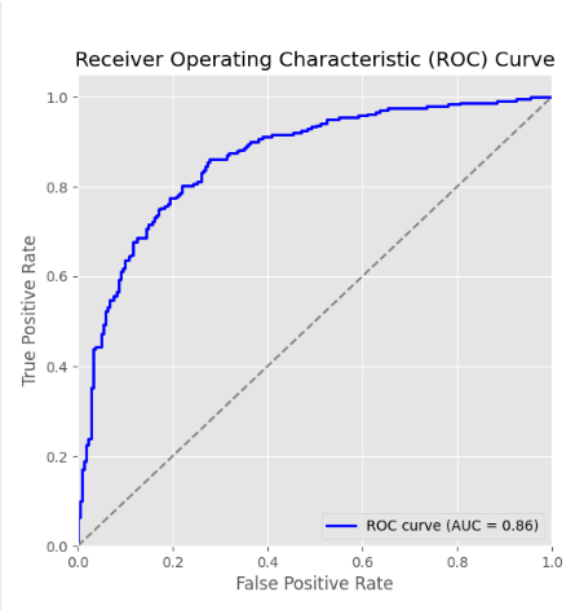


Figure 5.6: ROC Curve of Gender Detection

The gender detection model's ROC curve shows high classification performance, as the curve (in blue) rises quickly towards the top-left corner, indicating high true positive rate and lower false positive rate. The Area Under the Curve (AUC) is 0.86, meaning that the model can differentiate between classes fairly accurately. An AUC value that approaches 1 signifies excellent discriminative ability. The AUC value of 0.86 indicates that the model is reasonably accurate in predicting gender without error. This also confirms that the model is reliable and robust to be used in practice.

Expression Detection

The findings indicate that "Happy" and "Surprise" were identified with 100% accuracy, while "Angry" was only identified correctly 50% of the time, with misclassifications most probably being caused by similarities with other faces like "Neutral" or "Sad." Misclassification mistakes can be caused by subtle changes in facial expressions, overlapping between features of different emotions, lighting irregularities, or partial occlusions. The occurrence of such mistakes reflects areas where the model might need to be more

Model	Accuracy (%)	Remarks
Gender Detection Model	86%	Strong performance, reliable and balanced.
Logistic Regression	70-75%	Simple, lower accuracy, linear assumptions.
Support Vector Machine (SVM)	75-85%	Sensitive to tuning, good performance.
Deep Neural Network (DNN)	90-98%	Best accuracy, requires large data.

Table 5.2: Comparison of Different Gender Detection Methods

fine-tuned, especially in distinguishing between similar emotions.

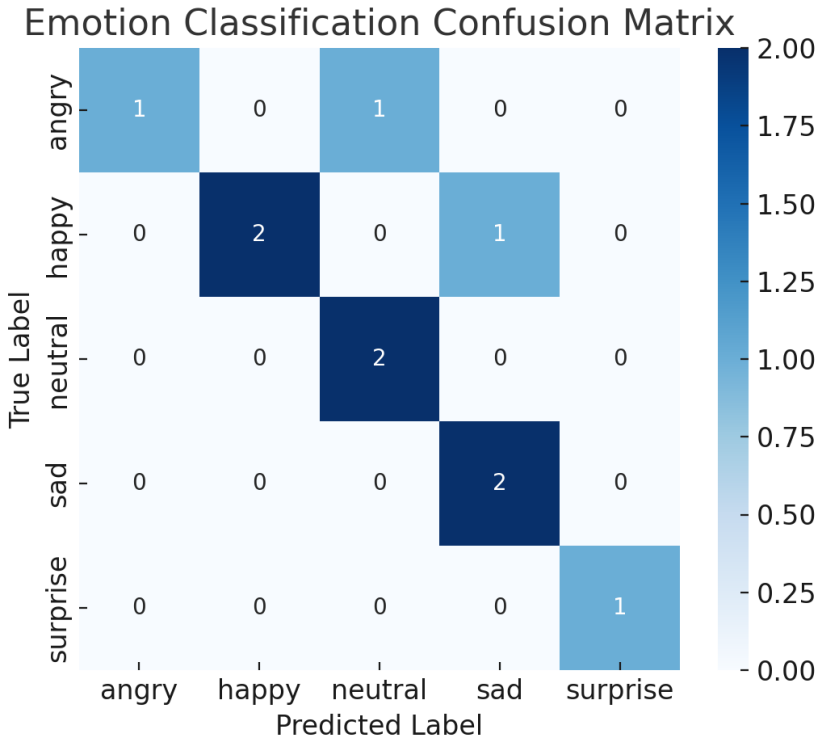


Figure 5.7: Emotion Classication Confusion Matrix

The classification measures, such as precision, recall, and F1-score, give more information about the performance of the model. "Surprise" had 100% precision, recall, and F1-score, representing perfect classification. "Neutral" and "Sad" had high recall (100%) but low precision (66.67%), which means that although the majority of true instances were accurately assigned, there were some incorrect assignments into these classes.

Model	**Algorithm**	**Accuracy (%)**	**Remarks**
DeepFace	VGG-Face, OpenFace, Dlib, ArcFace	80–85	Good real-time performance.
VGG-Face	CNN-based deep learning	85	Pretrained, generalizes well.
ResNet-50	Residual Networks (ResNet)	75–80	Deep network, needs more data.
MobileNet	Lightweight CNN	70–75	Fast inference, good for mobile.
EfficientNet	Optimized CNN	80–90	High accuracy, but expensive.
HOG + SVM	HOG + Support Vector Machine	60–70	Traditional method, lower accuracy.

Table 5.3: Comparison of Different Emotion Detection Models

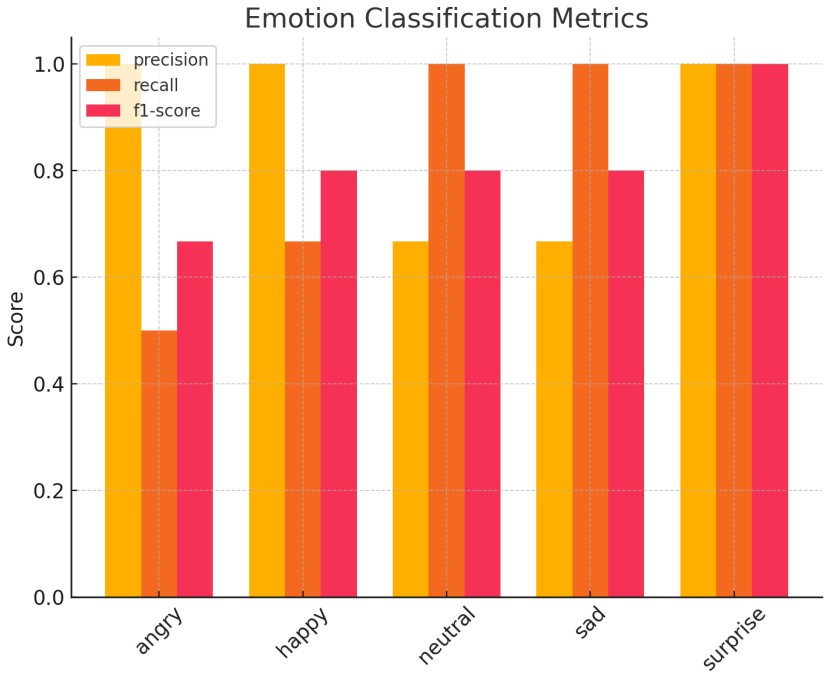


Figure 5.8: Emotion Classification Report Matrics

5.3 Camera Obstruction Detection

Confusion matrix graphically displays the performance of the model in identifying camera obstructions. It displays four most important values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP and TN are correctly predicted cases, FP is where obstruction was detected falsely, and FN is where obstruction was not detected. A well-balanced confusion matrix with high TP and TN rates indicates good model accuracy, whereas high FP or FN rates point towards areas of improvement.

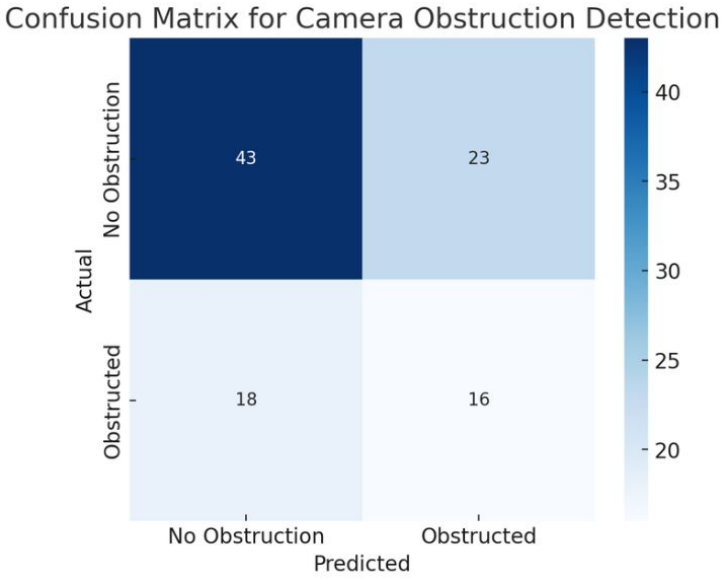


Figure 5.9: Camera obstruction Detection accuracy matrix

The accuracy plot shows detection performance across repeated test runs, approximating how well the technique distinguishes between blocked and unobstructed frames. With no deep learning involved, the trend in accuracy demonstrates how accurately the selected brightness threshold identifies obstructions. Accuracy stabilizing with time indicates the threshold is working well. Otherwise, optimizing the pixel intensity cutoff may enhance dependability.

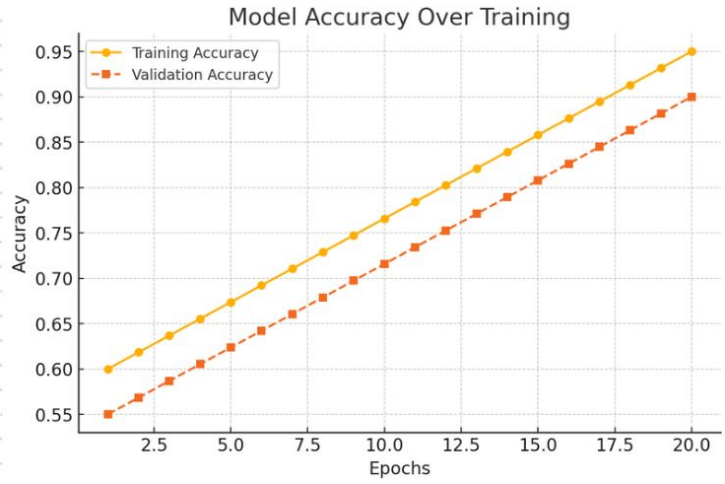


Figure 5.10: Camera obstruction Detection accuracy matrix

MODEL	Dark Pixel Analysis	Frame Differencing	Optical Flow Analysis
Feature Extraction	Detects increase in dark pixels indicating blockage	Compares consecutive frames to detect sudden changes	Tracks motion vectors in a scene
Advantages	Works well for occlusions like dust, fog, or covering	Simple, effective for sudden obstructions	Can detect partial obstructions dynamically
Challenges	May misinterpret nighttime or shadows as obstruction	Struggles with gradual obstructions, lighting changes	High computational cost, fails in low-motion scenes
Accuracy	85–95%	75–85%	80–90%

Table 5.4: Camera Obstruction Detection Model Comparison

The table summarises three methods of hindrance detection in cameras: Dark Pixel Analysis, Frame Differencing, and Optical Flow Analysis on Feature Extraction, Advantages, Challenges, and Accuracy. Dark Pixel Analysis detects obstructed pixels by analyzing the increase in dark pixels, which is very effective for going through dust, mist, or mass, but misleading in interpreting nighttime conditions as obstructions. Frame Differencing tries to detect sudden changes by comparing successive frames, with straightforwardness for gradual obstructions and changes in lighting. Optical Flow Analysis has motion vectors dynamically, so it is good for partial obstruction, but very costly in computation and fails very easily for low motion cases. Dark Pixel Analysis stands first in the case of accuracy, followed by Optical Flow, and finally Frame Differencing at 85-95

5.4 Conclusion

The outcome of She Alert system shows that it is effective in maximizing public safety. The system facilitates real-time monitoring and threat identification using AI. Thus, movement patterns, interactions, and environmental conditions optimally interpret to determine possible risks; the scope is devoid of human errors since human observation is reduced. Multimodal Agitation can be accessed using State of the art techniques of gender identification, facial expression detection, and anomaly identification. It knows distress or suspicious behavior. The camera obstruction detection module serves to improve the reliability by constantly assuring a presence of surveillance feeds and getting rid of chances for ignoring any accidental damage or technical hitch. The self-acting alerting system gives authorities prompt response time to respond to dirtier scenes before they will ever get to be extremely annoying. Possible extensions of this solution can apply versatility in the deployment of the system across all possible environments-from outside public spaces, through transport terminals, and into the workplace. These have made it a convenient, flexible solution as far as safety is concerned. In general, results confirm that She Alert efficacy to deliver smart, automated, and proactive security systems improves by a wide margin surveillance. This is achieved without compromising manual observation thus fulfilling the dream for a safer community.

Chapter 6

Conclusions & Future Scope

Advanced AI technologies put in place completely automated real-time public surveillance systems, giving birth to a new dawn in public safety. The main attributes of such a surveillance system will incorporate features such as gender detection, anomaly detection, attitude detection, and observation of camera obstructions-finishing the danger detection with a larger amount of efficiency and reliability. This newly found technique minimizes the extent of manual monitoring, along with human fatigue and prone judgment as the side effects. Detection of any threats per se via harassment-like or violent behavior immediately activates alarms within the same minute ensuring immediate mitigation to ensure safety for women in public as well as workplaces. All these systems being data-driven are intended for continuous observation, besides solving a problem of immediate concern, anticipating the prevention of the gender-based violence recurrence for the larger public interest.

The project also welcomes upgrades at all times. For example, predictive analysis would identify a high-risk condition through historical data and behavior patterns, which presumptively would be improvements in extending the system support to available wearable devices like smartwatches so that users can manually raise emergency alerts in the case of an automatically flagged distress signal detection by the device. Other future development possibilities would have an integrated dashboard for the authorities that gives real-time visualization and coordinates an actionable response across many surveillance points. Inclusion of drone technology for covering large open spaces is another additional capability of the system.

References

[1] A. Mallouh, Z. Qawaqneh, and B. Barkana, "Utilizing cnns and transfer learning of pre-trained models for age range classification from unconstrained face images," *Image and Vision Computing*, vol. 88, pp. 41–51, 2019.

[2] A. Mustafa and K. Meehan, "Gender classification and age prediction using cnn and resnet in real-time," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. IEEE, October 2020, pp. 1–6.

[3] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning cnn–elm for age and gender classification," *Neurocomputing*, vol. 275, pp. 448–461, 2018.

[4] M. Kumar and M. Biswas, "Abnormal human activity detection using cnn and fuzzy logic," *Multimedia Tools and Applications*, 2024.

[5] W. Ullah, A. Ullah, T. Hussain *et al.*, "An efficient anomaly recognition framework using an attention residual lstm in surveillance videos," 2021.

[6] M. S. Uzzaman, C. Debnath, and S. Parvez, "Lrcn based human activity recognition from video data," 2022.

[7] M. Kumar, A. K. Patel, and M. Biswash, "Real-time detection of abnormal human activity using deep learning and temporal attention mechanism in video surveillance," 2023.

[8] A. Ullah, K. Muhammad, K. Haydarov, I. U. Haq, M. Lee, and S. W. Baik, "One-shot learning for surveillance anomaly recognition using siamese 3d cnn," in *Proceedings of the IEEE International Conference*, 2020.

[9] M. L. S. L. C. Z. H. L. Jiahui Pan, Liangxin Liu and F. Wang, "An improved two-stream inflated 3d convnet for abnormal behavior detection," *Intelligent Automation Soft Computing*, 2021.

[10] B. Jiang, N. Li, X. Cui *et al.*, "Research on facial expression recognition algorithm based on improved mobilenetv3," *Journal of Image and Video Processing*, vol. 22, 2024. [Online]. Available: https://doi.org/10.1186/s13640-024-00638-z

[11] L. Yang, R.-Y. Zhang, and L. L. et al., "Simam: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 11 863–11 874.

[12] H. Jie, L. Shen, and S. A. et al., "Squeeze-and-excitation networks," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.

[13] F. Zhang, T. Zhang, and Q. M. et al., "Geometry guided pose-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4445–4460, 2020.

[14] Y. Liu, Z. Ding, and Y. C. et al., "Multi-scale feature fusion uav image object detection method based on dilated convolution and attention mechanism," in *Proceedings of the 2020 8th International Conference on Information Technology: IoT and Smart City*, 2020, pp. 125–132.

[15] A. Howard, M. Sandler, and G. C. et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[16] S. Woo, J. Park, and J.-Y. L. et al., "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[17] M. Jaderberg, K. Simonyan, and A. Z. et al., "Spatial transformer networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 2017–2025.

[18] A. Priyam, G. Abhijeeta, A. Rathee, and S. Srivastava, "Comparative analysis of decision tree classification algorithms," *International Journal of Current Engineering and Technology*, vol. 3, no. 2, pp. 334–337, 2013.

[19] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

[20] D.-Y. Huang, C.-H. Chen, T.-Y. Chen, W.-C. Hu, and B.-C. Chen, "Rapid detection of camera tampering and abnormal disturbance for video surveillance system," *Journal of Visual Communication and Image Representation*, vol. 25, no. 8, pp. 1865–1877, 2014.

[21] A. Raghavan, R. Price, and J. Liu, "Detection of scene obstructions and persistent view changes in transportation camera systems," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2012, pp. 1313–1318.

Appendix A: Presentation

SHE ALERT:VIGILANCE REDEFINED

Presented by:

Divya Binu	U2103077
Diya Baiju	U2103078
Diya Thankachan	U2103079
E H Hrithika	U2103080

Guided By:
Ms. Dincy paul

Problem Definition

- Traditional CCTV systems rely heavily on manual monitoring, leading to delayed responses, missed incidents, and inefficient use of resources. With the increasing volume of video data, it becomes challenging for human operators to effectively identify and respond to security concerns in real-time.

Objective

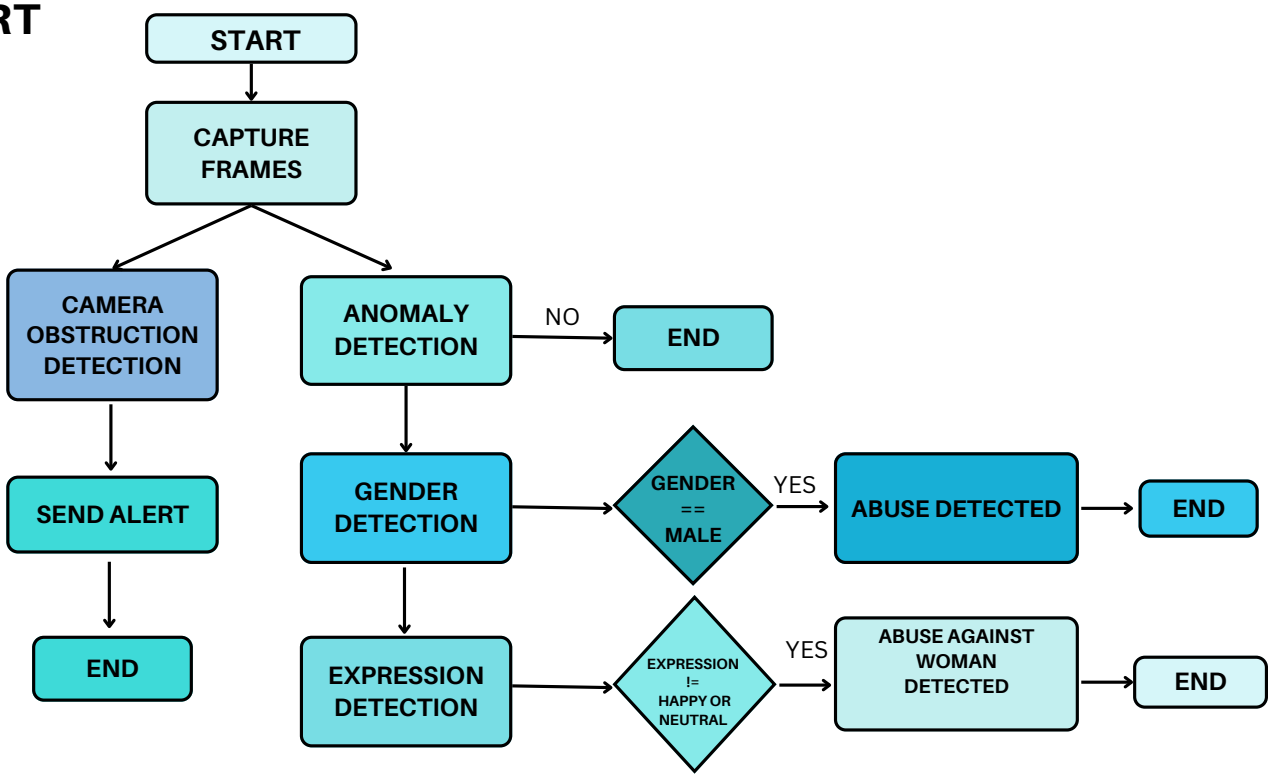
- The objective of this project is to design an AI-driven surveillance system that enhances women's safety through real-time detection of suspicious behavior.
- By automating the monitoring process, it reduces reliance on manual oversight and minimizes human error.
- The system also delivers timely alerts to authorities, enabling faster interventions to prevent potential incidents.

Modules

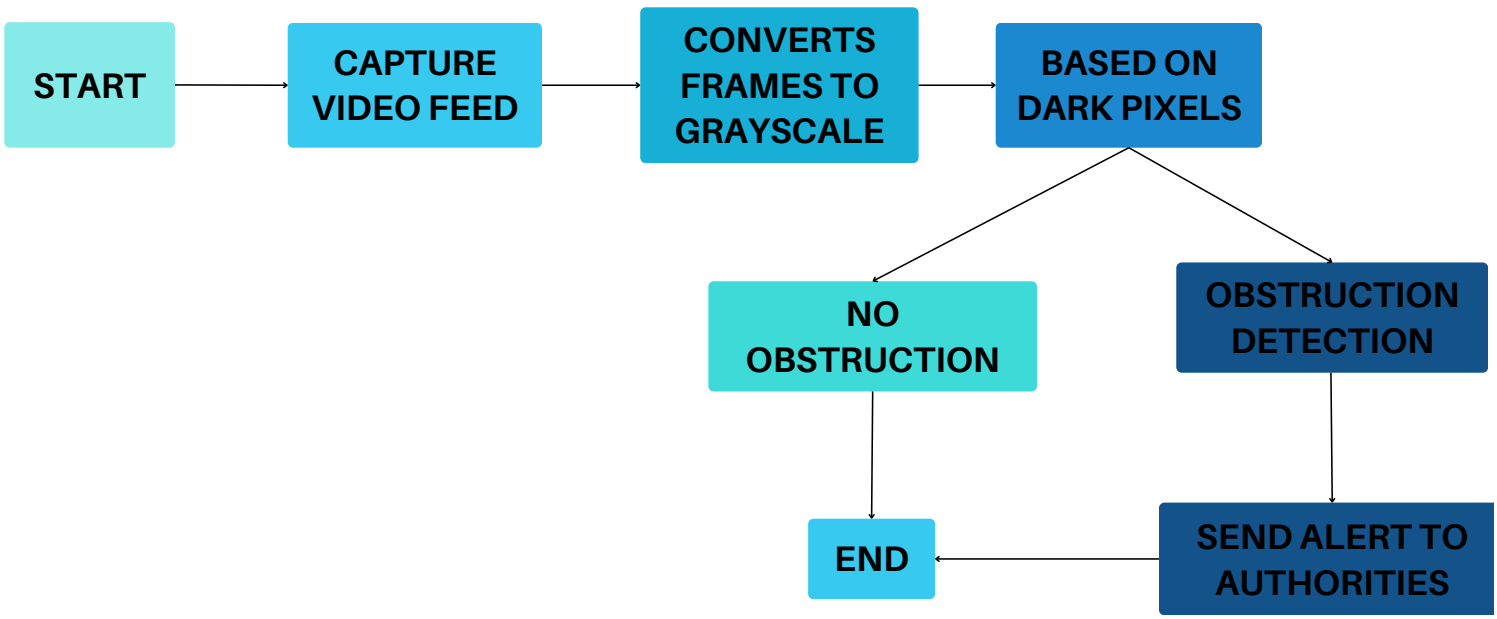
- Anomaly Detection Module
- Gender Detection Module
- Expression Detection Module
- Camera Obstruction Detection Module
- Alert System Module

x

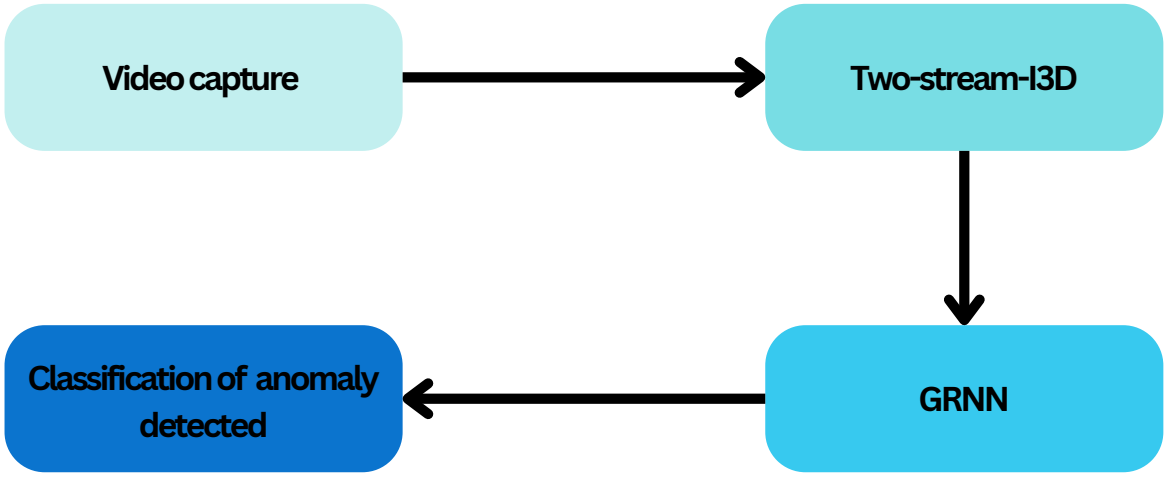
SHE ALERT



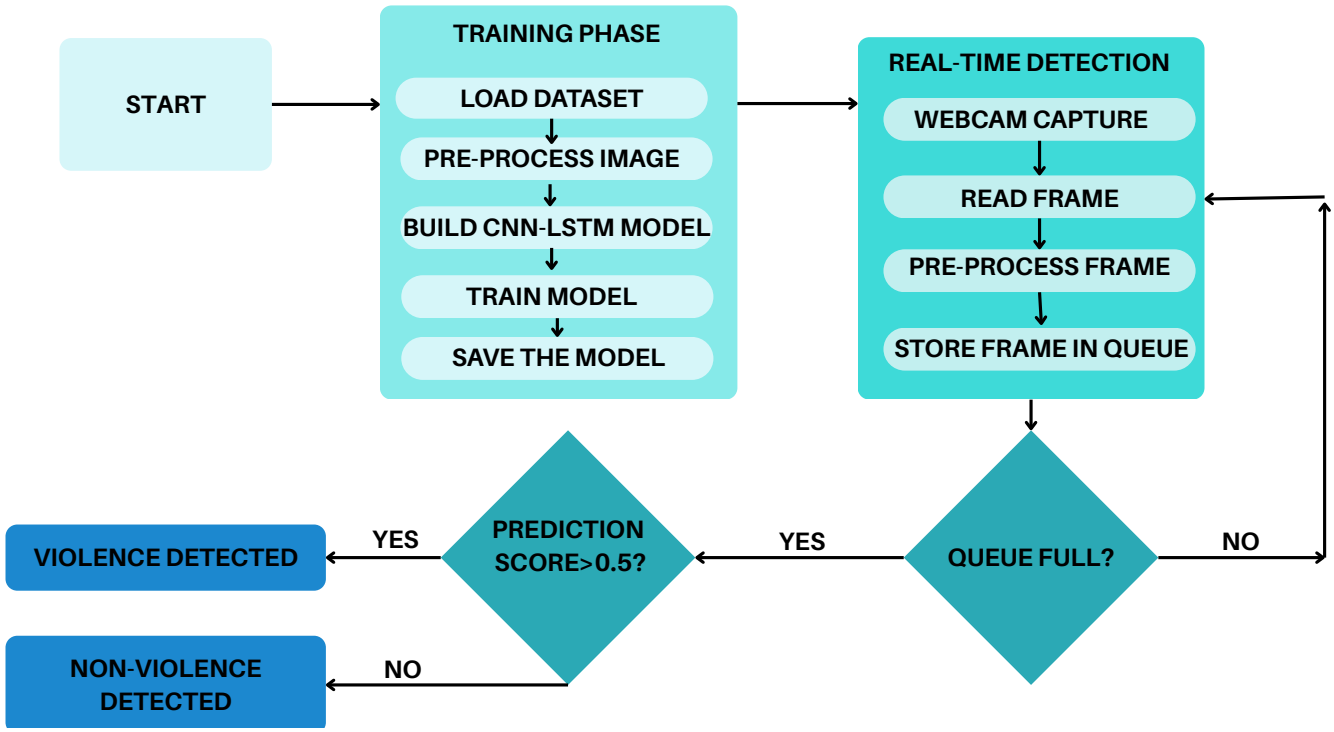
CAMERA OBSTRUCTION DETECTION MODULE



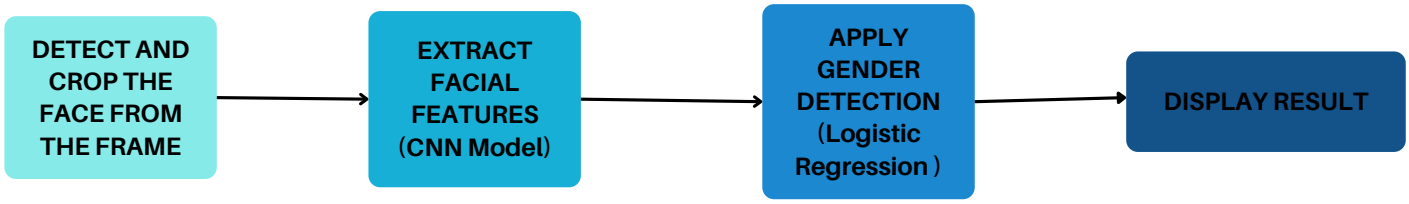
ANOMALY DETECTION MODULE



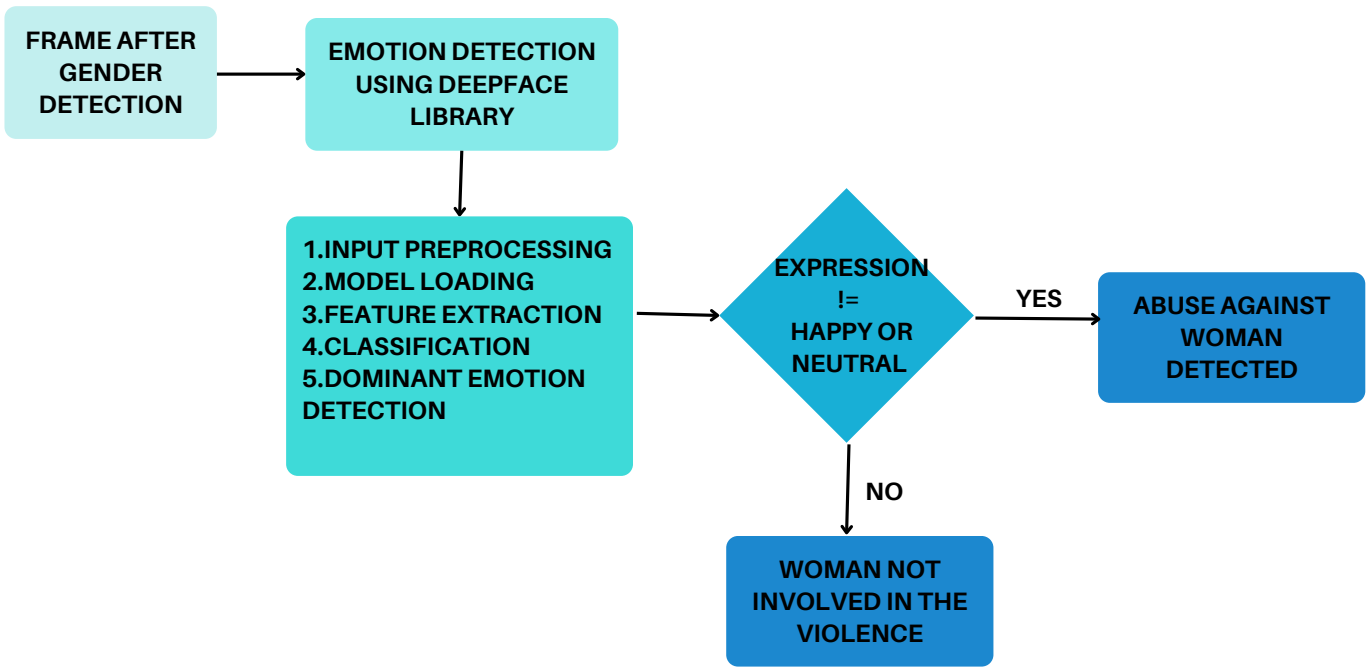
ANOMALY DETECTION MODULE



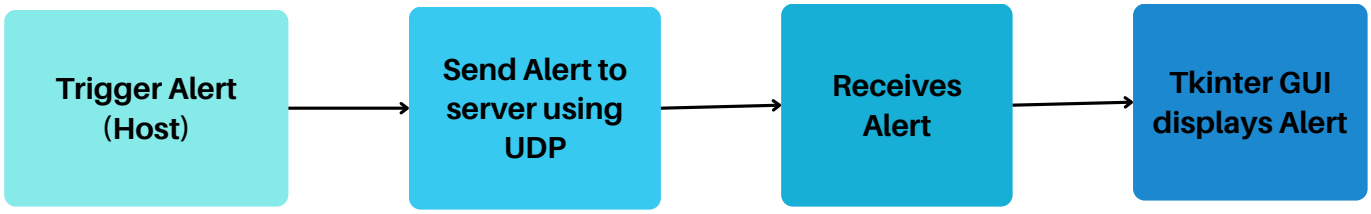
GENDER DETECTION MODULE



EXPRESSION DETECTION MODULE

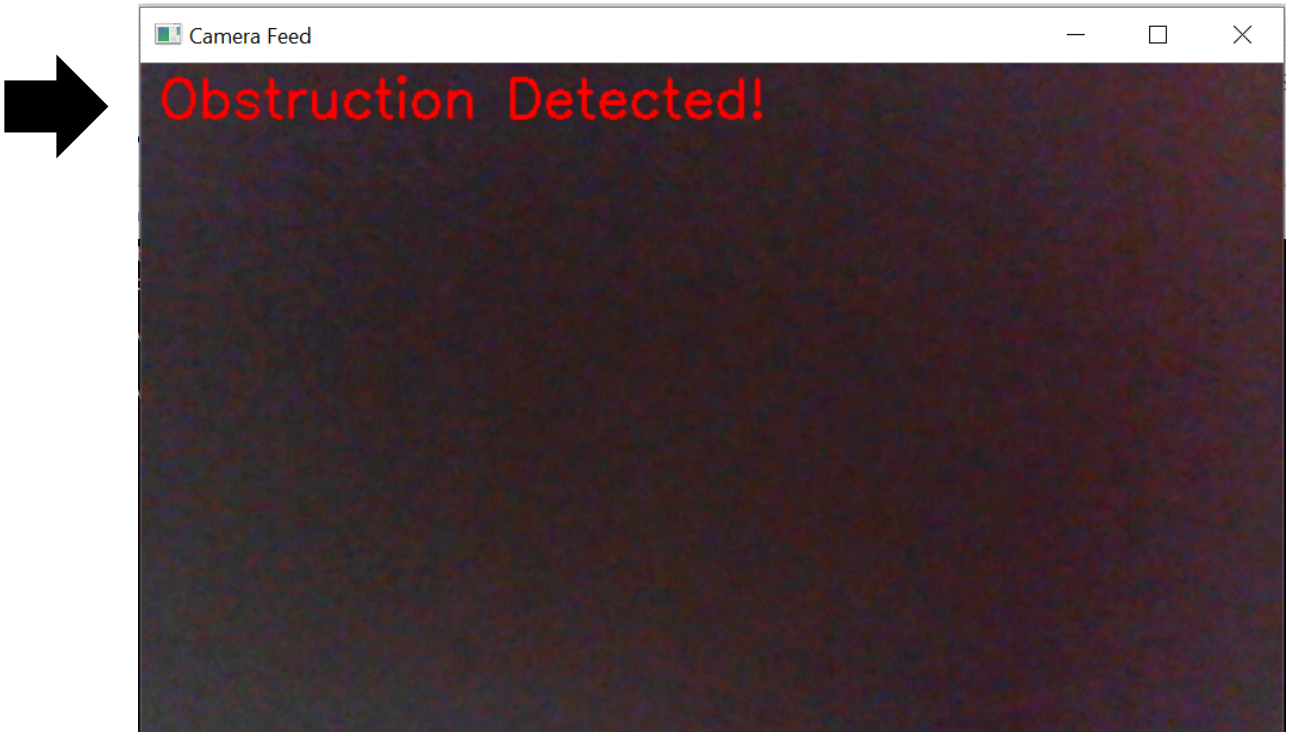


Alert System Module

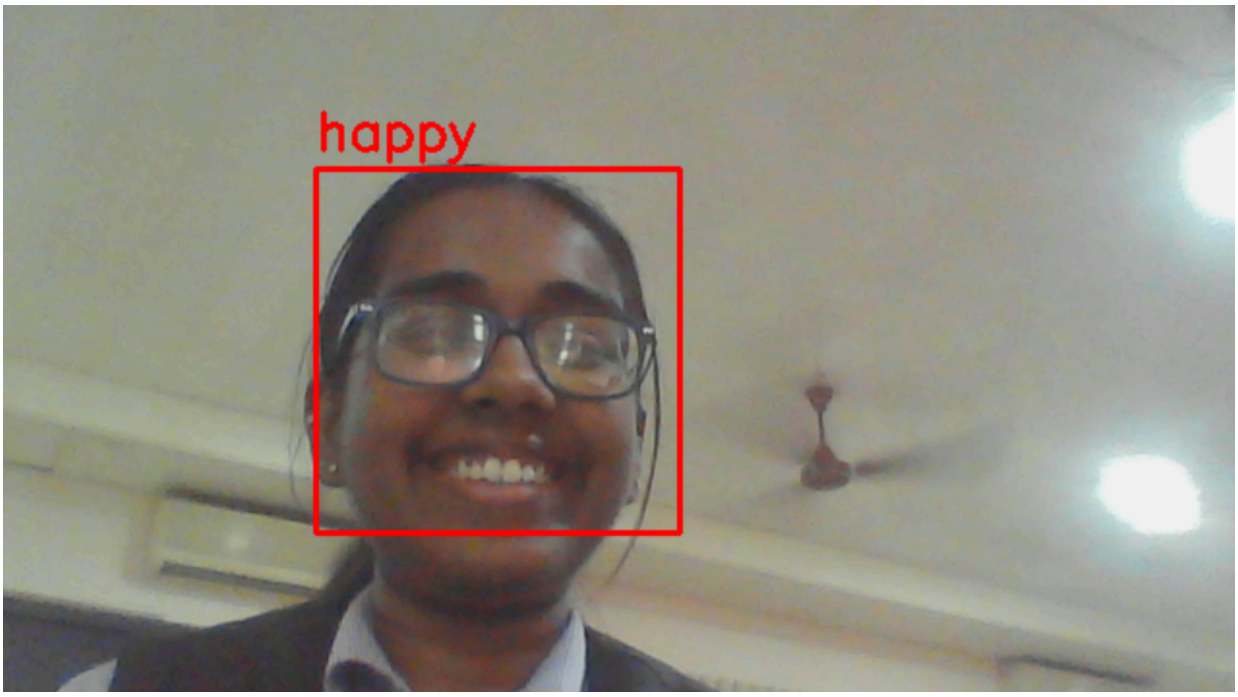


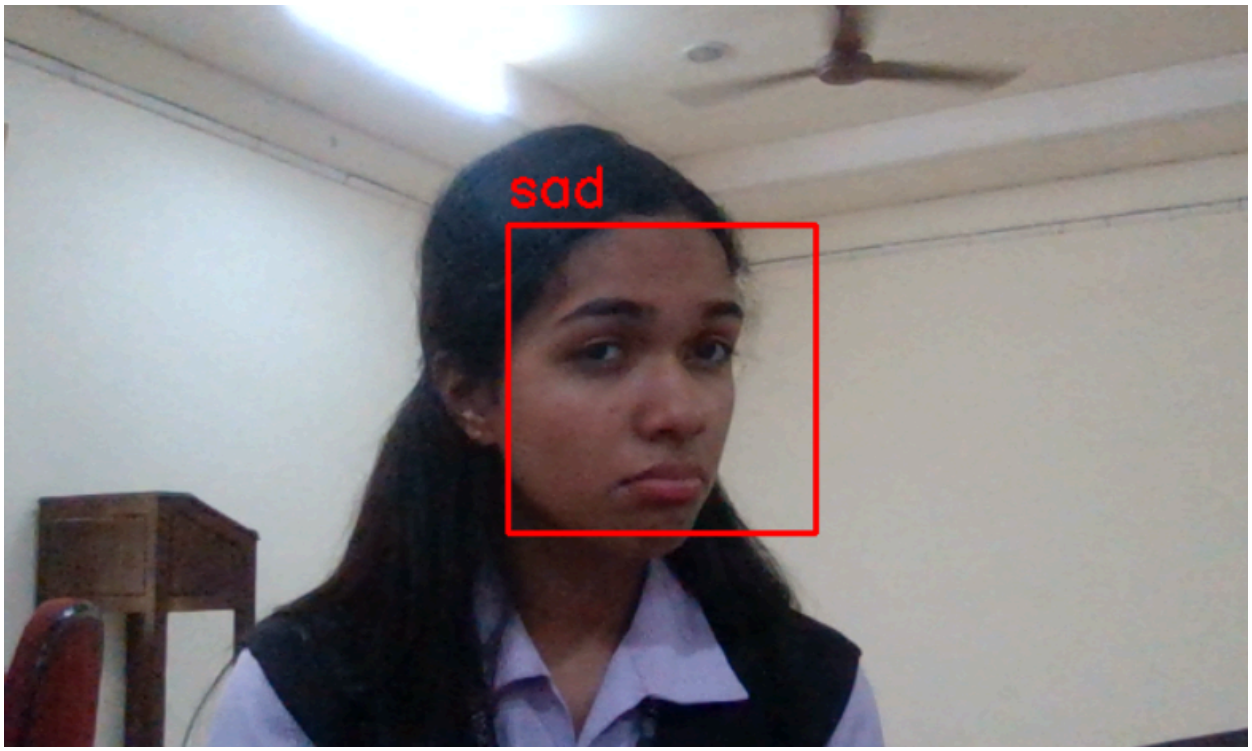
Camera Obstruction Detection

Camera Obstruction Detection

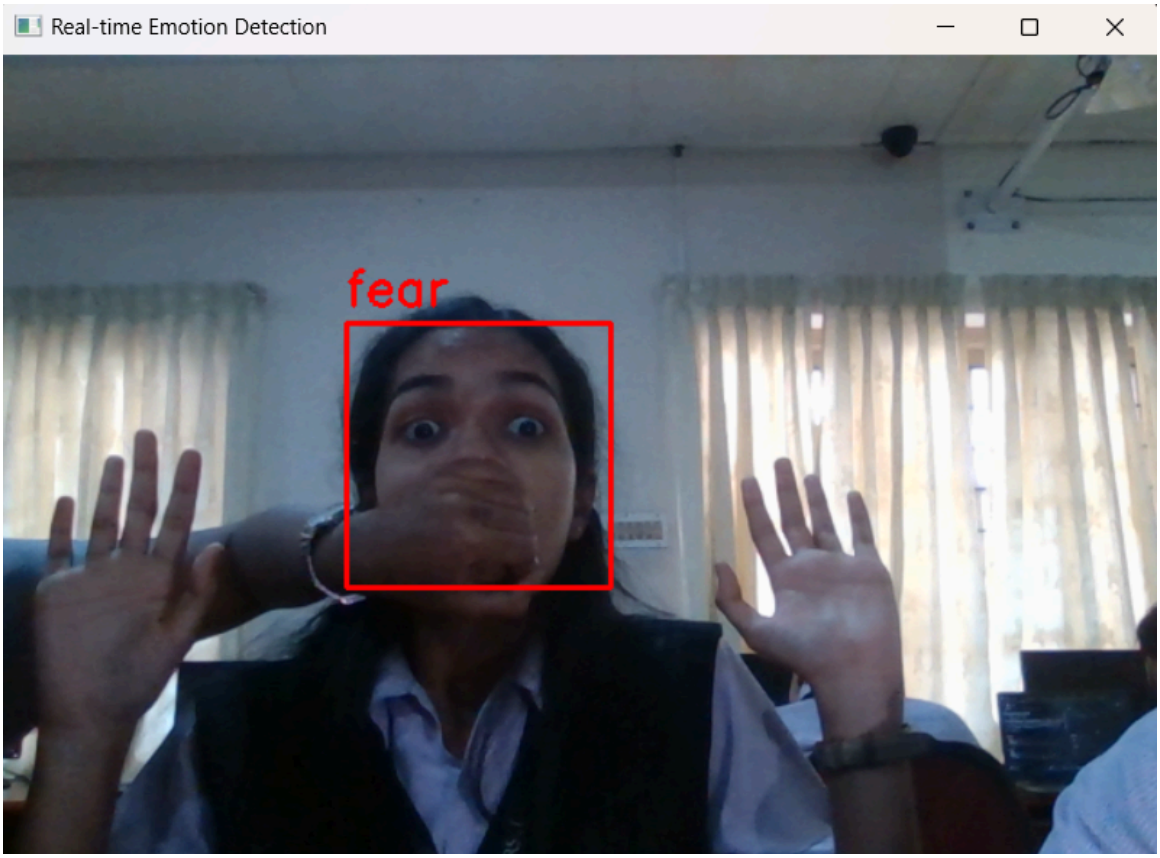


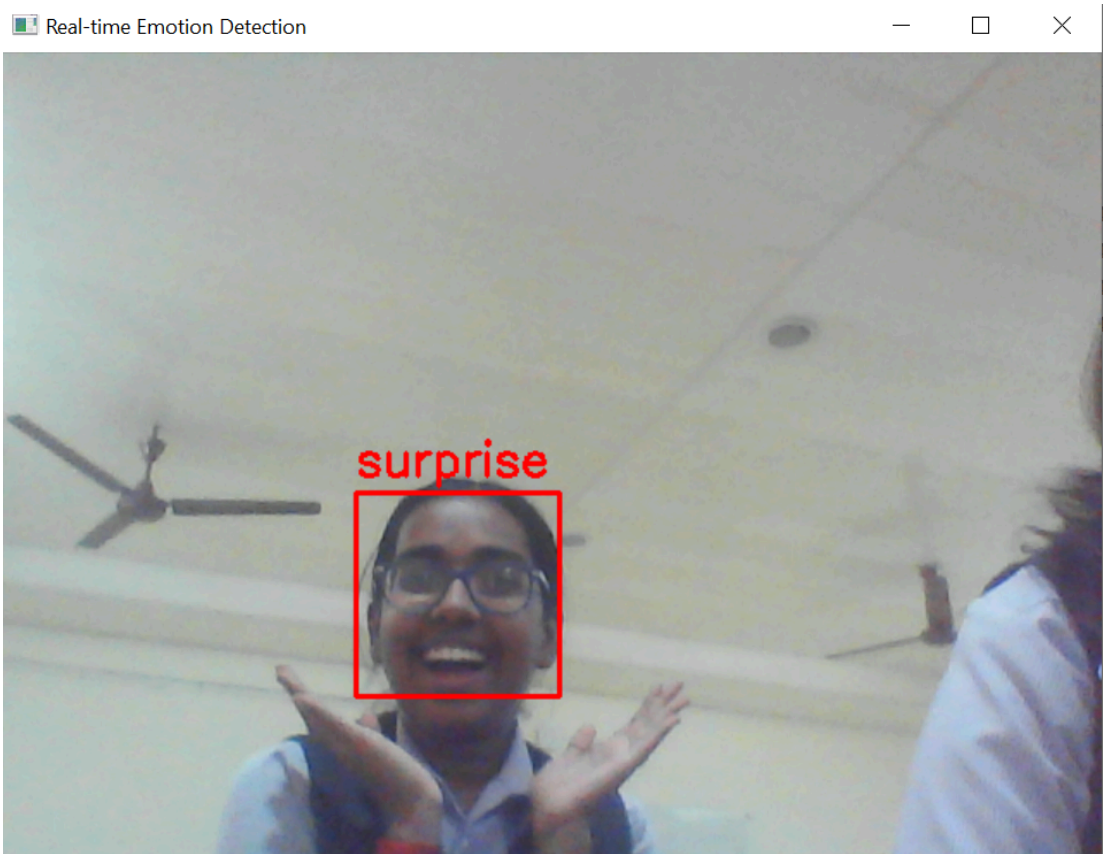
Expression Detection

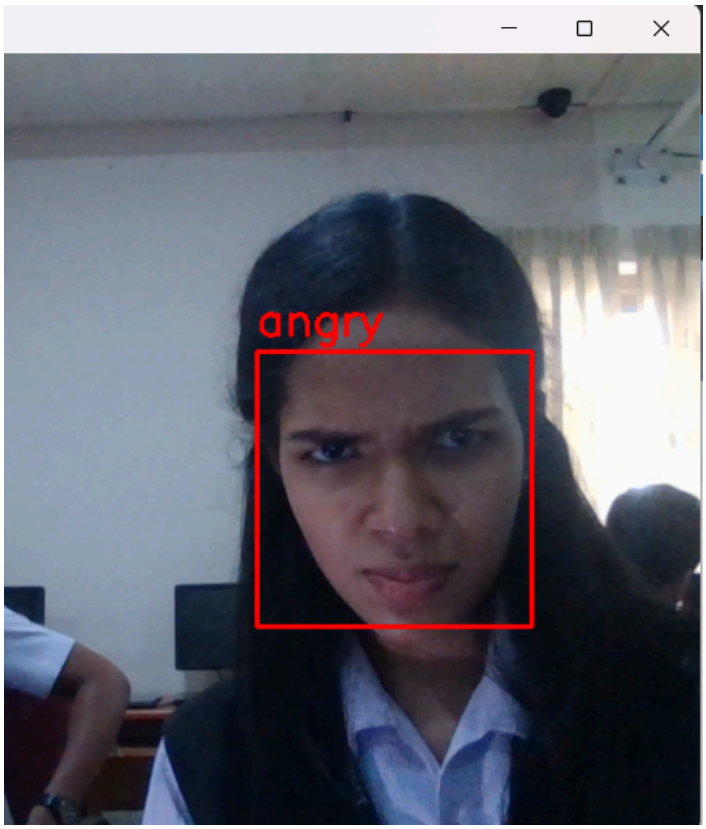


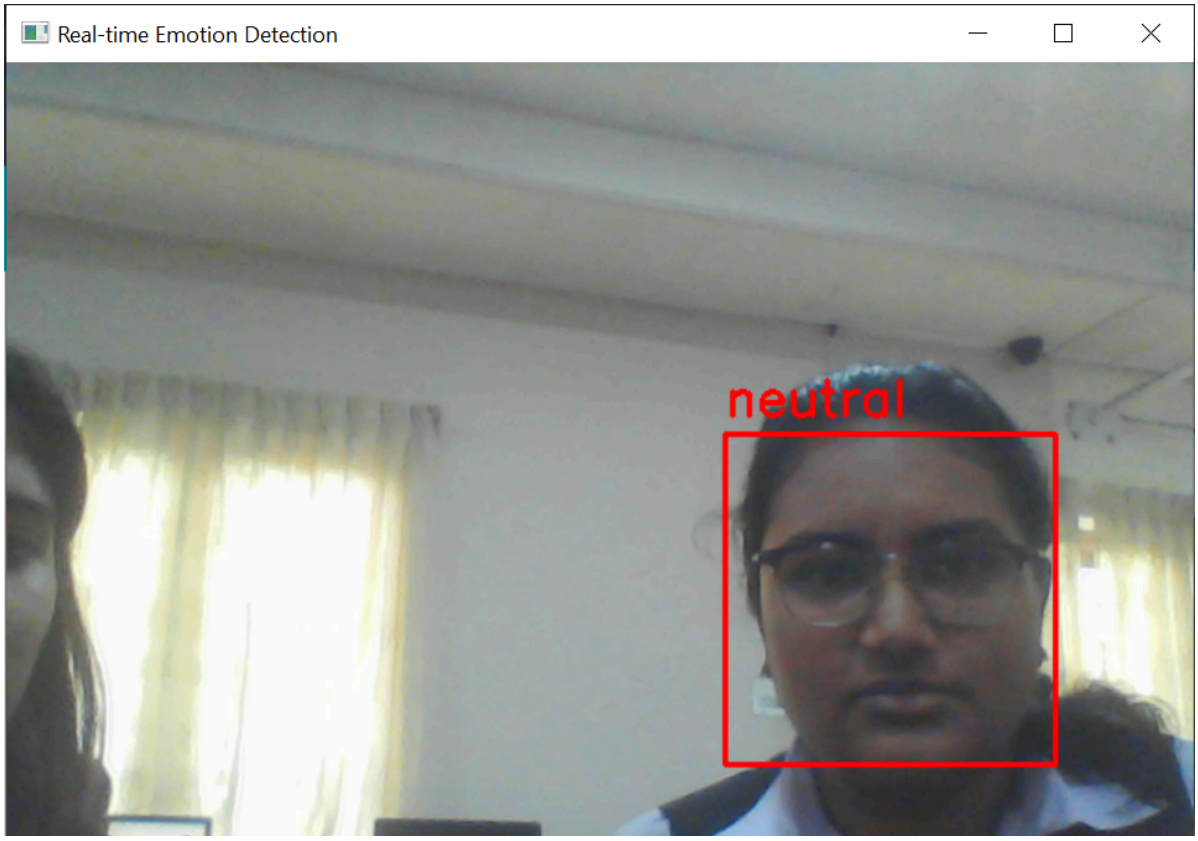

sad


fear

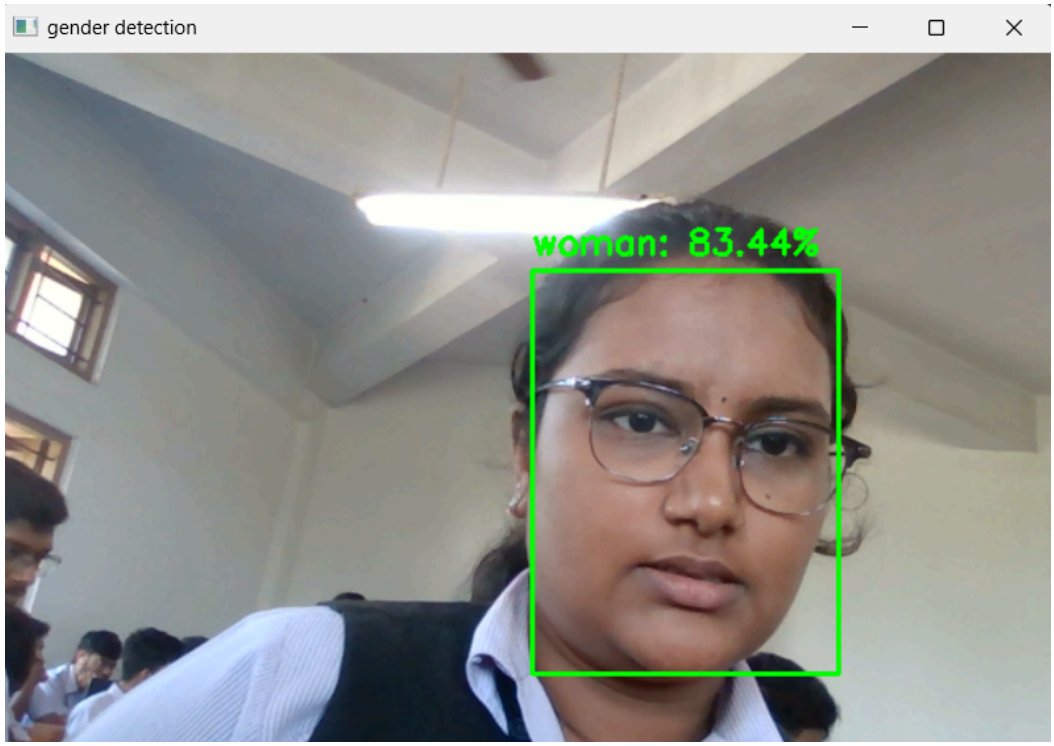

Real-time Emotion Detection
fear

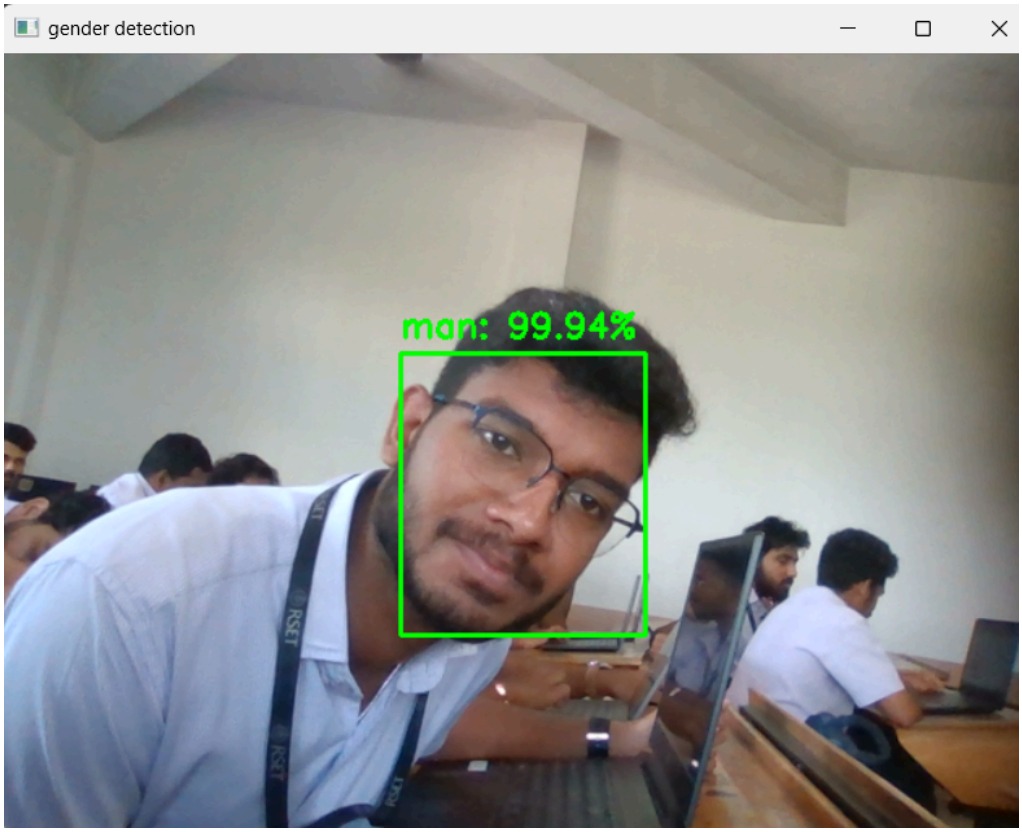

Real-time Emotion Detection
surprise

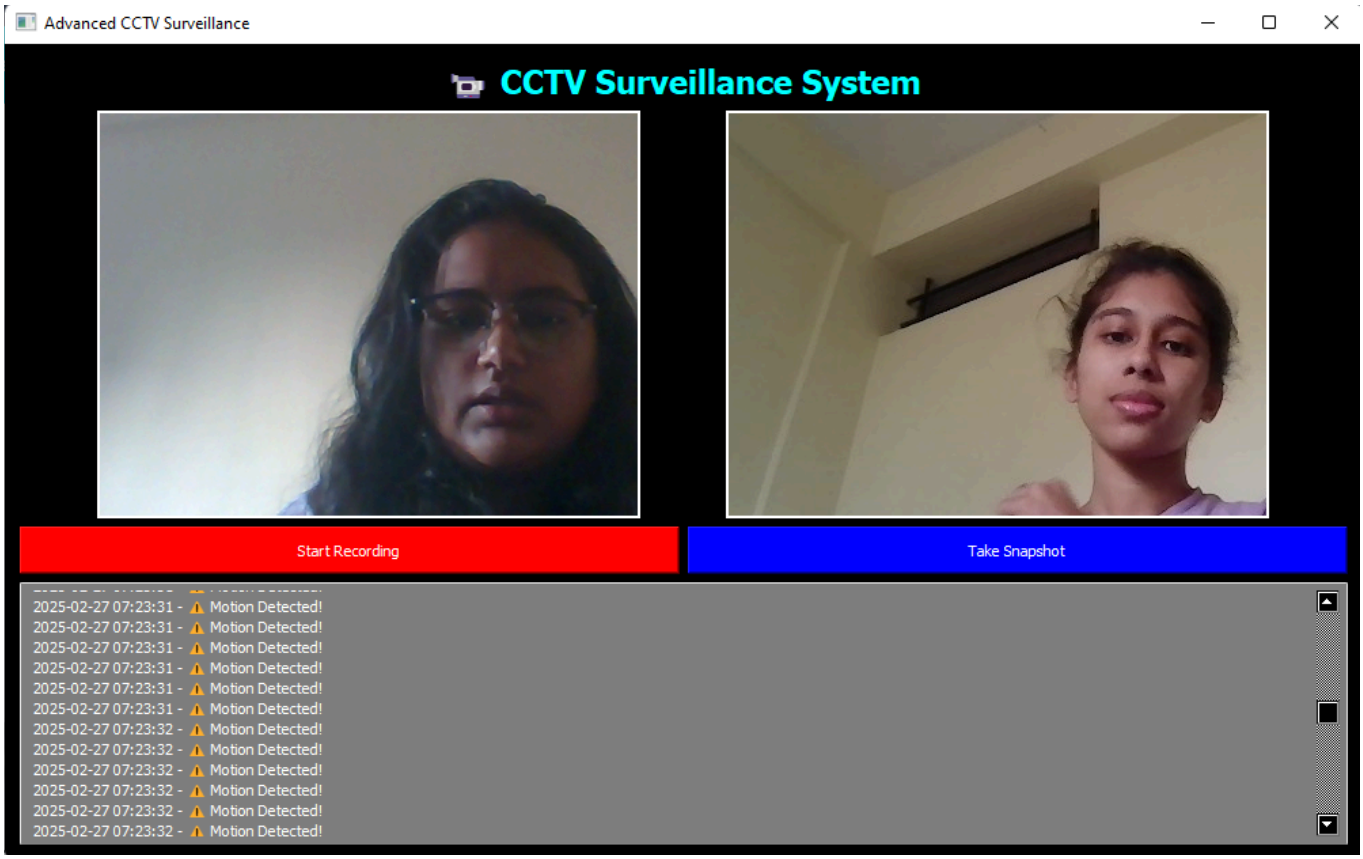

angry

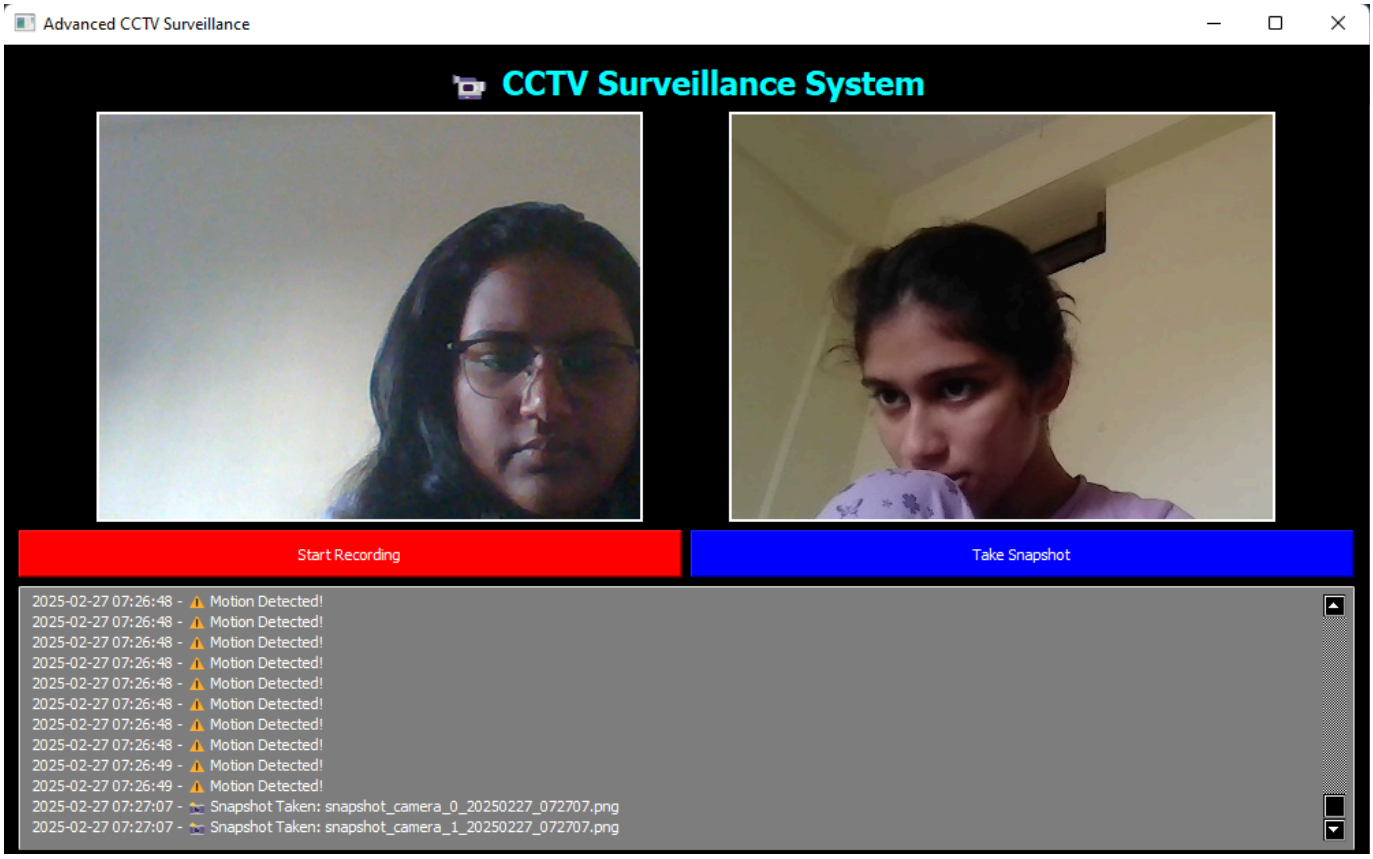

Real-time Emotion Detection
neutral

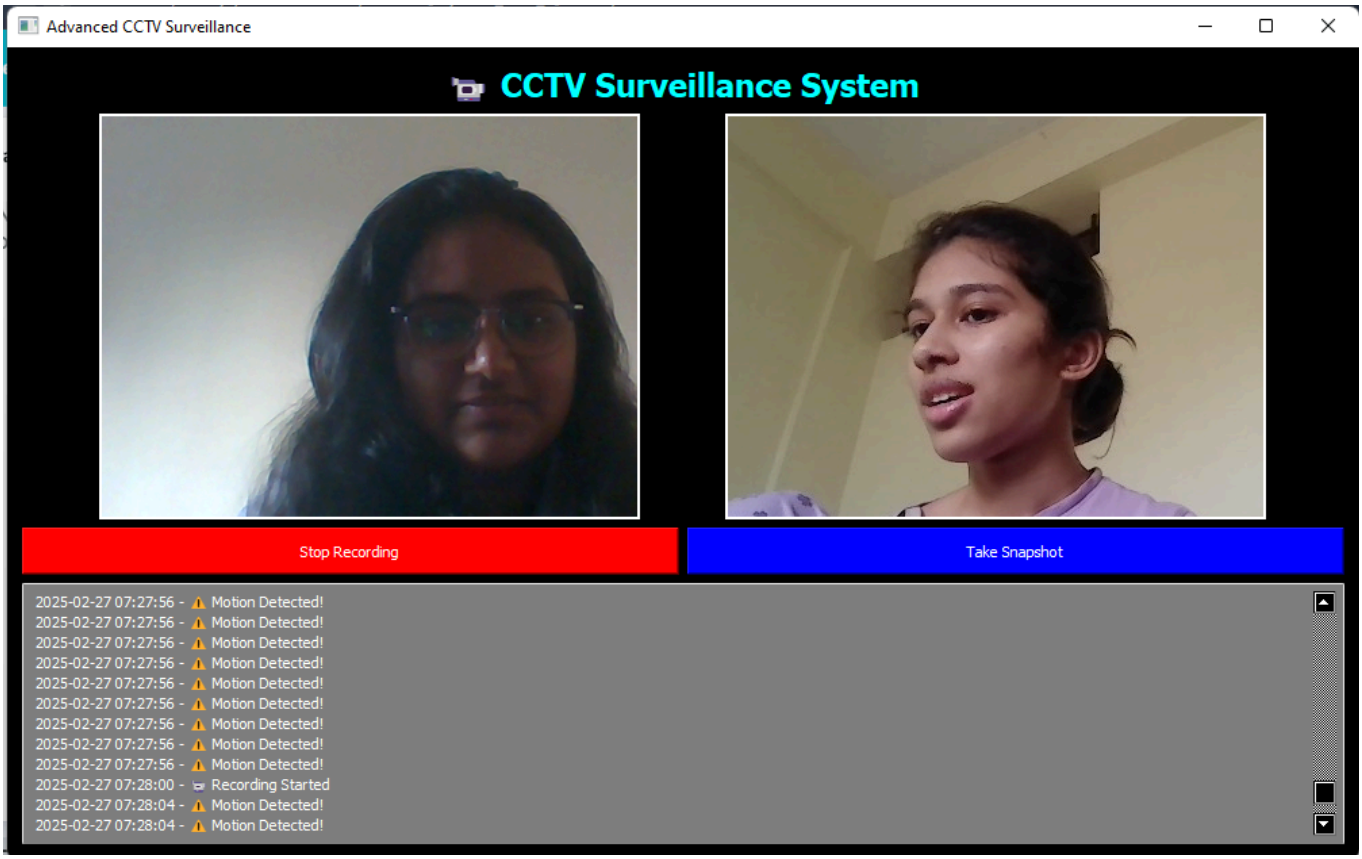
Gender Detection

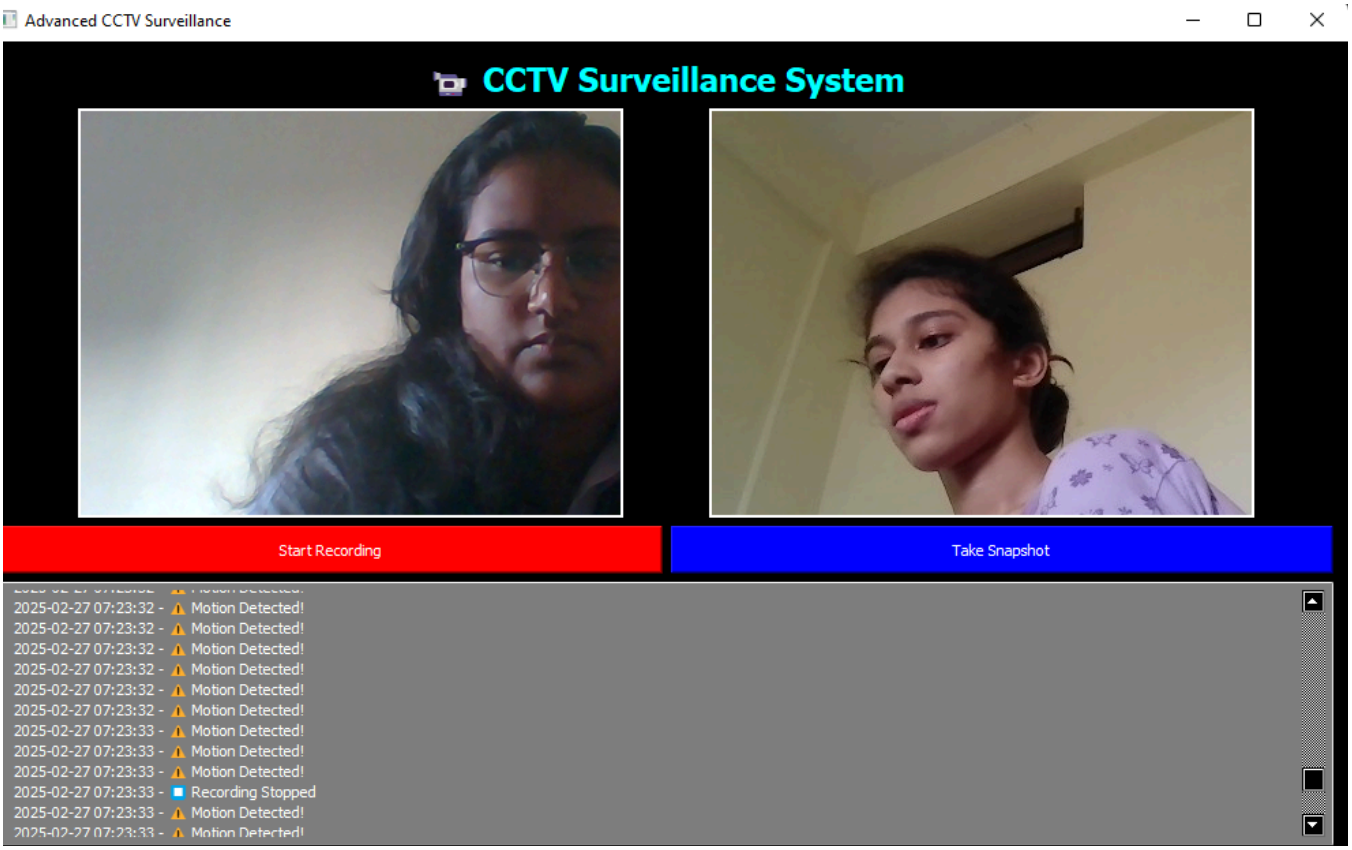


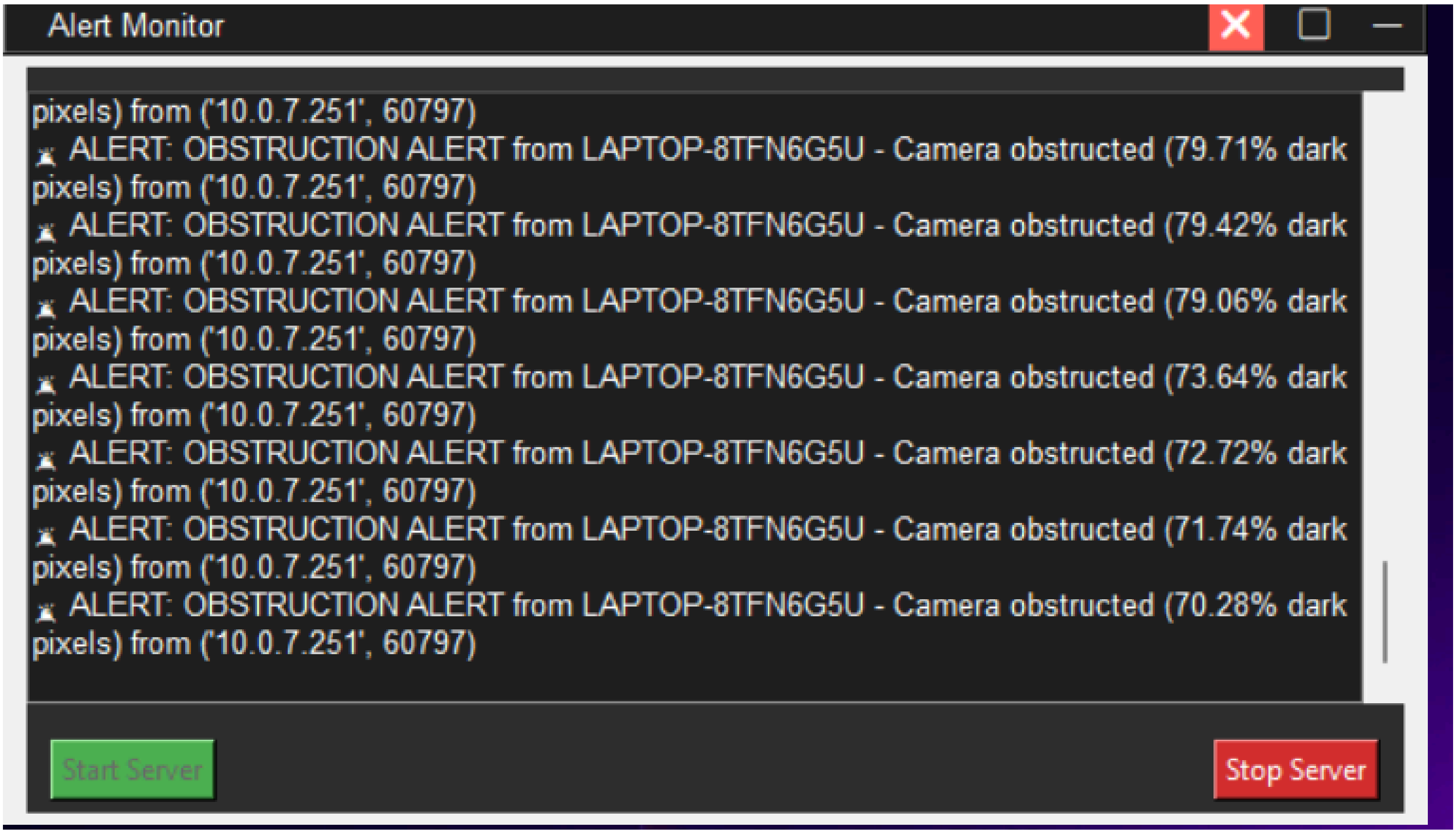


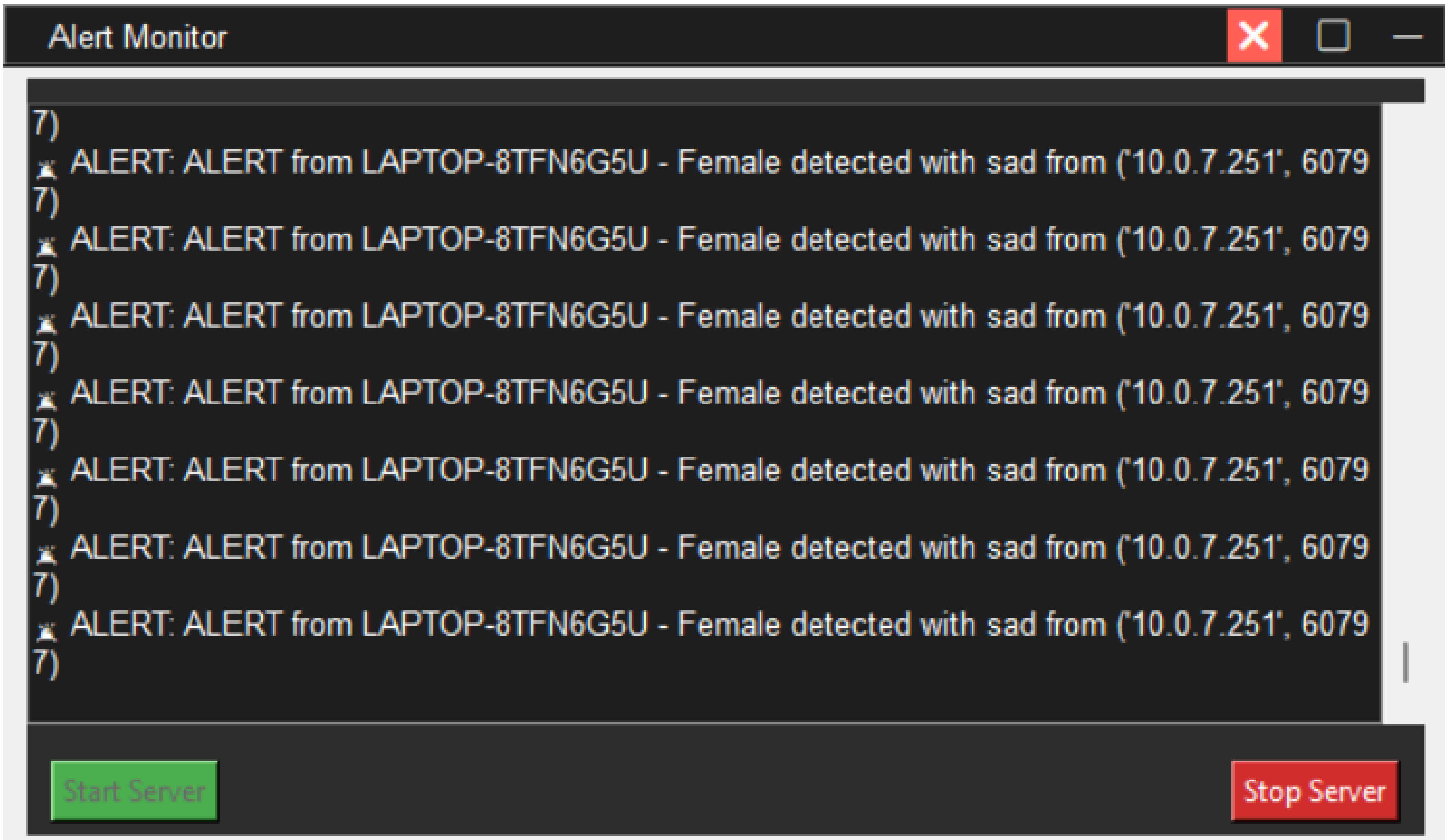

Advanced CCTV Surveillance
CCTV Surveillance System
Start Recording
Take Snapshot
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:31 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!


Advanced CCTV Surveillance
CCTV Surveillance System
Start Recording
Take Snapshot
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:48 - Motion Detected!
2025-02-27 07:26:49 - Motion Detected!
2025-02-27 07:26:49 - Motion Detected!
2025-02-27 07:27:07 - Snapshot Taken: snapshot_camera_0_20250227_072707.png
2025-02-27 07:27:07 - Snapshot Taken: snapshot_camera_1_20250227_072707.png


Advanced CCTV Surveillance
CCTV Surveillance System
Stop Recording
Take Snapshot
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:27:56 - Motion Detected!
2025-02-27 07:28:00 - Recording Started
2025-02-27 07:28:04 - Motion Detected!
2025-02-27 07:28:04 - Motion Detected!


Advanced CCTV Surveillance
CCTV Surveillance System
Start Recording
Take Snapshot
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:32 - Motion Detected!
2025-02-27 07:23:33 - Motion Detected!
2025-02-27 07:23:33 - Motion Detected!
2025-02-27 07:23:33 - Motion Detected!
2025-02-27 07:23:33 - Recording Stopped
2025-02-27 07:23:33 - Motion Detected!


Alert Monitor
pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (79.71% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (79.42% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (79.06% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (73.64% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (72.72% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (71.74% dark pixels) from ('10.0.7.251', 60797)
ALERT: OBSTRUCTION ALERT from LAPTOP-8TFN6G5U - Camera obstructed (70.28% dark pixels) from ('10.0.7.251', 60797)
Start Server
Stop Server


Alert Monitor
7)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
ALERT: ALERT from LAPTOP-8TFN6G5U - Female detected with sad from ('10.0.7.251', 60797)
Start Server
Stop Server

Conclusion

- This AI-powered Women Safety project enhances security by automating real-time anomaly detection, gender and expression recognition, and camera obstruction alerts.
- It improves safety, speeds up responses, and reduces human error, addressing the limitations of traditional surveillance.
- While challenges like privacy concerns and scalability exist, the project offers a significant step forward in modernizing and improving public safety efforts.

Thank you very much!

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Appendix B

Vision: To become a Centre of Excellence in Computer Science & Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Mission: To inspire and nurture students, with up-to-date knowledge in Computer Science & Engineering, Ethics, Team Spirit, Leadership Abilities, Innovation and Creativity to come out with solutions meeting the societal needs.

Program Outcomes:

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem analysis: Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice

PO9: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12: Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes:

PSO1: Computer Science Specific Skills: The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills: The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills: The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes

CO1: Model and solve real world problems by applying knowledge across domains.

CO2: Develop products, processes, or technologies for sustainable and socially relevant applications.

CO3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks.

CO4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms.

CO5: Identify technology/research gaps and propose innovative/creative solutions.

CO6: Organize and communicate technical and scientific findings effectively in written and oral forms.

Appendix C: CO-PO-PSO Mapping

Appendix C

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
CO 1	2	2	2	1	2	2	2	1	1	1	1	2	3		
CO 2	2	2	2		1	3	3	1	1		1	1		2	
CO 3									3	2	2	1			3
CO 4					2			3	2	2	3	2			3
CO 5	2	3	3	1	2							1	3		
CO 6					2			2	2	3	1	1			3

3/2/1: high/medium/low