Smart Surveillance: Ensuring Women's Safety Using Deep-Learning Algorithms

Veerraj Satish Chitragar
School of Computer Science and Engineering
KLE TECHNOLOGICAL UNIVERSITY
HUBLI-580031, KARNATAKA.
01fe22bcs164@kletech.ac.in

Avinash Valu Nayak
School of Computer Science and Engineering
KLE TECHNOLOGICAL UNIVERSITY
HUBLI-580031, KARNATAKA.
01fe22bcs204@kletech.ac.in

Yallavva Pujar
School of Computer Science and Engineering
KLE TECHNOLOGICAL UNIVERSITY
HUBLI-580031, KARNATAKA.
01fe22bcs148@kletech.ac.in

Mohammed Umar
School of Computer Science and Engineering
KLE TECHNOLOGICAL UNIVERSITY
HUBLI-580031, KARNATAKA.
01fe22bcs165@kletech.ac.in

Shashank Hegde
School of Computer Science and Engineering
KLE TECHNOLOGICAL UNIVERSITY
HUBLI-580031, KARNATAKA.
shashank.hegde@kletech.ac.in

***Abstract*—Our proposed work aims to improve women's safety by utilizing advanced machine learning methods for the detection and evaluation of surveillance video footage. The suggested system utilizes the YOLO11 (You Only Look Once) module for accurate individual detection, guaranteeing precise tracking in monitoring environments. A Huggingface model is used for gender classification to recognize women among the identified individuals. To tackle safety issues, the system employs a ViT-B/32 (Vision Transformer) model to identify occurrences of violence or possibly hazardous actions. Our multi-step model reached an overall accuracy of 72.9%, indicating the collective effectiveness of all consecutive elements. This extensive framework is intended to function in real-time, consistently tracking and assessing safety hazards. Its scalability enables smooth incorporation into current surveillance systems and adjustment to various environments, enhancing safety protocols. The system is evaluated using live video feeds, sending alerts through a chatbot when threats are detected. Alerts with essential details—like the count of males and females, emotional conditions, and signs of aggressive actions—were transmitted to a chatbot.**

***Index Terms*—Surveillance footage, Women safety, ViT-B/32, YOLO, Person detection, Gender detection, Huggingface model, Violence detection, Real-time monitoring, Deep learning, Machine learning.**

I. Introduction

Securing the safety of women in both public and private areas continues to be a major worldwide issue. With the rise of urbanization and increasing population density in cities, traditional security methods are frequently inadequate in dealing with the increasingly complex safety issues. Surveillance systems have always been a key tool for monitoring public areas; however, the large amount of data collected makes it extremely difficult to analyze manually. This has resulted in an increasing need for smart, automated systems that can monitor and analyze in real-time. Advances in machine learning and computer vision have led to new opportunities to improve surveillance systems by accurately detecting potential threats.

Our proposed work uses advanced machine learning techniques to create a reliable monitoring system focused on improving safety measures for women. Our methodology centers around combining various cutting-edge AI techniques through a multi-model approach. The initial stage consists of identifying people in surveillance videos, using the YOLO (You Only Look Once) module [4]. YOLO is a popular framework for object detection in deep learning that is recognized for its high speed and precision, making it perfect for applications that require real-time performance [11]. With the use of YOLO, our system is able to effectively identify and pinpoint people in intricate surroundings, establishing a strong basis for additional examination [17]. After identifying individuals, the system uses a Huggingface model to conduct gender classification [5]. This stage is essential in narrowing the focus of the analysis to women, enabling the system to specifically concentrate on evaluating their safety in public or controlled spaces.

While the initial analysis involves detecting person and gender, the main difficulty is in recognizing situations that could be harmful. In order to tackle this issue, we integrate a Vision Transformer model, namely the ViT-B/32 (Vision Transformer Base with 32x32 patches) [6], which has proven to be highly successful in different computer vision assignments. The ViT-B/32 model is employed to identify violent actions, a crucial sign of possible dangers to women's well-being. Vision Transformers provide a strong foundation for image categorization assignments, and their capability to analyze intricate correlations in visual information renders them especially well-suited for detecting violence. This model examines the identified people and their interactions, pinpointing potential threats through aggressive or suspicious behavior [2].

Our proposed work does not limit itself to only technical improvements in surveillance technology but solves an important social prob- lem by providing a proactive response to the safety of women. With reports of harassment and violence persisting in various environments, such as streets and schools, there is a pressing requirement for smart systems that can aid in identifying and stopping these incidents early. Our goal is to address this requirement through creating a solution that utilizes up-to-date AI tools and offers guidance on the ethical and effective use of such technologies to safeguard vulnerable populations. In Our proposed work , we imagine a future in which surveillance systems go beyond just recording events passively and instead actively contribute to ensuring the safety and security of everyone, with a specific focus on women. Our proposed work discusses the process, execution, and assessment of our suggested system, emphasizing the effectiveness of utilizing YOLO for person identification [15], a Huggingface model for gender categorization, and the ViT-B/32 model for detecting violence [19]. Our findings show that combining these models in one framework greatly enhances the ability to detect safety risks, opening up possibilities for advanced AI-based surveillance systems [8].

The paper is divided into the following sections. Section2 explains existing surveillance technologies and carries out a comparative analysis of deep learning-based methods for person detection, gender classification, and violence detection. It further goes on to explain advancements in YOLO, Vision Transformers, and other main components utilized in the proposed system. Section3 discusses the limitations of traditional surveillance systems, such as high computational requirements and false positives, and introduces a multi-model framework that combines YOLO, Huggingface models, and Vision Transformers to improve reliability and accuracy. Section4 is related to the proposed methodology that describes the pipelines involved for person detection, gender classification, facial expression recognition, pose estimation, and violence detection, finally putting together these models for unified use in real-time monitoring. Section5 will focus on the experimental results regarding the performance of the presented system compared to other already existing methods in terms of accuracy, latency, and robustness, with validation of the effectiveness of various components and their combined impacts on ensuring women's safety. Finally, Section6 concludes the paper and summarizes key findings and discusses potential future directions for improving the scalability, precision, and ethical considerations of AI-based surveillance systems.

II. Literature Survey

Ensuring the safety of women in public spaces like urban areas has been a concern globally. Traditional safety measures such as police patrolling, and personal safety devices are foundational but often reactive, relying on manual interventions will result in delayed responses. There are some wearable devices which need user activation which may lead to delay in responses and not feasible in high-risk situations. A smartphone application that uses machine learning to identify threats and deliver real-time notifications was proposed by Shankar et al. [20]. Though it mostly concentrates on mobile-based interventions, which still depend on user-triggered activities, this system incorporates automation to speed up reaction times. Surveillance systems have also used machine learning. By examining behavioral patterns and spotting questionable activity in real time, Gupta et al. [9] showed how video analytics may be used to identify threats against women.

Jimbo [10] investigates the progress made in real-time object detection with the launch of YOLO11, emphasizing its potential to enhance surveillance and security functions. The article examines how YOLO11 improves both the accuracy and speed of object detection, especially highlighting its efficacy in detecting and counting people in challenging settings. The incorporation of YOLO11 into surveillance systems presents considerable opportunities for instantaneous analysis, rendering it a useful resource for applications like crowd observation, security monitoring, and hazard identification.

Their research supports the use of AI in early violence detection. Advanced AI technologies have greatly affected public safety. Naved et al. [14] presented a real-time AI-driven system that incorporates facial recognition, behavior analysis, and automated alerts to improve the security of women. Their method bridges the gaps in proactive interventions by linking AI tools with the police. Similarly, Aktı et al. [1] introduced a vision-based fight detection system based on gesture and behavior recognition from surveillance cameras that can identify violent activities.

Recent developments in wearable tech for women's safety have incorporated AI and IoT to improve security. An important instance is the Safe-Guard device, utilizing AI for detecting threats, geofencing, voice identification, and live location tracking to offer prompt help during emergencies. Nonetheless, issues like sensor precision, device loss, and possible technical malfunctions emphasize the necessity for additional enhancements in these systems Devarakonda et al. [16].

Michelle et al. [13] investigated the application of deep learning in video analysis for intelligent surveillance systems focused on women's safety. Their method highlights immediate identification of threats via sophisticated neural networks that evaluate body language, gestures, and interactions. By enhancing detection precision and minimizing false positives, their system provides prompt notifications to authorities, allowing it to be suitable for various settings such as busy city areas or remote locations. This advancement underscores the transformative power of AI in improving conventional surveillance for proactive safety initiatives.

Violence detection systems are improving with multi-modal data integration. Vijeikis et al. [21] showcased the efficacy of video and audio cues for surveillance footage when detecting aggressive behavior considerably lowering false positives. Dosovitskiy et al. [6] also proposed Vision Transformers (ViTs), which adopt the self-attention mechanism within it, to realize image recognition. Its scalability and adaptability make it apt for processing complex surveillance footage in real time and therefore to enhance anomaly detection. Despite these advancements, existing systems face challenges, such as high false positive rates, limited contextual understanding, and significant computational demands. This research addresses these gaps by proposing an integrated AI-based safety system. Combining real-time video surveillance, person detection, gender classification, violence detection and automated alerts, the system aims to provide both preventive and proactive measures.

III. Methodology

The method utilizes a comprehensive strategy that begins with processing input videos to detect individuals utilizing the YOLO model. Identified individuals are subjected to gender identification and head counting, as well as movement monitoring. The system evaluates if an individual is isolated, examines their environment, and monitors pursuit actions. Additionally, it assesses hazardous scenarios like time zones, fear-related emotions, and possible violence employing the ViT-B/32 model. When a threat is detected, the system produces alerts, records frames, and notifies authorities for prompt response, ensuring an efficient real-time monitoring system.

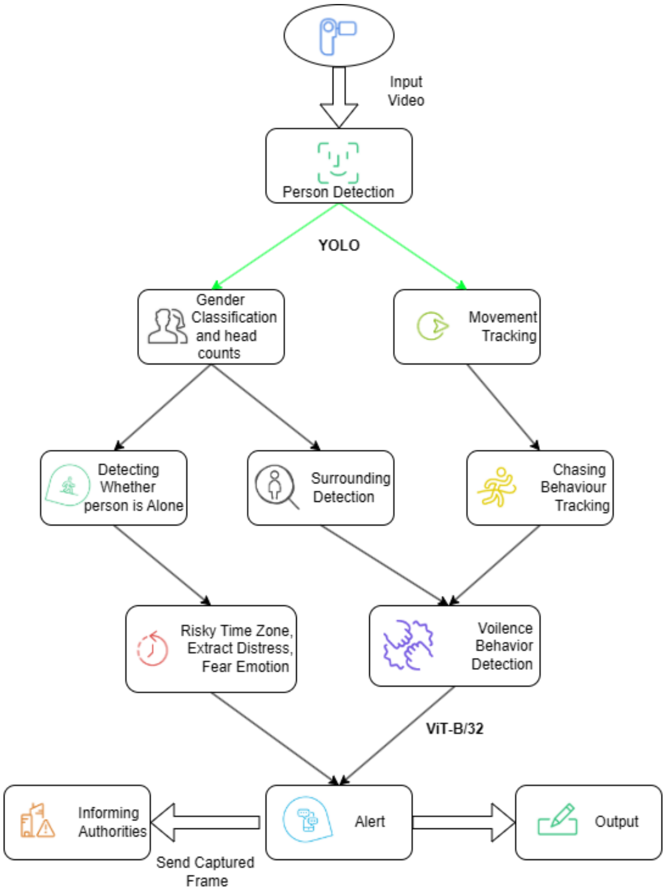
Input Video
Person Detection
YOLO
Gender Classification and head counts
Movement Tracking
Detecting Whether person is Alone
Surrounding Detection
Chasing Behaviour Tracking
Risky Time Zone, Extract Distress, Fear Emotion
Voilence Behavior Detection
ViT-B/32
Informing Authorities
Alert
Output
Send Captured Frame

Fig. 1. Workflow of proposed methodology

A. Person Detection

The initial stage in the system's process is identifying and pinpointing individuals in the surveillance video. To achieve this goal, we utilize the YOLO11 model, a sophisticated real-time object detection algorithm renowned for its exceptional accuracy and speed. YOLO11 analyzes video frames to recognize and pinpoint individuals, generating bounding boxes around each person detected shown in fig2. This stage is essential for monitoring movement in the surveillance setting and serves as the base for all future analyses. The ability of YOLO11 to function effectively on live video feeds makes it ideal for constant surveillance in changing environments.

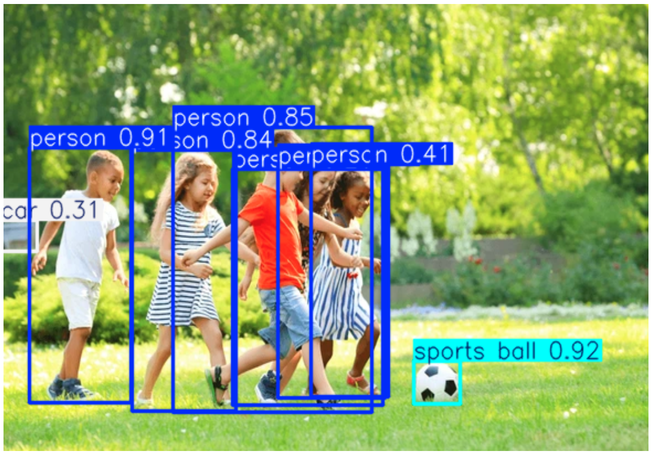

Fig. 2. Real-time object detection with YOLO11 accurately identifies several individuals and objects.

B. Gender Classification

After identifying individuals, the system carries out gender classification to distinguish between males and females. This stage involves using a pre-trained Huggingface model that is specifically designed for detecting gender. The model analyzes each identified person, using visual features to forecast their gender. This classification of gender helps the system concentrate on women in order to facilitate focused examination of women's safety. Moreover, the system also tracks the number of males and females in the observed area to enable demographic analysis of the crowd. This information is helpful in spotting situations with potential safety hazards, like locations with an uneven distribution of male and female individuals.

C. Facial Expression Recognition

The system includes a facial expression recognition module to understand the emotional state of identified females. This module is created to assess the facial expressions of recognized females, offering information on their emotional condition—like happy, fear, anger, or distress depending on the thresholds defined shown in fig3. Recognizing the emotional background is crucial in evaluating the well-being of people, as specific facial cues may signal uneasiness or suffering, prompting timely intervention. The model for recognizing facial expressions utilizes a classifier based on deep learning that has been trained on a varied dataset of facial emotions to guarantee accuracy in various ethnicity and situations.

$$\text{Expression} = \begin{cases} \text{Happy} & \text{if Mouth Width} > 0.05 \text{ and} \\ & \quad \text{Smile Curve} < 0 \\ \text{Surprise} & \text{if Eyebrow-Angle} > 20 \text{ and} \\ & \quad \text{Mouth Openness} > 0.03 \\ \text{Distress} & \text{if Eyebrow-Angle} > 10 \text{ and} \\ & \quad \text{Mouth Openness} > 0.02 \\ \text{Neutral} & \text{otherwise} \end{cases}$$

Expression Detection Formula (Threshold Logic) [22]

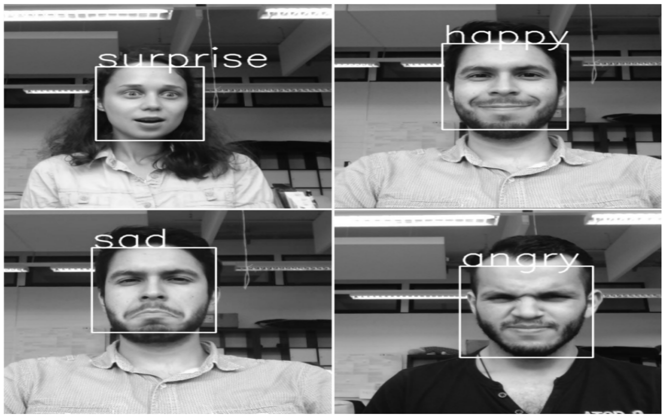

Fig. 3. Emotion-Recognition using huggingface model [3]

D. Pose Detection

After analyzing emotions, the system evaluates the posture and body language of identified females to better understand possible risks. We make use of the MediaPipe Pose Library, which is a tool for estimating body pose in real time, to analyze the body pose of every woman in the scene. MediaPipe Pose identifies crucial body landmarks as shown in fig4, enabling the system to evaluate both posture and movement such as sitting position shown in fig5. This examination is able to detect indicators of stress, like defensive stances or unpredictable actions, that could signal possible danger. Pose estimation offers more information to the analysis of facial expressions, allowing for a more complete evaluation of an individual's safety.

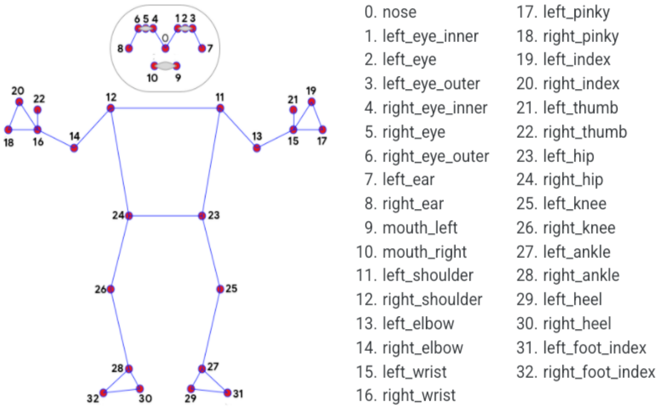

Fig. 4. Pose estimation utilizing keypoints to map the human skeletal framework for analyzing movement [12].

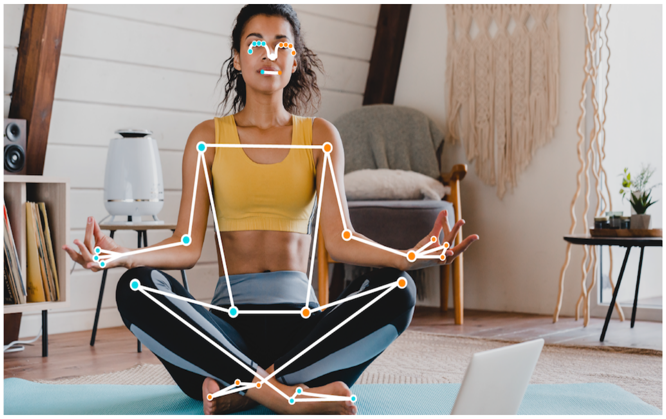

Fig. 5. Sitting posture analysis generated by the mediapipe pipeline model [18]

E. Violence Detection

The last stage of the system's analytical process is to identify potentially aggressive behavior. We utilize the ViT-B/32 model, which is a cutting-edge deep learning model known for its strong image classification abilities. The ViT-B/32 model analyzes visual information to detect aggression or violence, which are important signs of potential safety risks. The Vision Transformer model excels in analyzing intricate relationships in video frames, making it perfect for detecting both overt and nuanced indicators of violence.

Patch Embedding: The image is divided into non-overlapping patches, each flattened and linearly embedded. This is represented as:

$$\mathbf{z}_i = \text{Flatten}(\mathbf{p}_i) \cdot W_e + \mathbf{e}_i \tag{1}$$

Where in equation 1, $\mathbf{p}_i$ is the i-th image patch, W_e is the patch embedding matrix , $\mathbf{e}_i$ is the positional encoding for each patch.

Self-Attention Mechanism: The self-attention mechanism enables the model to capture dependencies between patches by computing relationships using the following formula:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} \tag{2}$$

Where in equation2, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices derived from the patch embeddings, d_k is the dimensionality of the queries and keys.

Final Classification: After passing through multiple layers of attention and transformation, the output embeddings are classified using a fully connected layer:

$$\text{output} = \text{softmax}(W_f \cdot \mathbf{z}_{\text{final}} + b_f) \tag{3}$$

Where in equation 3 [7], W_f is the weight matrix for classification, b_f is the bias term, $\mathbf{z}_{\text{final}}$ is the final embedding output after processing through the transformer layers.

Through the examination of the behaviors and engagements of recognized individuals, the system is able to rapidly pinpoint situations that could potentially endanger the safety of women.

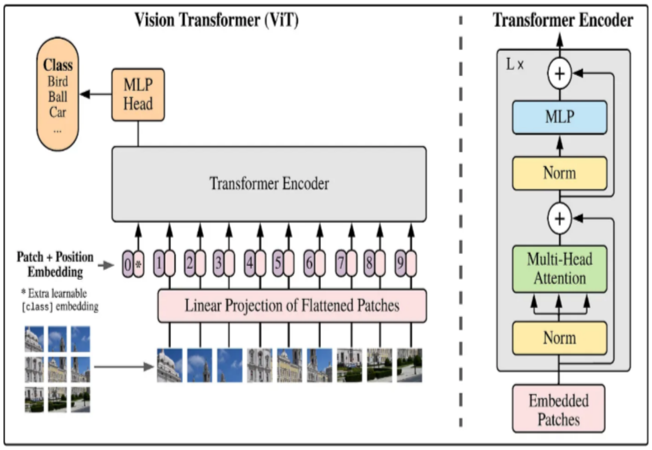

Fig. 6. ViT-B/32 architecture [6]

F. Integration and Monitoring

The whole system is combined in one framework built for monitoring in real-time. Individuals are first detected by YOLO11 as video frames are input into the pipeline. The Huggingface model is utilized for gender classification, with a separate count kept for males and females. For women, additional analysis is performed by the system, such as recognizing facial expressions and estimating poses using MediaPipe. At the same time, the ViT-B/32 model keeps constant watch for any indications of violent behavior in the scene. All assessments are carried out simultaneously to guarantee low delay, enabling quick identification of possible risks.

G. Alert Signal

When a danger is identified—drawing from various visual indicators like fear/distress in facial expressions, defensive body pose, or the occurrence of violent actions—the system produces an automatic notification. This notification contains comprehensive details, including the gender demographics, identified emotions, and pertinent pose analysis information. The notification can be directed to security staff or shown on a monitoring dashboard, allowing quick action against possible safety risks. Moreover, the system gathers analytical data over time, enabling trend analysis and deeper understanding of behavioral patterns that might reveal ongoing safety risks.

H. Evaluation and Validation

The proposed system is evaluated using a comprehensive dataset that includes diverse surveillance footage representing different scenarios and environments. The effectiveness of each component—YOLO11 for person detection, the Huggingface gender classification model, facial expression analysis, pose estimation, and ViT-B/32 for violence detection—is assessed individually and collectively. Metrics such as precision, recall, F1-score, and processing latency are used to gauge the performance of the system. A series of experiments are conducted to validate the system's accuracy in real-time, ensuring its reliability and robustness in various contexts.

The approach is split into various crucial stages, with each handling a distinct part of the surveillance analysis process as shown in fig1. The methodology is explained in great detail below.

IV. Results

The proposed system was efficiently tested on video in real-time, as shown in Fig.7, issuing safety notifications via a chatbot when potential dangers were identified. Table I presents the accuracy of the different models integrated into the system:

TABLE I
ACCURACY OF INDIVIDUAL MODELS FOR EACH SPECIFIC OBJECTIVE

Model	Objective	Accuracy
YOLO11	Person Detection	95.6%
Huggingface	Gender Classification	92.3%
MediaPipe	Facial Expression	89.7%
Vit-B/32	Violence Detection	90.4%

The table I demonstrates the system's ability to accurately and reliably recognize individuals, genders, emotional states, and violent behaviors, achieving high accuracy across all modules with a low latency of 1.8 seconds, making it suitable for real-time applications.The model attains a final accuracy of 72.9%, showcasing its capability to accurately identify and categorize safety-related incidents with a satisfactory degree of precision. Although it is effective, challenges persist, such as possible performance decline in crowded or dimly lit settings, biases in facial expression and gender models resulting from dataset constraints, and the requirement for additional testing to verify scalability with high video input loads. These elements underscore possibilities for future enhancement to increase resilience and relevance across various situations.

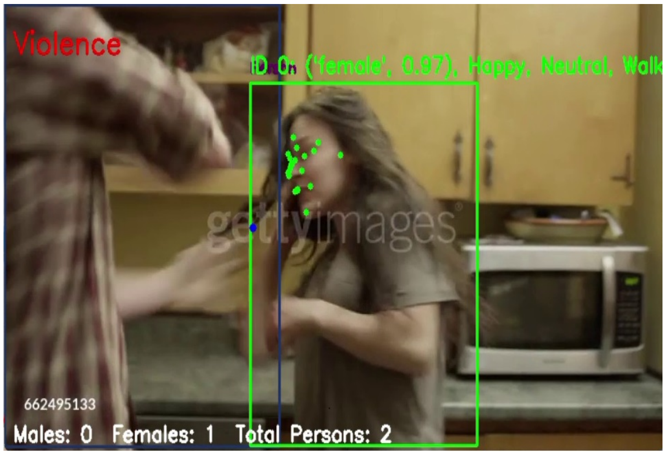
Violence
ID 0: ('female', 0.97), Happy, Neutral, Walk
gettyimages
662495133
Males: 0 Females: 1 Total Persons: 2

Fig. 7. Ultimate output demonstrating the combined effectiveness of all models for precise detection, classification, and behavior assessment.

V. Conclusion and Future Work

Our proposed work introduces a thorough AI-based monitoring system designed to improve the safety of women by analyzing video footage in real-time. Utilizing sophisticated machine learning models like YOLO11, Huggingface, MediaPipe, and ViT-B/32, the system employs a comprehensive method to detect and evaluate possible threats. By combining these methods, it is possible to quickly and accurately detect hazardous situations, which enables proactive reactions to any incidents that may occur. Our findings show that utilizing this combined system can greatly enhance the dependability and efficiency of safety evaluations, proving to be a valuable asset in bolstering public safety. This study not only enhances surveillance technology but also promotes the broader social aim of ensuring women's safety. Future studies will prioritize improving model precision, investigating intricate behavioral trends, and guaranteeing ethical and privacy-oriented integration in various practical situations.

Upcoming efforts will aim to improve model precision to minimize false positives in difficult situations such as crowded or dimly lit settings. Moreover, examining intricate behavioral trends and guaranteeing scalability for various environments will be emphasized. Ethical factors and privacy-protecting steps will be incorporated to ensure the system aligns with practical uses while effectively tackling social issues.

References

[1] Şeymanur Aktı, Gözde Ayşe Tataroğlu, and Hazım Kemal Ekenel. Vision-based fight detection from surveillance cameras. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2019.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[3] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017.

[4] Aadesh Guru Bhakt Dandamudi, Gorrepati Vasumithra, Gangisetti Praveen, and C.V Giriraja. Cnn based aerial image processing model for women security and smart surveillance. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1009–1017, 2020.

[5] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification, 2020.

[6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Alexey Dosovitskiy, Hakan Bilen, Michael Cogswell, Andrew Zisserman, and Rob Fergus. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1734–1747, 2016.

[8] Pallavi Sharda Garg, Samarth Sharma, Archana Singh, and Nitendra Kumar. Ai-based surveillance systems for effective attendance management: Challenges and opportunities. *Mathematical Models Using Artificial Intelligence for Surveillance Systems*, pages 69–89, 2024.

[9] Chirag Gupta, Tejas Vinchurkar, Sarthak Chavan, and PG Waware. Channelizing machine learning towards early threat detection and prevention against women using surveillance cameras.

[10] Jimbo. A first look at yolov11: Pushing the boundaries of real-time object detection, 2023. Accessed: 2024-12-21.

[11] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024.

[12] Chen Kuan-Yu, Jungpil Shin, Md Al, Hasan Md Mehedi, Jiun-Jian Liaw, Yuichi Okuyama, and Yoichi Tomioka. Fitness movement types and completeness detection using a transfer-learning-based deep neural network. *Sensors*, 22:5700, 07 2022.

[13] W. Irene Michelle, M. Z. Mohamed Ashik, N. Achyut, T. Nitya, Deepa Jose, and Jerold Kingston Gnanasekaran. Video analysis using deep learning in smart gadget for women saftey. In Nikhil Kumar Marriwala, Sunil Dhingra, Shruti Jain, and Dinesh Kumar, editors, *Mobile Radio Communications and 5G Networks*, pages 165–174, Singapore, 2024. Springer Nature Singapore.

[14] Mohd Naved, Awab Habib Fakih, A Narasima Venkatesh, P Vijayakumar, Pravin Ramdas Kshirsagar, et al. Artificial intelligence based women security and safety measure system. In *AIP conference proceedings*, volume 2393. AIP Publishing, 2022.

[15] Aleksander Radovan, Leo Mršić, Goran ambić, and Branko Mihaljević. A review of passenger counting in public transport concepts with solution proposal based on image processing and machine learning. *Eng*, 5(4):3284–3315, 2024.

[16] Mhaske Pragati Sambhaji, Patil Sakshi Balraje, Sable Aishwarya Ramhari, and Sawant Dipali Bhagwat. Women's security and smart surveillance. 2024.

[17] Akhilesh Sharma, Vipan Kumar, and Louis Longchamps. Comparative performance of yolov8, yolov9, yolov10, yolov11 and faster r-cnn models for detection of multiple weed species. *Smart Agricultural Technology*, 9:100648, 2024.

[18] R. Singh. Human pose tracking with mediapipe: Rerun showcase. *Towards Data Science*, Nov 2023. Accessed: 2024-12-22.

[19] Sanskar Singh, Shivaibhav Dewangan, Ghanta Sai Krishna, Vandit Tyagi, Sainath Reddy, and Prathistith Raj Medi. Video vision transformers for violence detection, 2022.

[20] M Tamilselvi, K Suresh Kumar, G Devi, P Shahad, P Sharmila, and S Lakshmisridevi. Unveiling ai patterns: An experimental evaluation of machine learning based women's safety prediction strategy. In *2024 2nd World Conference on Communication & Computing (WCONF)*, pages 1–5. IEEE, 2024.

[21] Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis. Efficient violence detection in surveillance. *Sensors*, 22(6):2216, 2022.

[22] Guojun Zhao and Yafei Jiang. Facial expression recognition using robust local binary patterns. *International Journal of Computer Vision*, 111(3):242–256, 2016.