AI-Driven Classroom Monitoring Using Edge Computing: A Multimodal Approach for Real-Time Emotion and Violence Detection

[1] Dr. A.K. Ashfauk Ahamed, [2] Bhuvaneswari L

[1] Associate Professor, Department of Computer Applications, BS Abdur Rahman Crescent Institute of Science and Technology, Vandalur, India.

[2] MCA Student, Department of Computer Applications, BS Abdur Rahman Crescent Institute of Science and Technology, Vandalur, India.

ABSTRACT

Disruptive behavior in classrooms impacts both student safety and the overall learning environment. Traditional classroom monitoring systems are often ineffective, making it crucial to develop systems capable of automatically detecting the violence and alert the higher authorities. This study introduces an AI-powered classroom monitoring system using Edge Computing to instantly identify students' emotion and physical aggression. The proposed framework integrates facial emotion recognition and human action recognition to identify emotional distress and violent behavior. A Convolutional Neural Network(CNN) is trained on the FER-2013 dataset is used to analyze facial expressions and a lightweight MobileNetV2-based model combined with Long Short-Term Memory (LSTM) networks is trained on Real-Life Violence Situations (RLVS) dataset is used to detect physical aggression. All processing is done locally on a Raspberry Pi to guarantee low latency and privacy, which makes the system appropriate for real-time deployment in educational settings. A decision-making mechanism combines outputs from models that identify emotional reactions and acts of violence, generating actionable intelligence and triggering alarms only when the probability of abuse exceeds a pre-set critical threshold. Secondly, the system logs incidents in a cloud-based dashboard to enable administrators to view incidents later.

Keywords: Classroom monitoring, Edge computing, Facial emotion recognition, Violence detection, Deep learning

LIST OF ABBREVIATIONS

The list of abbreviations are mentioned in Table 1.

Abbreviation	**Full Form**
AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolutional Neural Network
FER	Facial Emotion Recognition
FPS	Frames Per Second
HAR	Human Action Recognition
LSTM	Long Short-Term Memory
RGB	Red Green Blue (Color Model)
RLVS	Real-Life Violence Situations (Dataset)
SMTP	Simple Mail Transfer Protocol

Table 1. List of abbreviations

INTRODUCTION

The classroom environment plays a crucial role in the emotional development of the students. However, disruptive behaviors, including emotional distress and physical misconduct, can negatively impact both individual students and the overall learning experience. Traditional classroom monitoring systems rely on manual supervision by the administration which is inefficient and lacks real-time monitoring. As classrooms become more technologically advanced, the need for an automated, real-time system to monitor student behavior becomes essential.

Recent advancements in deep learning and computer vision have enabled automated systems to detect student emotions and behaviors with high accuracy. Studies have shown that facial emotion recognition can be effectively utilized for student engagement analysis and classroom behavior monitoring (Fakhar et al., 2022). Additionally, real-time behavior recognition systems have been employed to identify students' attention levels and aggressive actions, contributing to safer and more interactive learning environments (Trabelsi et al., 2023). However, existing solutions do not offer real-time detection and alert system using multimodal approach to optimize the violence detection which leads to false positives.

This paper presents an AI-driven classroom monitoring system that integrates facial emotion recognition and misconduct detection using Edge Computing on Raspberry Pi. The system uses deep learning and computer vision to detect emotions and aggressive physical actions, such as hitting or pushing in real-time. Local data processing on low-power edge device reduces latency and minimizes reliance on cloud-based processing. Edge computing approaches have demonstrated effectiveness in low-latency AI applications, including violence detection and real-time object recognition (Saad et al., 2024; Zekovic, 2024).

The main contributions of this paper are as follows:

- Multimodal behavior monitoring: We integrate facial emotion detection and physical aggression recognition to comprehensively analyze student behavior (Jaiswal et al., 2020; Ranganathan et al., 2016).
- Edge-based real-time processing: We implement the system on Raspberry Pi, allowing on-device computation to reduce network dependency and latency (Saad et al., 2024).
- AI-driven violence detection: Our model classifies aggressive actions (e.g., hitting, pushing) using deep learning-based action recognition techniques (Fatima Kiani and Kayani, 2022).

RELATED WORK

A. Facial Emotion Detection

Several studies have explored facial emotion detection as a means of assessing student engagement and classroom dynamics. (Jaiswal et al. 2020) introduced a deep learning-based facial emotion recognition system capable of identifying emotions with high accuracy, demonstrating the potential of AI-driven emotion analysis in educational environments. Similarly, (Ranganathan et al. 2016) proposed a multimodal approach integrating facial expressions and voice signals to enhance emotion classification accuracy.

Beyond emotion detection, classroom monitoring systems have also leveraged behavior recognition to assess student engagement and detect disruptive actions. (Fakhar et al. 2022) developed a real-time facial expression recognition system for smart classrooms, providing valuable insights into students' emotional states.

B. Agression Detection

In addition to emotion recognition, detecting aggressive behavior is crucial for maintaining a safe classroom environment. (Kiani and Kayani, 2022) implemented a deep learning-based violence detection system, demonstrating how real-time analysis of aggressive behavior can contribute to security applications. Their approach effectively identifies violent actions, such as physical fights, using video-based deep learning models. However, their system is primarily designed for broader security use cases and does not focus on classroom settings. Combining this with emotion recognition can improve classroom monitoring.

C. Edge Computing for Real-Time Processing

While cloud-based AI models provide high accuracy, their reliance on internet connectivity and cloud processing introduces latency and privacy concerns. Recent studies have explored the use of edge computing to address these challenges. (Saad et al. 2024) compared different deep learning models for real-time inference on Raspberry Pi 4, demonstrating that optimized models can run efficiently on edge devices without relying on cloud infrastructure. Similarly, (Zekovic, 2024) explored the integration of Edge Impulse and deep learning models on Raspberry Pi, highlighting the feasibility of real-time video analysis on low-power hardware. These studies emphasize the effectiveness of edge-based AI models but do not focus on multimodal integration of emotion detection and aggressive behavior recognition in an educational setting.

PROPOSED METHODOLOGY

Our proposed system integrates facial emotion detection and physical aggression recognition using deep learning and computer vision techniques, with real-time processing on a Raspberry Pi device. Unlike traditional surveillance methods that rely on manual monitoring or cloud-based processing, our system provides an automated, intelligent, and decentralized solution that can function effectively in classroom environments without excessive computational overhead. The primary goal is to detect aggressive behavior alongside emotional cues, ensuring that alerts are sent only in significant cases.

A. System Overview

The proposed system consists of a camera module, a deep learning-based facial emotion detection model, a human action recognition model, and an alert mechanism to notify administrative authorities upon detecting concerning behavior. The camera continuously captures classroom activities, feeding frames to both detection models. Each model independently processes the data, and a decision module determines if the detected action requires intervention which depicts the system architecture in Fig. 1.

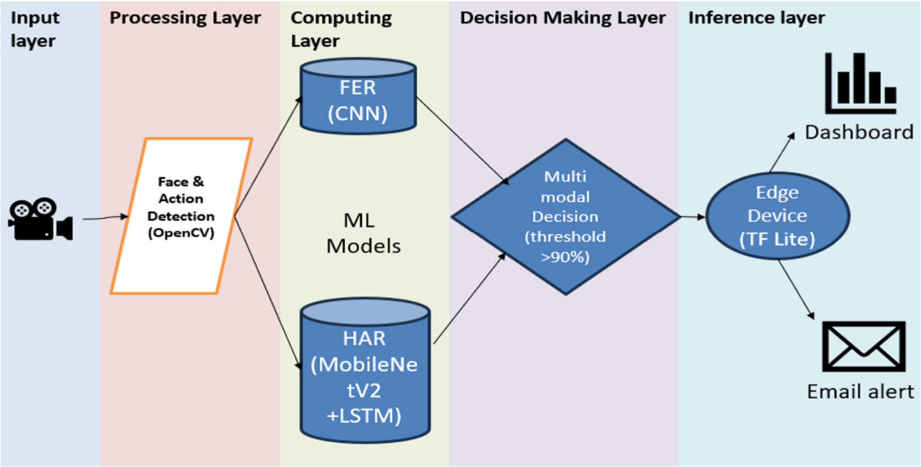

Fig. 1. System Architecture

B. Facial Emotion Detection

The facial emotion recognition model classifies student expressions into five categories: Angry, Happy, Neutral, Sad and Surprise. A CNN trained on the FER2013 dataset is used for real-time emotion detection. Although the JAFFE(Japanese Female Facial Expression) dataset gives more accuracy, we use FER2013 dataset because the JAFFE(Japanese Female Facial Expression) dataset lacks number of subjects, leading to overfitting problems.

The system does not generate alerts based on emotions alone, as momentary frustration or sadness does not necessarily indicate misconduct. Instead, emotion detection results serve as supporting evidence for behavior analysis. If anger or distress is detected, the system assigns a risk factor that contributes to the final misconduct evaluation.

C. Physical Agression Detection

To recognize aggressive physical actions such as hitting, pushing, or throwing objects, a Human Action Recognition (HAR) model is employed. The model is based on a hybrid architecture combining MobileNetV2 for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal dependencies across video frames. It is trained on the RLVS dataset. By analyzing sequences of frames rather than individual images, the model effectively captures motion patterns, reducing false positives caused by non-aggressive gestures like hand-raising or sudden movements.

D. Decision-Making and Alert System

To minimize false positives, the system employs a fusion-based decision module that combines results from both facial emotion detection and physical aggression recognition. Alerts are triggered only if both conditions meet predefined thresholds:

- Emotion Analysis: If a student is classified as Angry or Distressed with a high confidence score (>80%), the risk factor increases.
- Action Recognition: If an aggressive motion is detected with a probability above 85%, it contributes significantly to the final misconduct score.

- Context Awareness: The system considers the duration and frequency of aggressive actions within a time window (2-3 seconds of continuous aggression strengthens the misconduct classification).

- Final Decision: If the combined misconduct probability surpasses a predefined threshold of 90%, an E-mail alert is generated and sent to administrative authorities.

E. Edge Computing with Raspberry Pi

The entire system is deployed on a Raspberry Pi 4 Model B which is displayed in Fig. 2. leveraging edge computing for real-time processing. This setup ensures:

- Low latency, as video frames are processed locally instead of being transmitted to a cloud server.

- Improved privacy, as student data remains within the local network.

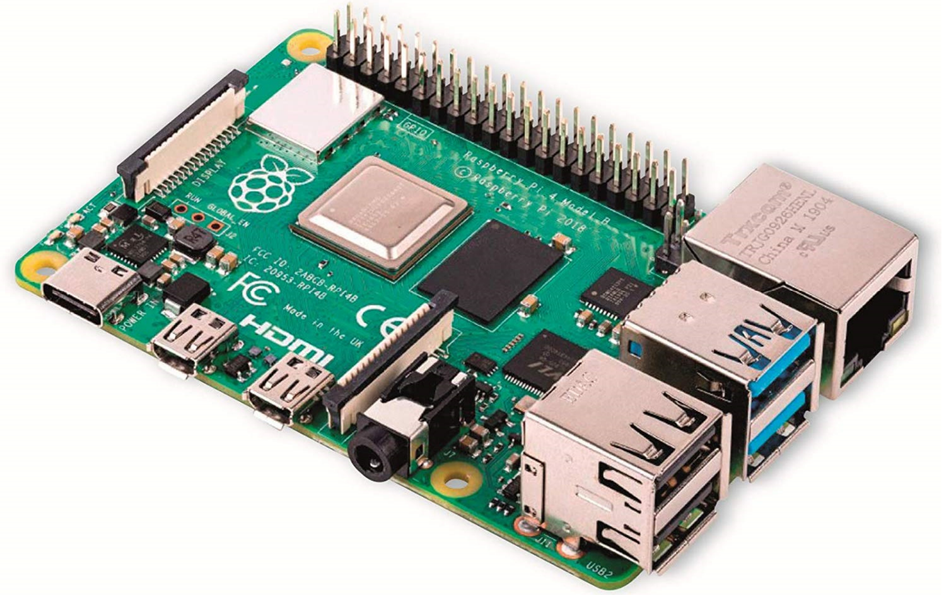

Fig. 2. Raspberry Pi 4 Model B

For computational efficiency, the models are optimized using TensorFlow Lite to run effectively on low-power edge devices without sacrificing accuracy. Raspberry Pi sends the final outputs to cloud database along with the timestamps which is directly accessed by the web dashboard hosted on a local machine which logs real-time events and displays monitoring status for administrative authorities.

IMPLEMENTATION

The implementation process consists of multiple stages including dataset preparation, model training, system integration and deployment. Implementation of the system as various stages are explained in depth in this section.

A. Facial Emotion Recognition

FER plays a crucial role in this multimodal system. In this work, FER is implemented using CNN which is trained on the FER2013 dataset which provides a vast and diverse range of facial expression which includes individuals across various demographics which helps in the model efficiency. This dataset consists of 48×48 pixel grayscale images which are categorized into five emotion classes such as Angry, Happy, Neutral, Sad and Surprise. The following sections describe the detailed process of implementing FER using CNNs.

Preprocessing is crucial in preparing the input data to be fed into the model. The following steps are performed:

- Grayscale Conversion: The original RGB images are converted to grayscale. Grayscale images reduce the complexity of the data while preserving important features of the face, allowing the model to focus on structural information instead of color variations.

- Histogram Equalization: To improve the contrast of the image, histogram equalization is applied. This step redistributes pixel intensity values, enhancing the visibility of important facial features under various lighting conditions, which helps the model to distinguish emotions more accurately.

- Noise Reduction: Noise from environmental factors or camera imperfections can interfere with the recognition process. Gaussian blurring or other noise reduction techniques are applied to smooth the image and remove high-frequency noise, allowing the model to focus on relevant features.

- Face Alignment and Normalization: Images are resized to a fixed 48×48 pixels, and facial landmarks are aligned (usually based on the eyes or mouth) to minimize the impact of head tilts and orientations. This step ensures that the facial features are consistently represented across all images.

Face detection is necessary to locate and crop faces from the background of each frame. Accurate face detection ensures that only relevant facial data is passed to the model, improving the system's performance. Haar Cascade Classifier, which is a machine learning-based approach that detects faces by scanning the image with a sliding window. This method is fast and effective for detecting frontal faces, particularly in controlled environments.

Once the face is detected, it is cropped and aligned to ensure only the facial region is processed by the emotion recognition model. The core of the facial emotion recognition system is the CNN, a deep learning model that excels in processing and classifying image data. The CNN architecture used in this study is as follows:

- Input Layer: The input to the model consists of pre-processed grayscale images of size 48×48 pixels, representing a face.

- Convolutional Layers: The network starts with several convolutional layers that apply small filters to the image to extract essential features such as edges, textures, and shapes. These layers use the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and enable the network to learn complex patterns.

- Max-Pooling Layers: After each convolution operation, max-pooling layers are applied to reduce the spatial dimensions of the feature maps, effectively downsampling the image and retaining the most important features. This also helps reduce computational complexity and prevent overfitting.

- Fully Connected Layers: The CNN consists of one or more fully connected (dense) layers that combine the extracted features and perform the final classification. These layers are responsible for mapping the extracted features to the corresponding emotion classes.

- Softmax Layer: The final layer is a softmax classifier that outputs a probability distribution over the five emotion classes: Angry, Happy, Neutral, Sad, Surprise. The emotion with the highest probability is selected as the predicted emotion.

B. Violence Detection

In this system, physical aggression is detected using HAR based on deep learning techniques that analyze real-time video feeds to classify actions as either violent or non-violent. The model is built based on MobileNetV2 and LSTM architecture. This hybrid model leverages the spatial feature extraction capability of MobileNetV2 and the temporal sequence learning capability of LSTM to detect aggression effectively.

The model is trained on RLVS dataset, which contains real-world surveillance videos labelled as violence and non-violence. Each video is segmented into short clips of 2-5 seconds for better model training. The video frames are resized to 128×128 pixels, normalized into fixed length sequences of 10 frames as this ensures consistent input to the model which also allows the model to learn temporal dynamics over time offering efficiency.

The model architecture consists of the combination of MobileNetV2 and LSTM layers. MobileNetV2 servers as the feature extractor which captures spatial details from each video frame. It's lightweight design makes it suitable for deploying our system in Raspberry Pi. The features from MobileNetV2 are passed into an LSTM network, which learns the temporal dependencies across the sequence of frames. The model is capable of detecting patterns by understanding how motion evolves over time. A final dense layer with a sigmoid activation function given in the below equation which provides a binary classification labelling input as violent or non-violent.

Sigmoid Activation Function $\sigma(x) = 1/(1+e$^-x)

Here, x is the input to the function, e is Euler's number(approximately 2.718) and the function squashes the input x into a value between 0 and 1. This function maps the input x to a probability value in the range [0,1], indicating the likelihood of aggressive behavior.

The real-time system continuously captures the video frames and process them in batches of ten. Each batch is passed through the trained model to generate a prediction. A majority voting mechanism is applied over consecutive predictions to improve the stability. If violence is detected over several frame windows, the action is flagged as violent. The system also uses probability threshold mechanism where the output is evaluated as aggression if only >90% which improves reliability and help reduce false positives and misclassification.

C. Decision Making and Alert System

The decision making module integrates the output from both the facial emotion recognition and the violence detection models to provide a more comprehensive assessment of student behavior. The system can more accurately detect potential misconduct by combining these two sources of input which minimizes false positives ensuring that only significant incidents trigger alerts.

Using a weighted fusion approach with the confidence of facial emotion recognition and the probability of aggression detected in the physical violence recognition model are combined into a final misconduct probability score. FER model derives the emotion score which reflects the intensity of emotions such as anger or distress and the action recognition model derives the aggression score indicating the probability of violent behavior.

Based on its significance, each score is assigned a weight. For example, if a student shows strong expression of anger(>80%), the emotion score is given higher weight. Similarly if the aggression detection model shows high probability(>85%) for violent behavior, it also gives strong influence on the final decision.

An alert is triggered only when the combined probability exceeds the 90% threshold. This ensures that the system remains robust and improves the accuracy. To further improve accuracy, the system incorporates context-aware filtering. For example, a brief arm swing or a fleeting angry expression lasting less than 2 seconds is considered a normal reaction and an alert is not triggered. Alert is sent only if the aggression persists for more than 2-3 seconds alongside a high emotion score. The alert includes the timestamp of the incident, detected emotion, violence status and a final decision. The alert is sent via email to the school authority using SMTP service configured through the Gmail API without any manual intervention.

E. Edge Computing Deployment on Raspberry Pi

With minimal power consumption and low latency, the entire real-time system is deployed on a Raspberry Pi Model B, serving as an edge device. Using TensorFlow Lite, the facial emotion recognition model and the physical aggression detection model are optimized to reduce the computational footprint while retaining accuracy. The system avoids the need of continuous cloud connectivity ensuring greater privacy, faster response times and improved energy efficiency, especially in a classroom setting where reliable real-time detection is essential.

F. Web Dashboard and Data Logging

The system includes a web-based dashboard developed using Flask allowing the school administration to monitor the events in real-time. Alerts generated by the decision-making module is logged into MongoDB Atlas and this data is used to feed the dashboard which displays the live updates from the decision-making system, event logs chronologically listed with an exact timestamp, search and filter capabilities and visual representation of the frequency of the detected emotions mentioned in Bar Graphs.

RESULTS AND DISCUSSION

The performance of the proposed AI-driven classroom monitoring system was evaluated with two primary modules: FER and violence detection. Both models were trained independently and are executed real-time on Raspberry Pi 4 Model B using TensorFlow Lite. Both modules processed frames at the same time with a live video input streaming from a camera.

The FER model which is trained based on CNN was designed to classify five emotion categories: Angry, Happy, Neutral, Sad and Surprise. The model achieved an overall accuracy of 89% with precision and recall across all classes. “Happy” and “Surprise” classes demonstrated the strongest performance while “Angry” and “Sad” which are most relevant to the violence detection also resulted with robust F1-scores of 0.87 and 0.88 respectively and this indicates the model’s effectiveness in recognizing emotional cues for violence. The detailed classification metrics for the FER model are presented in the Table 2.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Angry	0.88	0.87	0.87	250
Happy	0.91	0.90	0.90	250
Neutral	0.89	0.88	0.88	250
Sad	0.88	0.89	0.88	250
Surprise	0.90	0.91	0.90	250
Accuracy	**0.89**			
Macro Avg	0.89	0.89	0.89	
Weighted Avg	0.89	0.89	0.89	

Table 2. Classification Report for Facial Emotion Recognition Model

The violence detection model which utilizes hybrid MobileNetV2 + LSTM architecture trained on RLVS dataset, classifies short video clips of 10 frames each into either “Violence” or “Non-Violence”. The model achieved overall accuracy of 91%, with balanced precision and recall for both classes and an F1-score of 0.91 confirming the model's strong ability to detect physical aggression. . The detailed classification metrics for the violence detection model are presented in the Table 3.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Violence	0.92	0.90	0.91	150
Violence	0.90	0.92	0.91	150
Accuracy	**0.91**			
Macro Avg	0.91	0.91	0.91	
Weighted Avg	0.91	0.91	0.91	

Table 3. Classification Report for Violence Detection Model

A decision-making module was integrated to combine the outputs of both FER and aggression models. A weighted fusion mechanism is used to compute the final violence probability score and an alert is triggered only if the violence probability exceeds 90%. This threshold-based logic reduces false positives improving the overall reliability of the system. During testing, the integrated decision-making system achieved an accuracy of 93% confirming the combined analysis of both of the models. A context-aware filtering was also applied to suppress the brief or isolated movements ensuring only meaningful incidents are flagged. The web dashboard successfully displayed real-time alerts and historical logs confirming the effectiveness of the decision making module. The detailed Performance of the Integrated Decision-Making Module are presented in the Table 4.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Alert	0.94	0.91	0.92	140
No Alert	0.92	0.95	0.93	160
Accuracy			**0.93**	
Macro Avg	0.93	0.93	0.93	
Weighted Avg	0.93	0.93	0.93	

Table 4. Performance of the Integrated Decision-Making Module

CONCLUSION

This paper presented a real-time classroom monitoring system using multimodal AI techniques to detect both emotional states and physical aggression. By integrating a CNN-based FER module with a MobileNetV2+LSTM-based violence detection mode, the system effectively enhances classroom safety and awareness. Deployment of the system on a Raspberry Pi ensures low latency, privacy and energy efficiency. Future work will focus on refining the multimodal decision-making accuracy and expanding the system to detect a wider range of student behaviors for more comprehensive approach.

REFERENCES

A. Jaiswal, A. Krishnama Raju and S. Deb, "Facial Emotion Detection Using Deep Learning," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154121.

H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477679.

S. Fakhar, J. Baber, S. U. Bazai, S. Marjan, M. Jasinski, E. Jasinska, M. U. Chaudhry, Z. Leonowicz, and S. Hussain, “Smart classroom monitoring using novel real-time facial expression recognition system,” Appl. Sci., vol. 12, no. 23, p. 12134, 2022.

S. K. Vishnumolakala, V. S. Vallamkonda, S. C. C, N. P. Subheesh and J. Ali, "In-class Student Emotion and Engagement Detection System (iSEEDS): An AI-based Approach for Responsive Teaching," 2023 IEEE Global Engineering Education Conference (EDUCON), Kuwait, 2023, pp. 1-5, doi: 10.1109/EDUCON54358.2023.10125254.

Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, “Real-time attention monitoring system for classroom: A deep learning approach for student’s behavior recognition,” Big Data Cogn. Comput., vol. 7, no. 1, p. 48, 2023.

A. SAAD, A. AIDI and A. El Hassan BENYAMINA, "Optimized Deep Learning Models For Edge Computing: A Comparative Study on Raspberry PI4 For Real-Time Plant Disease Detection," 2024 4th International Conference on Embedded & Distributed Systems (EDiS), BECHAR, Algeria, 2024, pp. 273-278, doi: 10.1109/EDiS63605.2024.10783415.

A. Zekovic, "Computer Vision on Edge: Using Edge Impulse and Deep Learning Models on Raspberry Pi," 2024 32nd Telecommunications Forum (TELFOR), Belgrade, Serbia, 2024, pp. 1-4, doi: 10.1109/TELFOR63250.2024.10819144.

G. e. Fatima Kiani and T. Kayani, "Real-time Violence Detection using Deep Learning Techniques," 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS), Karachi, Pakistan, 2022, pp. 1-8, doi: 10.1109/ICONICS56716.2022.10100551.