Grenze International Journal of Engineering and Technology, June Issue

A Multi-Modal Framework for Emergency Vehicle Prioritization using Visual Detection and Acoustic Classification

Palak Sharma[1], Abhishek Kumar[2], Viraj Gautam[3], Ravindra Ray[4], Nitesh Kumar[5] and Charu Awasthi[6]
[1-6]Department of Information Technology, JSS Academy of Technical Education, Noida, India
Email: aapalaks@gmail.com, mr.abhi9974@gmail.com, virajgautam07.08@gmail.com, ravindraray1574@gmail.com, niteshkumarhrickld@gmail.com, charu.awasthi@jssaten.ac.in

***Abstract*— Traffic jams in cities often make it tough for emergency services to reach people in need, especially during the golden hour, when every minute counts. This issue arises from long distances and slow-moving vehicles. Most of the technologies that are currently used for managing traffic struggles to address this problem effectively because it typically depends on just one type of sensor. Challenges like blocked views, poor lighting, and background noise can make this approach unreliable.**
In our study, we introduced a new method that combines different types of data to give priority to emergency vehicles. We use a visual detection system and an audio classification system together, which has proven to be more reliable, especially in busy urban environments. Specifically, we have used a computer vision model called YOLOv8 to spot emergency vehicles through cameras placed on traffic signals. Alongside that, we have employed a sound recognition model named PANNs CNN14, which can pick up sirens and differentiate that from other sounds.
Both visual and audio data are processed simultaneously and blended together smartly. This allows us to quickly recall either sound or sight information in critical situations. The Use of regular traffic cameras and microphones instead of expensive equipment makes it easier to use and more reliable, which is the best part.

***Index Terms*— YOLOv8, PANNs CNN14, Computer Vision, Emergency Vehicle Prioritization, Audio Classification, Multi-Modal Fusion, Intelligent Transportation Systems.**

I. INTRODUCTION

In crowded cities such as Delhi, Mumbai, and Bengaluru, significant traffic congestion and large gatherings of people prevent emergency vehicles from traveling quickly. Ambulances, fire trucks, and police cars often get stuck on the road, wasting the "Golden Hour"—the first hour after an injury when medical treatment is most likely to save a life.
According to reports from the World Health Organization (WHO) and the National Crime Records Bureau (NCRB) [1], the main reason for many trauma-related deaths in India is the delay in ambulances reaching the hospital. Smart road systems have progressed a lot, but most traffic lights still use simple timers. These timers are rigid and cannot adjust to changing traffic or emergencies.

Grenze ID: 01.GIJET.12.2.172_20

Other ideas use GPS or special tags to control the lights [2], but these have practical problems. They need constant rule-following, specific equipment, and regular maintenance. These problems hinder its widespread adoption, particularly in locations with limited resources.

A fresh approach introduces the use of existing city infrastructure, rather than in-vehicle equipment, to address the limitations of current emergency detection methods. This system combines visual and acoustic sensors to improve reliability [3]: a camera system helps locate the general area of an emergency, and an audio detection system can "hear" an emergency vehicle's siren even when the vehicle is out of the camera's line of sight, or the view is blocked. By merging camera and audio information, the system is better at reliably identifying emergencies, particularly in busy and often chaotic city traffic, where conditions might be less than perfect.

A. Problem Definition

Three key challenges hinder EV detection in urban traffic [4]:

- Occlusion: Other cars and buildings often hide EVs from view.
- Ambient noise: The many sounds of a busy city make it hard to hear sirens clearly.
- Single-modality reliance: Current systems for detecting EVs use sight and sound separately, without combining the information to get a complete picture.

B. Research Objectives

The primary objectives of this research are:

- To develop a real-time visual EV detection module using YOLOv8 [5].
- To design an adaptive fusion logic that integrates visual and acoustic outputs to reduce false detections.
- To implement a robust acoustic siren detection module based on PANNs CNN14 [6].
- To simulate a dynamic "Green Wave" traffic signal control strategy for EV prioritization [2].

II. Related Works

TABLE I. Comparative Analysis of State-of-the-Art Emergency Detection Systems

Ref.	Author (Year)	Methodology	Performance	Limitations & Gaps
[1]	Sharma et al. (2019)	YOLOv3 Visual Detection	92.4% Accuracy	Poor performance in low light/occlusion; heavy GPU resources
[2]	Li & Huang (2024)	CNN-based Audio Class.	~95% Accuracy	Degrades in high-decibel urban noise; lacks visual verification
[3]	Kong et al. (2020)	PANNs (CNN14)	43% mAP (Audio Set)	Heavy model; high inference latency on edge devices
[4]	Park & Kim (2020)	CNN + RNN Fusion	18% gain vs. video-only	High implementation complexity; increased hardware costs
[5]	Gupta et al. (2020)	Deep Q-Learning (RL)	27% wait reduction	Simulation-only (SUMO); lacks real-world deployment
[6]	Chen & Luo (2021)	Night-Time YOLO	90.6% Accuracy (Night)	Specialized for night; less effective for daytime
[7]	Patel et al. (2022)	Deep CNN (Noise Filter)	94% Recognition	Dependent on directional microphones; struggles with wind
[8]	Ahmed & Rao (2023)	YOLOv8 Detection	96% mAP @ 80 FPS	Dependent on expensive server-side GPUs
[9]	Lee & Park (2019)	LiDAR + Camera Fusion	25% reliability boost	LiDAR sensors prohibitively expensive at scale
[10]	Reddy & Srinivas (2020)	DL Preemption	91% Detection Rate	Lacks modern acoustic integration; purely visual
[11]	Bhatia et al. (2021)	YOLOv3 Smart Traffic	89% Accuracy	Limited dataset; no audio-visual fusion
[12]	Chouhan et al. (2022)	YOLOv7 + Audio Analysis	High Recall	Audio analysis amplitude-based, not spectral
[13]	Zhang & Wang (2019)	RL Signal Optimization	Improved Flow	Focuses on flow theory, not vehicle detection
[14]	Zohaib et al. (2024)	ATSN Multimodal Fusion	3.8% Misdetection Rate	Complex architecture; unsuitable for edge devices
[15]	Bieker-Walz (2019)	Green Wave Simulation	30% Time Reduction	Relies on V2I data not yet ubiquitous

In recent years, there has been a large increase in the analysis of identifying and prioritizing emergency vehicles (EV). Over the past few decades, technology used for this purpose has advanced from simple RF signal-based and rules-based systems to more complex data-based systems utilizing advanced algorithms, such as those found in deep learning [7]. The bulk of the research to date can be categorized into one of the following categories: (i) Vision-Based Detection – Methods that utilize cameras and image processing to visually identify Emergency Vehicles [8]. (ii) Audio-Based Siren Recognition – Methods that are designed to detect the specific sounds associated with sirens [9]. (iii) Multi-Modal Frameworks – Methods that incorporate both visual and audio data to provide more accurate identification of Emergency Vehicles [10].

Vision-based methods generally use convolutional neural networks with road cameras. Techniques using the YOLO (You Only Look Once) family show robust detection accuracy in ideal scenarios [11]. However, their performance substantially decreases when faced with challenges like: obstacles blocking the view (stoppage), poor lighting, and bad weather conditions [12]. Further, some analyses require powerful, exclusive graphics cards (GPUs).

Some systems use sound to recognize emergency sirens by analyzing audio patterns or using AI models [13]. While advanced AI models (like PANNs CNN14) are better at recognizing sounds in different situations [6], they are slower to process information, which causes a delay.

Recent research has focused on multi-modal fusion approaches to enhance the robustness of detection [3]. Although these systems show measurable improvements in performance over unimodal approaches, they can be associated with an increase in architectural complexity.

Table I provide a comparison of current, top-tier systems designed to spot emergency vehicles and give them priority on the road. The table highlights the trade-offs between how accurately a system detects vehicles, how much computing power it requires, what kind of sensors it uses, and how practical it is to install in real-world settings. The current study shows a significant gap: current systems struggle to balance highly accurate detection with practical, easy-to-deploy infrastructure. This imbalance is the main reason we have developed the new, multi-modal framework presented in this paper.

III. METHODOLOGY

The proposed system is developed as a closed-loop emergency vehicle prioritization system, including four modules: input acquisition, feature extraction, multi-modal fusion, and signal control. Visual and acoustic data streams are emergencies, with prioritization decisions being induced based on fused detection outputs.

A. Visual Detection Using YOLOv8

The YOLOv8 one-stage object detector is valid in the visual module due to it is fast and reliable. It uses a specialized design called the C2f backbone to enhance how data flow, helping it identify objects more precisely, even under partial occlusion. YOLOv8 predicts object locations directly without acknowledging on predefined anchor boxes, handle different object shapes. To further enhance accuracy, particularly for fast-moving or blurry vehicles in emergency situations, it uses advanced techniques called Complete IoU (CIoU) loss and Distribution Focal Loss (DFL) to improve the predicted object boundaries.

B. Acoustic Siren Classification Using PANNs CNN14

Audio detection is developed as a classification task in the time-frequency domain. Inward audio signals are sampled at 32 kHz and are transformed using Short-Time Fourier Transform (STFT). Log-Mel spectrograms extracted from the magnitude spectra are fed to the CNN14 architecture from the PANNs framework as inputs.The model is structured to extract features from images; these include convolutional layers, global average pooling and global max pooling to capture both the transient nature of sirens and the background noise that is present continuously in urban areas.

C. Adaptive Multi Modal Fusion

Detection performance is obtained by adaptive fusion engine that fuses together the image-based output and audio-based output of the system and provides temporal consistency.

- Instantaneous Decision Logic: the binary outputs of the visual $v(t)$ and acoustic $a(t)$ detectors respectively at time t:
  - $v(t) = 1$ if $\max(Conf_{YOLO}) > \tau_{vis}$
  - $a(t) = 1$ if $P(Siren) > \tau_{aud}$

Two operation modes exist:

- OR Logic (High Sensitivity): $D_{inst}(t) = v(t) \lor a(t)$
- AND Logic (High Precision): $D_{inst}(t) = v(t) \land a(t)$

The standard configuration implements OR logic to avoid missing detections since avoiding false negatives is especially where important in emergency response applications where safety is priority.

- Temporal smoothing: A sliding window smoothing method is used to reduce transient detection failures. Let W be a window of N frames. The final decision $D_{smooth}(t)$ is computed as:

$$D_{smooth}(t)=\{1, if\ \sum_{0}^{N-1} \quad Dinst(t-i) \geq K\ 0, Otherwise$$

where K is the activation threshold. This method acts as a digital hysteresis filter by enhancing decision reliability and stability.

- Signal Control Logic: Algorithm1 demonstrate the fusion-driven signal control process for controlling signals utilizing the fusion method to triggering a priority message to emergency vehicles for prioritized access to intersections.

Algorithm 1: Safety-Critical Fusion Algorithm

Input: Video Stream, Audio Stream

Parameters: Thresholds τ_v, τ_a

- Initialize Queue ← [0] × 10
- While the system is running do
- Frame ← Video_Stream.read()
- Audio ← Audio_Stream.read()
- Conf_v ← YOLOv8(Frame)
- Conf_a ← CNN14(Audio)
- If (Conf_v > τ_v OR Conf_a > τ_a) then
- Vote ← 1
- Else
- Vote ← 0
- End If
- Queue.enqueue(Vote)
- Queue.dequeue()
- If Sum(Queue) ≥ 4 then
- Signal State: GREEN (Priority)
- Else
- Signal State: RED (Normal)
- End If
- End While

IV. System Design and Architecture

The system architecture framework is flexible and decentralized for low-latency emergency vehicle prioritization. The architecture is structured into four functional layers: input acquisition, feature extraction, fusion and decision, and control and visualization. All the layers work independently yet communicate using clear interfaces, thereby providing them with scale-out and real-time deployment capabilities.

A. System Overview

The system runs on edge devices, which means all processing happens locally. It continuously receives both video and audio data from vehicles at traffic lights in real time. Each video frame is analyzed in real-time using the YOLOv8 model, allowing us to easily detect any emergency vehicles. At the same time, the audio stream is divided into small clips of equal length, which helps in examining these clips using a PANNs-based model. This is done to check whether a siren sound is present or not. The system processes both video and audio at the same time so that both data can help each other.

B. Data Flow Analysis

The working of the system is explained using Data Flow Diagrams (DFDs). These diagrams show how information is moving between different parts of the system. In the Level-0 DFD, we show the system as an emergency-vehicle detection unit that interacts with external elements such as the traffic controller and the road environment over there at road side. The Level-1 DFD explains the system in more detail by breaking it into smaller internal parts, such as the signal control unit, the inference system, video and audio pre-processing, and data fusion. This process shows that even the sensing, decision-making, and action are separate steps, but they work together and in coordination with each other.

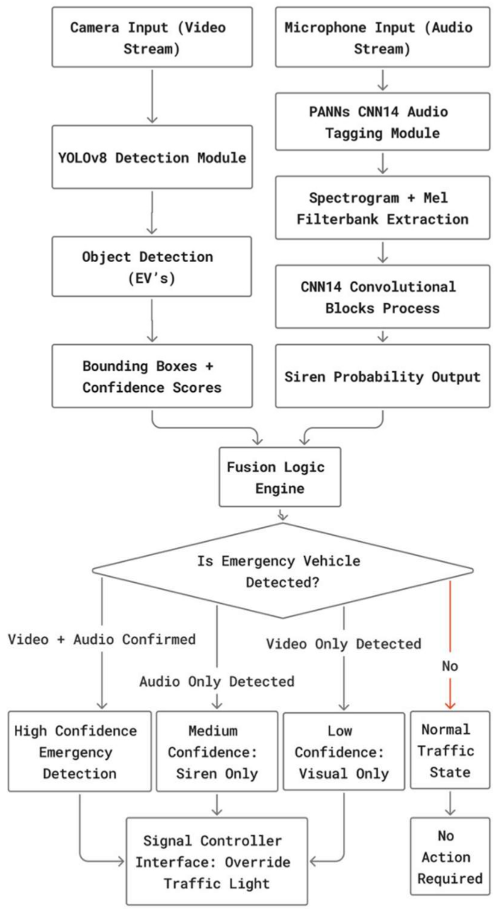

Figure 1. System Architecture

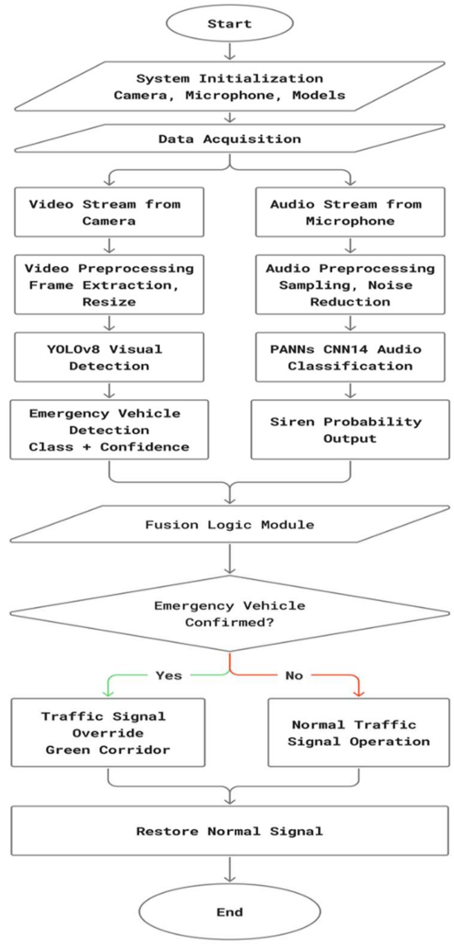

Figure 2. Methodology Flow

C. Methodology Flow

The end-to-end operational logic of the system is shown using a flowchart that outlines the overall sequence from data acquisition to traffic signal control. The flowchart highlights how sensing, inference, multi-modal fusion, temporal smoothening, and control actions are carried out step by step, offering a clear view of how the system functions in practice.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

Edge computing hardware is commonly used as the basis for developing proposed systems that support real time processing of information. For this study, experiments were conducted using a variety of different types of traffic conditions to assess how well each component of the system worked independently and how well they worked together through fusing the two different forms of data from the two different sensor sources.

A. Dataset Description

Visual detection uses a custom version of Roboflow that includes over 2,500 images, 1,500 images show emergency vehicles (ambulances, fire trucks, police vehicles), 1,000 images show non-emergency vehicles. The total number of images has been divided equally among three groups with the first group being used for training, the second group used for validating, and the third group used for testing. Each group provides a similar amount of images showing emergency vehicles versus non-emergency vehicles.

Audio classification uses a subset of Google's AudioSet called AudioSet, which has about 2.1 million audio recordings of 527 different sounds. The AudioSet allows the CNN14 model of the PANNs CNN14 to be able to identify siren sounds in complex environments with many background noises.

B. Performance Metrics

A custom set of synchronized videos and audio recordings representing typical urban traffic conditions have been developed for evaluating the overall performance of the system. The visual-based detection of emergency vehicles using the YOLOv8 model achieved a mAP@0.5 of 93.8%, and therefore demonstrates that it is possible to achieve a very high rate of detecting emergency vehicles regardless of the density of traffic around them. The acoustic-based module using the PANNs CNN14 achieved a score of 0.97 for detecting sirens, which demonstrates that it can detect sirens with a high degree of reliability in a noisy environment or when there are other competing sounds present.

C. Effectiveness of Multimodal Fusion

When the two modalities are fused together, the ability to detect emergency vehicles becomes apparent in difficult operating conditions. When there is significant visual obstruction, the fused system still detected emergency vehicles at a level of 93.8%, whereas the pure visual system dropped to a level of 15%. On the flip side, when there is a lot of noise in the environment and the acoustic system starts to degrade, the visual system compensates by providing accurate visual cues that enable continued detection. The results of these tests demonstrate that the fusion process does allow the two modalities to take advantage of their strengths and weaknesses.

Additionally, the logical OR-based fusion process ensures that the system will continue to function in a way that is predictable and reliable, even if one of the sensor channels fails.

D. System Latency

The entire pipeline runs in end-to-end real-time on the edge platform where a 25 fps video and a 1-second window audio is processed with less than 120 ms per-frame latency, enabling online emergency vehicle detection and traffic signal prioritization.

VI. CONCLUSION

In this paper, we can see how a multi-modal emergency vehicle system is effective in prioritizing EVs. Here, we have used YOLOv8 for real-time detection and PANNs CNN14 for siren classification together. This proposed system addresses all the limitations associated with unimodal Intelligent Transportation Systems, particularly in situations where the camera's vision gets blocked or in noisy conditions. Here, we have used the adaptive fusion strategy, combined with temporal smoothing, which will help the system to make a reliable decision and reduce unnecessary signal activations. Experimental results confirm that the detection performance of the framework is very high and reliable in different and adverse traffic conditions. Furthermore, the successful deployment of this system on edge hardware shows that advanced multimodal intelligence can be achieved without any dependence on high-cost infrastructure. This supports the easy and practical adoption of this system in an urban environment. Even though the performance of the system is too good, its effectiveness may be influenced by dataset diversity, extreme urban noise conditions, and variations in siren usage, which are planned to be addressed in the future.

A. Future Works

Several extensions are planned to further enhance system capability and scalability:

- V2X Integration: Adding a vehicle to the infrastructure communication through GPS data for verification and detecting EVs at an early stage.
- Directional Audio: Using multiple microphones to detect the exact direction of an upcoming vehicle through audio analysis.
- City-Scale Analytics: Using cloud-based storage and analysis to improve traffic management across the city and help in making better policy decisions.

ACKNOWLEDGEMENT

The authors wish to thank the faculty and technical staff of the Department of Information Technology, JSS Academy of Technical Education, Noida, for their support and guidance throughout this work.

References

[1] World Health Organization and National Crime Records Bureau, "Road traffic injuries in India," 2022.
[2] L. Bieker-Walz, "Green Wave Simulation," SUMO Proc., 2019.
[3] J. Chen, R. Hu, and B. Wang, "Audio-visual fusion-based multi-modal traffic monitoring system," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 6, 2020.
[4] P. Kumar and R. Gupta, "Multi-sensor fusion for intelligent transportation systems: A survey," IEEE Intell. Transp. Syst. Mag., vol. 11, no. 2, 2019.
[5] A. Silva and J. Torres, "Real-time vehicle detection and tracking using YOLOv8 for intelligent transportation," Sensors, vol. 23, 2023.
[6] Q. Kong et al., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 28, 2020.
[7] P. V. Reddy and K. Srinivas, "Deep Learning Preemption," Int. J. Adv. Comput. Sci. Appl., 2020.
[8] S. Wang, Y. Li, and H. Zhang, "YOLO-based emergency vehicle detection in urban traffic," in Proc. ICCT, 2019.
[9] R. Patel et al., "Siren Recognition in Noisy Cities Using Deep CNN," in Proc. IEEE ICASSP, 2022.
[10] M. Zohaib et al., "ATSN Multimodal Fusion," Mathematics, 2024.
[11] S. Bhatia et al., "YOLOv3 Smart Traffic," in Proc. IEEE ITSC, 2021.
[12] L. Zhang and H. Su, "Night-time emergency vehicle recognition using optimized deep neural networks," IEEE Trans. Veh. Technol., vol. 70, no. 9, 2021.
[13] P. Nguyen and A. Kumar, "Deep learning framework for emergency siren classification in real-world conditions," in Proc. IEEE ICASSP, 2024.
[14] J. F. Gemmeke et al., "AudioSet: An ontology and human-labeled dataset for audio events," in Proc. IEEE ICASSP, 2017.