False Positive Reduction in Emergency Vehicle Detection Using a Multimodal Edge-Based System

Indra Kristiawan
Master of Mechanical Engineering Study Program, Swiss German University, Tangerang, Indonesia | Divisi Kecerdasan Buatan, PT. Piranti Kecerdasan Buatan, Tangerang, Indonesia
indra.kristiawan@student.sgu.ac.id

Maulahikmah Galinium
Information Technology Department, Swiss German University Tangerang, Indonesia
maulahikmah.galinium@sgu.ac.id

Dwi Ahmad Dzulhijjah
Department of Cyber Physical System, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia | Divisi Kecerdasan Buatan, PT. Piranti Kecerdasan Buatan, Tangerang, Indonesia
dwiahmad@pasca.student.pens.ac.id

Kusrini
Informatics Postgraduate Program, Universitas AMIKOM Yogyakarta, Indonesia | Fundacion para la Investigacion y Desarrollo Tecnologico de la Sociedad del Conocimiento, Murcia, Spain
kusrini@amikom.ac.id

Bima Sena Bayu Dewantara
Department of Cyber Physical Systems, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia
bima@pens.ac.id

Henry Nasution
Swiss German University, Tangerang Indonesia | Teknologi Rekayasa Energi Terbarukan, Fakultas Teknologi Industri, Universitas Bung Hatta, Padang, Indonesia
henrynasution@bunghatta.ac.id (corresponding author)

Received: 24 July 2025 | Revised: 9 September 2025 | Accepted: 15 September 2025

 DOI: https://doi.org/10.48084/etasr.13619

ABSTRACT

Traditional vision-based emergency vehicle detection systems used in Adaptive Traffic Signal Control (ATSC) suffer from high False Positive (FP) classifications that compromise traffic flow efficiency and system reliability. This study proposes a multimodal detection framework that integrates a YOLOv5 visual detector with an ESP32-S3 acoustic analyzer using Mel-Filterbank Energy (MFE) features, with dual-modality confirmation achieved through AND-gate fusion. Field evaluation conducted over 11 days and comprising 1,380 detection events demonstrated a complete FP elimination from 900 to 0 cases, while preserving detection capability and achieving a precision of 100%, a specificity of 100%, a recall of 87.1%, a standard accuracy of 99.4%, and a preemption rate of 87.1%, corresponding to a 55.3 percentage point/% improvement over the vision-only baseline. These results confirm the effectiveness of selective acoustic confirmation in reducing detection ambiguities, maintaining real-time responsiveness, and enhancing the robustness of emergency vehicle detection in urban traffic management systems.

Keywords-Adaptive Traffic Signal Control (ATSC); computer vision; audio detection; emergency vehicle detection; sustainable cities and communities

I. INTRODUCTION

Emergency situations involving ambulances, fire trucks, and police vehicles require immediate and prioritized passage through traffic signal preemption to ensure rapid response. This is because, according to World Health Organization (WHO), emergency medical service response time is a key determinant of patient survival, with delays substantially increasing mortality risk [1-3]. The Area Traffic Control Systems (ATCSs), have therefore gained increasing attention as an enabling technology, with recent systems incorporating Computer Vision (CV) for automated detection and prioritization of emergency vehicles [4-6]. However, vision-based detection remains prone to significant limitations, with the most persistent issue being FP. For instance, it is common for non-emergency vehicles to be misclassified as ambulances due to headlight glare, reflective surfaces, or the presence of inactive emergency vehicles [7, 8]. Thereby, such misclassifications compromise detection reliability and hinder accurate traffic-signal preemption.

Existing ATCS implementations in Indonesia, including deployments in Jakarta [5], Balikpapan [9], Bandar Lampung [10], Batam [11], and Semarang [12], as well as broader reviews of domestic practice [13], largely rely on rule-based or schedule-driven control without integrated CV or acoustic confirmation. This reliance on conventional coordination highlights a gap between existing deployments and modern Artificial Intelligence (AI)-enabled traffic management systems.

Advanced CV models [14], such as YOLOv8, report precision levels exceeding 98% under benchmark conditions [15]; however, their performance often degrades in real-world environments where reflections, strobe effects, or ambiguous shapes are common [16]. Consequently, relying solely on vision-only approaches results in high False Positive Rates (FPRs) that undermine the credibility of ATCS.

To improve reliability, acoustic detection has been explored as a complementary modality. Recent deep learning approaches demonstrate promising results, including ensemble models for emergency siren recognition [17], public vehicle sound detectors [18], large-scale datasets incorporating sirens and road noise [19], and comprehensive surveys of acoustic sensors for transportation [20]. Nevertheless, these methods face challenges when deployed in uncontrolled urban traffic. For instance, traditional audio techniques, such as Fourier-based analysis [21] and Mel-Frequency Cepstral Coefficients (MFCC)-based classifiers [22], also perform well in controlled environments but exhibit reduced robustness in noisy city environments [23, 24]. In addition, continuous audio monitoring increases computational load and energy consumption on edge devices.

To address the limitations of unimodal detection, multimodal fusion strategies have been proposed [25]. For example, vision-based strobe detection can be more effective at night, whereas audio-based detection performs better during daytime conditions [26]. However, despite parallel and continuous fusion methods [27, 28] improving robustness, they often introduce additional latency and high resource consumption, limiting their suitability for real-time deployment.

This study proposes a robust and computationally efficient multimodal architecture that integrates vision and audio using asymmetric fusion logic. The framework employs YOLOv5 for visual detection and an ESP32-S3 module for selective siren verification based on MFE features [29]. Fusion is implemented using AND-gate logic [30], so acoustic verification is activated only after a visual candidate has been detected. Unlike prior works that use parallel fusion [8, 25] or audio-dominant pipelines [28], the proposed system prioritizes vision for spatial localization and leverages audio purely for confirmation. This design reduces computational overhead, improves energy efficiency, and enforces strict cross-modal validation, thereby substantially minimizing FP while preserving high recall in real-world ATCS deployments.

II. METHOD

A. Multimodal Overall Framework

The overall architecture of the proposed model is illustrated in Figure 1.

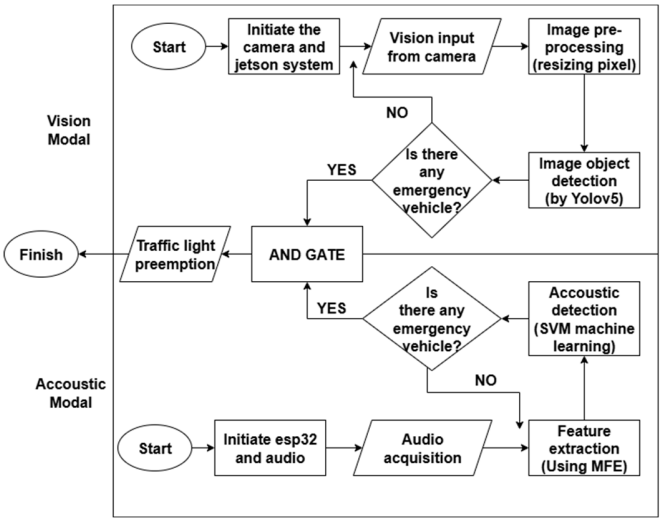

Fig. 1. Proposed multimodal framework for emergency vehicle detection.

The overall system operates as follows:

- Input Section: A dedicated narrow-angle lens streams raw frames to the vision module, while the acoustic module remains inactive initially to conserve computational resources and power.
- Emergency Vehicle Detection (Vision): The YOLOv5 algorithm identifies potential emergency vehicles through the vision module. When a candidate is detected, an EV_CANDIDATE signal is raised to trigger the acoustic verifier.
- Selective Audio Verification: The ESP32-S3 acoustic module is activated on-demand to record a 640 ms audio window, extract MFE features, and perform siren

classification using a lightweight Convolutional Neural Network (CNN).

- AND-Gate Fusion and Actuation: Preemption occurs only when both vision and acoustic modules confirm an active emergency vehicle, suppressing FP while ensuring reliable actuation.

B. Computer Vision Module: YOLOv5 for Emergency Vehicle Detection

The visual detection process utilizes an optimized YOLOv5 model, specifically enhanced for emergency vehicle detection on edge devices. System specifications include:

- Hardware: Jetson Nano 4GB (1.4 GHz Central Processing Unit (CPU) / 921 MHz Graphics Processing Unit (GPU), ≈4.1 W under load).
- Model Specifications: YOLOv5-s with pruning and quantization-aware training (v4q), 112 convolutional layers after optimization, input resolution 640×352 pixels, precision 82% on held-out urban-night subset, and FPR ≈18% (motivating multimodal expansion).

Additionally, the enhanced model incorporates adaptive Non-Maximum Suppression (NMS) and perspective-aware Region of Interest (ROI) masking to improve detection accuracy while maintaining real-time performance on resource-constrained edge devices.

C. Audio Processing Module: ESP32-S3 Acoustic Modal

The acoustic verification system operates as a selective confirmatory mechanism, activated only upon visual detection events to minimize computational overhead and power consumption. The hardware configuration included:

- Microcontroller: ESP32-S3-WROOM-1 of 240 MHz, and 512 kB Static Random Access Memory (SRAM).
- Microphone: INMP441 I^2S digital Micro-Electro-Mechanical-System (MEMS) microphone.
- Power Consumption: 118 mW (active state).
- Enclosure: Shielded with 45° tilt orientation.

Moreover, the signal processing pipeline included:

- Audio Acquisition: 640 ms recording window triggered by EV_CANDIDATE signal.
- Windowing: 512-sample Hamming window at 32 kHz sampling rate.
- Feature Extraction:
  a) 512-point Short-Time Fourier Transform (STFT).
  b) 32-band MFE computation.
  c) Log-compression for dynamic range optimization.

The classification architecture included:

- Network Type: 6-layer depthwise-separable CNN.
- Parameters: 13,952 total parameters (14.2 kB flash memory).
- Architecture: Conv1D → ReLU → Conv1D → GlobalMaxPool → Dense(softmax).
- Decision Threshold: P(siren) ≥ 0.80 → AUDIO_PASS, else AUDIO_FAIL.

Finally, the performance specifications included a feature extraction latency of 3.9 ms, an inference latency of 4.6 ms, and a total processing time of 8.5 ± 0.7 ms.

D. Edge Deployment and Optimization

The proposed system is optimized for edge deployment with the following characteristics:

- Resource Optimization: Vision processing leverages Jetson Nano GPU acceleration; acoustic processing runs independently on ESP32-S3, with selective activation reducing computational load by ~80%.
- Real-Time Performance: Total latency <150 ms (within traffic signal actuation requirements), memory footprint 24 kB Random Access Memory (RAM) / 33 kB flash for acoustic module, and low power consumption due to selective activation.
- Performance Targets: FPR <5% (safety threshold for traffic signal actuation); precision >95% for reliable emergency vehicle preemption; FP reduction ~80% improvement over vision-only baseline.

This asymmetric multimodal architecture balances detection reliability and computational efficiency, ensuring rapid emergency vehicle prioritization suitable for urban traffic signal control applications.

E. Dataset Description

The vision model (YOLOv5) was initially trained using the AMBULANCE DETECTION dataset [31], a publicly available dataset hosted on Roboflow Universe comprising 2,036 annotated images with a single target class: ambulance. The dataset was partitioned into 88% training (1,782 images), 8% validation (168 images), and 4% testing (86 images). Prior to training, all images underwent preprocessing, including auto-orientation and resizing to 640×640 pixels using stretch interpolation. To improve model generalization, data augmentation techniques were applied, generating three augmented outputs per training sample. These augmentations included horizontal and vertical flipping, random rotations within ±15°, grayscale conversion applied to 15% of samples, and Gaussian blur with a kernel size of up to 2.5 pixels.

The trained model was subsequently evaluated on 282,479 frames extracted from continuous ATCS surveillance camera recordings in Magelang, Central Java, Indonesia, during July-August 2024. Among these frames, 9,137 contained positive ambulance instances.

The acoustic classification model was developed using an audio dataset hosted on the Edge Impulse platform (Project ID: 699816) [32]. The dataset consists of 28 audio samples, each with a duration of 10 s, recorded at a sampling rate of 16 kHz,

totaling 4 min and 40 s of labeled audio data. Two classes are defined: i) ambulance, consisting of recordings of active Indonesian ambulance sirens (10 samples), and ii) noise, containing urban background sounds such as traffic, vehicle horns, construction activity, and general environmental acoustics (18 samples). The dataset was split into 24 training samples (9 ambulance-class and 15 noise-class) and 4 test samples (1 ambulance-class and 3 noise-class). Only ambulance sirens were treated as the positive class, while all other urban sounds were categorized as noise. When quantized to INT8 for on-device deployment on the ESP32, the model required 962 ms inference latency, 62.5 KB peak RAM, and 426.7 KB flash memory. In addition to this dataset-based evaluation, a standalone audio-only field evaluation was conducted separately using 100 real-world activation events to assess the model's deployment performance in isolation.

For the multimodal field evaluation, detection data were collected over an 11-day period from July 3-13, 2025, near Tidar Hospital junction, Magelang. The vision module, running on a Jetson Nano 4GB, was connected to a narrow-angle traffic camera mounted on a traffic pole at approximately 2 m above road level, facing the incoming traffic lane. The acoustic module, comprising an ESP32-S3 microcontroller with an INMP441 MEMS microphone housed in a shielded enclosure with 45° tilt orientation, was installed at approximately 1 m from the road edge. The system operated daily from 07:00 to 21:00 local time (GMT+7), covering both peak and off-peak traffic conditions, including morning rush hours, midday periods, and evening congestion. During this period, the system recorded 1,380 detection event timestamps during daily operational hours. Ground truth annotation was established through post-hoc review of synchronized video and audio recordings, identifying 70 true emergency events, defined as instances where both a visible emergency vehicle and an audibly active siren were present.

III. RESULTS AND DISCUSSION

A. Preliminary Vision-Based Performance

Initial experiments using a vision-only approach with YOLOv5 achieved a mean Average Precision (mAP) of 79% under benchmark conditions, evaluated on a held-out test split of the AMBULANCE DETECTION dataset comprising 86 images. However, during real-world deployment, performance was re-evaluated using the precision metric:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

where TP represents the number of True Positives (correctly detected emergency vehicles) and FP denotes the number of False Positives (non-emergency vehicles misclassified as ambulances).

Based on real-time deployment results on the complete set of 282,479 frames, the system recorded 8,950 TP and 3,584 FP, yielding a precision of 71.41%, a recall of 97.95%, and an F1-score of 82.9%. The discrepancy between benchmark mAP and field precision highlights inherent limitations of the vision-only configuration when exposed to unconstrained environmental conditions. Accuracy is intentionally excluded from this evaluation as the dataset is severely imbalanced, as only 9,137 positive frames exist out of 282,479 total (~3.2%). Under such conditions, a model that never detects anything would still achieve 96.8% accuracy, rendering the metric statistically meaningless for this application.

In addition, error analysis indicates that FP were primarily driven by visual ambiguities arising from real-world variability. According to Figure 2, these include:

- A reflective white vehicle that was misidentified as an ambulance due to similarities in shape and surface reflection.
- Low-light glare at night, where overexposed headlights created visual artifacts resembling emergency lighting.
- A long-range detection scenario in which the vehicle's appearance became ambiguous under reduced illumination.
- A visually complex urban scene affected by harsh lighting contrast and dense traffic elements, which introduced further inconsistencies in the classification process.

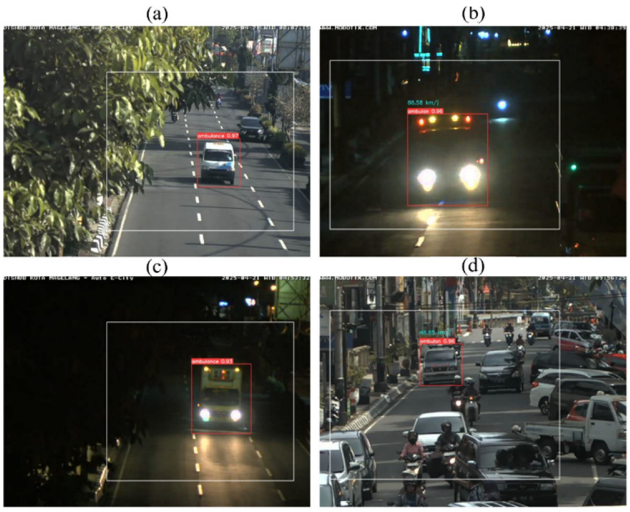

Fig. 2. FP cases.

B. Audio-Based Performance

The audio classification model, implemented on the ESP32-S3 module, demonstrated promising results under controlled laboratory conditions. Under full-precision (float32) validation on the Edge Impulse platform [32], the MFE-based model achieved 100% accuracy with perfect class separation on the 28-sample dataset. Following INT8 quantization for edge deployment on the ESP32-S3, accuracy was measured at 82% in recognizing ambulance sirens, a reduction expected from the quantization process, but acceptable for real-time embedded operation. Additionally, the model achieved a True Negative (TN) rate of 98%, effectively filtering out non-emergency environmental sounds such as vehicle horns and construction noise, while during real-world evaluation, the system achieved a TP detection rate of 91% for active siren cases. However, in practical deployment scenarios, precision becomes a more appropriate metric to assess reliability. Based on 100 recorded

activation events, only 68 corresponded to actual ambulance sirens, resulting in an estimated field precision of 68%. This degradation relative to laboratory performance was not attributable to spoofing (e.g., recorded siren playback), but rather to environmental acoustic complexity, as overlapping urban noise and inconsistent signal propagation frequently masked or mimicked siren patterns.

The audio signal processing using the MFE extraction method is visualized in Figure 3. The upper panel depicts the temporal-spectral energy distribution of an ambulance siren, where dominant frequency bands, typically between 1000 and 4000 Hz, serve as distinguishing features relative to background urban sounds. The lower panel shows the corresponding Fast Fourier Transform (FFT) bin weighting applied during feature extraction.

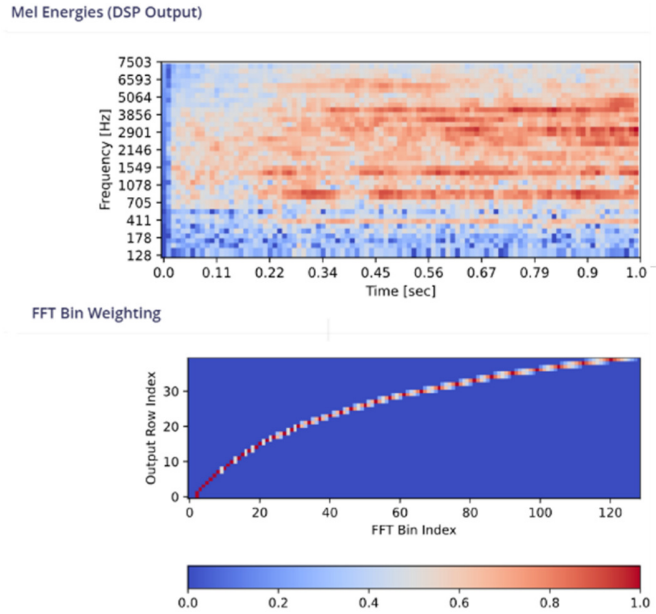

Fig. 3. DSP results of the ESP-32-S3 sense module.

The deployed processing pipeline introduces an end-to-end latency of approximately 290 ms, primarily driven by Digital Signal Processing (DSP) configuration parameters (window length: 2000 ms; hop length: 1000 ms; 32 filter banks). Although this exceeds the ideal responsiveness target of 100 ms for traffic signal preemption, the implementation remains highly memory-efficient, requiring only 24 kB of RAM and 33 kB of Read-Only Memory (ROM) on the ESP-32-S3 sense module. These results confirm the module's suitability for edge deployment while also underscoring the limitations of relying solely on audio sensing in complex urban environments.

C. Field Deployment and Multimodal System Evaluation

Across the deployment period, 1,380 detection event timestamps were recorded. The vision module produced 1,321 detections, including: i) 421 instances where an emergency vehicle was present but without an active siren, and ii) 900 FP where no emergency vehicle was present. An additional 50 ambulance instances were missed entirely by the vision module.

In contrast, the multimodal system, employing AND-gate fusion of visual and acoustic signals, successfully confirmed 61 of the 70 true emergency events, while nine events were missed, three due to acoustic interference from high ambient noise and six due to visual occlusion by other vehicles. No ghost detections (audio-only triggers without visual confirmation) and no failed acoustic confirmations (cases where both visual presence and siren were present, but audio failed) were observed.

A detailed summary of detection events is presented in Table I, while the comparative performance results between the vision-only and multimodal systems are summarized in Table II.

TABLE I. SUMMARY OF DETECTION EVENTS (11 DAYS, 07:00-21:00)

Criteria	Count
Total timestamps (event windows)	1,380
Vision detections (total)	1,321
Vision detected, true emergency vehicle (no siren active)	421
FP (vision detected, but no emergency at all)	900
Ambulance present but missed by vision	50
True emergency events (vehicle + siren) as ground truth	70
Multimodal detected (confirmed by both vision and acoustic modules)	61
Multimodal missed (3 missed by audio, 6 missed by vision)	9
Emergency events detected by vision, siren active, and audio confirmed	61
Vision detected + siren active, but NOT confirmed by acoustic module	0
Ghost detections: audio-only triggered when no vision or emergency vehicle was present	0

From Table II, the vision-only system exhibited high sensitivity but poor specificity. Specifically, although recall reached 87.1%, precision was only 6.35%, as approximately 68% of detections were false alarms. Additionally, specificity was similarly low at 31.3%, reflecting the system's limited ability to correctly reject non-emergency events. Standard accuracy was 34.1% and the F1-score 11.8%, rendering the system unreliable for autonomous traffic signal actuation.

In contrast, the multimodal system achieved 100% precision while preserving an identical recall of 87.1%, resulting in an F1-score of 93.1%, an improvement of 81.3%. Furthermore, standard accuracy increased to 99.4%, reflecting the complete elimination of FP. These results confirm that dual-modality confirmation through AND-gate fusion effectively addresses the specificity limitation inherent in vision-only detection and achieves a balanced trade-off between sensitivity and precision, suitable for real-world traffic signal preemption.

TABLE II. COMPREHENSIVE PERFORMANCE FIELD STUDY RESULTS (1,380 DETECTION EVENTS)

Performance metric	Vision-only system	Multimodal system	Improvement
TP (emergency events)	61	61	Equal coverage
TN	410	1310	+900
FP (emergency false alarm)	900	0	>90% reduction
False Negatives (FN) (missed emergencies)	9	9	Equal
Accuracy	34.1%	99.4%	+65.3%
Precision	6.35%	100%	+93.65%
Recall (sensitivity)	87.1%	87.1%	0%
F1-score	11.8%	93.1%	+81.3%
FPR	68.7%	0%	−68.7%
Preemption rate	31.8%	87.1%	+55.3%

D. Discussion and Implications

The poor performance of the vision-only system is consistent with known limitations of vision-based detectors in uncontrolled urban environments, where reflective surfaces, headlight glare, and inactive emergency vehicles are indistinguishable from active ones at the feature level.

The introduction of acoustic verification through AND-gate fusion changed the system's decision boundary. Rather than attempting to resolve visual ambiguity through more complex vision models, which would increase computational cost and still struggle with certain edge cases, the proposed architecture introduces a second, independent modality that directly addresses the root cause of FP: the absence of an active siren. This design reflects a key insight that visual similarity alone is insufficient for emergency vehicle preemption, but acoustic confirmation is highly discriminative since active sirens produce a distinctive spectral signature that passive vehicles cannot replicate.

As a result, the multimodal system achieved a much more competitive all-around performance which is able to be robust and consistent under diverse real-world conditions, including daytime, nighttime, rush-hour congestion, and varying weather.

Analyzing the nine missed detections revealed two independent failure modes. Three were caused by acoustic interference in high-noise environments, where siren frequencies were masked by overlapping urban sound sources, while six resulted from visual occlusion, where the emergency vehicle was partially or fully blocked from the camera field of view. These findings suggest that future improvements could target each modality separately; for example, through noise-robust audio preprocessing and multi-camera coverage or trajectory prediction.

From a practical standpoint, achieving 100% precision carries significant operational value. In traffic signal preemption, FP disrupt traffic flow unnecessarily and erode public trust in automated systems. By eliminating FP, the proposed system meets the reliability threshold required for unsupervised deployment. Furthermore, the asymmetric activation design, in which the acoustic module is triggered only after visual detection, reduces continuous computational load by approximately 80%, supporting long-term edge deployment without excessive power consumption.

Overall, these findings affirm that selective acoustic confirmation is not merely an incremental improvement over vision-only detection, but a qualitative shift in system reliability.

IV. CONCLUSION

This study presents a robust and resource-efficient multimodal detection framework for emergency vehicle identification in Adaptive Traffic Signal Control (ATSC) systems. By integrating a YOLOv5-based vision module with an ESP32-S3-based acoustic analyzer using Mel-Filterbank Energy (MFE) features and employing asymmetric AND-gate fusion logic, the proposed system effectively addresses the long-standing challenge of False Positives (FP) in vision-only detection. Deployed over an 11-day field study with 1,380 detection events, the multimodal system improved preemption rate from 31.8% to 87.1% and standard accuracy from 34.1% to 99.4%, while completely eliminating all 900 FP. The framework preserved high recall (87.1%) and achieved 100% precision and 100% specificity, validating the effectiveness of dual-modality confirmation for real-world traffic control. Unlike continuous or parallel multimodal fusion approaches, the proposed asymmetric architecture activates the acoustic module only upon visual trigger, ensuring efficient computation and low power consumption, suitable for edge deployment.

Overall, these findings demonstrate that incorporating selective audio verification into vision-based ATSC significantly enhances both the precision and robustness of emergency vehicle detection, offering a practical and scalable solution for intelligent urban traffic management.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support and facilities provided by PT. Piranti Kecerdasan Buatan (PIKEBU), which greatly contributed to the realization of this study.

REFERENCES

[1] M. S. Peelam, Naren, M. Gera, V. Chamola, and S. Zeadally, "A Review on Emergency Vehicle Management for Intelligent Transportation Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 15229–15246, Nov. 2024, https://doi.org/10.1109/TITS.2024.3440474.

[2] T. S. Geetha, C. Chellaswamy, and T. Kaliraja, "Pre-emption system for emergency medical service vehicles: a deep learning approach," *International Journal of General Systems*, pp. 1–33, Mar. 2025, https://doi.org/10.1080/03081079.2025.2481898.

[3] A. Mohanty, A. G. Mohapatra, and S. K. Mohanty, "Real-Time Traffic Monitoring with AI in Smart Cities," in *Internet of Vehicles and Computer Vision Solutions for Smart City Transformations*, A. Abraham, S. Prasad, A. Alhammadi, T. Lestable, and F. Chaabane, Eds. Cham: Springer Nature Switzerland, 2025, pp. 135–165.

[4] E. Warni, A. A. P. Alimuddin, A. E. U. Salam, M. Fachri, and M. Rizal H., "Enhancing Traffic Counting in Rainy Conditions: A Deep Learning Super Sampling and Multi-ROI Pixel Area Approach," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20095–20101, Feb. 2025, https://doi.org/10.48084/etasr.9515.

[5] N. Tania and R. Rachmawati, "Area Traffic Control System (ATCS) for Supporting Urban Traffic Management in DKI Jakarta," in *2022 7th International Conference on Electric Vehicular Technology (ICEVT)*, Bali, Indonesia, Sept. 2022, pp. 103–108, https://doi.org/10.1109/ICEVT55516.2022.9924924.

[6] C. Subba Rao, C. Chellaswamy, T. S. Geetha, and S. Arul, "Deep Learning Based Decision Support Framework for Dead Reckoning in Emergency Vehicle Preemption," *International Journal of Intelligent Transportation Systems Research*, vol. 22, no. 1, pp. 117–135, Apr. 2024, https://doi.org/10.1007/s13177-023-00384-y.

[7] M. Ashkanani, A. AlAjmi, A. Alhayyan, Z. Esmael, M. AlBedaiwi, and M. Nadeem, "A Self-Adaptive Traffic Signal System Integrating Real-Time Vehicle Detection and License Plate Recognition for Enhanced Traffic Management," *Inventions*, vol. 10, no. 1, Feb. 2025, Art. no. 14, https://doi.org/10.3390/inventions10010014.

[8] A. Mecocci and C. Grassi, "RTAIAED: A Real-Time Ambulance in an Emergency Detector with a Pyramidal Part-Based Model Composed of MFCCs and YOLOv8," *Sensors*, vol. 24, no. 7, Apr. 2024, Art. no. 2321, https://doi.org/10.3390/s24072321.

[9] N. L. W. R. Kurniati, "Optimizing the Performance of the Area Traffic Control System (ATCS) in Balikpapan City," *Jurnal Penelitian Transportasi Darat*, vol. 21, no. 2, pp. 155–164, June 2020, https://doi.org/10.25104/jptd.v21i2.1567.

[10] A. Efendi and M. N. Juwita, "Effectiveness of the Area Traffic Control System (ATCS) Program in Improving Traffic Flow in the City of Bandar Lampung (A Study of the Bandar Lampung City Transportation Agency)," *Triwikrama: Jurnal Ilmu Sosial*, vol. 3, no. 3, pp. 61–70, Mar. 2024.

[11] E. F. D. Ngasi and F. A. Wulandari, "Performance Analysis of the Batam City Transportation Agency in Implementing the Area Traffic Control System (ATCS) Program in Batam City in 2019", *Aufklarung: Jurnal Pendidikan, Sosial Dan Humaniora*, vol. 3, no. 3, pp. 216–227, Sep. 2023.

[12] D. H. Sunyoto, F. Ramadhan, and R. Ruktiningsih, "Study on the Implementation of Area Traffic Control System (ATCS) at Several Intersections in Semarang City," *G-SMART,* vol. 3, no. 1, Feb. 2021, Art. no. 35, https://doi.org/10.24167/gs.v3i1.1766.

[13] S. K. I. Aini, S. A. B. Putri, and I. Darmawan, "Implementation of Area Traffic Control System (ATCS) in various regions in Indonesia as an application of smart mobility," *Jurnal Multidisiplin Ilmu Akademik*, vol. 1, no. 6, pp. 428–439, Dec. 2024.

[14] A. Patel, S. Degadwala, and D. Vyas, "Enhancing Traffic Management with YOLOv5-Based Ambulance Tracking System," in *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, Sept. 2023, pp. 528–532, https://doi.org/10.1109/CCECE58730.2023.10288751.

[15] A. Noor *et al.*, "A Cloud-Based Ambulance Detection System Using YOLOv8 for Minimizing Ambulance Response Time," *Applied Sciences*, vol. 14, no. 6, Mar. 2024, Art. no. 2555, https://doi.org/10.3390/app14062555.

[16] C. W. Adi, B. N. Santoso, A. Sonhaji, and I. Kristiwan, "Active Strobe Detection Device for Priority Vehicles Based on Artificial Intelligence Integrated with Traffic Light Controls," 2021/SID/02696, Nov. 01, 2021.

[17] U. Mittal and P. Chawla, "Acoustic Based Emergency Vehicle Detection Using Ensemble of deep Learning Models," *Procedia Computer Science*, vol. 218, pp. 227–234, 2023, https://doi.org/10.1016/j.procs.2023.01.005.

[18] V. Sathya, H. Naveen Kumar, V. Saiganesh, and J. Sakthivel, "Siren Detector in Public Vehicles," in *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Chennai, India, Oct. 2024, pp. 1–6, https://doi.org/10.1109/ICPECTS62210.2024.10780249.

[19] M. Y. Shams, T. Abd El-Hafeez, and E. Hassan, "Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset," *Expert Systems with Applications*, vol. 249, Sept. 2024, Art. no. 123608, https://doi.org/10.1016/j.eswa.2024.123608.

[20] H. Parineh, M. Sarvi, and S. A. Bagloee, "Acoustic Sensors and Audio Signal Processing in Intelligent Transportation Systems: A Survey," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 8, pp. 4171–4190, Aug. 2025, https://doi.org/10.1109/TIV.2024.3476475.

[21] B. Fatimah, A. Preethi, V. Hrushikesh, A. Singh B., and H. R. Kotion, "An automatic siren detection algorithm using Fourier Decomposition Method and MFCC," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, July 2020, pp. 1–6, https://doi.org/10.1109/ICCCNT49239.2020.9225414.

[22] D. C. Chinvar, M. Rajat, R. L. Bellubbi, S. Sampath, and K. Guddad, "Ambulance Siren Detection using an MFCC based Support Vector Machine," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, Karnataka, India, Dec. 2021, pp. 1–5, https://doi.org/10.1109/ICMNWC52512.2021.9688340.

[23] M. Usaid, M. Asif, T. Rajab, M. Rashid, and S. I. Hassan, "Ambulance Siren Detection using Artificial Intelligence in Urban Scenarios," *Sir Syed University Research Journal of Engineering & Technology*, vol. 12, no. 1, pp. 92–97, June 2022, https://doi.org/10.33317/ssurj.467.

[24] M. Cantarini, L. Gabrielli, and S. Squartini, "Few-Shot Emergency Siren Detection," *Sensors*, vol. 22, no. 12, June 2022, Art. no. 4338, https://doi.org/10.3390/s22124338.

[25] M. Zohaib, M. Asim, and M. ELAffendi, "Enhancing Emergency Vehicle Detection: A Deep Learning Approach with Multimodal Fusion," *Mathematics*, vol. 12, no. 10, May 2024, Art. no. 1514, https://doi.org/10.3390/math12101514.

[26] K. Choroś, "Automatic Detection of Ambulance Vehicles in Day and Night Conditions in Surveillance Videos," in *Advances in Computational Collective Intelligence*, vol. 2166, N.-T. Nguyen, B. Franczyk, A. Ludwig, M. Nunez, J. Treur, G. Vossen, and A. Kozierkiewicz, Eds. Cham: Springer Nature Switzerland, 2024, pp. 327–337.

[27] S. Gopinathan, B. Abishek, G. Kathiravan, P. B. Roshith, and V. Bharath, "Smart Ambulance Traffic Sensing using Artificial Intelligence and Internet of Things," in *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai, India, Apr. 2024, pp. 1–5, https://doi.org/10.1109/IC3IoT60841.2024.10550228.

[28] S. Shilaskar *et al.*, "Multimodal detection of ambulance using Jetson nano," in *International Conference on Innovations in Computer Science, Electronics & Electrical Engineering-2022*, Andhra Pradesh, India, 2023, Art. no. 030003, https://doi.org/10.1063/5.0130361.

[29] S. Hymel *et al.*, "Edge Impulse: An MLOps Platform for Tiny Machine Learning." arXiv, Apr. 2023, https://doi.org/10.48550/arXiv.2212.03332.

[30] X. Gao, B. Cao, P. Zhu, N. Wang, and Q. Hu, "Asymmetric Reinforcing against Multi-modal Representation Bias." arXiv, 2025, https://doi.org/10.48550/ARXIV.2501.01240.

[31] *AMBULANCE DETECTION Dataset*. (2024), AmbulanceDetection. [Online]. Available: https://universe.roboflow.com/ambulancedetection/ambulance-detection-u4ao4.

[32] *Ambulance Siren Detector*. (2025), I. Kristiawan and D. A. Dzulhijjah. [Online]. Available: https://studio.edgeimpulse.com/public/699816/live.