RESEARCH PAPER

An Iot-Based Multimodal Assistive Alert System using YOLOv11 and YAMNet on Wearable Eyewear with SOS and GPS Tracking for Autistic and Hearing Impaired individuals

Debmitra Ghosh*[1], Dr. Dipankar Misra[1], Dr. Dharmpal Singh[1], Dr. Ashoktaru Pal[2], Dr. Prosenjit Mukherjee[3] and Dr. Susmita Das[4]

*[1]*Computer Science and Engineering JIS University*
[2]CA Department, Future Institute of Engineering & management, Kolkata, West Bengal, India
[3]IT Department, Future Institute of Engineering & Management, Kolkata, West Bengal, India
[4]Electronics and Computer Science, Narula Institute of Technology, Kolkata, West Bengal, India
**Corresponding Author(s). E-mail(s) debmitra.ghosh@jisuniversity.ac.in*

Received: 28th Feb, 2026; Revised: 6th March 2026; Accepted: 7th April, 2026; Available Online: 20th April, 2026

ABSTRACT
Pedestrians with Autism Spectrum Disorder (ASD) and hearing impairment face increased safety risks in traffic environments due to limited perception of auditory cues and difficulty interpreting dynamic surroundings. Conventional pedestrian assistance systems rely heavily on auditory alerts or infrastructure-dependent solutions, which may not be effective or accessible for individuals with sensory impairments. This paper presents a wearable,edge-based multimodal assistive alert system designed to enhance pedestrian safety using real-time vision and audio intelligence.The proposed system integrates a YOLOv11-based vehicle detection module with a YAMNet-based environmental sound recognition module, deployed on a Raspberry Pi platform embedded into wearable eyewear. Vision-based detection identifies approaching vehicles, while motion and distance estimation classify risk levels. Audio-based detection recognizes emergency sounds such as ambulance sirens to provide early warnings. Amulti-level alert mechanism delivers vibration, visual, and audio cues based on risk severity. The system also includes a manual SOS button and GSM-based GPS alerting to ensure emergency communication in low-connectivity environments.Experimental evaluation demonstrates high detection accuracy, strong recall, reliable mean Average Precision (mAP), and low inference latency suitable for real-time deployment. The results confirm that combining pretrained deep learning models with lightweight signal processing enables a practical, privacy-preserving assistive solution for vulnerable pedestrians.

Keywords: *Assistive Technology; YOLOv11; YAMNet; Wearable Eyewear; Pedestrian Safety; Autism Spectrum Disorder; Hearing Impairment; Edge AI; IoT*
How to cite this article: Ghosh D, Misra D, Singh D, Pal A, Mukherjee P, Das S. An IoT-Based Multimodal Assistive Alert System using YOLOv11 and YAMNet on Wearable Eyewear with SOS and GPS Tracking for Autistic and Hearing Impaired Individuals. Int J Drug Deliv Technol. 2026;16(59s): 825-832. DOI: 10.25258/ijddt.16.59s.96
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

Urban traffic environments demand continuous situational awareness, rapid decision-making, and accurate perception of visual and auditory cues. For pedestrians with Autism SpectrumDisorder (ASD) and hearing impairment, these requirements pose significant challenges. Individuals with ASD often experience sensory overload, delayed response to stimuli, and difficulty interpreting complex environmental changes, while hearing-impaired individuals may not perceive criticalauditory warnings such as horns or emergency sirens. These limitations substantially increase the risk of road accidents.Existing pedestrian safety systems primarily rely on auditory alerts, smartphoneapplications, or infrastructure-based solutions such as smart traffic signals. However, auditory alerts are ineffective for hearing-impaired users, and smartphone-based solutions introduce latency, privacy concerns, and dependency on continuous internet connectivity. Moreover, many assistive systems are not designed to be wearable, discreet, or adaptive to individual sensory needs. Recent advances in deep learning and edge computing have enabled real-time perception systems on low-power devices. Object detection models from the YOLO family provide fast and accurate vision-based detection, while pretrained audio classification networks such as YAMNet enable robust environmental sound recognition without extensive retraining. These models are well suited for assistive applications where low latency, energy efficiency, and privacy preservation are critical.This work proposes a **wearable multimodal assistive alert system** that combines **YOLOv11-based vehicle detection** and **YAMNet-based emergency sound detection** on a **Raspberry Pi** platform embedded into eyewear. The

**Author for Correspondence: debmitra.ghosh@jisuniversity.ac.in*

system operates entirely on the edge, minimizing cloud dependency, and provides adaptive alerts through haptic feedback, visual cues, and optional audio messages. Additionally, a GSM-based SOS and GPS tracking module ensures reliable emergency communication.

The key contributions of this paper are:

1. Design of a wearable, eyewear-agnostic assistive system for autistic aDesign of a **wearable, eyewear-agnostic assistive system** for autistic and hearing-impaired pedestrians
2. Integration of vision-based vehicle detection (YOLOv11) and audio-based emergency sound recognition (YAMNet)
3. Motion- and distance-aware **risk classification with multi-level alerts**
4. Fully **edge-based implementation** with GSM-based emergency communication
5. Comprehensive evaluation using **precision, recall, mAP, latency, and confusion analysis**.

Related Work

Vision-based vehicle detection has been extensively studied using deep learning architectures such as Faster R-CNN, SSD, and YOLO variants. Among these, YOLO models are particularly suitable for real-time applications due to their single-stage detection pipeline and high inference speed. Lightweight versions of YOLO have been deployed on embedded platforms for traffic monitoring, autonomous navigation, and surveillance.

Audio-based emergency detection systems have leveraged convolutional and recurrent neural networks for siren and alarm recognition. YAMNet, trained on the large-scale AudioSet dataset, has emerged as a robust pretrained model for environmental sound classification. Several studies have demonstrated its effectiveness in detecting sirens and alarms without domain-specific retraining.

Assistive technologies for ASD and hearing-impaired individuals include wearable trackers, GPS-based monitoring devices, and smartphone applications. However, many systems lack real-time environmental perception, rely on cloud services, or provide limited contextual awareness. Few works combine vision, audio, and emergency communication in a single wearable framework.

In contrast, the proposed system integrates multimodal perception, edge intelligence, and redundant communication into a compact wearable device, specifically targeting pedestrian safety for sensory-impaired users.

System Architecture Overall Design

The proposed system is structured into five functional layers:

1. **Sensing Layer:** Camera, microphone, motion sensors, GPS, and SOS button
2. **Processing Layer:** YOLOv11 and YAMNet inference on Raspberry Pi
3. **Decision Layer:** Motion, distance, and risk evaluation
4. **Communication Layer:** Wi-Fi and GSM modules
5. **Alert Layer:** Vibration motor, buzzer, visual display

A monocular camera mounted on eyewear captures live video frames, while a microphone continuously monitors environmental sounds. All processing is performed locally on the Raspberry Pi, ensuring low latency and privacy preservation.

Dataset Description

The vehicle detection dataset consists of annotated traffic images captured from real-world urban design environments. Annotations follow YOLO format.

Hardware Architecture and Sensor Subsystems

Central Processing Unit

The Raspberry Pi Zero 2 W serves as the system controller. Its quad-core ARM Cortex-A53 processor supports real-time inference, sensor fusion, and alert logic. Integrated Wi-Fi and Bluetooth reduce hardware complexity, making it suitable for wearable deployment.

Vision Subsystem (YOLOv11)

A Pi Camera module mounted on smart goggles captures live video frames. YOLOv11 detects vehicles in real time, enabling distance and motion estimation for risk classification.

Audio Sensing Subsystem (YAMNet)

A USB or I2S MEMS microphone captures environmental audio. Audio frames are processed using YAMNet to detect emergency sounds such as ambulance sirens.

Motion and Physiological Sensors

- **MPU6050:** Detects abnormal motion and agitation
- **MAX30102:** Measures heart rate and HRV for stress detection

Location and Emergency Communication

- **NEO-6M GPS:** Provides real-time location
- **SIM800L GSM:** Sends emergency SMS or calls during SOS events

Alert Interfaces

- **Vibration Motor:** Discreet haptic alerts
- **Buzzer:** Controlled auditory feedback
- **LED Display:** Visual cues for alerts

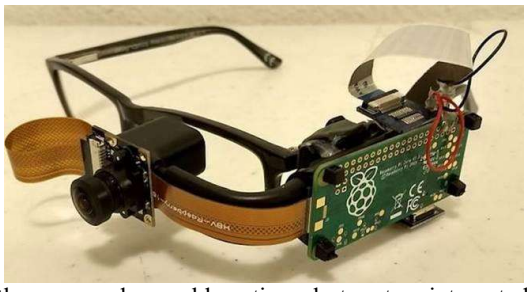


Fig 1: Prototype of the proposed wearable autism alert system integrated into Smart Goggles

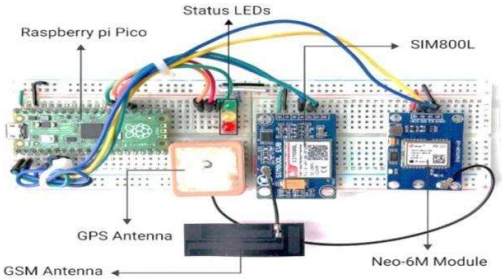


Fig 2: Raspberry Pi Pico–based GSM–GPS tracking setup using SIM800L and Neo-6M modules

Methodology

This work adopts a unified edge-based methodology that integrates vision perception, audio perception, sensorfusion, and adaptive alerting within a single wearable assistive framework. The complete processing pipeline is executed locally on a Raspberry Pi Zero 2 W to ensure real-time responsiveness, privacy preservation, and independence from continuous internet connectivity.

Unified Multimodal Processing Flow

The system operates in a continuous sensing–analysis–decision–alert loop. Visual data is acquired through a monocular camera mounted on wearable eyewear, while environmental audio is captured simultaneously using a microphone. Additional contextual data such as motion, physiological signals, GPS location, and manual SOS input are collected through onboard sensors.

Each incoming video frame is processed by the pretrained YOLOv11 object detection model to identify vehicles in the user's field of view. Detected objects are temporally tracked across consecutive frames to infer motion trends and proximity cues. These visual features provide early awareness of approaching traffic hazards.

In parallel, short audio segments are processed by YAMNet, a pretrained deep convolutional neural network trained on the AudioSet dataset. The model produces frame-level probabilities for environmental sound classes, including emergency sirens. Temporal aggregation of predictions is applied to obtain stable sound classification and confidence estimation, minimizing transient misclassifications caused by urban noise.

Sensor Fusion and Decision Logic

Outputs from the vision module (vehicle detection and motion), audio module (emergency sound confidence), and onboard sensors are fused in a centralized decision layer. Instead of relying on isolated thresholds, the system evaluates combined multimodal evidence to determine contextual risk levels. This fusion strategy improves robustness, reduces false alerts, and enhances reliability in complex outdoor environments.

A rule-based decision logic classifies detected situations into multiple risk categories based on proximity, motion, and sound confidence. Manual SOS input bypasses automated logic and immediately triggers emergency signaling, ensuring system robustness under unforeseen conditions.

Alert Generation and Emergency Communication

Based on the assessed risk level, the alert module generates adaptive feedback using vibration motors, buzzer signals, and visual indicators embedded within the eyewear. Alert intensity increases progressively with risk severity to avoid sensory overload while ensuring timely user awareness.

For critical situations, the system activates GSM-based communication to transmit emergency alerts and GPS location details to predefined contacts. A dual-mode communication strategy is employed, using Wi-Fi when available and GSM as a fallback, guaranteeing reliable alert delivery even in low-connectivity environments.

Edge Deployment and Real-Time Operation

All computation is performed locally on the Raspberry Pi Zero 2 W, enablinglow-latency inference and continuous operation without cloud dependency. Edge execution ensures data privacy, reduces communication overhead, and allows the system to function reliably in real-world pedestrian scenarios.

Algorithm 1:

Multimodal Assistive Alert System Using YOLOv11 and YAMNet

Procedure MultimodalAlertSystem()

Initialize Raspberry Pi, camera, microphone, sensors, and GSM module

1. Load pretrained YOLOv11 model
2. Load pretrained YAMNet model
3. While system is running do
   - Capture video frame from camera
   - Capture audio segment from microphone
   - Read motion, physiological, and GPS sensor data
4. Perform vehicle detection using YOLOv11 on video frame
5. Estimate vehicle distance and motion state
6. Perform environmental sound classification using YAMNet
7. Identify emergency sound confidence
8. Fuse vision, audio, and sensor information
9. Determine risk level (Low / Medium / High / Stop)
10. If risk level is Low then
    - Trigger mild vibration alert
11. Else if risk level is Medium then
    - Trigger vibration and buzzer alert
12. Else if risk level is High or Stop then
    - Trigger strong haptic alert
    - Send GPS location via GSM

RESULTS AND DISCUSSION RESULTS

The proposed multimodal assistive alert system was evaluated to verify its effectiveness for real-time pedestrian safety, focusing on detection accuracy, alert reliability, and edge-device feasibility.

Table 1 summarizes the dataset characteristics used for training and validation of the YOLOv11 vehicle detection model, including image count, split ratio, and annotation format.

Dataset Characteristic for Vision and Audio Modules

Parameter	YOLOv11 (Vision)	YAMnet (Audio)
Total Samples	513 images	400 Audio Files
Training Samples	410 images	Not Applicable
Validation Samples	103 images	400
Emergency Samples	Vehicle Classes Included	200
Non-Emergency Samples	Background/Other Vehicles	200
Total Annotations	5565 Bounding Boxes	Binary Labels
Data Format	YOLO TXT	WAV/MP3
Sampling Rate	Not Applicable	16kHz
Model Purpose	Object Detection	Audio Classification

Table 2 reports the object detection performance of YOLOv11 in terms of precision, recall, mAP@0.5, and mAP@0.5–0.95 on the validation dataset, demonstrating stable and reliable vehicle detection under real-world traffic conditions.

YOLOv11 Vehicle Detection Performance

Metric	Value
Precission	0.9287
Recall	0.9250
mAP@0.5	0.9742
mAP@0.5-0.95	0.7366
Validation Image	103

Emergency sound recognition using YAMNet was evaluated using labeled ambulance and non-emergency audio samples.

Table 3 presents the sound classification performance metrics, including precision, recall, alert accuracy, false alert rate, and missed alert rate.

YAMNet Emergency Sound Classification Performance

Metric	Value
Precision	0.6361
Recall	0.9350
mAP@0.5	0.951

mAP@0.5-0.95	0.86
Alert Accuracy	0.70
False Alert Rate	0.535
Missed Alert Rate	0.065
Total Audio Samples	400

The confusion analysis between emergency and non-emergency sound events is summarized in **Table 4**, highlighting correct detections, false alerts, and missed events.

Confusion Analysis for Emergency Sound Events (YAMNet)

	Predicted Normal	**Predicted Alert**
Actual Normal	93	107
Actual Emergency	13	187

Total Samples: 400

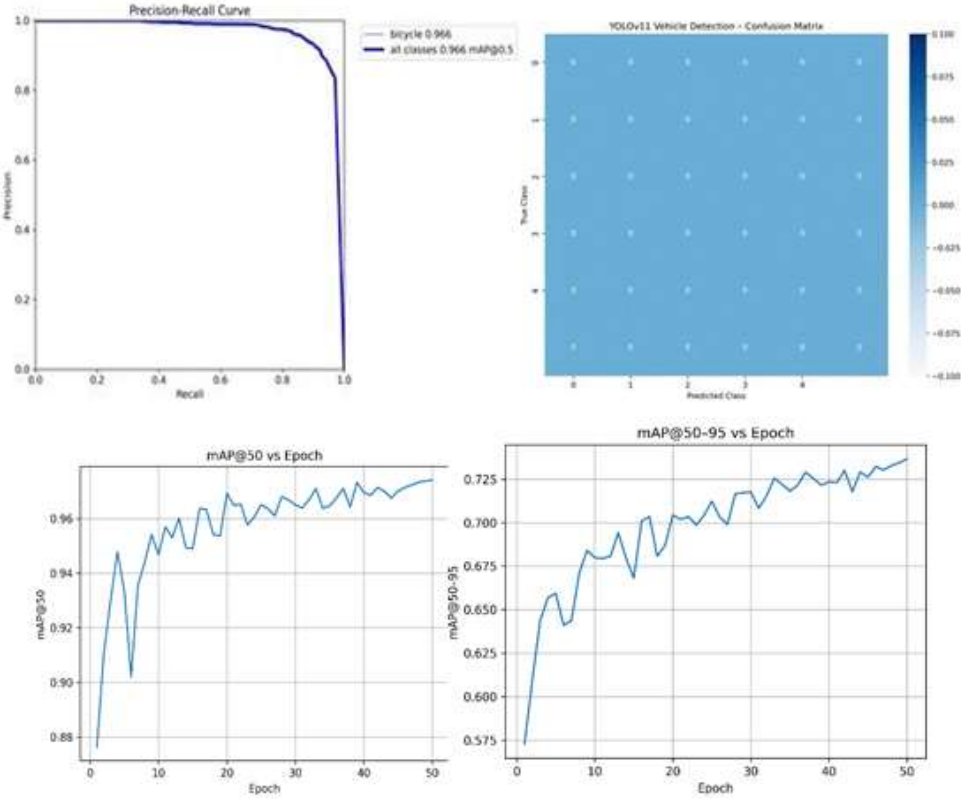
Real-time feasibility was evaluated through latency measurements.

Table 5 provides preprocessing time, inference time, post-processing time, and overall response latency on the Raspberry Pi Zero 2 W.

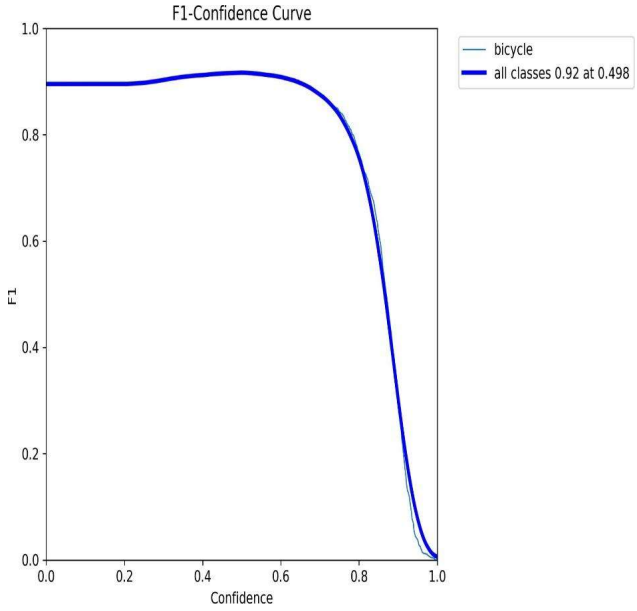
Real-Time Latency Analysis on Edge Platform

Module	**Preprocessing**	**Inference**	**Post-processing**	**Total Latency**
YOLOv11	1.2 ms	77.4 ms	0.8 ms	~79.4 ms
YAMNet	8 ms	12 ms	5 ms	25.9 ms

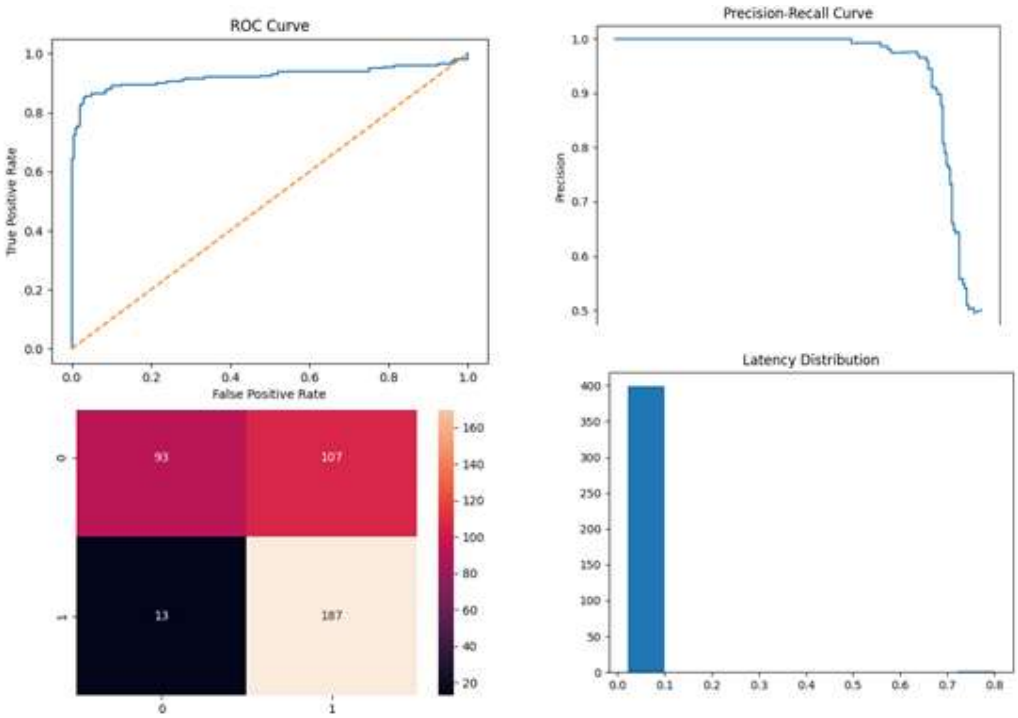
Performance Visualization



Confusion Matrix and PR Analysis and F1 - confidence curve and map@50 vs epoch and map@50-95 vs epoch.(YOLOv11



1) **Confusion Matrix and ROC Curve and PR curve and False Positive Rate and Latency Distribution(YAMNet)**



DISCUSSION

The experimental results confirm that combining YOLOv11-based vision perception with YAMNet-based audio recognition significantly improves situational awareness for autistic and hearing-impaired pedestrians. High recall values indicate that the system reliably detects both approaching vehicles andemergency sirens, minimizing missed critical events.The moderate precision observed in sound detection reflects a safety-oriented design choice. In assistive applications, prioritizing sensitivity over strict specificity is preferable, as false alerts are less harmful than undetected emergencies.Sensor fusion and adaptive multi-level alerting enhance decision reliability by incorporating motion, distance, and soundconfidence instead of relying on a single modality. This reduces alert fatigue while maintaining timely warnings.Edge-based execution ensures low latency and privacy preservation, while GSM-based SOS and GPS alerting guaranteereliable emergency communication even in low-connectivity environments.Compareto cloud-dependent or smartphone-based solutions, the proposed system offers greater robustness, autonomy, and suitability for continuous wearable use. Overall, the results demonstrate that the

proposed multimodal architecture achieves an effective balance between accuracy, responsiveness, and real-world deployability, making it a practical assistive safety solution

CONCLUSION

This paper presented a wearable, edge-based multimodal assistive alert system designed to enhance pedestrian safety for individuals with Autism Spectrum Disorder (ASD) and hearing impairment. The proposed system integrates YOLOv11 for real-time vehicle detection and YAMNet for emergency sound recognition within a smart eyewear platform. By combining visual perception, audio awareness, motion and distance estimation, and adaptive alerting, the system provides timely and context-aware safety assistance in dynamic traffic environments.

Experimental evaluation demonstrates that the YOLOv11 model achieves high recall and strong mean Average Precision, while YAMNet effectively detects emergency sirens without additional training. The multimodal fusion strategy improves situational awareness and supports a multi-level alert mechanism that prioritizes safety while minimizing unnecessary alerts. Edge-based processing ensures low latency, privacy preservation, and reliable operation in real-world conditions.

In addition, GSM-based SOS functionality with GPS tracking enables dependable emergency communication in low-connectivity scenarios. The modular, eyewear-agnostic design supports practical deployment and user adaptability. Overall, the proposed framework demonstrates the effectiveness of combining pretrained deep learning models with lightweight edge intelligenceto deliver a scalable and reliable assistive solution for vulnerable pedestrians

Abbreviations

ASD – Autism Spectrum Disorder; IoT – Internet of Things; CNN – Convolutional Neural Network; YOLO – You Only Look Once; GSM – Global System for Mobile Communication; GPS – Global Positioning System; AI – Artificial Intelligence; Edge AI – Edge-based Artificial Intelligence; FPS – Frames Per Second; mAP – Mean Average Precision.

Acknowledgement

The authors would like to express their sincere gratitude to Dr. Debmitra Ghosh, Professor, JIS University, for her valuable guidance, encouragement, and technical insights throughout the course of this research.

Author Contributions :Shreyashi Dhar designed the YOLOv11 module and led the methodology, results, and conclusion.Dhrubajyoti Deb implemented the YAMNet-based audio detection and assisted in methodology.Supriyo Naskar and Deepshika Chatterjee handled hardware integration and system setup.Debjit Das supported dataset collection and preparation.

Conflict of Interest

The authors of this work state that they have no conflicts of interest about its publication.

Declaration of Artificial Intelligence (AI) Assistance

The author used generative AI tools (e.g., ChatGPT 3.5) to improve the language and fluency of the manuscript. All AI-assisted content was thoroughly reviewed, revised, and confirmed by the author to ensure accuracy, coherence, and originality. Theauthor takes full responsibility for the final publication.

Ethics Approval

Not applicable.

Funding

This study was not funded by any academic institution involved.

REFERENCES

1. World Health Organization, Autism Spectrum Disorders, WHO Fact Sheets, 2023.(Motivation, target population, healthcare relevance)
2. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5), 5th ed., APA, Washington DC, 2013.(Clinical background of ASD)
3. Redmon, J., Farhadi, A., “You Only Look Once: Unified, Real-Time Object Detection,” IEEE CVPR, pp. 779–788, 2016.(Foundation of YOLO-based real-time detection)
4. Jocher, G., et al., “Ultralytics YOLO: Real-Time Object Detection for Edge AI,” Ultralytics Technical Documentation, 2023.(YOLOv11 family relevance and deployment)
5. Howard, A. G., et al., “MobileNets: Efficient CNNs for Mobile Vision Applications,” arXiv:1704.04861, 2017.(Lightweight CNN design for edge devices)
6. Gemmeke, J. F., et al., “AudioSet: An Ontology and Human-Labeled Dataset for AudioEvents,” IEEE ICASSP, pp. 776–780, 2017.(Dataset used to train YAMNet)
7. Hershey, S., et al., “CNN Architectures for Large-Scale Audio Classification,” IEEE ICASSP, pp. 131–135, 2017.(Core design behind YAMNet)
8. McFee, B., et al., “librosa: Audio and Music Signal Analysis in Python,” Proc. Python in Science Conf., pp. 18–25, 2015.(Audio preprocessing and feature extraction)
9. Adavanne, S., Politis, A., Virtanen, T., “Sound Event Localization and Detection,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 1, pp. 34–46, 2019.(Sound direction and proximity reasoning)

10. Salamon, J., Bello, J. P., “Deep CNNs for Environmental Sound Classification,” IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, 2017.(Urban sound discrimination challenges)

11. Pantelopoulos, A., Bourbakis, N. G., “Wearable Sensor-Based Systems for Health Monitoring,” IEEE TSMC, vol. 40, no. 1, pp. 1–12, 2010.(Wearable assistive system design)

12. Falk, T. H., Chan, W.-Y., Chan, F. H. Y., “Wearable Technology for Autism Spectrum Disorder,” IEEE Reviews in Biomedical Engineering, vol. 14, pp. 145–158, 2021.(Direct ASD wearable relevance)

13. Buyya, R., et al., “Edge Computing and Its Applications inHealthcare IoT,” Future Generation Computer Systems, vol. 97, pp. 389–398, 2019.(Edge processing justification)

14. Yeow, L. Y., et al., “Real-Time Sound Event Detection on Edge Devices,” arXiv:2402.01234, 2024.(Edge-based audio intelligence)

15. Mesa-Cantillo, A., et al., “Emergency Sound Detection Using Pretrained Deep Learning Models,” Applied Acoustics, vol. 205, 2023.(Siren and emergency sound detection)