DRISTII: A Multi-Modal IoT Assistive System for Real-Time Object Detection and Navigation Support for the Visually Impaired

Nishant Shah, Ansh Bajpai, Shrivaths S. Nair, and Manas Verma K
Department of Computer Science and Engineering, Amrita School of Computing,
Amrita Vishwa Vidyapeetham, Amritapuri, India
Email: {am.sc.u4aie23166, am.sc.u4aie23108, am.sc.u4aie23148, am.sc.u4aie23126}@am.students.amrita.edu

Abstract*— Visually impaired individuals face persistent challenges in perceiving their surroundings, particularly in dynamic environments where both object recognition and spatial awareness are critical. Traditional assistive solutions often rely on single modality, such as often vision-based recognition or ultrasonic ranging, which limits reliability in real-world conditions. Therefore, an integrated, multi-modal system is required to deliver accurate object detection, distance estimation, and intuitive feedback in real-time. This paper proposes DRISTII, a Raspberry Pi 5-based assistive framework that combines lightweight YOLOv8-nano model for efficient object detection using COCO labels, ultrasonic distance measurement, dynamic face recognition, contextual image captioning, and voice-haptic feedback within an MQTT-enabled architecture. Experimental validation in indoor environments demonstrates an average latency of $25\,\mathrm{ms}$ and detection precision of $98.42\,\%$. The system further enhances situational awareness by issuing tactile alerts for obstacles within a $25\,\mathrm{cm}$ range and delivering audio-based contextual descriptions of detected objects and faces. The results confirm that DRISTII provides a practical, low-cost, and real-time assistive platform that improves the safety and autonomy of visually impaired users through the integration of multiple sensing and feedback modalities

Index Terms—**Internet of Things (IoT), Visual Impairment,YOLOv8, Ultrasonic detection, Object Detection, Face Recognition, Voice Assistance**

I. Introduction

Visual impairments affects millions of individuals worldwide, restricting independence and creating safety risks in daily navigation. Assistive technologies have emerged to support mobility and environmental awareness, ranging from canes and wearables to AI-enabled smart devices. With the rapid evolution of IoT and computer vision, there is increasing potential to deliver portable, low-cost, real-time solutions that integrate multiple sensing modalities.

Visually impaired users require fast and accurate recognition of surrounding objects. Many prior works employ deep learning models such as SSD-MobileNet or faster R-CNN, which either sacrifice accuracy for speed or are computationally too heavy for portable devices.

Identifying objects alone is insufficient without knowing their proximity. Some systems employ depth cameras or LiDAR. which are costly and power-hungry, while others use only ultrasonic sensors, which lack contextual awareness.

Even when detection and ranging are performed, feedback to the user often remains limited to either audio or vibration. Existing works typically focus on one channel, leading to cognitive overload or missed cues in noisy environments.

This paper addresses the lack of an integrated, low latency assistive system that combines object recognition, distance estimation, and intuitive multi-modal feedback for visually impaired users, deployable on a portable, low-cost platform.

Current approaches are often limited by high computational requirements, dependence on cloud connectivity, or reliance on single-modality sensing. Such constraints reduce usability in real-world dynamic environments where offline operation, fast interference, and robust feedback are essential.

Unlike prior works limited to single-modality sensing or cloud-dependent inference, **DRISTII** integrates vision, ranging, face recognition, captioning, and multimodal feedback in a portable, real-time, on-device system, achieving both high accuracy and low latency. The main contributions of this work are as follows:

- **A unified multi-modal assistive framework:** Integration of vision-based object detection, ultrasonic ranging, dynamic face recognition, captioning, and multimodal (voice + haptic) feedback in a single low-cost wearable platform.
- **Edge-optimized deployment of YOLOv8-nano:** Efficient fine-tuning and quantization of YOLOv8-nano on Raspberry Pi 5, achieving $25\,\mathrm{ms}$ latency

and $98.42\,\%$ precision in real-time indoor evaluation.

- **Dynamic and lightweight face recognition:** Implementation of on-device KNN-based face recognition with support for incremental enrollment of new users, enabling personalized interaction without cloud dependence.
- **Contextual awareness through captioning:** Extension of object detection with lightweight image captioning and voice synthesis, providing richer situational descriptions beyond object labels.

II. Related Works

In recent years, we have witnessed a rapid development of technologies enhancing situational awareness for visually impaired individuals. Facial recognition plays a crucial role in various applications, as it tackles challenges such as changes in lighting, position, and facial expressions.

A. Object Detection Models for Assistive Applications

Object detection is central to enabling visually impaired individuals to perceive their surroundings. The COCO dataset [9] and the YOLO model family [10] provide strong foundations for object detection tasks. Krishnan et al. [3] conducted a comparative study of YOLO models on edge devices, demonstrating the trade-off between accuracy and inference latency. Their results support the choice of light weight detectors such as YOLOv8-nano for real-time assistive systems. Masud et al. [5] proposed a smart assistive system integrating object detection and classification to support obstruction avoidance, showing improved navigation safety. Similarly. Rahman and Sadi [8] proposed an IoT-enabled system that performed object recognition for visually impaired users, but their dependence on cloud connectivity restricted usability in offline conditions. These works confirm the utility of camera-based object detection, but highlight the challenge of achieving high accuracy and low latency on portable devices.

B. Distance Measurement and Ultrasonic-Based Assistive Systems

For visually impaired users, awareness of obstacle proximity is as important as object identification. Roy and Banerjee [4] introduced an ultrasonic sensor-based smart stick that provided vibration cues for obstacle detection. Such systems are low cost and energy efficient, yet they only measure distance and do not provide contextual information about the obstacle. Conversely, camera-based vision systems [3], [5], [8] employing deep-learning models can recognize and classify objects but cannot reliably estimate distance without depth sensors or additional ranging modules. Combining ultrasonic ranging with vision-based detection therefore offers a holistic assistive solution, enabling both semantic understanding and spatial awareness. However, most existing systems adopt only one modality, reducing robustness in real-world navigation.

C. IoT-enabled Assistive Communication Frameworks

IoT technologies provide modularity and lightweight communication for wearable assistive devices. Light [7] compared communication protocols for wearables and concluded that MQTT offers minimal overhead, making it suitable for real-time assistive feedback. Rahman and Sadi [8] applied IoT to connect recognition modules for blind navigation, but their reliance on cloud processing introduced latency and limited offline usability. These findings suggest that IoT is vital for scalability, but robust assistive systems should emphasize on-device processing, using IoT protocols mainly for modular integration and communication.

D. Vision-language and Multimodal Feedback

For visually impaired assistance, feedback must be intuitive and context-rich. Li et al. [6] proposed BLIP(Bootstrapping Language-Image Pretraining), which achieved strong results in image captioning and multimodal understanding tasks. Captioning allows the system to describe objects beyond labels, improving accessibility. However, such models are computationally large and unsuitable for embedded deployment. This highlights the need for lightweight captioning approaches that can run locally while still enriching user awareness.

E. Face Recognition in Assistive Systems

Recognizing familiar individuals is an important assistive capability. Gupta et al. [2] developed an ensemble-based face recognition method using voting and bagging to increase robustness against pose and lighting variations. While this improves accuracy, ensemble methods are computationally expensive and impractical for potable assistive devices. This creates a gap for edge-optimized face recognition that supports real-time, on-device enrollment and recognition for visually impaired users.

Although object detection, ultrasonic ranging, IoT frameworks, captioning models, and face recognition have each advanced, there is a need of a system that provides a unified, lightweight and offline-capable assistive platform that combines semantic recognition, distance estimation, user identification, and multimodal feedback in real-time. This paper addresses this gap through the development of DRISTII, a Raspberry Pi 5-based framework designed specifically for visually impaired navigation.

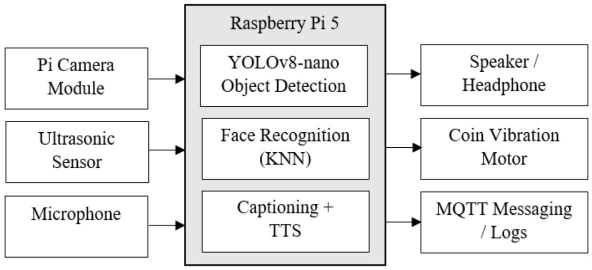

Fig. 1: Block diagram showing sensor inputs, Raspberry Pi 5 processing modules and multimodal outputs.

III. Methodology and System Design

The proposed system, DRISTII, is a wearable assistive platform built on Raspberry Pi 5 that combines on-device AI inference with multimodal sensing. The design emphasises low latency, offline operation, and real-time responsiveness.

A. System Architecture

The architecture of DRISTII (Fig. 1) comprises five key modules: object detection (YOLOv8-nano for lightweight real-time detection), distance measurement(ultrasonic sensing for proximity estimation), captioning with voice interface (contextual description of scenes), haptic feedback (vibration alerts for close-range obstacles), and face recognition(dynamic recognition of known individuals). These modules communicate via MQTT, while the Raspberry Pi acts as the central controller. Visual and sensor inputs are captured in real-time, processed locally, and translated into feedback(speech, haptics, or alerts). If an obstacle is within $25\,\mathrm{cm}$, the coin vibration motor activates for tactile warning. Text outputs are converted to audio through a TTS engine and delivered via earphones or bone-conduction devices.

B. System Workflow

The operational pipeline of DRISTII is illustrated in Fig. 2. The system proceeds through the following stages:

1) **System Initialization:** Boot the Raspberry Pi, load pretrained models, and initialize sensors (camera, ultrasonic, and IMU).
2) **Real-Time Data Capture:** Acquire continuous image frames and auxiliary sensor streams from the environment.
3) **Data Processing:** Apply YOLOv8-nano and sensor-fusion routines to detect obstacles, classify objects, and recognize known individuals.
4) **Decision Making:** Determine contextual actions such as announcing a person's name, describing surroundings, or remaining idle.

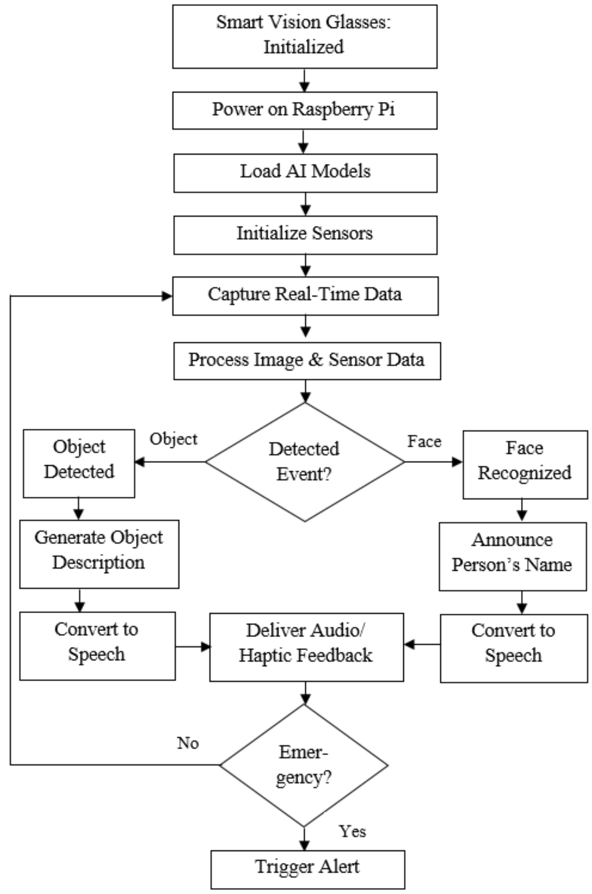

Fig. 2: Flow Diagram of DRISTII operational pipeline

5) **Feedback Generation:** Convert textual outputs into speech via a TTS engine and deliver them through audio interfaces.
6) **Emergency Detection:** Trigger haptic vibration if an obstacle is detected within $25\,\mathrm{cm}$.
7) **Continuous Loop:** Return to data acquisition, enabling uninterrupted adaptive assistance.

C. Implementation

1) Hardware Components:

- **Raspberry Pi 5:** Quad-Core Cortex-A76 @2.4 GHz with up to 8 GB LPDDR4X RAM; chosen for its balance between edge AI performance and portability.
- **Pi Camera Module (5 MP):** Captures 1080p video and high-resolution stills via CSI port for object detection and face recognition.
- **HC-SR04 Ultrasonic Sensor:** Sonar-based proximity sensor (2–400 cm range, 0.3 cm resolution) for obstacle detection.
- **Coin Vibration Motor (10 mm ERM):** Provides tactile alerts when obstacles are detected within $25\,\mathrm{cm}$.

- **Microphone and Headphones:** Sony WH-1000XM4 used for audio input/output, providing noise-cancelled speech feedback via Bluetooth or wired connection.
- **Wired Power Bank:** Portable 5 V lithium-ion supply for mobility and field testing.
- **MicroSD Card:** Stores OS, datasets, pretrained weights, user profiles, and logs.

2) Software Stack:

- **YOLOv8-nano (PyTorch, Ultralytics):** Used for real-time object detection.
- **Python + OpenCV:** For image capture, preprocessing, and KNN-based face recognition.
- **MQTT:** Lightweight protocol for inter-module communication.
- **TTS + Speech Recognition:** Provides audio-based user interaction.

3) YOLOv8-Nano Model Training: To ensure reliable real-time object detection under varied indoor conditions, the YOLOv8-nano model was fine-tuned on a targeted subset of the **COCO (Common Objects in Context)** dataset [9]. The selected categories—*person*, *bottle*, *chair*, *cup*, and *door*—reflect common obstacles and objects encountered by visually impaired users. The Ultralytics framework was used for training, incorporating data augmentation techniques such as rotation, scaling, and brightness/contrast adjustment to enhance generalization.

The model was fine-tuned over 4–5 iterations using transfer learning with pre-trained weights. Post-training, the model was quantized and exported in TorchScript format for efficient inference on edge devices. YOLOv8-nano was selected due to its optimal trade-off between accuracy and latency, making it well-suited for resource-constrained platforms like the Raspberry Pi.

4) Face Recognition Using K-Nearest Neighbors: Face recognition in DRISTII is implemented using a lightweight K-Nearest Neighbors (KNN) classifier via OpenCV's `cv2.ml.KNearest` module. Detected faces are encoded into 128-dimensional embeddings using the `face_recognition` library. These embeddings, paired with identity labels, are stored in a local database.

During inference, the embedding of a new face is computed and compared to stored embeddings. The classifier assigns the label based on the majority vote among the k nearest neighbors in the embedding space. This approach enables fast, on-device face recognition without complex training, and supports easy updates by adding new faces. It is efficient, scalable, and suitable for deployment on low-power platforms like the Raspberry Pi, particularly in controlled indoor environments.

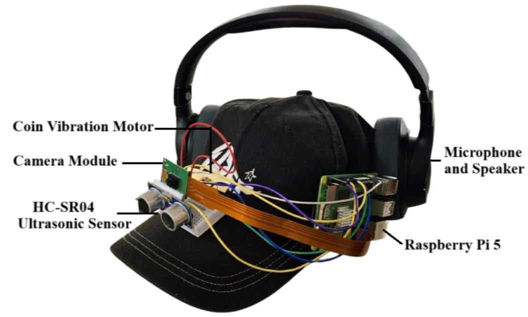

Fig. 3: Prototype of the DRISTII assistive system showing integrated components: Raspberry Pi 5, camera module, ultrasonic sensor, vibration motor, and headset.

TABLE I: System performance metrics (indoor evaluation)

Metric	Value
Detection precision (PPV)	98.42%
Inference latency (end-to-end)	25 ms
False positive rate (FPR)	1.3%
False negative rate (FNR)	0.8%
Ultrasonic MAE (distance)	$\pm$0.8 cm
Haptic trigger threshold	25 cm

IV. Experimental Results

A. Evaluation Setup

All experiments were conducted in indoor corridors and rooms under stable ambient lighting. The Raspberry Pi 5 executed all inference locally (no cloud). The camera was positioned at eye level on the wearable mount; the ultrasonic sensor was front-facing at a height of 1.5 m. For distance tests, planar targets were placed on stands along the optical axis at 25 cm to 200 cm in 25 cm increments. For object detection, images contained COCO-subset categories used in training (person, bottle, chair, cup, door). Face tests used a gallery of enrolled identities and distractor faces.

B. Quantitative Results

Table I summarizes system-level performance. DRISTII achieves **25 ms** average end-to-end latency and **98.42%** detection precision (positive predictive value) on the indoor evaluation set. The ultrasonic ranging error is $\pm$**0.8 cm** (MAE) across 25 cm to 200 cm. The haptic alert is triggered at $\leq$25 cm.

C. Qualitative Results

Fig. 4 illustrates unknown-face rejection; Fig. 7 shows correct identification of an enrolled person. Fig. 5

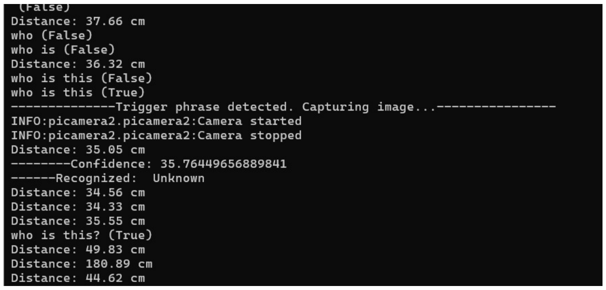

Fig. 4: Unknown Face Detection

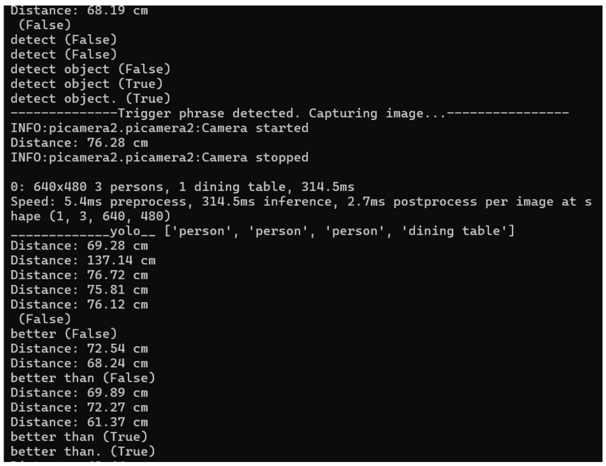

Fig. 5: Object Detection

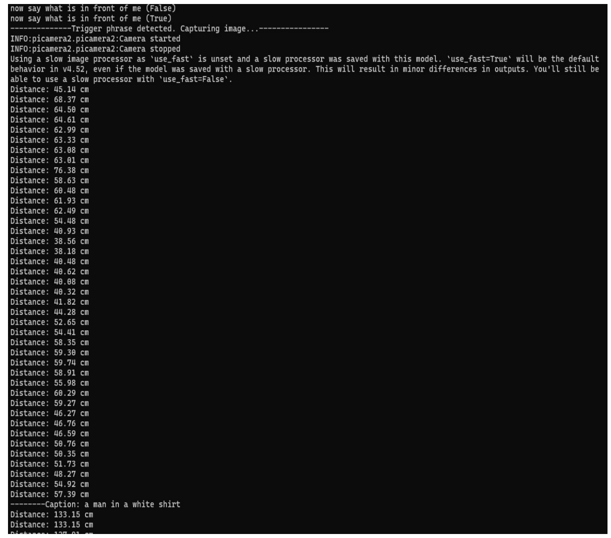

Fig. 6: Image Captioning

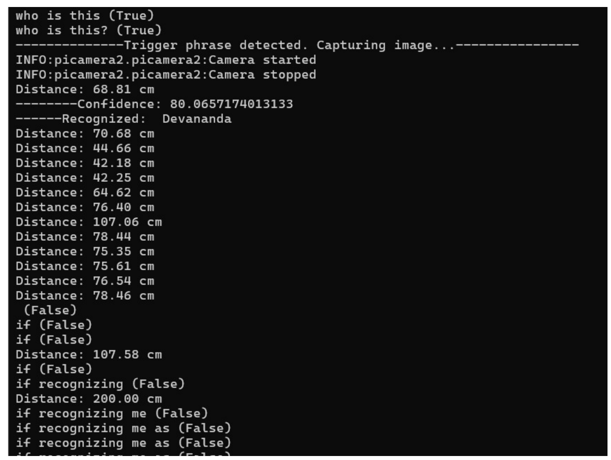

Fig. 7: Known Face Detection

demonstrates real-time YOLOv8-nano detections on the COCO-subset categories. Fig. 6 shows generated captions that add contextual descriptions beyond single labels. Fig. 8 confirms obstacle ranging with the $25\,\mathrm{cm}$ haptic threshold. Fig. 9 provides an OCR example supporting text accessibility.

D. Comparison with Existing Works

To highlight the improvements of DRISTII over prior systems, Table II summarizes a feature-level and performance comparison with Rahman *et al.* [8] and Masud *et al.* [5]. Unlike these approaches, DRISTII achieves real-time operation with YOLOv8-nano on Raspberry Pi 5, integrates multimodal feedback (voice and haptics), and supports dynamic face recognition and contextual awareness. Table II demonstrates the comparison of proposed system with existing systems.

E. Discussion

Despite its promising performance, DRISTII exhibits certain limitations and practical considerations that guide future improvements. The YOLOv8-nano model required multiple retraining cycles to enhance precision, during which user interaction was temporarily halted, limiting real-time usability. Face recognition accuracy degraded in low-light environments, suggesting the need for adaptive preprocessing or low-light data augmentation during enrollment. The voice interface processed multiple commands sequentially without filtering, sometimes producing redundant responses; this can be mitigated by applying rate-limiting or intent de-duplication. Although on-device processing removed cloud latency, energy consumption remains a challenge, and strategies such as duty-cycling the camera and ultrasonic modules can reduce power use during idle periods. The system also lacked mechanisms to verify the authenticity of user-provided information (e.g., names during registration), which may leave it vulnerable to false inputs. Addressing these aspects will enhance the robustness, efficiency, and usability of the system in real-world deployment.

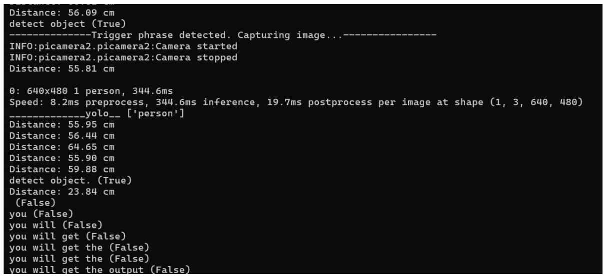

Fig. 8: Ultrasonic Obstacle Detection

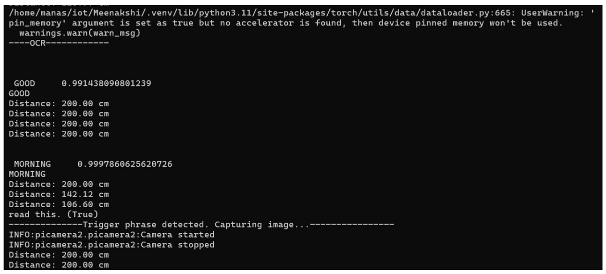

Fig. 9: Optical Character Recognition (OCR)

V. Conclusion and Future Work

This paper proposes DRISTII, a multi-modal assistive system that integrates YOLOv8-nano based object detection, ultrasonic ranging, face recognition, captioning, and multimodal feedback on a Raspberry Pi 5. Unlike prior single-modality or cloud-dependent systems, the system achieves $98.42\,\%$ detection precision with $25\,\mathrm{ms}$ latency and $\pm 0.8\,\mathrm{cm}$ distance accuracy in real-time indoor evaluation. The prototype demonstrates that low-cost, on-device AI can enhance safety and autonomy for visually impaired users. Future work will address illumination robustness, efficient retraining, and secure user input verification, as well as extend functionality with object tracking, GPS/LiDAR modules, and caregiver dashboards. These enhancements will further strengthen DRISTII as a practical and scalable solution for assistive navigation.

TABLE II: Comparison of DRISTII with existing assistive systems

Metric	DRISTII (proposed)	Rahman et al. [8]	Masud et al. [5]
Modality	Multi-modal (vision + ultrasonic + voice + haptics)	Vision-only (IoT-enabled)	Vision-only (object classification)
AI Model	YOLOv8-nano	SSD+MobileNet	Viola–Jones
Inference Time	25 ms	Not specified	Not specified
Detection Accuracy	98.42%	99.31%	91%
Voice Commands	Supported	No	No
Tactile Feedback	Coin vibrator	No	No
Face Recognition	Dynamic registration	No	Basic detection
IoT Protocol	MQTT	Basic connectivity	Limited
Hardware Platform	Raspberry Pi 5	Raspberry Pi 4 (8 GB)	Raspberry Pi 4B (4 GB)
Proximity Threshold	25 cm (precise)	150 cm	Not specified

References

[1] B. M. Gopal, Y. N. Reddy, C. B. Siddhartha, S. Rashmi, V. R. Pandi, V. Valsan, S. Sathyan, and K. Suresh, "IoT Based Safety Monitoring and Communication System for Underground Coal Mines," in *Advanced Network Technologies and Intelligent Computing*, A. Verma, P. Verma, K. K. Pattanaik, S. K. Dhurandher, and I. Woungang, Eds. Cham: Springer, 2024, pp. 69–80.

[2] A. Gupta, S. Sriram, and V. Nivethitha, "Harnessing Diversity in Face Recognition: A Voting and Bagging Ensemble Approach," in *Proc. 2024 Int'l Conf. on Automation and Computation (AUTOCOM)*, Chennai, India, 2024.

[3] P. Krishnan, S. Kattamuri, G. R. Prabhu, and M. Rashmi, "Assistive Eye: A Comparative Analysis of YOLO Object Detection Models on Edge Devices," in *Proc. ACM*, New York, NY, USA, 2024.

[4] S. Roy and T. Banerjee, "Assistive Ultrasonic Sensor-Based Smart Stick for the Visually Impaired," *International Journal of Engineering Research and Technology (IJERT)*, vol. 12, no. 1, pp. 50–54, 2023.

[5] U. Masud, T. Saeed, H. M. Malaikah, F. U. Islam, and G. Abbas, "Smart Assistive System for Visually Impaired People Obstruction Avoidance Through Object Detection and Classification," *IEEE Access*, vol. 10, pp. 13428–13441, 2022.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *arXiv preprint arXiv:2201.12086*, 2022.

[7] E. Light, "IoT Communication Protocols for Wearables: A Comparison of MQTT and CoAP," *Sensors*, vol. 23, no. 4, pp. 2123, 2023.

[8] M. A. Rahman and M. S. Sadi, "IoT Enabled Automated Object Recognition for the Visually Impaired," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100015, 2021.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," *Proc. European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48

[10] G. Jocher et al., "YOLO by Ultralytics," GitHub repository, 2023. [Online].