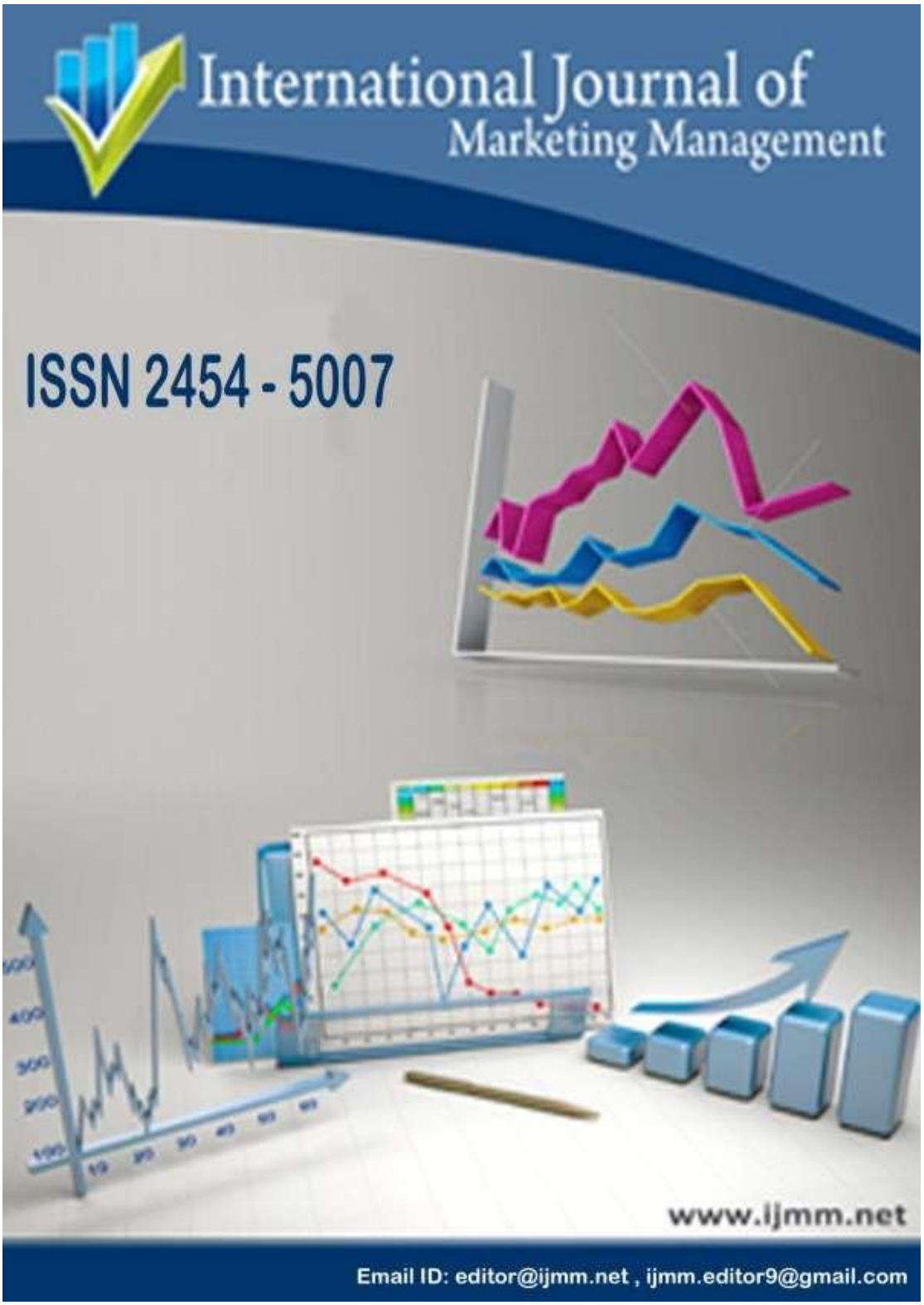
International Journal of
Marketing Management
ISSN 2454 - 5007
www.ijmm.net
Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

A Deep learning voice Assisted System for Intelligent Indoor Navigation of the Visually Impaired

[1] Joy Sangeeth Raj.G,[2] Pulugari Maheshwari,[3] Sakinala Sandhya,[4] Nagarthi Shashank Reddy,[5] Vuyuru Srilakshmi Priya,

[1]Assistant Professor, Department of ECE, Narsimha Reddy Engineering Collage, Maisammaguda(V), Kompally, Telangana.

[2,3,4,5] Student, Department of ECE, Narsimha Reddy Engineering Collage, Maisammaguda(V), Kompally, Telangana.

***Abstract*—** For the visually impaired, navigating inside spaces remains a major challenge, limiting their freedom of movement and autonomy. Guide dogs and white canes are helpful, but they can't convey complex geographical information. Navigational aids have come a long way in the last ten years, thanks to developments in computer vision and artificial intelligence. This research presents an artificial intelligence (AI) system for room and object detection that may help the visually impaired and those who are blind navigate interior spaces more easily. Our system sorts rooms according to the things they include using a deep learning model—specifically, YOLO for object recognition. This study also includes real-time audio output and depth detection to help users navigate better by measuring the distances of things near them. Additionally, the system incorporates voice narration to provide verbal explanations, warn users of difficulties, and show them alternative ways. A well-structured dataset optimized for indoor scene categorization with top-notch accuracy is also presented in this research. The goal of the solution is to provide visually impaired persons with a practical and easy-to-use gadget that bridges the gap between existing navigation help and full autonomous support systems. Object recognition, computer vision, room categorization, and deep learning are some of the terms used to describe assistive technology.

I. INTRODUCTION

Assisting visually impaired people with movement inside buildings has been a persistent problem in the field of study. White canes and other mechanical aids have been used in the past to raise environmental awareness, but only to a limited extent. Additionally, guide dogs have been widely used; however, they come with a hefty price tag and need extensive training. Early electronic navigation aids made use of digital technology, ultrasonic sensors, and radio frequency identification (RFID)-based systems to assist people detect impediments and locate pre-set places. However, these methods have limitations in terms of spatial awareness and their dependence on external infrastructure. More sophisticated navigation systems are now feasible thanks to recent breakthroughs in computer vision, deep learning, and artificial intelligence. Modern approaches for real-time object identification and classification make use of MobileNet, convolutional neural networks (CNNs), and object detection algorithms like YOLO. The addition of audible input from AI-powered voice assistants and wearable electronics has further simplified their usage. There are still some issues with real-time performance, accuracy varies across different interior situations, and integration into user-friendly user interfaces, but these technologies have greatly improved the mobility independence of visually impaired people. To get over these restrictions, this study presents an AI-powered room and item identification system. In order to effectively categorize interior areas, our method employs YOLO for object detection and generates structured data. To facilitate navigation, voice technology is used to real-timely name items and room kinds. The use of depth sensing in this work allows for more precise object distance estimate and navigation. If the system alerts the user two meters away from a detected door, for instance, they will feel more comfortable navigating the space. Thanks to voice or auto-detection-activated picture capture, the gadget may be used hands-free. To further develop item classification capabilities in contexts dependent on regions, this study makes use of a unique dataset that is centered on interior scenes in India. This study is an important first step in creating a real-time

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

navigation system that can assist those with visual impairments.

II. RELATED WORK

In order to identify impediments in real-time, an intelligent assisted navigation system was developed using YOLOv8. The device also incorporates moisture sensors and ultrasonic sensors to evaluate distance and surface conditions. An app for smartphones is at the heart of the system, providing users with the ability to make vocal commands and get both visual and tactile feedback as they navigate. It was able to identify humans with a 92.4% success rate.[1] However, its performance is limited to the items included in the YOLOv8 dataset and might need some more enhancements to handle unforeseen circumstances. The method's integration of camera-based detection with voice-guided navigation makes it relevant to the present investigation. A blind stick with many Internet of Things (IoT) functions was suggested; it would use ultrasonic sensors and GPS to navigate in real-time and identify obstructions. For outdoor navigation, the device provides audio feedback to let users know about surrounding obstructions. The study provides modest real-time analysis of the object identification system, although it is effective in basic functions. This study is brought up to date in terms of research emphasis with the use of video sensors to identify obstructions. The Internet of Things (IoT) and deep learning object detection-based approaches were used to create a smart guide system that helps the visually impaired with navigation and safety. The incorporation of deep learning enhances detection accuracy, and caregiver communication facilitates remote help. The research does not include the real-world performance test. This is quite relevant to the present investigation, especially when it comes to using YOLOv11 and other deep learning models for navigational aids. To help the visually handicapped navigate inside spaces, a system was developed that uses radio frequency identification (RFID), ultrasonic sensors, and Raspberry Pi to provide automated navigation. In interior environments, the technology provides voice guiding and efficient, highly accurate obstacle detection. [4] Due to the lack of functionality in outside environments, the device is only suitable for use inside. The aims of the current study are fully compatible with the incorporation of voice-guided navigation and obstacle detection. Combining artificial intelligence-powered facial recognition with camera-based object identification and ultrasonic sensor-based obstacle detection resulted in a highly intelligent system.[5] Even while it puts security first rather than complete navigation, its real-time facial recognition and obstacle avoidance are quite accurate. If person identification elements are to be included, the facial recognition aspect might be very valuable for this research. A Raspberry Pi 4 and a camera installed on a hat formed a wearable visual assistance device. [6] The device uses a beeping mechanism to provide aural feedback and uses large vision-language models (LVLMs) for real-time item identification. The user-friendliness of the system is enhanced by the ease with which additional items or individuals may be added for future identification. To further improve safety and freedom for visually challenged users, a distance sensor also transmits warnings to prevent accidents. To provide audio explanations of the user's surroundings in real-time, the Vocal Eyes system integrates a quantized Florence-2 vision-language model with a light text-to-speech engine. It provides context-aware aural feedback and analyzes live video feeds to identify objects, pedestrians, and barriers. It runs on hardware like the NVIDIA Jetson Orin Nano. [7] Users with visual impairments will have better situational awareness because to the system's emphasis on low latency and customized audio output. DISHA's outdoor navigation paradigm is based on a low-energy edge-deployed transformer. Battery life and computation are both preserved as the algorithm accurately recognizes sidewalks and traversable routes using a novel trimming mechanism.[8] By addressing the issue of energy restriction in wearable assistive technology, experimental findings demonstrate a significant improvement in accuracy and a longer device life. People with vision impairments may now access the world around them with the aid of DRISHTI, a wearable assistive device that is both affordable and effective. It has a camera module, speakers, Bluetooth, and an ESP32 CPU. [9] The technology is able to transform the user's motion data into audible output thanks to artificial intelligence processing. In order to discover solutions that are both efficient and cost-effective, this strategy blends mobility independence with both. The built-in camera of smartphones and Raspberry Pi devices, together with deep learning algorithms and large language models (LLMs), allowed for the construction of an interior navigation system.[10] The technology takes pictures of the surroundings in real time and uploads them to a database, where they are transformed into instructions that humans can understand. A door-to-door navigation system that doesn't rely on GPS signals may be delivered with this setup, which lowers the processing needs of user devices.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

III. METHODOLOGY

Fig. 1 shows an indoor navigation system that combines computer vision, depth estimation, and real-time audio input to help the visually impaired identify items and grasp their spatial relationships in their surroundings. To eliminate the need for additional devices, this system is designed to function on mobile devices, allowing users to engage just via voice instructions.

The core methodology is divided into five main components:

Preprocessing and Image Acquisition—Improving the accuracy of the model by acquiring higher-quality photos using a mobile camera. One application is object detection and classification, which involves utilizing the YOLO11 deep learning model to identify common home items and then categorizing different sorts of rooms according to those things. Estimating the distance of objects from the user using a monocular depth estimation model (ZoeDepth, DBT, or DIAMS) is part of depth estimation and distance calculation. Voice Feedback and Scene Description — Using LLaVA to generate scene descriptions and provide real-time audio explanations of identified items and their distances.

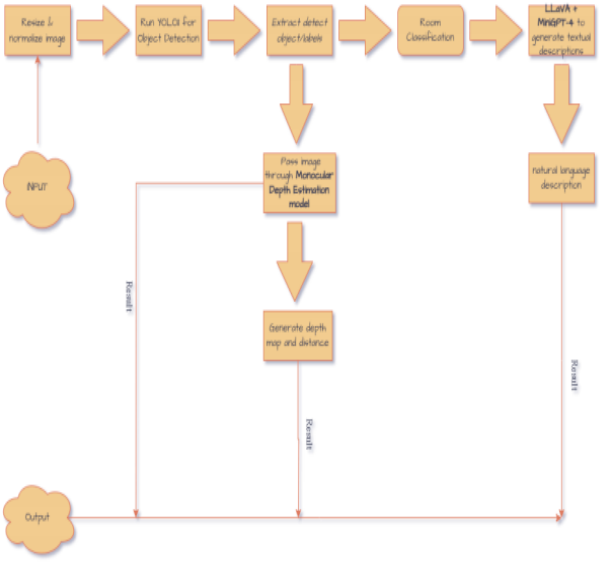

Fig. 1. Methodology of indoor navigation

To make sure the system works well in actual interior environments, image capture is an important stage. There are two ways in which the mobile app may take photographs of the user's surroundings: over 70 iterations over 16 independent batches, each with an input pixel resolution of 640 × 640. To expedite convergence, training was conducted in a GPU-accelerated environment. For object prediction, multiclass classification, and bounding box regression, the loss function maintained the compound structure of YOLO and binary cross entropy, as well as Complete Intersection over Union (CIoU) loss. The train and validation sets were appropriately divided, and the organizational architecture and pathways to the datasets were built utilizing YOLO principles. The setup made it easier to get good results with little inference.

1. Bounding Box Prediction

For object localization in interior settings, the formula for estimating bounding box coordinates is crucial:

$$b_x = \sigma(t_x) + c_x, b_y = \sigma(t_y) + c_y,$$
$$b_w = p_w \cdot e^{t_w}, \quad b_h = p_h \cdot e^{t_h}$$

Accurate guiding for visually impaired users is made possible by recognizing item placements using this equation, which finds the center (bx, by) and dimensions (bw, bh) of detected objects in relation to grid cells and anchor boxes. Mean Average Precision, abbreviated as mAP, Precise Mean Across All IoU Cutoffs (e.g., 0.5 to 0.95).An important emphasis for evaluating the system's performance in indoor contexts is this metric, which measures the model's correctness and dependability with an attained mAP@0.5 of 0.479. When assessing the model's Precision and Recall, the mAP formula is an essential tool:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

B. Fine-Tuning and Room Classification Approach YOLO11 for Indoor Navigation

For better item detection in everyday life, a curated custom dataset is used, which includes: Bedroom sets, dining room sets, couches, armoires, and tables Home appliances: air conditioners, stoves, ovens, and fans Where to Find Things: Doorways, Staircases, and Elevators Dangers: Unlocked cabinets, sticking

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

out items, and dangling light fixtures The system uses a rule-based categorization technique to identify the room type when items are detected: The presence of a refrigerator, stove, and dining table constitutes a kitchen. It is considered a bedroom if there is a bed, a wardrobe, and a light. Assume the room as a living room if you see a couch, TV, and coffee table there. It is considered a bathroom if there is a sink and a toilet. Once the items have been detected, the system uses rule-based reasoning to determine the room type. Figure 1 is one example.

Fig. 2. Dining Room

Fig. 3. Living Room

2. The presence of a dining table in Figure 2 indicates that it is a dining room, while the presence of a couch and television in Figure 4 indicate that it is a living room. By using object-context linkage, accurate room identification may be accomplished in this manner.

C. Depth Estimation & Distance Calculation

For monocular depth estimate, our system uses the Depth Anything V2 model because to its versatility and outstanding performance. Depth Anything V2 is perfect for real-time navigation on mobile devices with limited hardware because it uses a new architecture to improve the accuracy of depth predictions from single photos, unlike older models like ZoeDepth or DIAMS. We chose this model because it consistently produces accurate distance estimates in noisy settings and because it is resilient over a wide variety of interior situations. It can also handle complicated item arrangements. It strikes a good balance between speed and quality, and its lightweight design is in line with the need for efficient processing. As the model finds items and gives you their coordinates for their bounding boxes, you can use that information to get an estimate of their depth. Predicting relative depth values throughout the scene, Depth Anything V2 builds comprehensive depth maps. A convolutional neural

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

network (CNN) processes the input pictures, and the model refines depth estimates using a loss function that combines structural similarity (SSIM) and L1 loss. The central equation is:

$$[L = \alpha \cdot L_{L1} + (1-\alpha) \cdot (1 - SSIM(d_{pred}, d$$

In this context, L represents the overall loss, ($LL1$) is the average absolute error between the predicted and ground truth depths, $dpred$ is the predicted depth, dgt is the ground truth depth, SSIM is a measure of structural similarity, and () (usually 0.5) equalizes the contributions. A multi-scale technique is used to further enhance the depth prediction. The final depth map (D) is produced as:

$$[D = \sum_{s=1}^{S} w_s \cdot D_s] \qquad (5)$$

A number of scales (S), their weights (ws), and the depth map at each scale (s) are all part of this equation. By using environmental signals, roughly expressed as, a bespoke distance mapping algorithm converts these estimates into real distances (in meters).

$$\left[d_{real} = k \cdot \frac{1}{d_{pred}} + b\right] \qquad (6)$$

With (dreal) being the estimated real-world distance and (k) and (b) being calibration constants obtained from the scene's geometry.

Fig. 4. Object Distances

A real-life picture is augmented with distance annotations in this visual overlay. A sofa is indicated at 2.57 meters and a potted plant at 1.04 meters. It validates the model's practical value by reflecting its accuracy in converting depth data into usable distance measurements (Fig. 4). Figure 5 displays a depth map of a living room scenario that was built with Depth Anything V2. As an example of how well the model can depict spatial connections, the sofa is highlighted at about 2.57 meters and the shelves at about 1.44 meters by a color gradient that goes from purple (far away) to yellow (near).

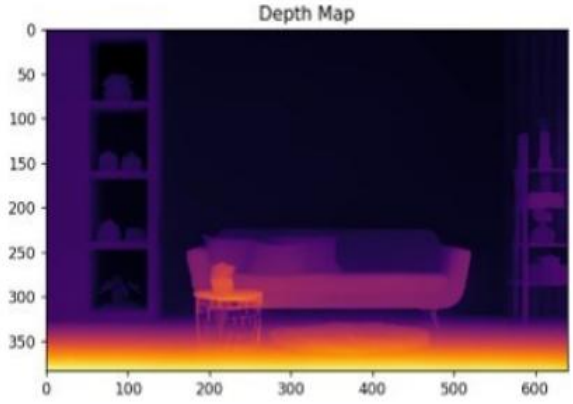

Fig. 5. Depth Map

D. Scene Description & Voice Feedback

The device incorporates a voice assistant that translates item detection findings into descriptions in normal language, allowing for a hands-free experience. Integrating LLaVA (Large Language and Vision Assistant) into the pipeline using a prompt-based, zero-shot inference technique allows for the production of scene descriptions that can be understood by humans from visual inputs. Prompt engineering guides LLaVA's multimodal reasoning rather than fine-tuning the model. An original picture was blended with the top N predicted class labels acquired after object recognition using YOLOv11 to produce structured inputs. To elicit varying degrees of semantic abstraction and clarity, three different prompt templates were developed: (1) "Describe this scene to a blind person, using only objects and where they are," (2) "Describe this indoor place clearly, including important objects but without excess," and (3) "Make a simple and straightforward description of what is shown in this picture." These instructions trained LLaVA to provide concise, location-aware, and context-based explanations. This modular design allows for versatile scene type customization without sacrificing the advantages of large-scale pretrained vision-language models or the additional training that comes with it. Processing of voice: tools like Google Translate, Amazon Polly, or Whisper by OpenAI are used to produce high-quality, realistic speech. Warning Tones: A warning tone is played if the object's proximity is low enough (less than 1 meter, for example). The system will provide

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

directions to the user (such as "Walk forward 3 meters to reach the door.") if it detects a door. Speech output is contextually adjusted to eliminate superfluous repetition, enhancing the user experience. Pictured above is the living room in question: An oak table, a white couch, and a rug of crimson make up the living room. On each side of the couch and table are two armchairs. A artwork adorns the wall, while a plant rests on a nearby surface. On top of that, the couch has some attractive cushions.

IV. RESULTS AND DISCUSSION

This model has successfully identified 0.615 items, demonstrating reliable object recognition. However, there is room for improvement in recognizing all relevant interior objects, as shown by the recall of 0.427 and mAP@0.5 of 0.479. Furthermore, by including depth estimates, spatial awareness is brought to interior surroundings. This is shown by visible depth maps that allow for precise direction and exact distance markings, such as a sofa at 2.57m and a potted plant at 1.04m. An ideal confidence threshold for balancing detection accuracy and coverage is roughly 0.2-0.4, according to confidence curves and precision-recall assessments. Raising recall, enhancing mAP, and using focus loss all contribute to better accuracy.

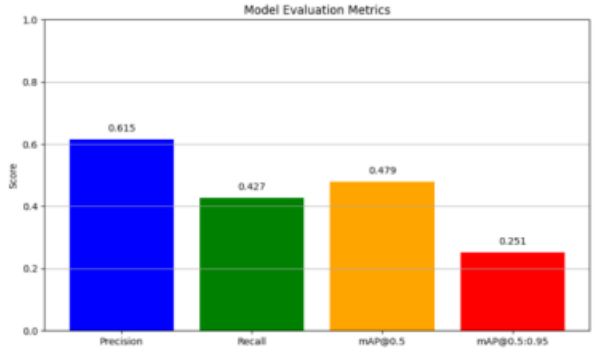

Fig. 6. Model Evaluation Metrics

The key performance indicators are shown in Figure 6: Achieving 0.615 for precision, 0.427 for recall, 0.479 for mAP@0.5, and 0.251 for mAP@0.5:0.95.To avoid the system providing incorrect directions to visually impaired users, precision(0.615) evaluates 61.5% of the recognized items as accurate. The model accurately recognizes 42.7% of all correlating items, with a recall of 0.427, suggesting there is room for improvement to ensure no crucial things are missed.The mean Average Precision at an IoU threshold of 0.5, denoted as mAP@0.5 (0.479), is a good indicator of moderate item identification accuracy and is appropriate for an early-stage prototype. The model's performance does not meet more rigorous assessment standards, as shown by mAP@0.5:0.95 (0.251), the average mAP across multiple IoU thresholds. This indicates that the model requires tuning to handle varied circumstances. Figure 7 shows that when the confidence criteria are increased, the average accuracy for all classes decreases. While the model's progressive drop indicates that it has to be tweaked to keep accuracy at higher thresholds as well, the initial high precision for low thresholds provides accurate object detection for the visually handicapped. Model confidence ratings may be fine-tuned using methods like temperature scaling to improve accuracy. Learning from difficult or misclassified data may also be emphasized using loss functions such as focused loss.

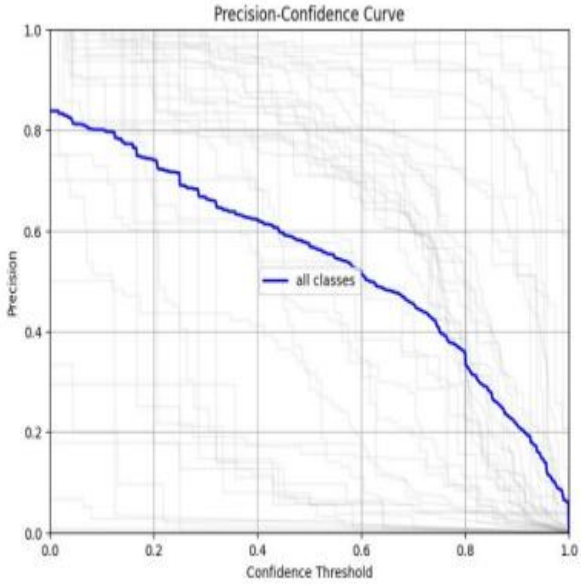

Fig. 7. Precision Confidence Curve

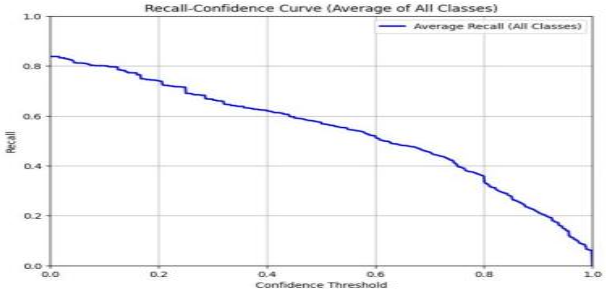

Fig. 8. Average Recall-Confidence Curve

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

For each given confidence level, we display the average recall on a graph in Figure 8. When the confidence level goes from zero to one, the recall goes down from almost one to zero. The model's high recall at low thresholds makes it easy to lead visually impaired individuals through congested areas; it can identify most objects on the first attempt. Because greater levels of confidence may lose out on some actual advantages due to the abrupt drop-off, a good balance of thresholds is necessary to maximize utilization.

V. CONCLUSION

An revolutionary software called Smart Assistive System for the Blind processes interior scenes using a smartphone and seeks to increase the freedom of the sight impaired. Although not finished yet, the effort has made good progress by enhancing the YOLOv11 object identification algorithm using a one-of-a-kind collection of commonplace interior objects, including furniture and home items connected to the mobility of blind users. Although there is room for improvement in the perception of all important interior items, the calibration is attained to a level of 0.615, indicating successful object identification. However, the recall is 0.427 and the mAP@0.5 is 0.479. In addition, a visible depth map and effective distance labeling (e.g., sofa at 2.57m, potted plant at 1.04m) allow for excellent guiding in interior contexts when depth estimation is used. When balancing detection accuracy with coverage, confidence curves and precision-recall plots suggest a confidence threshold of 0.2-0.4 as the sweet spot. In order to develop a robust, real-time assistance solution, future work will center on improving the model, adding more indoor objects to the dataset, and evaluating depth estimation in different indoor environments. The project shows promise in indoor scene analysis.

REFERENCES

[1] Khan, Sulaiman, Shah Nazir, and Habib Ullah Khan. "Analysis of navigation assistants for blind and visually impaired people: A systematic review." *IEEE access* 9 (2021): 26712-26734.

[2] Souza, Leandro Rossetti de, Rosemary Francisco, João Elison da Rosa Tavares, and Jorge Luis Victória Barbosa. "Intelligent environments and assistive technologies for assisting visually impaired people: a systematic literature review." *Universal Access in the Information Society* (2024): 1-28.

[3] Ramaraj, Kottaimalai, Varun Teja, Teja Venkata Sainath, Siva Reddy, and Purushotham Naidu. "Custom Voice Assistants: Enhancing Accessibility for the Visually Impaired." In *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pp. 1-6. IEEE, 2025.

[4] Azhaguraja, R., KS Ashvith Kumar, C. V. Paranthaman, and Arun Kumar. "Assistive Technologies for Blind and Visually Impaired Individuals–A Short Review." In *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp. 1-6. IEEE, 2025.

[5] Khan, Sulaiman, Shah Nazir, and Habib Ullah Khan. "Analysis of navigation assistants for blind and visually impaired people: A systematic review." *IEEE access* 9 (2021): 26712-26734.

[6] Saleem, Talal, and V. Sivakumar. "A Mobile Lens: Voice-Assisted Smartphone Solutions for the Sightless to Assist Indoor Object Identification." *EAI Endorsed Transactions on Internet of Things* 10 (2024).

[7] Madhu, B. M., Sanket Raj, and Sayan Das. "Vision Assistance System for Visually Impaired." In *2025 3rd International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pp. 1-5. IEEE, 2025.

[8] Hemavathy, J., A. Sabarika Shree, S. Priyanka, and K. Subhashree. "AI based voice assisted object recognition for visually impaired society." In *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pp. 1-7. IEEE, 2023.

[9] Muktha, D. S., G. Niveditha, Nikhil Anthony Pinto, and Somnath Sinha. "Enhancing Mobility: A Smart Cane with Integrated Navigation System and Voice-Assisted Guidance for the Visually Impaired." In *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 1124-1129. IEEE, 2024.

[10] Sindhu, B., B. Bhagya Preethi, S. Leela Venkata Lakshmi, B. Praveen Reddy, and S. Srinivas Kiran. "Voice-Assisted Artificial Intelligence Based Question Answering System for the Visually Impaired." In *Algorithms in Advanced Artificial Intelligence*, pp. 135-140. CRC Press, 2025