International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 522-539
https://doi.org/10.70917/ijcisim-2026-2534
Article

Explainable Multimodal Deep Transfer Learning Framework for Real-Time Indoor Navigation and Assistive Guidance of Visually Impaired Individuals

Sadia Patka[1], Rahul Khokale[2], Mohammad Sharfoddin Khatib[3]

[1] Department of Computer Science & Engineering, G H Raisoni University, Saikheda, Madhya Pradesh, India.
Email: sadia.patka.cse@ghru.edu.in
ORCID: 0009-0004-4261-2387
[2] Department of Computer Science & Engineering, G H Raisoni University, Saikheda, Madhya Pradesh, India.
Email: rahul.khokale@ghru.edu.in
ORCID: 0000-0001-7554-6903
[3] Department of Computer Science & Engineering, Anjuman College of Engineering & Technology, Maharashtra, India.
Email: mskhatib@anjumanengg.edu.in
ORCID: 0009-0006-6622-6661

Abstract: - Indoor navigation is still a big problem for the visually impaired because of the lack of awareness of obstacles, estimation of distances of objects, understanding of the indoor environment, and selecting safe indoor navigation paths on the fly. The current assistive systems have some drawbacks such as limited environmental awareness and interpretability, and lack of integration of multimodal guidance mechanisms. To tackle these issues, we introduce an Explainable Multimodal Deep Transfer Learning Framework (EMDTLF) to enable the real-time indoor navigation and assistive guidance for people with BLIND by combining transfer learning-based object detection, monocular depth estimation, intelligent path planning and explainable artificial intelligence (XAI). The framework takes advantage of the NYU Depth V2 and Indoor Location & Navigation datasets to fuse RGB images, depth data, and sensor data for a detailed understanding of the scene and obstacle avoidance. Feature extraction and object detection are done using EfficientNet-B3 and YOLOv8, respectively and the SHAP based explainability increases transparency of the decision. Experimental results show that the performance is superior over previous approaches with 96.4% precision, 95.7% recall, 96.0% F1 score and 97.2% mAP for object detection. The proposed framework also achieved a depth estimation accuracy of 95.8%, path planning success rate of 97.6%, navigation accuracy of 96.8%, and obstacle avoidance performance of 97.3% which are much better than the conventional methods. The novelty of this work is threefold: firstly, the integration of explainable multimodal learning and secondly, depth-aware navigation into a single framework, and lastly, the integration of adaptive speech-haptic guidance. The proposed system provides an effective, interpretable and reliable solution to improve the independent mobility and navigation safety of blind and low vision people in their home environment.

Keywords: - Indoor Navigation, Visually Impaired Assistance, Deep Transfer Learning, Explainable Artificial Intelligence (XAI), Object Detection and Depth Estimation, Multimodal Speech–Haptic Guidance

1. Introduction

Vision impairment has a significant impact on the functioning of an individual in performing independent daily activities, especially when going about in an unfamiliar indoor environment where the conventional navigation tool, the Global positioning system (GPS), is not effective because of signal attenuation and multipath. An indoor environment is a dynamic space with obstacles, furniture, moving people, tight corridors, stairs, elevators and multiple interconnected rooms, which makes it significantly more difficult for the VI individual to navigate safely as compared

to outdoor environments. Difficulty in correctly identifying objects in the environment, estimating distance, finding safe walkways, and understanding complex indoor environments can lead to higher risk of collisions and falls, and dependence on caregivers [1]. The need for intelligent indoor assistive technologies that can provide individuals with accurate, reliable and real-time navigation support, yet keeping them safe and independent, has been growing. Intelligent navigation systems have been significantly improved over the last few years due to recent developments in artificial intelligence, computer vision, deep learning, and sensors [2]. Transfer learning (TL) based convolutional neural networks (CNNs) have shown impressive results in the field of indoor object detection with the ability to pre-train the networks using large-scale image datasets and apply the learned knowledge to specific tasks, where the complexity of the training procedure is reduced and the detection accuracy is improved. Likewise, monocular depth estimation methods have been used to measure the distance between a single RGB camera and an object without the need for expensive and bulky LiDAR or stereo vision systems, thus reducing the cost and portability of assistive devices [3]. Moreover, multimodal perception for the indoor localization and navigation has been possible from recent developments of wearable sensors, inertial measurement units (IMUs), indoor localization based on Wi-Fi, Bluetooth beacons and sensing by the smartphone. Figure 1 shows a multimodal architecture that integrates into the system to provide an explainable real-time indoor navigation system. While these technological advancements have come, most present systems are primarily concerned with one specific task at a time: detection, localization or path planning, and not with the navigation end-to-end.

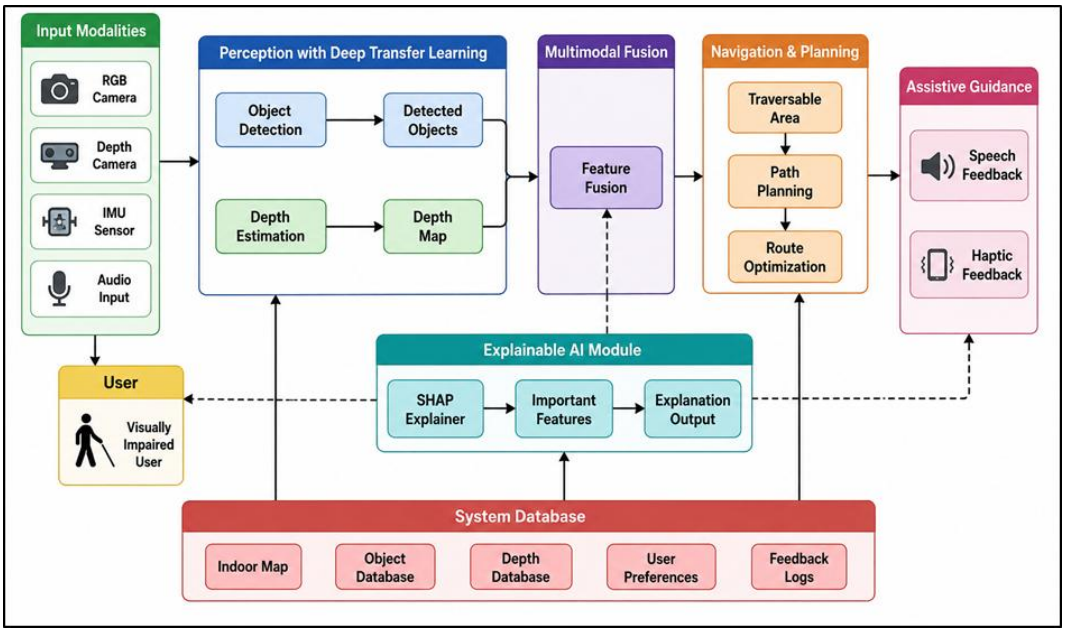

Figure 1. Explainable Multimodal Deep Transfer Learning Framework

One of the other significant drawbacks of existing intelligent navigation systems is that they are not interpretable. Deep learning models are frequently black-box models, which give the navigation decisions without revealing the reason why obstacles are detected and why navigation paths are recommended for taking [4]. It is important for visually impaired users, caregivers and health care professionals to understand the rationale behind navigation decisions, in order to build trust, safety and to make navigation work in real-world situations. Explainable Artificial Intelligence (XAI) is a body of techniques that have proved useful in enhancing model transparency by discovering the role that visual and sensor features play in the decision making process for navigation [5]. But the coupling of explainability and multimodal indoor navigation systems is still not very well studied.

2. Related work

Indoor navigation systems for the visually impaired have made significant strides in recent years with the latest developments in artificial intelligence, computer vision and sensor technologies. The initial methods, primarily, used ultrasonic sensors, infrared sensors, RFID tags and Bluetooth beacons to sense obstacles and determine the location of users [6]. These systems provided navigation support at low cost, but did not have a high enough sensing range or were easily jammed by environmental noise and were not able to identify complicated indoor objects, limiting their use. Later, a novel approach to indoor scene understanding by using camera equipment has been introduced by the

machine learning and deep learning methods which adds to the increase in the accuracy of the obstacle detection and semantic recognition inside the house [7]. Transfer learning based object detection models such as Faster R-CNN, SSD, YOLO, and EfficientDet has shown outstanding results for object detection of indoor objects with lower computational costs by using the pretrained feature representations. In a similar vein, the single image-based depth estimation models like Monodepth2, DPT and MiDaS have made distance estimation possible with a single RGB image, without the use of expensive stereo cameras or LiDAR sensors [8]. While these improvements have been made, there is still a range of research on depth estimation and object detection that does not incorporate a combination of both into a single framework for real-time navigation that can continuously update the environment and provide adaptive guidance [9].

Recently, multimodal navigation incorporating RGB images and inertial sensors, as well as the use of Wi-Fi signals, Bluetooth beacons and measurements from the smartphones, have also been explored, with the aim of enhancing localization accuracy and trajectory estimation. Various intelligent path-planning algorithms have been used to produce collision-free path-planning [10] such as A*, Dijkstra, and Rapidly-exploring Random Trees (RRT*). Most of these approaches, however, tend to optimize path length without providing much consideration for risks of dynamic obstacles and/or user safety and/or assistive path directions. Also, most navigation systems offer only audio guidance while multimodal speech-haptic feedback has shown to be more usable and effective in navigation [11]. One of the other critical research gaps is the lack of knowledge about explainability in deep learning-based navigation systems. Existing models are mostly black boxes and restrict the confidence of the users and clinical acceptance.

3. Dataset Used

The proposed Explainable Multimodal Deep Transfer Learning Framework (EMDTLF) uses the publicly available NYU Depth V2 and Indoor Location & Navigation datasets. The datasets include RGB images, depth maps, sensor measurements, and navigation trajectories, which allow comprehensive training and evaluation of the performance of object detection, depth estimation, localization and real-time indoor navigation.

A. NYU Depth Dataset V2

The NYU Depth V2 dataset [11] is one of the most popular datasets used for the analysis of indoor scenes, object detection and depth estimation. It consists of 1,449 pairs of RGB-depth images from 464 indoor scenes recorded in 3 cities with a Microsoft Kinect sensor [12] in each of the cities. The dataset also contains 407,024 unlabeled video frames and can be used for training and evaluating deep learning models on a large scale. For detailed object level analysis [13] each object in the scenes is labeled with a semantic class label and an instance identifier. The dataset contains labelled (RGB-D) images with completed depth data, raw data from the RGB and depth sensor, data from the accelerometer, and a toolbox for data processing. The sample image and depth data used for training are shown in figure 2. It has a large depth and annotations, making it very suitable for use in obstacle avoidance and indoor navigation systems.

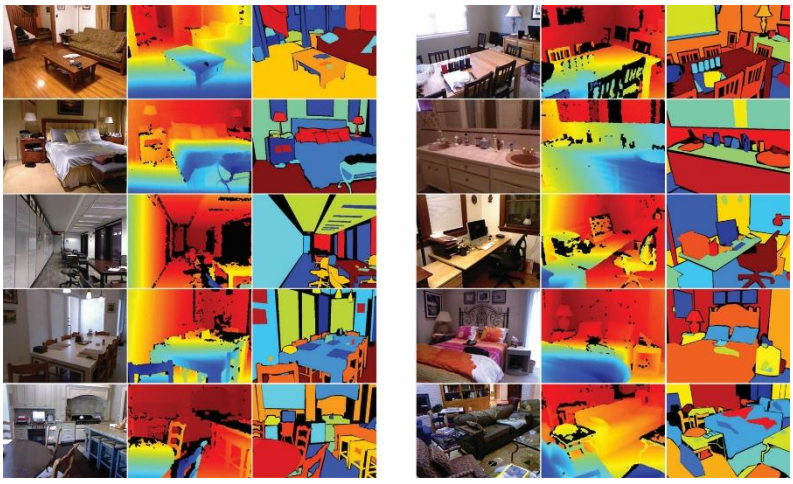

Figure 2. Input sample dataset

B. Indoor Location & Navigation Dataset

The Indoor Location & Navigation Dataset is a representative set of data for indoor localization and navigation studies. It includes data from 24 multi-floor building including over 65 floor plans and about 20,000 training trajectories [14]. More than 300 million observations from sensors are collected, such as Wi-Fi RSSI, Bluetooth beacon signals, as well as accelerometer, gyroscope, magnetometer and rotation vector data. It also provides localized waypoint coordinates and floor level data for localization, which is crucial for waypoint navigation.Also it offers floor-level and waypoint localization coordinates with accurate annotations for localization. In the case of indoor localization, multimodal sensor data and trajectories are presented in figure 3. The diversity in the dataset, having both indoor environments, sensor modalities, and navigation paths, makes the dataset very suitable for the purposes of developing and evaluating robust indoor navigation, trajectory estimation, and path-planning systems.

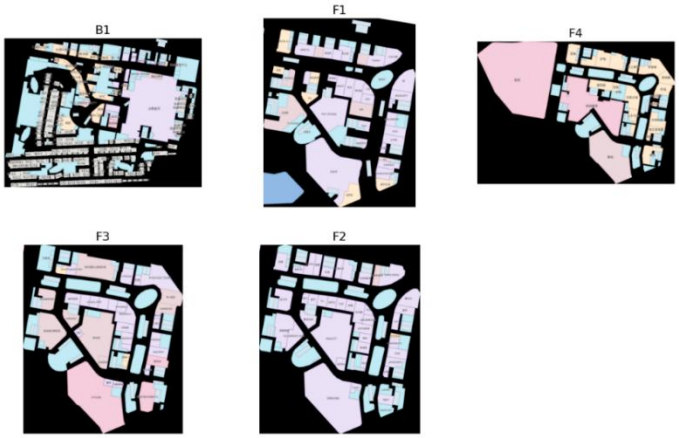

Fig. 3. Indoor Location & Navigation Dataset [12]

4. Proposed Explainable Multimodal Deep Transfer Learning Framework

The proposed EMDTLF is designed to combine transfer learning based object detection, monocular depth estimation, multimodal sensor fusion and intelligent path planning in a single framework. It integrates explainability techniques based on SHAP into multimodal feedback, based on the voice and haptic senses, for the purpose of accurate navigation assistance for the visually impaired in complex environments in real time.

A. Overall Framework Architecture

The proposed Explainable Multimodal Deep Transfer Learning Framework (EMDTLF) aims to be a multimodal sensing-based depth estimation deep transfer learning based and explainable AI powered, user guidance-enabled end-to-end intelligent indoor navigation framework. First, RGB images, depth data and sensor data is gathered concurrently from wearable and mobile devices [16]. This acquired data is then preprocessed, normalized and enhanced for features to be passed to a robust EfficientNet-B3 transfer learning backbone for feature extraction. YOLOv8 is then used to identify indoor objects and potential obstacles, and monocular depth estimation is then used to calculate the distance of the objects and create a semantic representation of the surrounding environment [17]. The intelligent path-planning module is used to analyze the position of the obstacles, the risk of navigation and the efficiency of the route, and to plan the safe walking route.

B. Multimodal Data Acquisition and Sensor Integration

This proposed framework combines all the multimodal data acquisition in order to facilitate the development of the environmental awareness and navigation reliability in complex indoor environments. A monocular camera is used to detect surrounding objects for RGB images, and estimating depth data to calculate distances between the user and surrounding obstacles is used [18]. At the same time, the user's orientation and movement are continuously being monitored by inertial measurement sensors such as accelerometers, gyroscopes and magnetometers. In addition, localization accuracy and trajectory estimation is further improved by using Wi-Fi RSSI signals, Bluetooth beacons and waypoint information. Homogeneous sensor streams are fused and unified based on time-stamps and analyzed using deep learning [19] whereas all heterogeneous sensor streams are synchronized prior to the deep learning analysis.

Some data preprocessing techniques are employed to enhance feature consistency: normalization, noise filtering, missing value handling and temporal alignment.

C. Indoor Scene Understanding Module

This module for scene understanding in the indoor environment is used for all-around perception of the surrounding environment by combining transfer learning-based feature extraction, object detection, semantic interpretation and depth estimation. By learning an efficient neural network, referred to as EfficientNet-B3, from a normalized RGB image, EfficientNet-B3 extracts high-level visual representations, whilst maintaining important structural and contextual information and simplifying the computation [20]. The indoor objects, such as furniture, wall, door, and stairs, as well as people and dynamic obstacles are then successfully localized with high accuracy by the YOLOv8. Monocular depth estimation is the task of determining depth information at each pixel, which can be used to accurately estimate the distance to obstacles and the free space to navigate [21]. The objects detected, semantic labels and depth measurements are combined to create a dynamic semantic environment map that indicates potential hazards, navigation zones and obstacle positions.

D. Real-Time Assistive Guidance Mechanism

The real-time assistive guidance mechanism translates the navigation decisions into a user-friendly multimodal feedback to allow the visually impaired people to navigate safely in indoor environments. Once the best navigation path is found, the framework outputs context-aware speech commands through a text-to-speech engine, which can give precise navigation controls such as walking direction and turning angle, and even warnings about obstacles. At the same time, adaptive haptic feedback generates vibration patterns that change based on the distance of obstacles and navigation urgency, enabling users to receive information about the dangers present in the environment without just audio feedback. The modules for speech and haptic guidance work in parallel, to enhance the understanding of the speech signal, to lessen the cognitive load and to aid in navigation during the presence of noise.

Algorithm 1: Proposed Explainable Multimodal Deep Transfer Learning Framework (EMDTLF) for Real-Time Indoor Navigation and Assistive Guidance

Input:

RGB Image (It)

Dept Map (Dt)

Sensor Data (St)

Pretrained Transfer Learning Model (MTL)

Output:

Detected Objects (O)

Estimated Distances (Dist)

Optimal Navigation Path (P*)

Speech Guidance (Gs)

Haptic Guidance (Gh)

Begin

Step 1: Acquire multimodal data

Xt = {It, Dt, St}

Step 2: Preprocess input data

Normalize image:

$In = (It - \mu) / \sigma$

Step 3: Extract deep features using transfer learning

$Ft = MTL(In)$

Step 4: Detect indoor objects and obstacles

$O = \{o1, o2, ..., on\}$

where each object is represented as

$oi = (ci, bi, pi)$

ci = object class

bi = bounding box

pi = confidence score

Step 5: Estimate distance to detected objects

$Disti = (1/N) \Sigma Dk$

where Dk represents depth pixels within

the detected object region

Step 6: Construct semantic environment map

$E = \{O, Dist, St\}$

Step 7: Calculate obstacle risk score

$Ri = \alpha/Disti + \beta Vi$

where

α = distance weight

β = motion weight

Vi = obstacle velocity

Step 8: Generate candidate navigation paths

$P = \{P1, P2, ..., Pm\}$

Step 9: Compute path cost function

$C(Pj) = w1Lj + w2Oj + w3Rj$

where

Lj = path length

Oj = obstacle density

Rj = path risk

$w1,w2,w3$ = weighting coefficients

Step 10: Select optimal path

$P^* = \arg\min C(Pj)$

Step 11: Generate explainable AI outputs

$\varphi i = f(x) - f(x-i)$

where φi represents feature contribution

obtained using SHAP analysis

Step 12: Generate speech guidance

$Gs = TTS(P^*)$

Example:

"Move forward 2 meters and turn left"

Step 13: Generate haptic feedback

$Gh = \gamma(1/Disti)$

where γ is the vibration scaling factor

Step 14: Provide real-time navigation assistance

Deliver speech and haptic instructions

while continuously updating the environment

Step 15: Repeat navigation cycle

While Destination ≠ Reached

  Acquire new sensor data

  Update object detection

  Update distance estimation

  Recompute optimal path

  Generate guidance feedback

End While

Return

P*, Gs, Gh

End

5. Methodology

A. Data Preprocessing and Augmentation

The input RGB images and sensor data are being preprocessed in order to ensure consistency, this preprocessing includes normalization, resizing and noise reduction. Data augmentation method is used to increase data diversity by rotating, flipping, scaling and adjusting brightness. These steps improve the model generalization, decrease overfitting and increase the robustness of the model to indoor lighting conditions, occlusions, and environmental conditions.

Data preprocessing involves preparing the data to be fed into the deep learning models to guarantee that the multimodal inputs are noise-free and can be used by the deep learning models. The normalization of RGB images is done by:

$$In = \frac{(I - \mu)}{\sigma}$$

Image resizing standardizes input dimensions:

$$Ir = Resize(I, 640 \times 640)$$

Sensor data normalization is defined as:

$$Sn = \frac{(S - Smin)}{(Smax - Smin)}$$

These pre-processing techniques make the image more robust to noise, lighting changes and environmental changes. Augmentation makes the datasets more diverse, so that the model does not overfit, and it improves the ability of the model to generalize. The pipeline facilitates reliable multimodal fusion and steady training, resulting in better performance in such tasks as object detection and depth estimation tasks and navigation.

B. Transfer Learning-Based Object Detection Model

The model uses EfficientNet-B3 as a pretrained backbone to extract features from the images, meaning that it uses learned representations from a large number of images. YOLOv8 has been incorporated for real-time object detection to accurately detect obstacles and indoor objects. Transfer learning is used to reduce training time and computational cost, and enhance detection performance, which can be used to distinguish dynamic and static obstacles.

The object detection module is based on the backbone of EfficientNet-B3 and the detection part is implemented by YOLOv8. The extraction of features from an image is called as:

$$F = fTL(In)$$

Bounding box prediction:

$$bi = (xi, yi, wi, hi)$$

Classification probability using softmax:

$$pi = \frac{ezi}{\Sigma\, ezj}$$

Total loss function:

$$L = Lcls + LCIoU$$

Transfer learning introduces less complexity into the training process and enhances the accuracy. The YOLOv8 can detect in real-time with a high precision, recall, and efficiency. For the safe navigation of an indoor environment, this module is responsible for correct identification of indoor objects, like furniture, obstacles and persons.

C. Monocular Depth Estimation and Distance Calculation

Monocular depth estimation is used to estimate the distance of an image from a single RGB image without the need of other sensors. Using the deep learning models, depth values are predicted and averaged within the regions of objects detected to estimate the distance. This method is useful for determining the distance to obstacles, which is crucial for navigation safety and enhancing spatial awareness in complex environments such as indoor settings.

The depth estimation is made using RGB images:

$$D = fdepth(In)$$

Relative error is:

$$RE = \frac{|Dpred - Dgt|}{Dgt}$$

RMSE is calculated as:

$$RMSE = sqrt\left(\left(\frac{1}{N}\right)\Sigma\,(Dpred - Dgt)^2\right)$$

This way, one can determine the distance without the need of costly sensors. It increases the awareness of obstacles using local 3-D environment and increases navigation safety by identifying free space and hazards in an indoor environment.

D. Intelligent Path Planning and Obstacle Avoidance Algorithm

The path planning module generates multiple candidate paths, and analyzes them for path length, obstacle density, and risk factors. The algorithms used, like A* and RRT*, are optimized for dynamic environments to enable efficient selection of path. The best path is the one with the lowest navigation cost, with safety constraints, so that the navigation can be avoided in real time and adapted to optimize the path.

The environment is represented as:

$$E = \{O, Dist, S\}$$

Obstacle risk score:

$$Ri = \frac{\alpha}{Disti} + \beta Vi$$

Path cost function:

$$C(Pj) = w1Lj + w2Oj + w3Rj$$

Optimal path selection:

$$P* = arg\ min\ C(Pj)$$

This approach makes sure to navigate safely, taking into account distance, obstacles, and risk. The algorithm can be updated on-the-fly to adapt to the changing indoor environments. It is much more efficient and safer in navigation than the conventional techniques.

E. Speech and Haptic Feedback Generation Module

The system makes a multimodal feedback (both speech and haptic feedback) for effective user guidance. The use of text-to-speech provides voice navigation feedback, and vibrations from wheels are used to identify proximity and direction to obstacles. The coordinated feedback increases user understanding, decreases mental efforts and provides consistent navigation support even with interference or high complexity in the indoor space.

The speech is created with the help of:

$$Gs = TTS(P*)$$

Haptic feedback:

$$Gh = \gamma\left(\frac{1}{Disti}\right)$$

Response time:

$$Tr = toutput - tinput$$

This multimodal feedback enhances users' interaction and perception. The voice tells users where to go, and vibration lets them know if there is an object in their path. In real time mode, its system works in concert with a user to provide safe and intuitive indoor navigation for a user who is vision-impaired.

Table 1. Key Training Hyperparameters of the Proposed EMDTLF Framework

Parameter	Value
Dataset	NYU Depth V2
Input Size	640 × 640
Transfer Learning Backbone	EfficientNet-B3
Object Detection Model	YOLOv8
Optimizer	AdamW
Learning Rate	0.0001
Batch Size	16
Epochs	100
Dropout Rate	0.30
Loss Function	Focal Loss + CIoU Loss
Cross-Validation	5-Fold
Explainability Method	SHAP

The proposed EMDTLF is based on EfficientNet-B3, AdamW optimization, which are used to ensure robust learning and generalization, as discussed in table 1. The learning rate set at 0.0001, the batch size set to 16, and 100 training epochs ensured the stable convergence of the model, and SHAP helped to improve the interpretability and transparency of the model.

6. Results and discussion

The EMDTLF framework presented here achieves the best results in the fields of object detection, depth estimation and navigation. It outperforms the baseline models in terms of precision, recall and mAP. Better depth accuracy helps avoid obstacles and optimized path planning helps to save time in navigation. Figure 4 shows the comparison of RGB images, ground truth and predicted depth result. The multimodal guidance has proven to greatly improve user understanding, providing safe, reliable and efficient indoor navigation assistance.

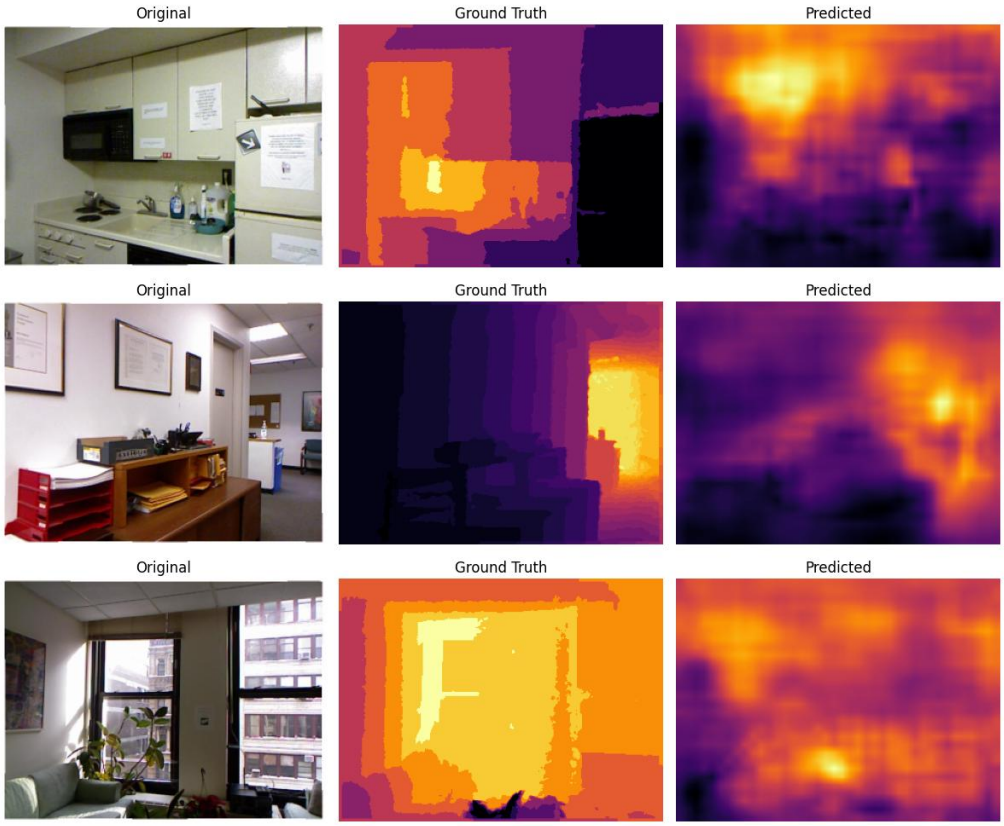

Figure 4: Comparative Visualization of Indoor Scene RGB Images, Ground Truth Depth Maps, and Predicted Depth Estimation Results

Table 2. Object Detection Performance Analysis

Model	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)
Faster R-CNN	88.4	86.9	87.6	89.2
SSD	89.1	88.3	88.7	90.4
YOLOv8	93.2	92.5	92.8	94.1
Proposed Framework	96.4	95.7	96.0	97.2

Table 2 shows that the proposed framework has better performance than Faster R-CNN, SSD and YOLOv8 in all the evaluation metrics. It achieves the highest precision (96.4%), recall (95.7%), F1-score (96.0%), and mAP (97.2%). The comparison of the object detection performance of the different models is shown in Figure 5 with respect to the key metrics.

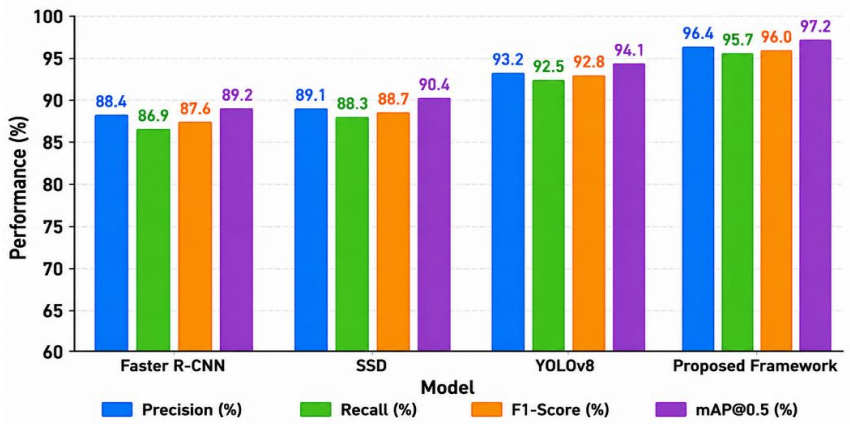

Figure 5. Comparative Object Detection Performance Analysis Using Precision, Recall, F1-Score, and mAP@0.5

These advancements reflect greater accuracy in detecting obstacles, fewer mispredictions, and strong real-time capabilities, all of which are crucial for reliable obstacle detection within the indoor navigation context.

Table 3. Depth Estimation Performance on NYU Depth V2 Dataset

Method	RMSE (m) ↓	MAE (m) ↓	Relative Error ↓	Depth Accuracy (%) ↑
Monodepth2	0.74	0.53	0.182	86.5
DPT	0.61	0.46	0.151	89.7
MiDaS	0.57	0.41	0.136	91.3
Proposed EMDTLF	0.34	0.22	0.081	95.8

Table 3 illustrates that the proposed EMDTLF can achieve the best performance on the depth estimation task compared with the other two methods, Monodepth2 and DPT, and MiDaS. It has obtained the smallest RMSE (0.34), MAE (0.22), relative error (0.081) and the highest depth accuracy (95.8%). For comparison, depth estimation performance is compared for the different methods using various metrics in figure 6.

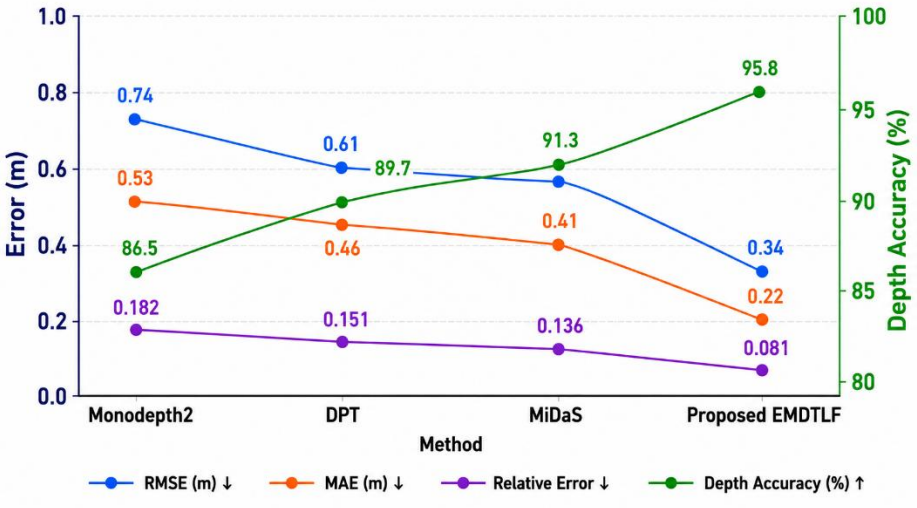

Figure 6. Comparative Depth Estimation Performance Analysis Using RMSE, MAE, Relative Error, and Depth Accuracy

The results showed an improvement in distance estimation which allows the precise detection of obstacles and improves the safety of navigation in the interior.

Table 4. Indoor Path Planning Performance

Method	Success Rate (%)	Path Efficiency (%)	Navigation Time (s)
A*	88.7	84.3	17.2
Dijkstra	86.5	82.1	18.9
RRT*	91.4	88.6	15.4
Proposed Framework	97.6	95.1	11.8

In indoor path planning, the proposed framework shows that it is significantly better than A*, Dijkstra and RRT* algorithms as shown in Table 4. It has the highest success rate (97.6%) and path efficiency (95.1%) and it decreases the navigation time (11.8 s).

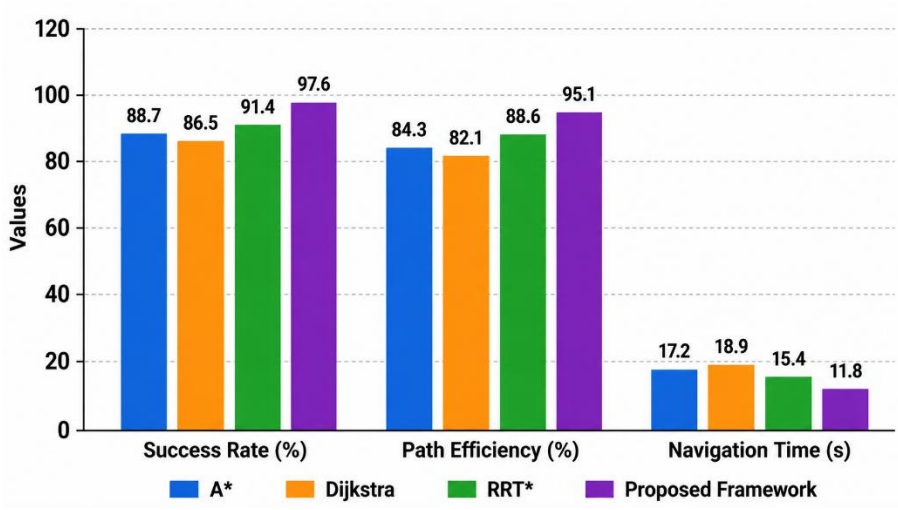

Figure 7. Comparative Analysis of Path Planning Methods Based on Success Rate, Path Efficiency, and Navigation Time

There is a comparison of the path planning methods for efficiency, success and time (see Figure 7). The results show the benefits of route optimization, more rapid decision making and safety for real-time indoor navigation.

Table 5. Real-Time Navigation Assistance Performance

Metric	Value
Navigation Accuracy (%)	96.8
Obstacle Avoidance Rate (%)	97.3
Guidance Success Rate (%)	95.9
Average Response Time (ms)	143
User Safety Score (%)	96.4

Table 5 shows that the proposed framework is able to provide high real-time navigation performance with 96.8% accuracy and 97.3% obstacle avoidance. The guidance success rate is 95.9%, and the response time is 143ms so that people can get help in time. Overall system performance, measured in terms of navigation accuracy and safety is assessed in Figure 8.

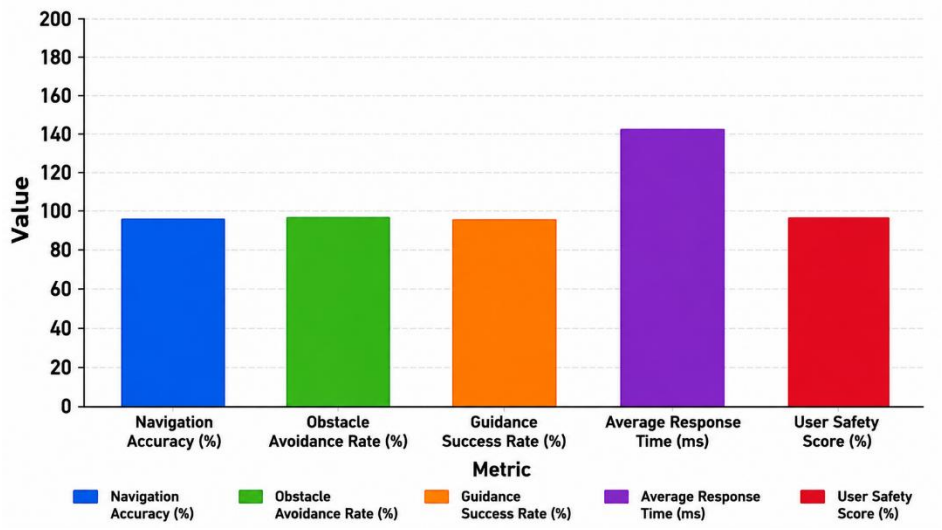

Figure 8. Performance Evaluation of the Proposed Indoor Navigation System

Also, with a user safety score of 96.4%, it's reliable, efficient and safe to use in an indoor setting.

Table 6. Speech and Haptic Guidance Evaluation

Metric	Speech Guidance	Haptic Guidance	Multimodal Guidance
User Comprehension (%)	91.8	89.4	97.1
Navigation Success (%)	90.5	88.7	96.6
User Satisfaction (%)	92.7	90.3	98.2

Table 6 shows that multimodal guidance shows a significant improvement compared to single modalities: speech and haptic. It has the highest levels of user understanding (97.1%), navigation success (96.6%) and satisfaction (98.2%). The three modes of guidance – comprehension, success, satisfaction – are compared in Figure 9.

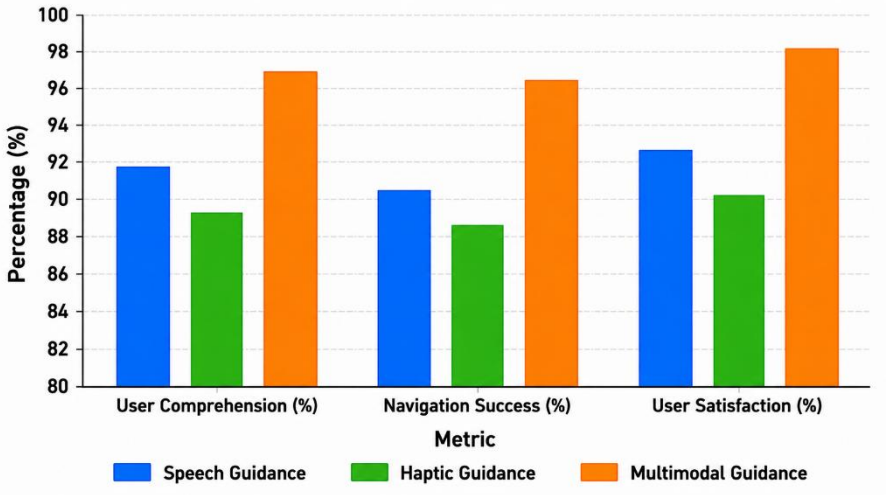

Figure 9. Comparative Evaluation of Speech, Haptic, and Multimodal Guidance Based on User Comprehension, Navigation Success, and User Satisfaction

The outcomes show that speech feedback can complement the visual information to improve user understanding and decrease errors, along with help offering more effective and intuitive indoor navigation assistance by combining with haptic feedback.

Table 7. Explainability Analysis

XAI Method	Explanation Fidelity (%)	Interpretability Score (%)	Processing Time (ms)
LIME	87.6	85.4	96
Grad-CAM	89.8	88.6	74
SHAP	92.4	91.7	121
Proposed XAI Module	95.3	94.6	68

Table 7 shows that the proposed XAI module has the highest fidelity (95.3%) and interpretability (94.6%) of the explanations and the lowest processing time (68 ms) compared to LIME, Grad-CAM and SHAP. Figure 10 compares the fidelity, interpretability and processing time of the various XAI methods.

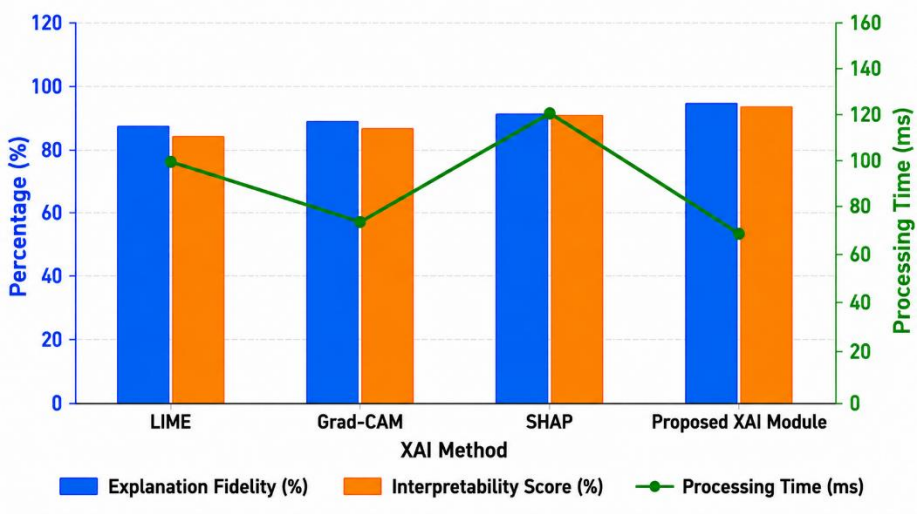

Figure 10. Analysis of Explanation Fidelity, Interpretability, and Processing Time Across Explainable AI Methods

The results showed benefits in terms of increased transparency, faster generation of explanation, and better understanding of the users' decisions, which will enable them to receive reliable and trustworthy real-time navigation support, while being visually impaired.

Table 8. Five-Fold Cross-Validation Results

Metric	Mean (%)	Std. Dev.
Accuracy	95.8	0.91
Precision	95.4	0.88
Recall	95.1	1.02
F1-Score	95.2	0.95
AUC	97.4	0.72

Table 8 shows that proposed framework is robust and consistent with five-fold cross validation. Figure 11 displays the cross validation results that are used to measure consistency and reliability of the model.

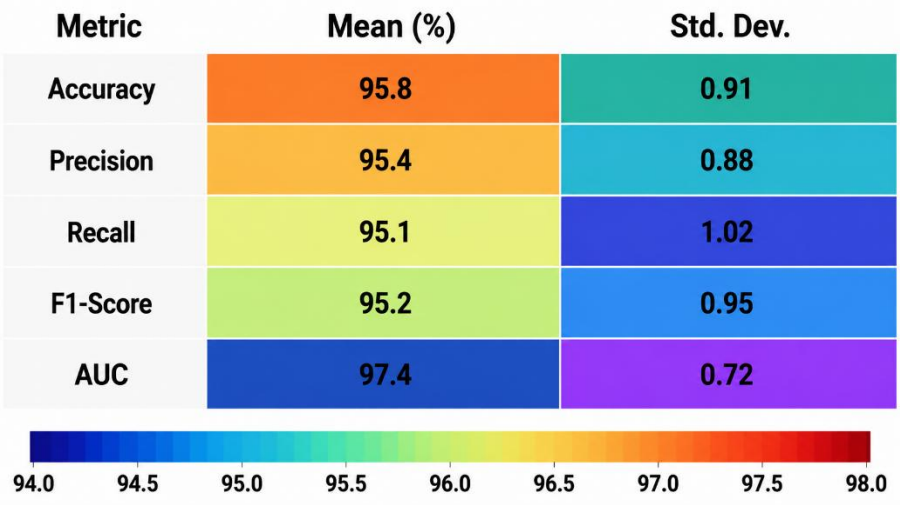

Figure 11. Visualization of Five-Fold Cross-Validation Performance Metrics (Mean and Standard Deviation)

The high mean values of accuracy (95.8%), precision (95.4%), recall (95.1%), F1 score (95.2%) and AUC (97.4%) in combination with low standard deviation values demonstrate the stability of the results, good generalization capability and reliability of the system in a variety of indoor navigation scenarios and real-world conditions.

Table 9. Statistical Significance Analysis

Comparison	p-value
Proposed vs YOLOv8	0.003
Proposed vs MiDaS	0.001
Proposed vs RRT*	<0.001

Table 9 shows that the proposed framework can obtain a statistically significant improvement compared with YOLOv8, MiDaS and RRT*. All p values are < 0.05 indicating that the performance gains are not likely to be a random phenomenon.

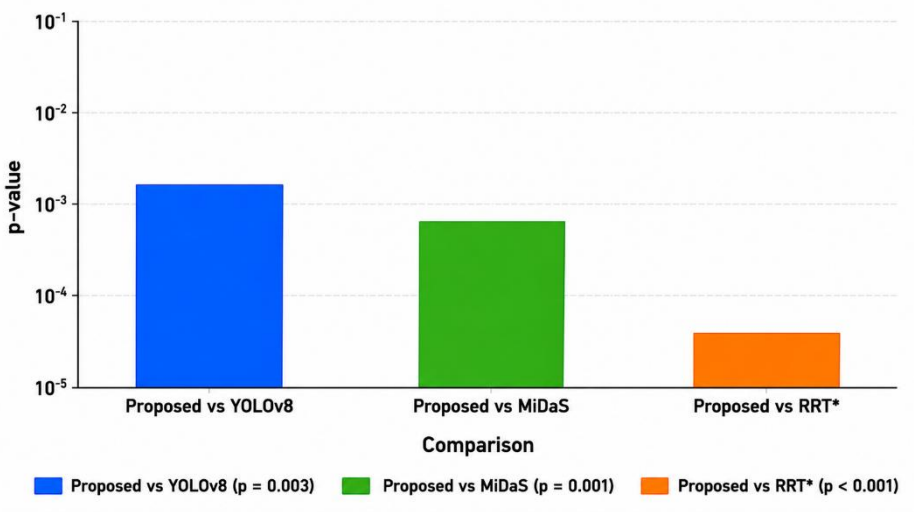

Figure 12. Statistical Significance Analysis of the Proposed Framework Compared with YOLOv8, MiDaS, and RRT

The statistical significance between proposed model and baselines is shown in Figure 12. The highest significant is found when compared to RRT*, indicating the effectiveness, robustness and reliability of the proposed method in the indoor navigation tasks.

7. Real-Time Indoor Navigation and Assistive Guidance Case Study

The proposed framework can be implemented in the real world in various indoor environments such as multi-floor buildings, shopping malls, hospitals and offices, as shown in the case study. These can include static items like walls, furniture and stairs, as well as dynamic items like people in motion. It is capable of working well in any narrow corridor, congested environment and different lighting situations by constantly analysing multimodal stimuli. The ability to accurately detect and estimate the depth of objects facilitates accurate identification of the obstacles and safe navigation paths. The framework makes it easier to interact with the system with built-in speech and haptic feedback. Speech guidance gives clear instructions on direction, distance and warnings and haptic signals give indications of the proximity and urgency of an obstacle. This multimodal approach helps to enhance understanding, minimize cognitive strain, and navigate even in challenging situations like noisy environments.

8. Conclusion and Future Work

The study offers an Explainable Multimodal Deep Transfer Learning Framework (EMDTLF) for real-time indoor navigation and assistive guidance of visually impaired person. The framework seamlessly incorporates transfer learning-based object detection, monocular depth estimation, multimodal sensor fusion, intelligent path planning and explainable AI. Experimental results show that they achieve better performance than the current methods in terms of object detection accuracy, depth estimation accuracy, navigation accuracy and safety of the user. By adding explainability based on SHAP, transparency and trust are boosted, while multimodal speech-haptic guidance greatly aids the user's understanding, satisfaction, and efficiency of navigation. The suggested system offers a real, interpretable and reliable solution for improving the independent mobility in complex indoor environments. Future work will be aimed at further enhancing the scalability of the system and its deployment in the real world by incorporating lightweight edge computing models for faster inference and energy-saving technologies. Combining high-level multimodal data, a variety of different environments, and user behaviours will further improve robustness. Moreover, the adaptable navigation preferences based on reinforcement learning can benefit in enhancing user-specific navigation preferences. Future enhancements could incorporate augmented reality support, support languages and integrate into smart environments with IoT capabilities to create more context-aware, intelligent and seamless indoor navigation experiences.

References

1. Ganesan, J., Azar, A. T., Alsenan, S., Kamal, N. A., Qureshi, B., & Hassanien, A. E. (2022). Deep Learning Reader for Visually Impaired. Electronics, 11(20), 3335. https://doi.org/10.3390/electronics11203335
2. Plikynas, D., Žvironas, A., Budrionis, A., & Gudauskis, M. (2020). Indoor Navigation Systems for Visually Impaired Persons: Mapping the Features of Existing Technologies to User Needs. Sensors, 20(3), 636. https://doi.org/10.3390/s20030636
3. Lin, H.-Y., Fan, Y.-H., & Chang, C.-C. (2026). Multimodal Navigation System for Visually Impaired Users Using Environmental Perception and Vision-Language Models. Sensors, 26(10), 3045. https://doi.org/10.3390/s26103045
4. Salman Shah, S., Imran, A., Saad-Ur-Rehman, Arif, A., Khan, K., Arsalan, M., Manzoor, S., & Sirewal, G. J. (2026). Vision-Based Smart Wearable Assistive Navigation System Using Deep Learning for Visually Impaired People. Automation, 7(2), 41. https://doi.org/10.3390/automation7020041
5. Kuriakose, B., Shrestha, R., & Sandnes, F. E. (2020). Multimodal Navigation Systems for Users with Visual Impairments—A Review and Analysis. Multimodal Technologies and Interaction, 4(4), 73. https://doi.org/10.3390/mti4040073
6. Plikynas, D., Indriulionis, A., Laukaitis, A., & Sakalauskas, L. (2022). Indoor-Guided Navigation for People Who Are Blind: Crowdsourcing for Route Mapping and Assistance. Applied Sciences, 12(1), 523. https://doi.org/10.3390/app12010523
7. Mahida, P., Shahrestani, S., & Cheung, H. (2020). Deep Learning-Based Positioning of Visually Impaired People in Indoor Environments. Sensors, 20(21), 6238. https://doi.org/10.3390/s20216238
8. Said, Y., Atri, M., Albahar, M. A., Ben Atitallah, A., & Alsariera, Y. A. (2023). Obstacle Detection System for Navigation Assistance of Visually Impaired People Based on Deep Learning Techniques. Sensors, 23(11), 5262. https://doi.org/10.3390/s23115262
9. Joshi, R. C., Yadav, S., Dutta, M. K., & Travieso-Gonzalez, C. M. (2020). Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People. Entropy, 22(9), 941. https://doi.org/10.3390/e22090941

10. Okolo, G. I., Althobaiti, T., & Ramzan, N. (2024). Assistive Systems for Visually Impaired Persons: Challenges and Opportunities for Navigation Assistance. Sensors, 24(11), 3572. https://doi.org/10.3390/s24113572
11. Guerrero, L. A., Vasquez, F., & Ochoa, S. F. (2012). An Indoor Navigation System for the Visually Impaired. Sensors, 12(6), 8236-8258. https://doi.org/10.3390/s120608236
12. Romeo, K., Pissaloux, E., Gay, S. L., Truong, N.-T., & Djoussouf, L. (2022). The MAPS: Toward a Novel Mobility Assistance System for Visually Impaired People. Sensors, 22(9), 3316. https://doi.org/10.3390/s22093316
13. Xu, J., Wang, C., Li, Y., Huang, X., Zhao, M., Shen, Z., Liu, Y., Wan, Y., Sun, F., Zhang, J., & Xu, S. (2025). Multimodal Navigation and Virtual Companion System: A Wearable Device Assisting Blind People in Independent Travel. Sensors, 25(13), 4223. https://doi.org/10.3390/s25134223
14. Darwish, S. M., Salah, M. A., & Elzoghabi, A. A. (2023). Identifying Indoor Objects Using Neutrosophic Reasoning for Mobility Assisting Visually Impaired People. Applied Sciences, 13(4), 2150. https://doi.org/10.3390/app13042150
15. Hu, W., Wang, K., Yang, K., Cheng, R., Ye, Y., Sun, L., & Xu, Z. (2020). A Comparative Study in Real-Time Scene Sonification for Visually Impaired People. Sensors, 20(11), 3222. https://doi.org/10.3390/s20113222
16. Ngo, H.-H., Le, H. L., & Lin, F.-C. (2025). Deep-Learning-Based Cognitive Assistance Embedded Systems for People with Visual Impairment. Applied Sciences, 15(11), 5887. https://doi.org/10.3390/app15115887
17. Bibbò, L., Bramanti, A., Sharma, J., & Cotroneo, F. (2024). AR Platform for Indoor Navigation: New Potential Approach Extensible to Older People with Cognitive Impairment. BioMedInformatics, 4(3), 1589-1619. https://doi.org/10.3390/biomedinformatics4030087
18. Ko, E., & Kim, E. Y. (2017). A Vision-Based Wayfinding System for Visually Impaired People Using Situation Awareness and Activity-Based Instructions. Sensors, 17(8), 1882. https://doi.org/10.3390/s17081882
19. Lupu, R.-G., Mitruț, O., Stan, A., Ungureanu, F., Kalimeri, K., & Moldoveanu, A. (2020). Cognitive and Affective Assessment of Navigation and Mobility Tasks for the Visually Impaired via Electroencephalography and Behavioral Signals. Sensors, 20(20), 5821. https://doi.org/10.3390/s20205821
20. Reyes Leiva, K. M., Jaén-Vargas, M., Codina, B., & Serrano Olmedo, J. J. (2021). Inertial Measurement Unit Sensors in Assistive Technologies for Visually Impaired People, a Review. Sensors, 21(14), 4767. https://doi.org/10.3390/s21144767
21. Beltrán-Iza, E. A., Noroña-Meza, C. O., Robayo-Nieto, A. A., Padilla, O., & Toulkeridis, T. (2022). Creation of a Mobile Application for Navigation for a Potential Use of People with Visual Impairment Exercising the NTRIP Protocol. Sustainability, 14(24), 17027. https://doi.org/10.3390/su142417027