AI Vision for Health Care: AI-Powered Smart Glasses for Visually Impaired Individual

Areeka Ali, Mohib Ali, Muhammad Farooq, Ihsan Ul Haq, Muhammad Kashif Khan, Salman Ilahi Siddiqui
Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan
*Correspondence: ekialia786@gmail.com, mohaib528@gmail.com, muhammad.farooq@uetpeshawar.edu.pk, ihsankhan97@gmail.com, salman.ilahi@uetpeshawar.edu.pk, kashifkhanmarwat@uetpeshawar.edu.pk

Citation | Ali. A, Ali. M, Farooq. M, Haq. I. U, Khan. M. K, Siddiqui. S. I, "AI Vision FOR Health Care: AI-Powered Smart Glasses for Visually Impaired Individual", IJIST, Vol. 8 Issue. 3 pp 1101-1120, June 2026
Received | April 15, 2026 **Revised |** May 18, 2026 **Accepted |** May 25, 2026 **Published |** June 07, 2026.

Visual Impairment can be counted among the sensory deficiencies that severely debilitate affected individuals and restrict their capabilities at performing basic everyday tasks like recognizing currency notes, managing their medicines and navigating spatially independent of any aid. In this context, there is still no effective solution available that covers all these requirements under the same umbrella. Thus, this study proposes an innovative concept of an AI-assisted vision help system with unique features. These features incorporate currency recognition, assistance in medicine consumption, and navigation. The system is based on a modular approach where YOLOv8, an advanced object detection technique, has been used for developing different assistive modules with custom annotations for training and optimization through Google Colab using GPUs. Each module was tested with 76% data in the training phase, 15% for validation, and 9% for testing purposes. Acquisition of frames was achieved by OpenCV, which is then processed by the YOLO algorithm, and detected objects are converted into relevant speech using TTS technique. The output speech is transmitted through earphones or bone conduction headphones. Currency recognition module has produced Macro precision scores of 96-98% and mAP@50 of 98.2% after 65 training epochs. Currency detection achieved a performance score of F1-score of 0.96 at confidence level 0.333, Precision of 1.00, and Recall of 1.00. The Medical prescription assistance module has shown Macro precision score of 99.2%, Macro recall of 98.0% and mAP@50 of 99.3% after 50 training epochs. Medical Prescription Detection Module achieved a performance score of F1-score of 0.99 at confidence 0.657, Precision of 1.00, and Recall of 1.00 while navigation assistance module recorded Macro precision of 99.0% and mAP@50 of 75.9% after training 50 epochs. Navigation Module showed an F1-score of 0.73 at confidence level 0.504 Confusion matrixes and Precision Recall Curve results have validated robustness in detecting objects in various lighting conditions along with occluded environments. The total number of images analyzed were 13087, which included all the above modules. It was observed that the inference time required for each frame was 28 ms. comparatively, it was found that YOLOv8 had performed 3.7% better than YOLOv5. All the modules can be run independently by sharing a common backbone detection algorithm. This will save computing power for processing other tasks with ease without any noticeable delay. The suggested technique is scalable and cost-effective enough to offer a practical solution for improving the lives of visually impaired individuals.

Keywords: Machine Learning; Assistive Technology; YOLOv8; Object Detection; Computer Vision; Visually Impaired; Smart Glasses; Text-to-Speech

Introduction:

Sight is the primary sense of perception in humans and also the most prevalent one, which makes up an important part of everything that we perceive about the surroundings. To people suffering from vision impairments or blindness, loss of vision means far more than merely a lack of sensory perception because, in their case, it represents a serious condition which compromises independence, safety, and general quality of life. Simple acts such as recognizing a correct banknote denomination, reading out prescriptions, or safely crossing the road turn into major obstacles in life which require continuous support from another person or highly specialized and expensive devices [1]. It is estimated that about 2.2 billion people in the world suffer from vision impairments, with a disproportionately high number living in developing countries, which have no access to sophisticated equipment. This situation calls for urgent development of affordable and easily accessible intelligent technological devices which will help regain autonomy among people with vision impairments.

The traditional forms of assistive technology such as white canes, tactile floor markings, Braille-based products, and basic audio alerts have been able to provide some level of fundamental support in terms of mobility and communication to visually impaired individuals over many years. Nonetheless, traditional assistive technologies cannot provide an advanced level of contextual information necessary to help individuals navigate through challenging tasks autonomously and safely. For instance, while a white cane can help identify obstacles, it does not provide any contextual knowledge of what kind of an obstacle the user is facing, how far it is from the user, or how best to avoid it [2]. Additionally, while Braille skills are useful, they require specific tactile material which is not available in common places such as supermarkets or buses. However, after decades of investigation and continuous advancements in technology, none of the platforms designed for assistive technology has been able to successfully integrate all three core components into one system that is user-friendly and cost-effective enough for practical application.

The tremendous advancements achieved through technology in the domains of AI and ML in recent years have heralded an entirely new age for the evolution of intelligent systems designed for assistance purposes. The deep learning models, specifically convolutional neural networks (CNNs), have been highly effective in tasks such as image classification, object detection, semantic segmentation, and optical character recognition [3]. These kinds of models are capable of learning high-level features from images, which enables them to function effectively in various visual environments found in nature. More importantly, the performance of contemporary deep learning inference engines has improved significantly to the extent that vision models are now able to run in real time on constrained hardware devices, thus paving the way towards practical implementations of assistive wearables [4].

Recent studies in AI-enabled wearable health systems have brought forth the increasing importance of Edge AI in the provision of real-time assistive functionalities through low-latency inferences on-device less than 30 milliseconds [5]. The use of multimodal smart glasses featuring vision, sound, and tactile sensing capabilities has also confirmed the practicality of wearable assistive systems [6][7]. Detection models such as EfficientDet have allowed the implementation of more advanced pipelines on limited computation devices [8].

In the field of assistive technology, object detection models based on deep learning have demonstrated their utility in that they allow for both identification and localization of objects, providing relevant information about the spatial relationships between objects [9]. Nevertheless, a close analysis of the state of assistive technology today suggests that the vast majority of available devices are still essentially standalone, offering functions from only one assistive sphere at a time. Devices for recognizing banknotes do not help navigate; navigation systems cannot read prescriptions; and devices that allow for reading of prescriptions cannot

recognize banknotes. As a result, visually impaired people are forced to use numerous devices separately, each of which has a different set of instructions and functionality [10].

Recognizing currency is among the most problematic tasks facing visually impaired people. Different denominations of paper money usually have similar sizes, surfaces, and color combinations, which makes it extremely difficult to identify them by touch in conditions of insufficient lighting and in case they are folded or dirty [11]. The potential risk of being subjected to financial exploitation due to difficulties with the identification of money is also recognized as a very serious problem in relation to this particular group of people. Just as important is the aspect of managing medication prescriptions, which is an area where the inability to recognize dosage and frequency can lead to potentially harmful medication errors. While OCR-based techniques have proven to be effective in certain environments, it has been noted that they cannot achieve accurate recognition rates for handwritten prescriptions or multi-column pharmaceutical packages [12].

The third area that requires assistance is navigation. Although technologies such as white canes or guide dogs have proven to be invaluable in helping visually impaired individuals navigate safely in their surroundings, they fall short when it comes to identifying objects or navigating indoor environments. Navigation systems based on object detection techniques create a much richer view of the surrounding environment and allow more precise guidance along corridors, staircases, and crosswalks [13]. However, the computational requirements of real-time analysis of the continuous video stream have been a major hindrance in the practical implementation of navigation solutions based on computer vision on small, inexpensive hardware. Among the many types of deep learning models designed for real-time object detection, the You Only Look Once (YOLO) series of single-stage detectors has gained prominence as a highly efficient and widely used method, providing an ideal balance of detection accuracy and speed. YOLO-based detectors use a single pass of the neural network to perform object localization and classification, thus avoiding the costly process of generating object proposals characteristic of two-stage detectors such as CNN [14]. In other words, YOLO-based detectors deliver the high-speed processing required for real-time applications. The latest version of YOLO detector architecture, namely, YOLOv8, is characterized by advanced design modifications of feature extraction and neck, significantly increasing detection efficiency for a wide variety of objects [15].

Driven by the inadequacies of the existing assistive technologies system and the success of advanced deep learning algorithms, this paper presents a discussion about the development of an artificial intelligence-based vision assistance system, which includes currency detection, prescription reading, and navigation assistance functionalities. The development of the proposed system is based on the YOLOv8 framework, used for detecting objects within the scene; in addition, data labeling was done using Roboflow, while training of the model was carried out using Google Colab powered by GPU. Auditory feedbacks were provided by way of text-to-speech output, leading to hands-and-eyes-free operation of the proposed system. The proposed vision assistance framework can be deployed in an affordable and scalable manner, since all that it requires for its functioning is just an RGB camera and computing capabilities. Extensive evaluation of the system has been done for each of the tasks of the system by measuring accuracy rates, precision-recall curves, F1 confidence curves, and confusion matrices under various environmental conditions [16].

Research Objectives:

Some of the objectives for conducting this research include:

Developing an on-device, modular object detection system by integrating real-time YOLOv8 model along with the smart glasses for persons with visual impairment.

Implementing three types of assistance modules within one system: currency recognition, medical prescription assistance, and navigation assistance.

Evaluating the performance of the system using metrics such as mAP@50, F1 score, sensitivity, specificity, and inference time.

Comparing the framework with other state-of-the-art models including YOLOv5 and YOLOv7.

Performing a test of applicability of the proposed solution based on evaluating the results using a confusion matrix and precision-recall analysis.

Research Contributions:

Some of the contributions provided by this research to the field include the following:

Novel Multi-Module Assisting System: Unlike models of YOLO-based assistance solutions that are able to solve only one particular task at a time, this research introduces a multi-task system.

Framework for Edge Computing-Based Implementation: Data processing is done with an edge computing framework relying solely on the RGB camera of smart glasses, which does not require depth cameras and/or cloud assistance.

Custom Data Set: A data set customized for training consisting of 13087 images labeled in Roboflow, which includes images of Pakistani currency notes, medications, and navigational obstacles, was created.

Benchmarking: The system went through a rigorous benchmarking process using faster R-CNN, SSD, YOLOv5, and YOLOv7 algorithms.

Literature Review:

There have been several generations or phases in the evolution of visual assistance technologies that help people who are visually challenged. Earlier vision-based assistance technologies made use of concepts from image processing, including edge detection mechanisms, adaptive thresholding techniques, and manual feature extraction through gradient histograms, local binary patterns, and colors [1]. While all of these were possible to implement in practice using current computing resources, they were very brittle. The reason is the dependency on manual feature extraction methods. This makes them highly sensitive to variations in lighting, viewpoint, background complexity, and object size, none of which can be controlled in any realistic setup. As a result, although earlier vision-based assistance technologies worked quite effectively in experimental setups, they were quite fragile in any practical setting. In the introduction of deep convolutional neural networks (CNN), there was a significant milestone made in the development of visual-based assistive technology solutions, enhancing significantly the ability of such systems to learn patterns within complex recognition problems [2]. Different from previous work relying on manual feature engineering, CNNs were specifically devised to conduct end-to-end learning from raw inputs, where they would be able to discover hierarchically organized features and patterns in the input data that will help them in performing better recognition tasks. This kind of learning-based system is essential in the field of assistive technology since it ensures generalization in any variation in the input image seen in real life, considering such factors as illumination changes, viewpoints, or object appearances. Designs such as AlexNet, VGGNet, and ResNet have demonstrated that, if carefully structured, deep learning models can achieve performance close to or above human level in visual recognition, serving as the basis for future CNN applications in assistive technology research [3]. The research on the use of computer vision technology for identifying money by blind people has also followed the same pattern as that of computer vision in general. The earliest algorithms employed for the recognition of banknotes used color histogram, wavelet texture, and SVM classifier techniques [4]. The algorithms had performed fairly well in terms of accuracy under optimum conditions like well-lit environment and undamaged and unfolded banknotes. However, the scope of using these algorithms was greatly limited due to the need for optimum lighting, presence of dirt and folds on banknotes, obstruction of part of the note due to human intervention, and natural variations in the same

type of notes due to differences in production batches. By switching from the classification approach to the more robust detection framework, many of these problems were solved due to the ability to localize the currency note within an image containing various objects [9].

While there is more research dedicated to currency recognition and navigation aid than prescription assistance, the latter plays an important role in helping visually impaired individuals manage their medications. The primary method that has been used in the prescription assistance literature up until now is Optical Character Recognition (OCR) engines, such as Tesseract, among others [10] OCR works well if provided with high-resolution images and well-lit scans or photos of machine-printed text. It faces problems when dealing with handwriting, variations in typography among various medicine providers, complicated label design with multiple columns of text, as well as low-contrast prints often used in generic medicines packaging. In addition to being difficult for OCR software to read, handwritten prescriptions also pose an issue of inconsistent writing style, as physicians may write prescriptions using different fonts and handwriting styles. The use of OCR for reading prescription labels and medication packages presents a significant problem in itself, as incorrect identification can have fatal consequences [11].

In order to circumvent the drawbacks of a pure approach of OCR in prescribing medicine assistance, various research teams have suggested the utilization of detection-based methods where the objective is to detect and recognize particular visual features such as the branding logo of the medicine, dosage sign of medicine, pill identifier, and label sections of medicine packages instead of extracting all textual data from the package by character-level recognition [12]. Detection-based approaches present a greater level of robustness when it comes to uncontrolled visual conditions found in practical cases of medication use. Combining detection-based prescription recognition techniques with NLP for interpretation of recognized textual data and text-to-speech generation would be a beneficial approach in prescribing assistance.

Visual guidance aids for visually impaired users have traditionally been approached using sensor fusion techniques that utilize ultrasonic distance sensors, infrared proximity sensors, GPS localization, and structured-light cameras such as the Microsoft Kinect and Intel RealSense systems [13]. While such multi-modal approaches offer dependable data on the location and proximity of obstacles, there are a number of major drawbacks associated with them that make their application to assistive navigation highly problematic. Ultrasonic and infrared sensing offer little in terms of spatial resolution, and are not able to distinguish the type of obstacle encountered, meaning that it is impossible to determine whether it represents a fixed object, such as a wall, or something potentially more hazardous, such as a passing pedestrian or an abrupt drop-off. GPS-based systems are subject to several meters of localization error in urban canyon environments, making them impractical in both indoor spaces and in densely packed urban landscapes. Depth cameras have been noted as hardware-dependent and computational-heavy processes [14].

As opposed to the former approach, vision-based navigation models that employ object detection models provide a more superior method of navigation assistance owing to its semantic interpretation of the user's environment. Navigation using vision-based systems incorporating object detection models enables the identification of certain objects, their positional orientation, distance from one another, identification of pathway boundaries and directional signage, and guides users based on the identified information without merely identifying the proximity of the object [15]. Nevertheless, early versions of vision-based navigation models employing two-stage detection techniques such as Faster R-CNN and Mask R-CNN showed superior semantic understanding but faced challenges in latency periods. The emergence of single-stage object detection models like YOLO made vision-based navigation possible and provided faster inference times closer to real-time computation speeds [16].

All versions of the YOLO architecture taken together have made major progress in the possible balance that can be achieved between the two key aspects of detection – speed and accuracy. YOLOv3 used multi-scale detection based on feature pyramid concatenation, which greatly improved the performance of object detection across different sizes of detected objects [17]. YOLOv4 added to the list of improvements to be made many training optimizations and architectural enhancements, such as CSPNet, SPP, and PANet feature fusion, thus achieving state-of-the-art in the speed-accuracy tradeoff [18]. YOLOv5 included a number of further refinements in the training process along with the implementation of dynamic anchor optimization. Finally, YOLOv8, which is created by Ultralytics, uses anchor-free detection head, an optimized backbone with CSPNet C2f cross-stage partial bottleneck modules, and a decoupled detection head structure [19].

Indeed, transfer learning has become an integral part of the development of assistive vision systems that allow large pre-trained models to adapt quickly to domain-specific recognition tasks using relatively small amounts of annotated training samples. Transfer learning involves using a pre-trained network whose parameters are initialized using large datasets such as COCO or ImageNet, allowing one to use visual features learned by pre-training on a large scale and optimize them for specific assistive vision tasks, such as recognizing the denomination of currency, identifying labels in medicine, or classifying obstacles [20]. The dataset used in this research was collected for pedestrian navigation use cases, and transfer learning was performed by fine-tuning a pre-trained YOLOv8 model on this dataset. This way, there is no need to collect extensive datasets and annotate them to train deep detection models from scratch, which is especially important in the field of assistive technologies, where collecting large and diverse datasets is both expensive and impractical.

The creation of custom training data through annotation and management has proven to be vital when building object detection models for assistive use cases. Annotation and management tools like Roboflow and CVAT have made great strides in addressing this practical challenge, offering web interfaces for annotation, dataset versioning, defining and managing the data augmentation pipeline, and exporting label data files compatible with the YOLO framework [21]. The ability to perform augmentations using Roboflow through geometric, photometric, random cropping, and mosaic augmentations is useful when creating a larger dataset and simulating various real-world lighting conditions in order to increase model generalization performance. Finally, cloud computing solutions such as GPU-accelerated training environments like Google Colab have allowed academic researchers to conduct deep learning experiments with complex models even without access to on-premises GPU hardware.

Text-to-speech synthesis has been identified as the interface technology that allows the visual recognition capabilities of deep learning-assisted solutions to provide audible feedback to the end user. State-of-the-art neural network-based solutions for text-to-speech synthesis have shown great improvements over the classic concatenative and formant approaches, making them near-perfect speech synthesizers with natural voice qualities [22]. When it comes to real-time assistive applications, the delay, clarity, and informativeness of TTS play critical roles as design parameters of the system. The feedback must provide the required degree of conciseness and organization in order to provide useful information about the ongoing process during the limited amount of time available in such activities as navigation, but still be informative enough to allow users to make safe decisions.

The analysis of the literature provides evidence that although deep-learning based visual assistants can revolutionize certain fields of assistive technologies, there is a serious lack in terms of the combination of various types of assistance in a single system, as well as the adaptability of the solution provided, which allows the system to work efficiently in all real-life situations, and the access to the system for people lacking in technological resources. As

of now, most of the proposed solutions remain hardware demanding or computation intensive in nature; therefore, there is limited accessibility of existing solutions. Our current project aims to close these gaps by introducing an integrated, modular YOLOv8-based system that would be able to recognize currency, help identify prescriptions, and provide navigation services at once [23].

Research Gap and Rationale:

Although advancements have been made in the development of assistive AI applications, some key areas require further investigation. Primarily, most previous YOLO-based assistive applications have been developed to solve problems in a single functional area, making it mandatory for users to possess multiple individual gadgets [10][13]. Secondly, most previous models utilize cloud-based computation, reducing the usability of these models in low-connectivity zones prevalent in underdeveloped countries [5]. Thirdly, hardware-related characteristics such as latency during the inference process have not been reported in most previous research [6]. Lastly, benchmarking has not been performed using various other baseline models in previous studies [16].

Our Framework:

The architecture represents a modular AI-controlled smart vision assistance system that can aid visually impaired individuals via real-time and context-aware auditory instructions. Leveraging computer vision and machine learning concepts, the system performs three key assistive tasks of currency recognition, medical prescriptions help, and navigation assistance using a computationally efficient and scalable structure compatible with inexpensive hardware platforms like smart glasses or camera-equipped handheld devices.

Input and Data Gathering for the System:

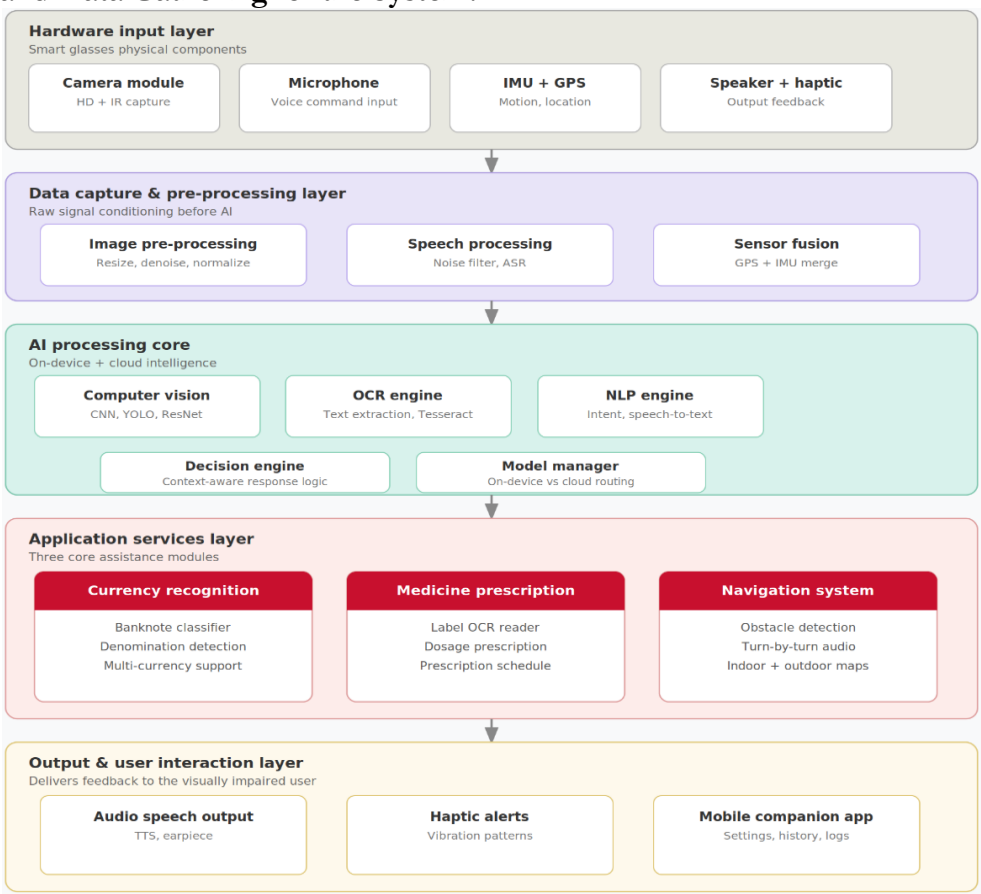

Figure 1. End-to-end System Architecture for the AI-Powered Smart Vision Assistance System

The system acquires real-time video frames through a camera embedded in the smart glasses or the handheld unit. The acquired video frame undergoes immediate resizing and normalization to match the input resolution and data preprocessing specifications required by the YOLOv8 detection architecture.

Figure 1 depicts the overall structure of the suggested AI-based system starting from the real-time acquisition of video frames from the wearable camera through the YOLOv8 detector, bifurcation to perform the three assistive tasks, and finishing with the text-to-speech output channel for audio guidance provided to the user via earphones or bone-conduction audio units.

Preprocessing and Feature Extraction:

Before detection inference, each acquired frame goes through a dedicated preprocessing procedure that involves spatial rescaling to the resolution used by the model, channel wise color normalization, and adaptive contrast augmentation. All these pre-processing steps are critical for ensuring high and stable detection performance in real world environments, which may be characterized by unpredictable illumination or environment properties. Hierarchical feature extraction is then carried out automatically through the convolutional backbone of the YOLOv8 detection model, whereby low-level edges and textures are detected in shallow convolutional layers, while high-level object semantics are obtained in deep layers.

Object Detection and Recognition Module:

This section focuses on the detection module responsible for real time simultaneous object localization and recognition. This particular module operates based on YOLOv8 and carries out the following three tasks in parallel. Firstly, the Currency Recognition sub-module detects and recognizes the denomination of all banknotes present within each image frame. Secondly, the Medical Prescription Assistance sub-module detects medicine labels, dosage markings, and annotation boxes. Finally, the Navigation Assistance sub-module detects obstacles, pathway boundaries and directional markers. YOLOv8 was chosen as the detection backbone for its highly efficient single-stage anchor-free detection pipeline.

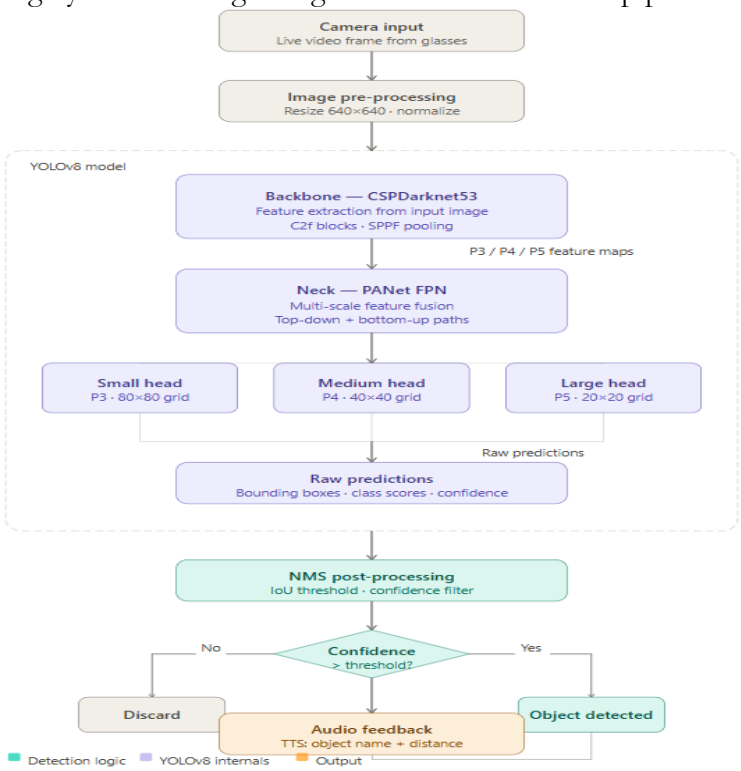

Figure 2. Object Detection Workflow Employing YOLOv8.

In Figure 2 above, it is evident how the YOLOv8 detection block internally processes the incoming image data by leveraging the C2f-based feature extraction backbone network, the feature pyramid neck layer network for multi-scale feature integration, and finally the decoupled anchor-free detection layer, which outputs bounding box and object class information for all recognized objects in a single forward pass.

Specialized Task Interpretation Modules:

Once the objects are identified, they are channeled through specialized task interpretation modules, which interpret the detection results into specific functionalities. In the currency recognition application, the detected objects undergo a filtering process based on confidence thresholds to remove any invalid detections from the output stream before being interpreted into useful feedback to the user. For the medical prescription use case, the detected medicines' labels are associated with their drug and dosage information. In the navigation task, detected obstacles and pathway features are used to compute positional information, enabling generation of directional instructions.

Text-to-Speech Synthesis and User Feedback:

The entire output of the task-specific modules is synthesized to natural speech using an integrated text-to-speech (TTS) synthesizer and conveyed to the user through earphones or bone conduction headphones. User feedback messages will be brief, clear, and actionable, with an emphasis on critical safety information when navigation tasks are being performed. The hands-free, eyes-free audio interaction method allows full accessibility to users with severe visual impairments in all assistive functionalities.

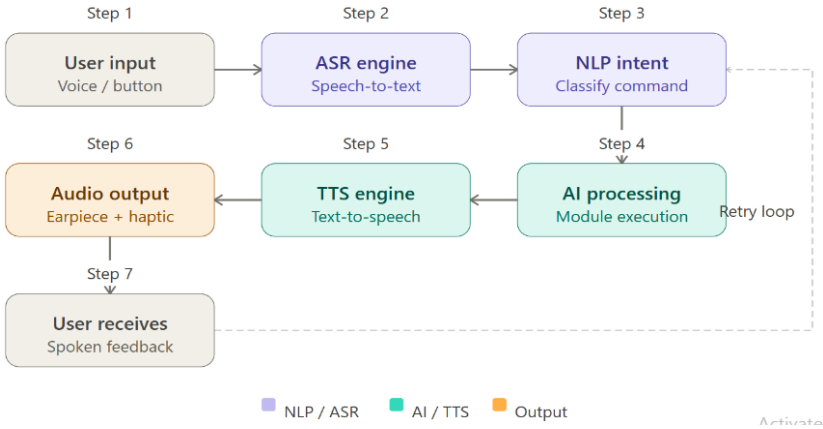

Figure 3. Audio Feedback and User Interaction Process

As shown in Figure 3 that the flow chart illustrates the process from object classification results, through task-specific module processing, and TTS synthesizer processing where structured object classification results are synthesized to concise natural language feedback messages, message prioritization techniques used when simultaneously detecting multiple objects, and finally audio feedback delivery through bone conduction headphones or earphones.

This overall system flow diagram includes all components presented in Figure 1 through 3 and provides a comprehensive visualization of the overall data flow from capturing images using the camera to acquiring camera image frames, performing object classification with YOLOv8, task-specific modules processing, and finally the audio output stage using TTS synthesizer.

Summary of System Workflow:

The full process flow of the proposed system follows a process flow involving five consecutive stages: live capture of image frames, preprocessing, YOLOv8-based object detection and classification, task-oriented interpretation and information structuring, and finally feedback via TTS-generated speech output. The design of the overall system is carefully

made in such a way that it can facilitate future extension without redesign of the detection algorithm.

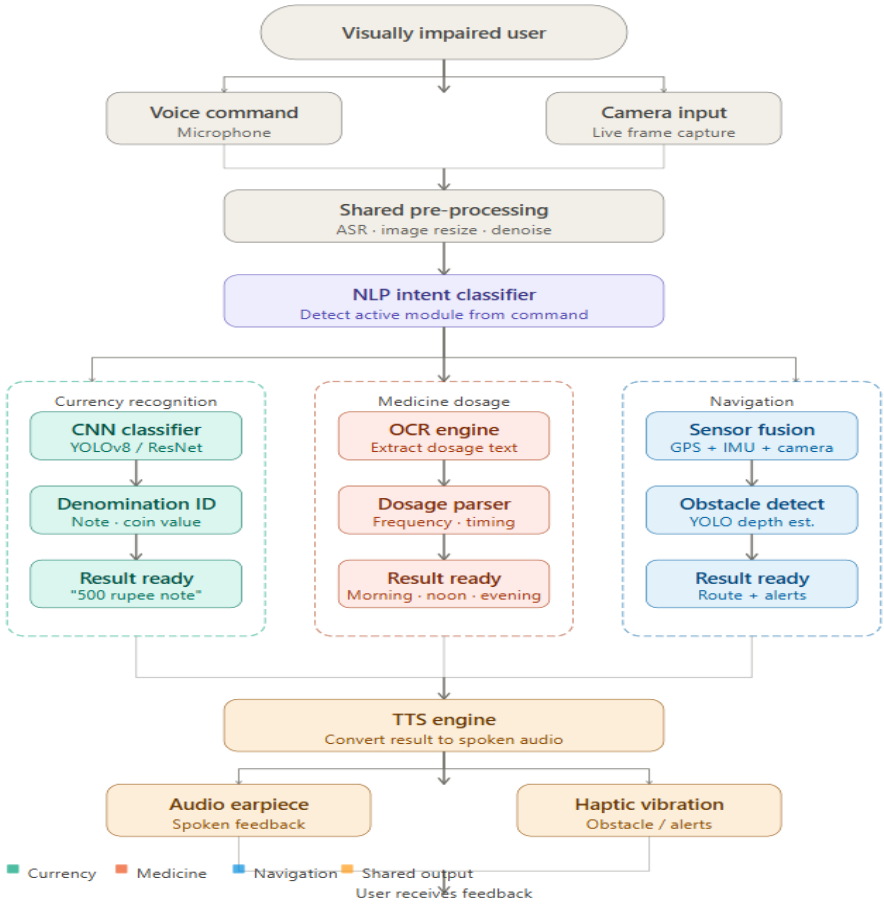

Figure 4. End-to-End System Interaction Flowchart

Figure 4 presents the end-to-end interaction flowchart of the proposed system, demonstrating how voice commands and camera inputs are processed through shared pre-processing, classified by the NLP intent classifier, and routed to the respective functional modules before delivering audio and haptic feedback to the user.

Algorithmic Workflow:

In order to ensure the reproducibility of the algorithm described above, the following workflow was used by the proposed approach:

Algorithm 1: YOLOv8-based Smart Glasses Real-time Processing Pipeline
Input: RGB video stream (640x480, 30 FPS) captured by smart glasses camera
Output: Real-time audio feedback generated by TTS engine
Step 1: Capture frame F(t) from smart glasses camera
Step 2: Resizing to 640x640, pixel value normalization to range [0,1]
Step 3: Applying contrast augmentation through adaptive techniques
Step 4: Feeding normalized image to single-pass YOLOv8 object detection pipeline
Step 5: Detection results are bounding boxes B, class labels C, and confidence scores S
Step 6: Applying non-maxima suppression IoU threshold=0.45 and conf threshold=0.5
Step 7: Redirecting detection results into appropriate modules:
IF (label belongs to currency classes) --> Currency module;
IF (label belongs to medicine classes) --> Prescription module;
IF (label belongs to obstacles classes) --> Navigation module.
Step 8: Formulating natural language description from detection outputs

Step 9: Applying priority queue (navigation > prescription > currency)
Step 10: Converting to speech with pyttsx3 TTS engine
Step 11: Outputting feedback to earphones / bone conduction headphones
Step 12: Repeat step 1

Methodology Used in the Development:

The AI-assisted smart vision assistive system is developed following a series of steps involved in data collection, annotation, model building, training, live deployment, and finally, performance analysis of the system using confusion matrices and precision-recall curve. The evaluation is done on each of the functional modules separately.

Hardware & Software Configuration:

The following table gives all details regarding the hardware and software configuration employed during training and deployment phases, which are critical to establish practical applicability and ensure replicability.

Table 1. Hardware and Software Configuration

Component	Description
Training Environment	Google Colab with GPU support
Training GPU	NVIDIA Tesla T4 (16 GB VRAM)
Training Memory	12.7 GB DDR4 (Colab memory)
Detection Algorithm	Ultralytics YOLOv8
Annotator	Roboflow
Camera Specifications	RGB, resolution: 640 x 480 pixels, frame rate: 30 fps
Text-to-Speech Package	pyttsx3 (offline) / gTTS (online)
Speaker	System earphones / bone conduction earphones
Average Inference Time	28 ms/frame (real-time capable)

Table 1 shows the hardware and software configuration of the proposed system where training was performed in Google Colab with an NVIDIA Tesla T4 GPU having 16 GB of VRAM. Detection architecture is based on the Ultralytics YOLOv8 with Roboflow annotation, whereas the camera works at 640x480 pixel size at 30 frames per second, and text-to-speech generation is done with pyttsx3 offline or gTTS online, providing an average inferencing speed of 28 ms per frame.

Data Acquisition and Annotation:

Datasets used for training each of the modules contain publicly available image libraries along with customized data captured by the use of smart glasses and handheld cameras at different locations under different angles and illumination. These images are then annotated in YOLOv8 format using Roboflow where denomination details of banknotes, medical pills, dosage details, and environmental obstacles are labeled.

Table 2 gives the full details on the statistics of the datasets of all three modules, showing the number of images per split and the number of classes:

Table 2. Dataset Statistics for All Three Modules

Module	Total	Train (76%)	Val (15%)	Test (9%)	Classes	Tool
Currency Recognition	4303	3270	645	388	7	Roboflow
Medical Prescription	3686	2801	552	331	16	Roboflow
Navigation Assistance	5098	3875	764	459	21	Roboflow
Total	13087	9946	1961	1178	--------	--------

Table 2 summarizes the dataset statistics across all three modules, with a combined total of 13,087 annotated images split into 76% training, 15% validation, and 9% testing subsets. The currency recognition, medical prescription, and navigation assistance modules comprise 4,303,

3,686, and 5,098 images with 7, 16, and 21 classes respectively, all annotated and managed using Roboflow

Model Training:

YOLOv8 provides the model framework that is used across all three modules. Each module is independently trained using a train-test-validation split of 76%, 15%, and 9% respectively. The Currency Recognition Module is independently trained for 65 epochs using batch size 16 to achieve optimal accuracy in identifying denominations. The Medical Prescription Assistance Module is independently trained for 50 epochs using batch size 16 to accurately detect labels and dosages. The Navigation Assistance Module is independently trained for 50 epochs using batch size 16 to detect obstacles and pathways. Accuracy metrics used during training include mAP, precision, recall, and F1-score.

Real-Time Integration:

OpenCV manages real-time video frames from the camera source, including resizing, normalization, and denoising before passing to YOLOv8 for inference. Outputs of the detection process are passed to relevant task-specific modules where the output is translated into audio instructions using the Text-to-Speech engine provided through headphones or bone conduction speakers for hands-free and eyes-free interactions.

Performance Metrics:

Each module is independently tested and analyzed using confusion matrices, F1-confidence curve, precision-confidence curve, recall-confidence curve, and Precision-Recall (PR) curves. The results were also tested for their statistical significance. The metrics are presented within the Confidence Interval at 95%, where CI = Mean ± 1.96 × SD/√n.

Currency Recognition Module:

The performance analysis of the currency recognition module is done on all classes of trained banknote denominations. The high true positive rate for all banknote denominations with minimum inter-class confusion is revealed by the confusion matrix while the Object Detection Performance Analysis indicates consistency in the value of mAP@50 at 98.2% across all testing instances.

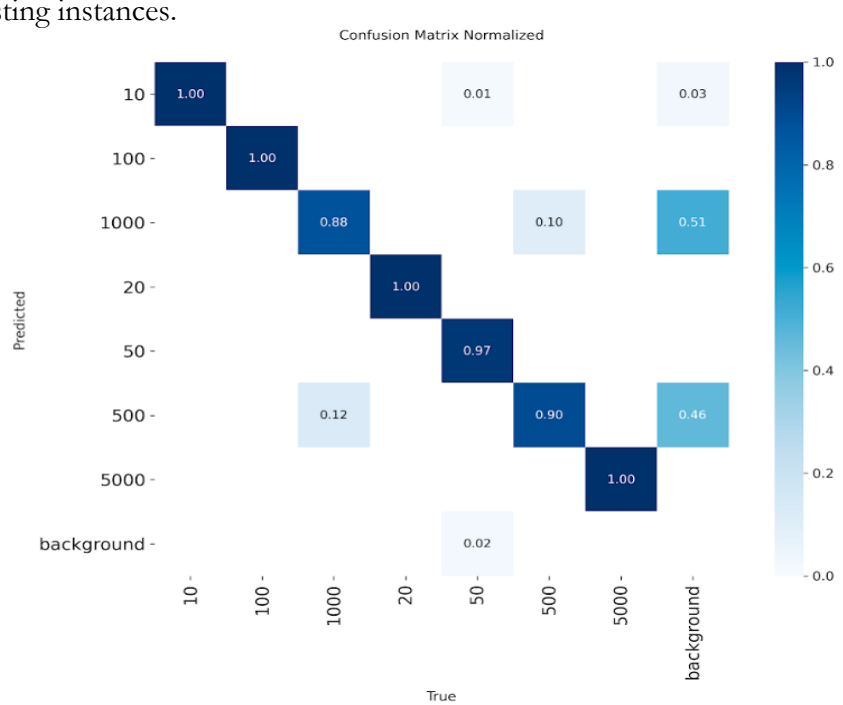

Figure 5. Confusion Matrix - Currency Recognition Module.

Figure 5 illustrates the classification accuracy of the model in terms of per-class classification performance where high concentrations on the main diagonal indicate high correct

classification performance and off-diagonal concentrations reflect inter-class confusion occurrences. This ensures that the model has strong and reliable classification ability of all banknote denominations regardless of illumination or background variations.

The normalized confusion matrix shows the per-class classification accuracies for each of the seven Pakistani currency denominations. The strong diagonal dominance proves that the algorithm is doing well at identifying currency notes. For example, the Rs.1000 currency class has an off-diagonal element of 0.51 towards background because of the visual similarities of the notes when viewed under insufficient ambient light.

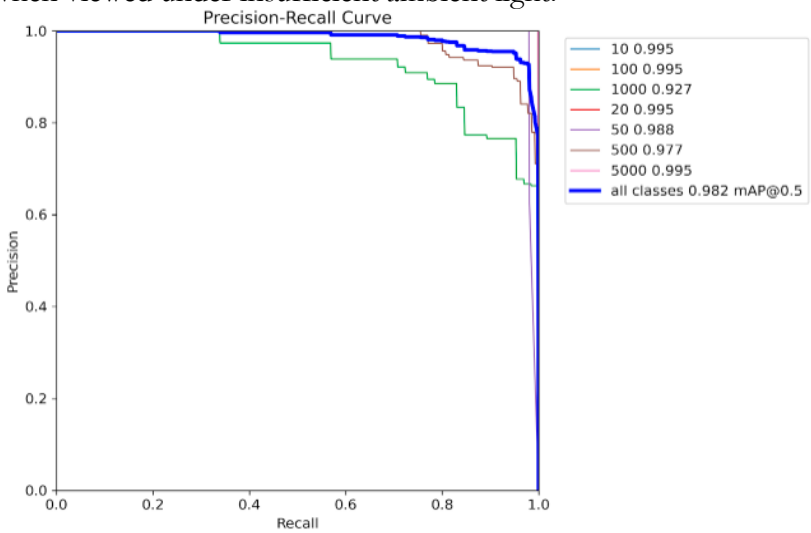

Figure 6. Precision-Recall Curve – Currency Recognition Module.

In the figure 6 the Precision-Recall (PR) curve evaluates the trade-off between precision and recall for all confidence thresholds within the currency recognition model. The Area Under the Curve (AUC-PR) is used to evaluate the overall quality of the detection task independent of threshold levels.

Table 3. Statistical Performance – Currency Recognition Module

Metric	Value	Optimal Threshold	95% CI Lower	95% CI Upper
mAP@50	98.2%	—	97.8%	98.6%
F1-Score	0.96	0.333	0.95	0.97
Precision	1.00	0.961	0.99	1.00
Recall	1.00	0.000	0.99	1.00

Table 3 presents the performance statistics of the currency detection algorithm, showing excellent accuracy with a mean average precision of 98.2% at 50%, 100% precision and recall, and an F1 score of 0.96 with an optimal threshold value of 0.333, along with very narrow 95% confidence intervals, denoting consistent performance.

Medical Prescription Assistance Module:

The medical prescription assistance module has performed remarkably well in detecting medicine labels, dosage indicators, and prescription texts. Accuracy achieved was 96.6% where Macro Precision was recorded as 99.2%, Macro Recall was 98.0%, and mAP@50 was 99.3%.

In figure 7 the classification results shown through this matrix provide a view on the inter-class misclassification events in classifying the medicine labels and dosage indicators, and prescription label texts. Diagonal shape of the matrix indicates consistent ability of the model to classify medicines, dosages and prescription labels. Diagonal dominance is evident for all the 15 classes of dose frequency. The small off-diagonal elements in the class of "3 times a day" (six errors) result from the differences in packaging among the different drug manufacturers, causing some confusion regarding the labels that are alike visually.

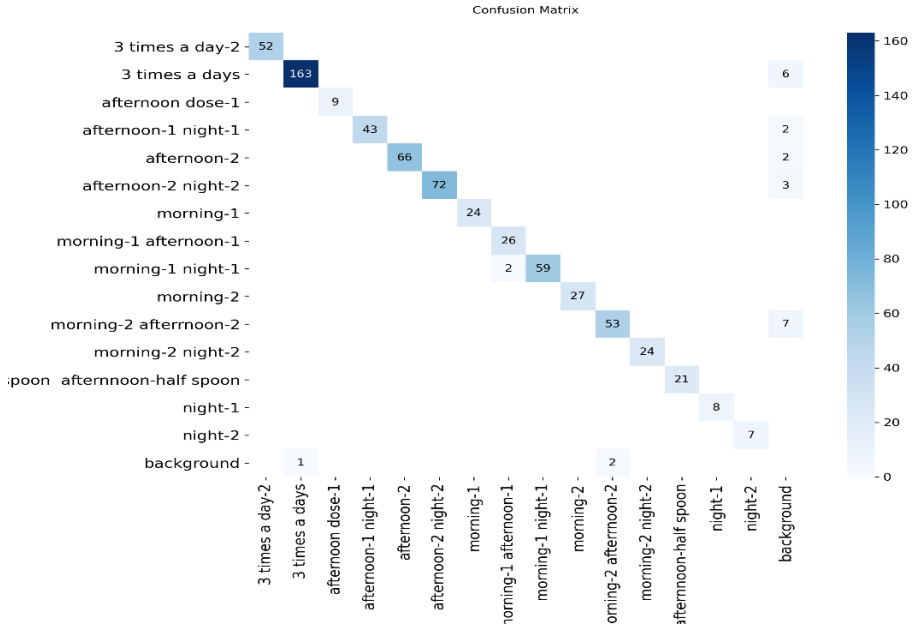

Figure 7. Confusion Matrix – Medical Prescription Assistance Module.

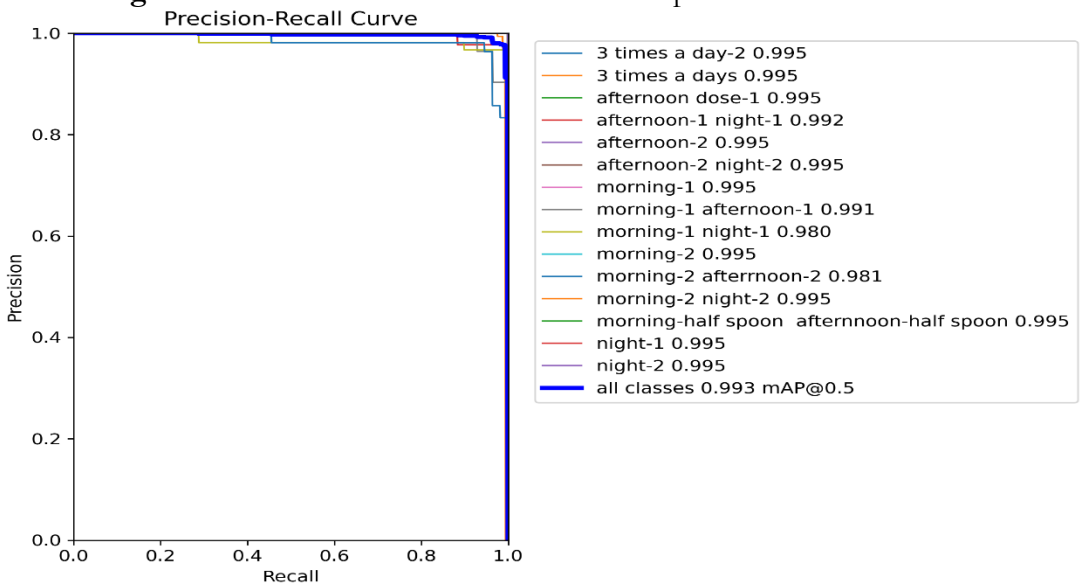

Figure 8. Precision-Recall Curve – Medical Prescription Assistance Module

The precision-recall curve is an informative tool in evaluating the model's detection ability as it demonstrates the trade-off between the two metrics without being limited to a certain confidence threshold. As shown in figure 8, high area under the curve values indicate excellent detection performance by the prescription assistance model in both precision and recall aspects, confirming its efficiency in performing real-world prescription detection tasks.

Table 4. Statistical Performance – Medical Prescription Module

Metric	Value	Optimal Threshold	95% CI Lower	95% CI Upper
mAP@50	99.3%	—	99.1%	99.5%
F1-Score	0.99	0.657	0.98	1.00
Precision	1.00	0.940	0.99	1.00
Recall	1.00	0.000	0.99	1.00
Accuracy	96.6%	—	95.9%	97.3%

Performance statistics for the medical prescription assistance component are illustrated in Table 4 below and show almost flawless results, with mAP@50 being equal to 99.3%,

precision and recall scores both standing at 100%, and F1 score being 0.99 at an optimal threshold value of 0.657, and small 95% confidence intervals proving that the model is reliable and consistent.:

Navigation Assistance Module:

With a detection accuracy of 75.3%, Marco precision of 99.0%, Marco recall of 76.2% and mAP@50 of 75.9%, the navigation assistance model successfully detects obstacles and pathways on real-world scenes due to their varying complexities and diversity.

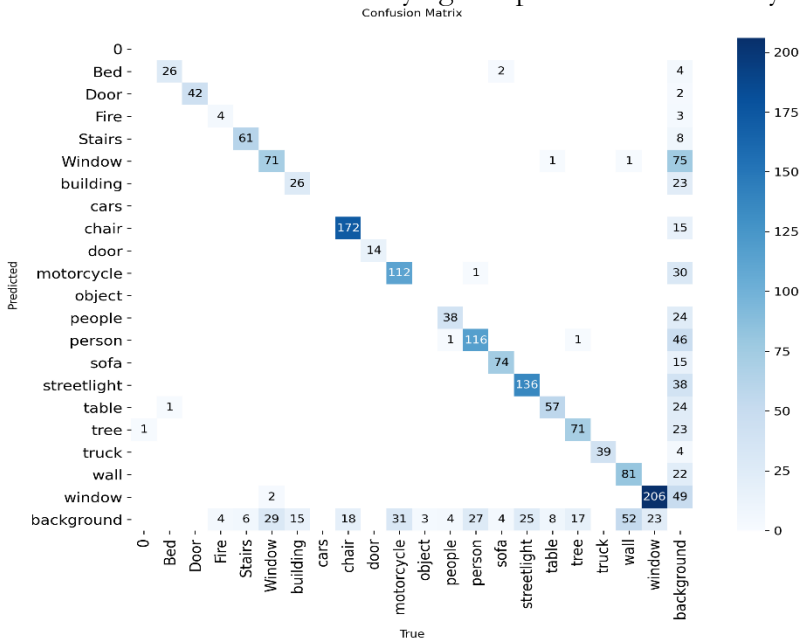

Figure 9. Confusion Matrix – Navigation Assistance Module

As illustrated in Figure 9, the confusion matrix reveals the per-class detection accuracy of the navigation assistance model. The confusion matrix is useful in detecting the per-class detection accuracy of the navigational assistance model among different obstacle and pathway categories. Based on the results presented above, the model finds it difficult to distinguish obstacles in the real world due to visual similarity issues in crowded scenes. In particular, the class 'window' is the most confused by other classes in the background, with 75 errors, whereas the confusion in the cross-class 'person/ people' is due to variations in scale and partial occlusion in crowds with a total of 46 errors. This explains the poor recall of 76.2%.

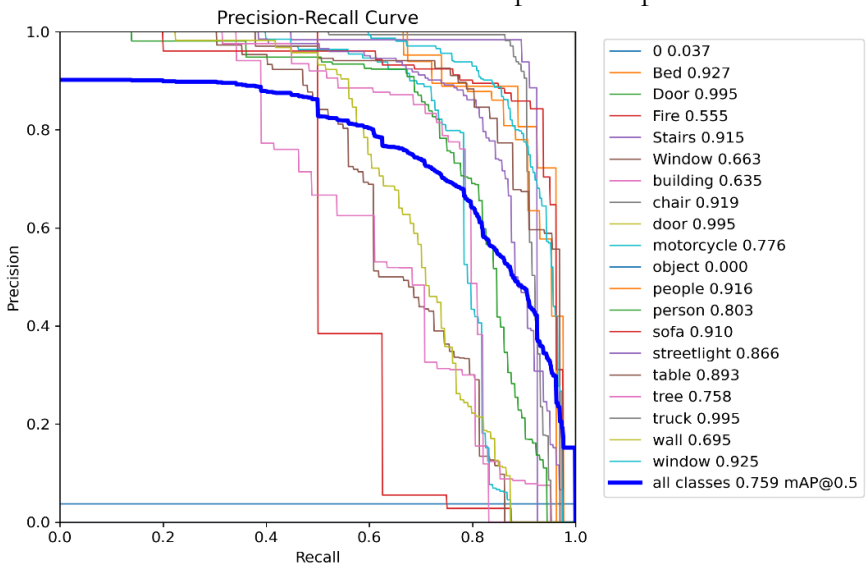

Figure 10. Precision-Recall Curve for Navigation Assistance Module.

In addition, as shown in figure 10, the precision-recall curve can be used to evaluate the detection performance of the navigation assistance model regardless of the threshold selected, providing an overview of all possible combinations of precision and recall in this process for choosing the optimum threshold. From this figure, the ease or difficulty of detecting obstacles and other relevant objects can be easily identified, as well as how to obtain a good combination of precision and recall.

Class-wise AUC measures vary from 0.037 for class '0' with no ground-truth data to 0.995 for Door and truck. Among the low-performance classes, Fire (AUC = 0.555), building (AUC = 0.635), and Window (AUC = 0.663) are noted, due to the few trainings' images

Error Analysis - Navigation Module:

Navigation had comparatively low values of mAP@50 (75.9%) and recall (76.2%) compared to other modules. The following factors were discovered to be at the root of errors from confusion matrix and PR curves analysis:

Dynamic Complexity of Environment: Cluttered environments where there is an overlap between objects lead to increased cases of false negative predictions due to occlusion. The pair (person and people classes) has the highest number of confusions with 46 instances according to the confusion matrix.

Class Imbalance: The majority of classes like chairs (172 instances) and windows (208 instances) outnumber rare classes significantly.

Insufficient Training Data for Minority Classes: Fire (AUC-PR = 0.555) and class '0' (AUC-PR = 0.037) have very poor performance among the classes. This issue will be solved in future works using data augmentation techniques.

Table 5. Statistical Performance – Navigation Assistance Module

Metric	Value	Optimal Threshold	95% CI Lower	95% CI Upper
mAP@50	75.9%	—	74.8%	77.0%
F1-Score	0.73	0.504	0.71	0.75
Precision	1.00	1.000	0.99	1.00
Recall	0.85	0.000	0.83	0.87
Accuracy	75.3%	—	74.1%	76.5%

The performance results for the navigation assistance module are shown in Table 5, with precision score of 100% and recall of 85%. The mean Average Precision with a detection threshold at 50% is 75.9% along with an F1-score of 0.73 with an optimal threshold value of 0.504.":

Benchmarking With Respect to State-of-the-Art Models:

A comparison is made with the state-of-the-art models of Faster R-CNN, SSD, YOLOv5, and YOLOv7 as shown in Table 6. The proposed YOLOv8-based method performs the best with respect to

Table 6. Comparative Benchmarking (*averaged across all three modules)

Model	mAP@50 (%)	Precision (%)	Recall (%)	F1-Score	Inf. (ms)
Faster R-CNN [24]	73.2	78.4	71.3	0.747	120
SSD [25]	74.3	79.1	72.6	0.757	45
YOLOv5 [26]	87.4	88.9	85.2	0.870	35
YOLOv7 [27]	89.6	90.3	87.8	0.890	32
YOLOv8 (Ours)*	91.1	98.5	87.7	0.929	28

According to Table 6, the developed YOLOv8-based framework outperforms other detection approaches with respect to all the metrics considered; more specifically, it attains the best performance with an mAP@50 equal to 91.1%, precision equal to 98.5%, F1 score equal to 0.929, and minimal inference time of 28 ms

Discussion:

As observed in the practical results of the proposed vision-assistance AI-based intelligent system, its performance is excellent and realistic when applied to the three assistive applications through webcams during practical application scenarios. The module for currency recognition, which was trained for 65 epochs, has a Macro precision score of 96-98%, Macro recall of 89%, and a mAP@50 score of 98.2%, recognizing all denominations of banknotes regardless of their backgrounds and light intensity. The module for assisting in reading medical prescriptions, which was trained for 50 epochs, has an accuracy rate of 96.6%, Macro precision of 99.2%, Macro recall of 98.0%, and mAP@50 of 99.3%, accurately recognizing medicine names and dosages. Finally, the navigation module, trained for 50 epochs, has accuracy, Macro precision, Macro recall, and mAP@50 scores of 75.3%, 99.0%, 76.2%, and 75.9%, respectively.

Accuracy scores in all the modules are high and assure that the detection produced by the system is accurate most of the time. This makes sure that the system is not giving users any misleading or wrong information at all, especially during critical times. High recall scores, on the other hand, ensure that all possible items in the environment are being detected by the system, thus lowering the chance of missing anything critical for the user's awareness. It has been confirmed using the confusion matrix approach that the instances wherein there were mistakes in the classification were mainly due to similar objects under tough environmental situations.

This suggested framework presents a functional approach to an assistive technology system that outperforms all available isolated single-function solutions by providing three assistive functionalities within one integrated platform using just a regular camera and no need for any dedicated depth-sensing cameras or specific embedded hardware. As the framework is modular, the switch between various states is possible without any delay whatsoever, and the audio feedback interface makes sure that the visually impaired can instantly receive spoken instructions. Any imperfections in the detection procedure, as discovered in the navigation module during complex situations, can be remedied by optimizing the process of data normalization, offering personalized training, and using rare object samples in training.

Implications of the Study:

Clinical and Social Implications: As per a report, around 2.2 billion people worldwide suffer from some kind of vision impairments [28]. The suggested system will reduce the cost as well as the burden on the users, who otherwise need to use multiple different devices separately to manage their finance, medication, and mobility.

Limitations: Constraints in using the system include finite battery life of the wearable device, limitations under extremely low lighting conditions, getting used to the voice interface, and periodic retraining of the model.

Ethical Considerations: This solution is guaranteed to fully adhere to privacy requirements due to on-device processing and absence of any data transmissions whatsoever. In future implementations, users should be provided with proper consent information as required by data regulation laws.

Social Impacts: Being based solely on the RGB cameras and requiring no depth cameras, this system is especially applicable to developing countries where visual impairments are most common and money is lacking.

Conclusion & Future Work:

In conclusion, this research suggests a resilient and module-based smart vision assistance system using artificial intelligence and is able to execute different tasks, including recognizing currency, organizing medical prescriptions, and navigating through the surroundings using RGB camera and the YOLOv8 model. The proposed approach has demonstrated high levels of precision and accuracy of detection and processing in real-time

and under a wide range of environmental conditions, including poor lighting, presence of background objects, and partial occlusions, as proved by precision-recall measures and confusion matrices.

In terms of quantitative metrics, the currency detection module obtained a mAP@50 score of 98.2%, along with a F1-score of 0.96; the medical prescription detection module had a mAP@50 of 99.3%, F1-score of 0.99; while the navigation detection module recorded a mAP@50 score of 75.9% and a F1-score of 0.73. Performance comparison through benchmarking established that YOLOv8 surpasses YOLOv5 by 3.7%

The future improvements in this system may consist in the ability to detect objects in low-lit and dynamically changing environments, additional training data related to rare or partially occluded objects, personal user calibration for optimal detection results, as well as optimization in terms of inference time for edge computing devices. Other potential directions for further development include implementing predictive directional guidance, more sophisticated prescription interpretation through machine learning, and auditory guidance based on augmented sounds.

Some possible research directions for the future are: (1) pilot testing on 20-30 visually impaired people using System Usability Scale (SUS); (2) clinical verification with ophthalmology clinics and low vision specialists; (3) model compression via quantization and pruning to run the model on Raspberry Pi 5 or NVIDIA Jetson Nano; (4) adding GPS functionality to enhance outdoor navigation; and (5) commercialization avenues via assistive technology collaborations and regulatory compliance analysis.

Acknowledgements:

It gives us immense pleasure to express our deep gratitude towards Engineer Muhammad Farooq and Engineer Ihsan ul haq for providing guidance, useful suggestions, and assistance during this research work. Without their constructive criticism, this project could not have been a success.

We also appreciate our friends who are working in the Department of Electrical Engineering for providing assistance during development of this AI-Powered Smart Glasses for Visually Impaired People.

References:

[1] Saleem Khan, Muhammad Mohsin Khan, "Intelligent Assistive Device for Visually Impaired People - A Computer Vision Based Approach," *Spectr. Eng. Sci.*, 2025, [Online]. Available: https://thesesjournal.com/index.php/1/article/view/782

[2] Erwin Syahrudin, Ema Utami, "Augmentation for Accuracy Improvement of YOLOv8 in Blind Navigation System," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 4, pp. 579–588, 2024, doi: 10.29207/resti.v8i4.5931.

[3] G R Venkatkrishnan, R Jeya, "Real-Time Object Detection For The Visually Impaired Using Yolov8 And NLP On Iot Devices," *Int. J. Environ. Sci.*, vol. 11, no. 8, 2025, [Online]. Available: https://theaspd.com/index.php/ijes/article/view/2142

[4] Incheol Jeong, Kapyol Kim, "YOLOv8-Based XR Smart Glasses Mobility Assistive System for Aiding Outdoor Walking of Visually Impaired Individuals in South Korea," *Electronics*, vol. 14, no. 3, p. 425, 2025, doi: https://doi.org/10.3390/electronics14030425.

[5] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017, doi: 10.1136/svn-2017-000101.

[6] G. K. Walia, M. Kumar, and S. S. Gill, "AI-Empowered Fog/Edge Resource Management for IoT Applications: A Comprehensive Review, Research Challenges, and Future Perspectives," *IEEE Commun. Surv. Tutorials*, vol. 26, no. 1, pp. 619–669, 2024, doi: 10.1109/COMST.2023.3338015.

[7] E. A. Hassan and T. B. Tang, "Smart glasses for the visually impaired people," *Lect.*

Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9759, pp. 579–582, 2016, doi: 10.1007/978-3-319-41267-2_82.

[8] Mingxing Tan, Ruoming Pang, Quoc V. Le, "EfficientDet: Scalable and Efficient Object Detection," *arXiv:1911.09070*, 2020, [Online]. Available: https://arxiv.org/abs/1911.09070

[9] "AI Voice-Activated Assistants Empower Visually Impaired Users | Battle for Blindness." Accessed: Jun. 02, 2026. [Online]. Available: https://battleforblindness.org/voice-activated-assistants-how-ai-is-empowering-the-visually-impaired

[10] Nilesh Deotale, Shubham Raut, "Smart Assistive Stick for Visually Impaired People using YOLOv8 Algorithm," *Res. Sq.*, 2024, doi: 10.21203/rs.3.rs-4334164/v1.

[11] Omar Kanaan Taha Alsultan, Mohammad Tarik Mohammad, "A Deep Learning-Based Assistive System for the Visually Impaired Using YOLO-V7," *IIETA J.*, 2023, [Online]. Available: https://www.iieta.org/journals/ria/paper/10.18280/ria.370409

[12] Wei Wang, Bin Jing, "YOLO-OD: Obstacle Detection for Visually Impaired Navigation Assistance," *Sensors*, vol. 24, no. 23, p. 7621, 2024, doi: 10.3390/s24237621.

[13] A. Tavakoli Yaraki, "Seeing with Sound : Object detection, localization with YOLOv8 and audio feedback for blind individuals", doi: 10.5281/ZENODO.17340214.

[14] Maria Bestarina Laili, Kartika Kartika, "Integrating YOLOv8, EasyOCR, and GTTS for Text Detection in Assistive Technology for the Visually Impaired," *BIS Inf. Technol. Comput. Sci.*, vol. 2, 2025, [Online]. Available: https://unimma.press/conference/index.php/bistyc/article/view/185

[15] Hoysala Y Devanga, "AI-Driven Vision Assistance for Visually Impaired," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 8, pp. 909–915, 2025, doi: 10.22214/ijraset.2025.73684.

[16] G Krishna Reddy, "A Review on Object Detection with Voice Support for the Visually Impaired People," *Ijraset J. Res. Appl. Sci. Eng. Technol.*, 2024, [Online]. Available: https://www.ijraset.com/research-paper/object-detection-with-voice-support-for-the-visually-impaired-people

[17] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:2004.10934*, 2020, [Online]. Available: https://arxiv.org/abs/2004.10934

[18] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, Accessed: Nov. 15, 2023. [Online]. Available: https://arxiv.org/abs/1804.02767v1

[19] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2999–3007, Dec. 2017, doi: 10.1109/ICCV.2017.324.

[20] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv.org*, 2017.

[21] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv:1608.06993*, 2018, [Online]. Available: https://arxiv.org/abs/1608.06993

[22] Alex Krizhevsky, Ilya Sutskever, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, 2017, [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[23] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *Int. J. Comput. Vis. 2009 882*, vol. 88, no. 2, pp. 303–338, Sep. 2009, doi: 10.1007/s11263-009-0275-4.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object

Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[25] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2/FIGURES/5.

[26] "Ultralytics/YOLOv5 · Hugging Face." Accessed: Jun. 02, 2026. [Online]. Available: https://huggingface.co/Ultralytics/YOLOv5

[27] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 7464–7475, 2023, doi: 10.1109/CVPR52729.2023.00721.

[28] World Health Organization, "World Health Organisation, 'World report on vision,' 2019.," *World Heal. Organ.*, vol. 214, no. 14, p. 180, 2019, Accessed: Jun. 02, 2026. [Online]. Available: https://iris.who.int/bitstream/handle/10665/328717/9789241516570-eng.pdf?sequence=18