YOLOv8n-Based Lightweight Object Detection model for People with Visual Impairment

Shruti Mallikarjun

Department of Computer Science Karnataka State Akkamahadevi Women's University, Vijayapura
Research Scholar, Vijayapura, Karnataka, India shrutipasargi290@gmail.com

Dr. Sheetalrani Kawale

Department of Computer Science Karnataka State Akkamahadevi Women's University, Vijayapura
Assistant Professor, Vijayapura, Karnataka, India sheetalrani@kswu.ac.in

Abstract

Vision is a crucial sense for living beings. A vast number of individuals worldwide have vision impairment. These individuals have challenges in moving autonomously and securely, including difficulty in obtaining information and communication. In this study, we address the challenge of object detection for visually impaired individuals by employing a YOLOv8n model to identify ten specific objects: Suitcase-Luggage, Switch-Box, Bottle, Dustbin, Mirror, Person, Staircase, Stove, Toilet, and Toothbrush. To achieve this, we developed a custom dataset using the Roboflow platform, ensuring comprehensive annotation for accurate detection. The dataset was meticulously divided into training, testing, and validation subsets to facilitate robust model evaluation. Following the training process, we analyzed the model's performance using confusion matrices, highlighting the precision, recall, and overall accuracy for each object category. The results demonstrate the model's effectiveness in detecting and distinguishing between the specified objects, underscoring its potential application in assistive technologies for the blind. This research contributes to the ongoing efforts to enhance accessibility and independence for visually impaired individuals through advanced machine learning techniques.

Keywords-object detection, visually impaired, You look only once (YOLO), annotation, bounding box.

1. Introduction

Visual impairment, a widespread disorder that affects millions of people worldwide, poses substantial difficulties in communication and autonomous navigation. Although there have been improvements in eye health, age-related disorders such as macular degeneration, cataracts, and glaucoma still play a significant role in the occurrence of visual impairment, especially among older individuals. [1]. Visually impaired persons have distinct difficulties while traversing outside areas, requiring them to depend on aural clues or helpful equipment such as white canes or guide dogs. Nevertheless, developing technologies provide encouraging options to improve spatial awareness and enable autonomous movement. Computer vision-based assistive devices, such as object detection glasses, have gained interest among these technologies [2]. These wearable

gadgets use sophisticated algorithms to identify items in the user's environment, offering immediate auditory or tactile responses. Developing the user interface and gesture controls is crucial for maximising the potential of technology and making it more usable and accessible for visually impaired users [3]. The emergence of computer vision systems has been driven by the need to comprehend the vast quantity of digital pictures from many fields. Computer vision is the scientific field that focuses on creating methods for enabling computers to see, comprehend, and analyse the visual information included in digital pictures, such as movies and images [4]. In order to get a comprehensive comprehension of digital photos or videos, it is inadequate to just concentrate on categorising various images. Instead, it is necessary to specify the objects' categories and positions inside each image [5]. Object detection, a prominent focus in the realm of computer vision, encompasses this particular problem. Object detection is the procedure of locating and classifying several items inside an image by using bounding boxes to indicate their positions and labels to define their categories [6]. Furthermore, object identification methods may be achieved by either using conventional machine-learning approaches or applying deep learning algorithms [7]. Object detection has been integrated into several fields, ranging from personal security to workplace efficiency, including autonomous driving, intelligent video surveillance, facial recognition, and numerous people-counting applications. Utilising current breakthroughs in deep learning, we have created a machine intelligence assistance system that can reliably identify things. Although computer vision techniques have been extensively used in research, there has been a lack of adequate development of deep learning components to assist those with visual impairments. Therefore, the contributions of this study as follows:

- Developing robust annotation tools for visually impaired individuals to assist in object detection involves combining accessible interfaces with advanced AI technologies. Roboflow provides robust annotation tools that facilitate precise and efficient labeling. For each image, bounding boxes were drawn around the target objects. The platform supports multiple annotators working simultaneously, ensuring a rapid yet thorough annotation process.
- The YOLOv8n model, when trained on a vast dataset, can accurately identify significant landmarks to aid visually impaired individuals in navigating interior spaces.

The subsequent sections of this article are delineated below. The literature survey is described in Section 2. The proposed resolution is outlined in Section 3. Section 4 covers the examination of performance and outcomes using comparative research. Section 5 concludes with a summary and suggestions for further study.

2. Related works

Object detection systems for visually impaired individuals are a growing area of research, leveraging advancements in computer vision, machine learning, and assistive technologies. This section aims to provide an overview of existing methodologies, and their effectiveness in assisting visually impaired users.

In [8] emphasised the need of using sophisticated, kitchen-oriented methods that use deep learning to enhance the precision of detection and provide immediate, interactive assistance via voice technology. The proposed approach, considerably increase the freedom and safety of visually impaired chefs, a major assistive technology advancement. The KERAS dataset was used to create a novel Binary Object Detection Pattern Model (BODPM) for face key point detection and recognition in [9]. The unmasked face was found using binary patterns and proximity detection. Objects that were disguised to blend in with their surroundings were identified within a range where the probability of their presence was at its highest, reaching 100%. In [10], a novel and precise bus detection model is presented, which is built around an enhanced iteration of the YOLOv5 model. The proposed model is not only lightweight but also achieves high accuracy in bus identification. Integrated the SimSPPF module into the YOLOv5 backbone as a replacement for the SPPF module. This enhancement improves computing efficiency and enhances the accuracy of object detection. Ultimately, we enhanced the original YOLOv5 structure to create a more efficient and quicker Slim scale detection model. This modification is crucial for real-time object identification applications. A model is developed in [11] to aid Visually Impaired Persons (VIP's) in recognising interior objects in their everyday lives. The model uses a recently developed optimisation method called Honey Adam African Vultures Optimisation (HAAVO). In this context, the process of identifying and distinguishing objects is accomplished by the use of Generative Adversarial Network (GAN) for object detection, and Deep Convolutional Neural Network (DCNN) for object identification. The technology presented in [12] involves the use of a Raspberry-Pi 4B, Camera, Ultrasonic Sensor, and Arduino, which are installed on the individual's stick. An ultrasonic sensor, affixed to a servomotor, was used to gauge the distance between the visually impaired individual and any obstructions. A deep CNN model with an accuracy of 83.3% is used in [13], while the dataset consists of over 1000 categories. Furthermore, a quantitative comparison study is conducted based on scores, using the supported characteristics of the devices. The restrictions are enumerated from the aforementioned techniques. In [14], implemented seven different YOLO object identification algorithms to assess their performance on images including ordinary things seen on highways and sidewalks. After a thorough analysis, it was determined that YOLOv8 is the most optimal model, achieving an accuracy rate of 80% and a recall rate of 68.2% on a well recognised Obstacle Dataset. In [15] RoboFlow uses the suggested weather-specific dataset, CVD. The final accuracy values have been enhanced to 73.26%, 72.84%, and 73.47%, which is very advantageous for both autonomous and conventional vehicle operations.

- Binary Object Detection Pattern Model (BODPM) may struggle with accuracy and reliability, especially under varying lighting conditions and object occlusion. The model’s performance can be inconsistent, leading to potential misidentifications or missed detections.
- YOLOv5 model demands can cause latency issues, especially on portable devices with limited computational power and battery life.

To overcome above limitations YOLOv8n ,model is adopted here. It provides high accuracy and speed in detecting and identifying objects, making real-time processing more efficient even on portable devices with limited computational power. YOLOv8n also supports integration with intuitive and accessible interfaces that provide clear audio or haptic feedback, ensuring the system is user-friendly.

3. Proposed methodology

To develop an efficient object detection model for visually impaired persons using YOLOv8n model's workflow is illustrated in figure 1. Once the images were uploaded, the annotation process was initiated. Roboflow provides robust annotation tools that facilitate precise and efficient labeling. For each image, bounding boxes were drawn around the target objects. The platform supports multiple annotators working simultaneously, ensuring a rapid yet thorough annotation process. After completing the annotation process and ensuring the dataset's quality, the annotated data was exported in a format compatible with the YOLOv8n model.

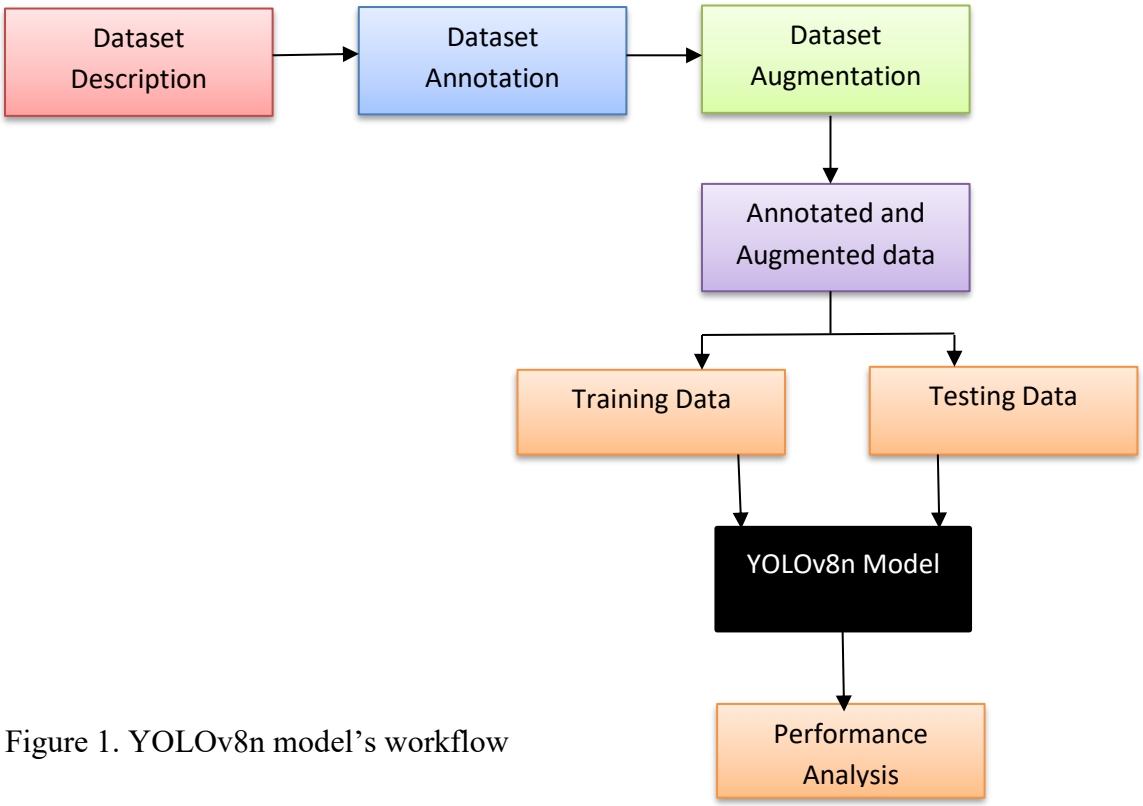

Figure 1. YOLOv8n model's workflow

Roboflow supports multiple export formats, including YOLO-specific annotations, making it straightforward to integrate the dataset with our training pipeline. The annotated data is forwarded to YOLOv8n model designed for high speed and low computational cost.

3.1 Dataset description

Collecting images for the ten target objects [16] (Suitcase-Luggage, Switch-Box, Bottle, Dustbin, Mirror, Person, Staircase, Stove, Toilet, and Toothbrush) using the Roboflow platform. Each image must contain at least one of the ten target objects. The objects should be clearly visible and identifiable in the images, without significant obstructions or occlusions. Under Indoor and Outdoor Settings, we can collect images from both indoor and outdoor environments to cover different use cases (e.g., suitcases in outdoor, staircases in buildings, toilets in homes).

3.2 Image annotation

Once the images were uploaded, the annotation process was initiated. Roboflow provides robust annotation tools that facilitate precise and efficient labeling. For each image, bounding boxes were drawn around the target objects. The platform supports multiple annotators working simultaneously, ensuring a rapid yet thorough annotation process. The image is divided, either automatically or manually, into segments that contain meaningful material. These segments include items along with their category, identification, and potential activity. The image's semantic categorisation, referred to as the semantic class, is documented as the top level of the hierarchical description structure. An image's A_I annotation I is made up of a series of keywords, $c_i, \dots c_n$, whose order is determined by whether the ideas c_i are present in the image. Furthermore, the sequence includes d implicit descriptors (imdescriptors) that define the meaning of visual contents that are understood by humans.

- Input: Set of training instances The set $T = \{t_1, t_2, \dots t_r\}$ consists of tuples $t_i = (F_I, A_I)$, where each tuple represents the low-level features F_I and the related annotation A_I of an image.
- Output: The appropriate pattern for annotating image I is to use a collection of keywords $c_i, \dots c_n$, arranged in order of their relevance to the image.

This vector may be regarded as an accurate representation of the general properties of images belonging to the same category. The i th component of a prototype vector p' for category c is calculated as $p'_i = \frac{1}{|c|}\sum_{x \in c} f_i(x)$, where f_i represents the i -th component of the feature vector of an image $x \in c$ and $|c|$ represents the number of pictures in category c.

In order to select a subspace of the feature components, it is necessary to provide weights to the components that are important for distinguishing between different categories. Typically, local and global criteria are merged to determine their relative importance. Assume that the image database contains N images, denoted as x_j, where j represents the index of the image. Let f_i represent a key characteristic that is highly indicative of a certain category of images or a particular class c_m. The weighting w_i of component i in prototype vector p' is calculated using the following formula:

$$w_i = freq(f_i, c_m) log\left(\frac{N}{occ(f_i, C)}\right)$$

Where the feature frequency $freq(f_i, c_m)$ indicates how often feature f_i appears in pictures of class c_m whereas $occ(f_i, C)$ depicts the presence of feature f_i in other classes. The set C is defined as the set of elements $C = \{c_1, \dots . c_{m-1}, c_{m+1}, \dots c_n\}$.

The relevant item type (e.g., Suitcase-Luggage, Bottle, Person) was assigned to each bounding box.

The interface provided by Roboflow facilitates the simple selection of pre-established labels, hence minimising the probability of mistakes occurring during the labelling process. Once the annotation process was finished and the dataset's quality was verified, the annotated data was exported in a format that is compatible with the YOLOv8n model. Roboflow offers many export formats, including YOLO-specific annotations, which simplifies the process of incorporating the dataset into this training workflow.

3.3 YOLOv8n Model

Object detection systems for visually impaired people using YOLOv8n offer several advantages. Technologically, YOLOv8n provides high accuracy and speed in detecting and identifying objects, making real-time processing more efficient even on portable devices with limited computational power. This model's improved performance helps reduce latency, enhancing the user experience. Usability is significantly improved with YOLOv8n's capability to work effectively in diverse and cluttered environments, offering reliable object detection under various lighting conditions and weather scenarios. The structure of the model we constructed includes a backbone, neck, and head, as seen in Figure 2. In the following sections, we provide the design principles for each component of the model's structure, as well as the modules inside each component.

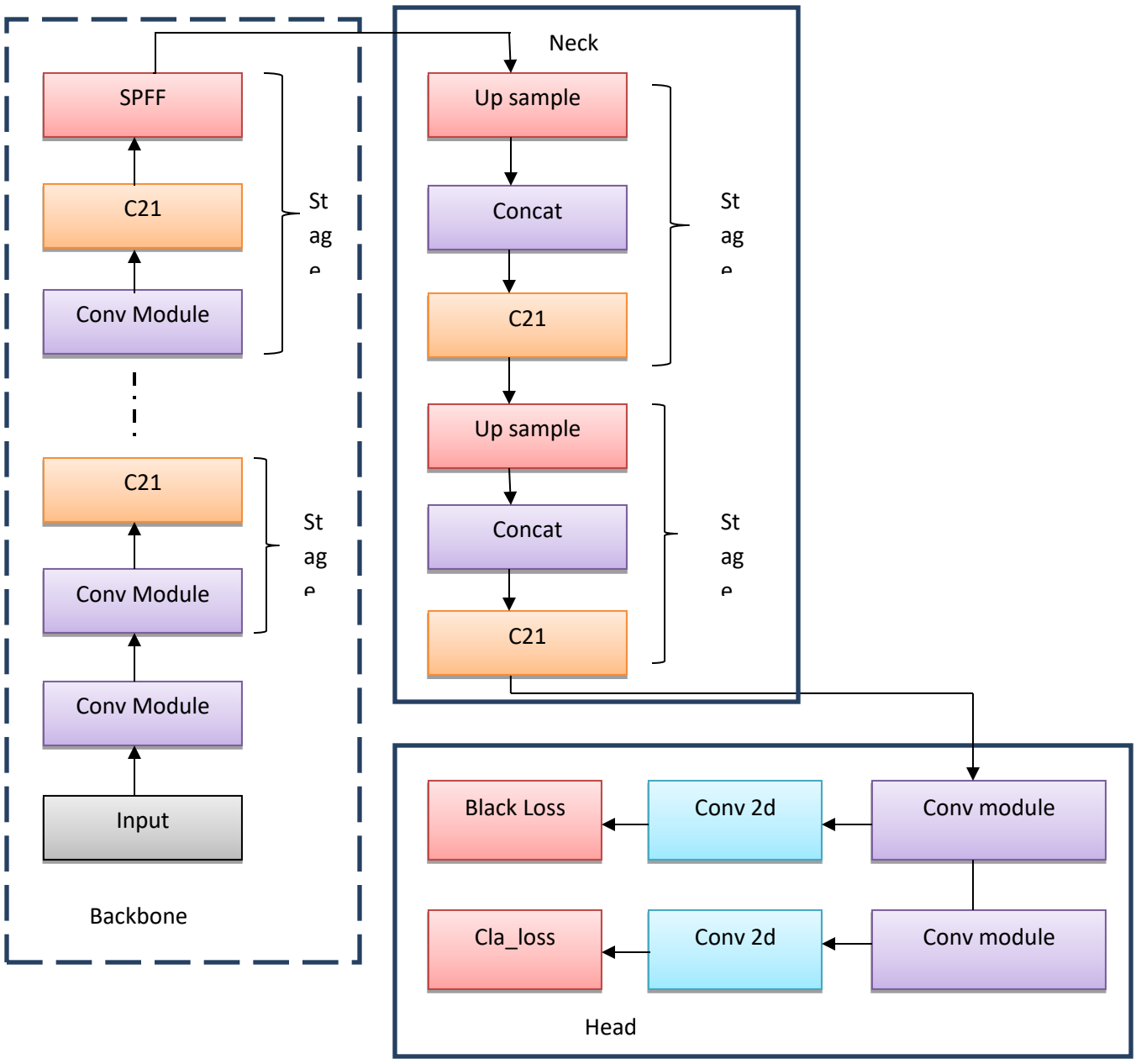

Figure-2 YOLOv8 model architecture

3.3.1 Backbone:

The primary framework of the model use the Cross Stage Partial (CSP) architecture to divide the feature map into two distinct sections. The first segment employs convolution techniques, while the subsequent segment is appended to the output of the preceding segment. The use of the CSP architecture enhances the CNNs' capacity to learn and diminishes the computational expenses associated with the model. The YOLOv8 model incorporates the C2f module, which combines the

C3 module with the idea from YOLOv7. This integration enables the model to get more comprehensive gradient flow information. The C3 module comprises 3 ConvModule and n DarknetBottleNeck components. The C2f module comprises 2 ConvModule and n DarknetBottleNeck components, which are coupled using Split and Concat operations. ConvModule layers are Conv-BN-SiLU, and n is the bottleneck count. In contrast to YOLOv553, this method uses C2f instead of C3. In addition, we reduce the number of blocks every step compared to YOLOv5 to reduce computational cost. This strategy reduces the number of blocks from 3 in Stage 1 to 6 in Stage 2, 6 in Stage 3, and 3 in Stage 4. Spatial Pyramid Pooling - Fast (SPPF), an upgraded form of SPP, speeds up model inference at Stage 4. We now have a model with better learning and shorter inference times.

3.3.2 Neck:

In general, networks with more depth acquire a greater amount of feature information, leading to improved dense prediction. However, deep networks reduce location information, and too many convolution processes reduce information for tiny objects. Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) architectures are needed for multi-scale feature fusion. Fig. 2 shows that our model architecture's Neck component integrates network data via multi-scale feature fusion. Higher layers absorb more information due to extra network layers, whereas lower layers retain location information because to fewer convolution layers. FPN uses upsampling to boost feature information in the bottom feature map, building on YOLOv5. PAN downsamples the top feature map for further information. The two feature outputs are combined to accurately anticipate multidimensional pictures. To reduce computational cost, we use the FP-PAN (Feature Pyramid-Path Aggregation Network) and avoid convolution operations during up sampling.

3.3.3 Head:

We utilise a decoupled head with distinct classification and detection heads, unlike YOLOv5. Figure 2 shows that this model preserves just classification and regression branches and eliminates the branch without an object. Anchor-Base calculates the regression object's four offsets from a large number of image anchors. Use anchors and offsets to position the object. The Anchor-Free method locates the object's centroid and estimates its distance from the enclosing box.

Experimental analysis

Experimental setup - The yolov8n.pt model, part of the YOLOv8 family, was selected for its balance between performance and computational efficiency, making it suitable for real-time object detection tasks. The training process spanned 100 epochs, allowing the model ample opportunity to learn and refine its understanding of the features and patterns associated with each object class. This extended training duration helped in improving the model's accuracy and

robustness. A batch size of 16 was used. This size strikes a balance between computational load and the ability to generalize well from the training data. It ensures that each training step is efficient, leveraging the parallel processing capabilities of the Tesla T4 GPU, while also providing enough variability in each batch to enhance learning.

Performance metrics- Model assessment is crucial for assessing the performance of a model and its conformity with the research goals. Various assessment measures may be used in bus object identification, including variables pertaining to the precision, velocity, and efficacy of the model. The primary emphasis on recognising bus objects is in evaluation measures pertaining to accuracy, since the model must possess the ability to recognise buses with a high degree of precision. Precision (P) and recall (R) are frequent evaluation criteria. Precision = true positives (TP) / true positives and false positives (TP + FP). The model's positive class prediction accuracy is measured by this metric. Recall, or sensitivity, is measured by dividing true positives (TP) by true positives and false negatives (TP + FN). This score measures the model's positive classification accuracy. Table 1 lists assessment criteria.

Table-1 Performance metrics

Performance matrix	**Formula**	**Description**
accuracy	$acc = \frac{TP + TN}{TP + TN + FP + FN}$	Where, $pt = true\ positive\ rate$ $pf = false\ positive\ rate$ $nf = false\ negative\ rate$ $nt = true\ negative\ rate$
recall	$re = \frac{pt}{pt + nf}$	
precision	$pr = \frac{nt}{nt + pf}$	
F1-score	$ppv = \frac{pt}{pt + pf}$	

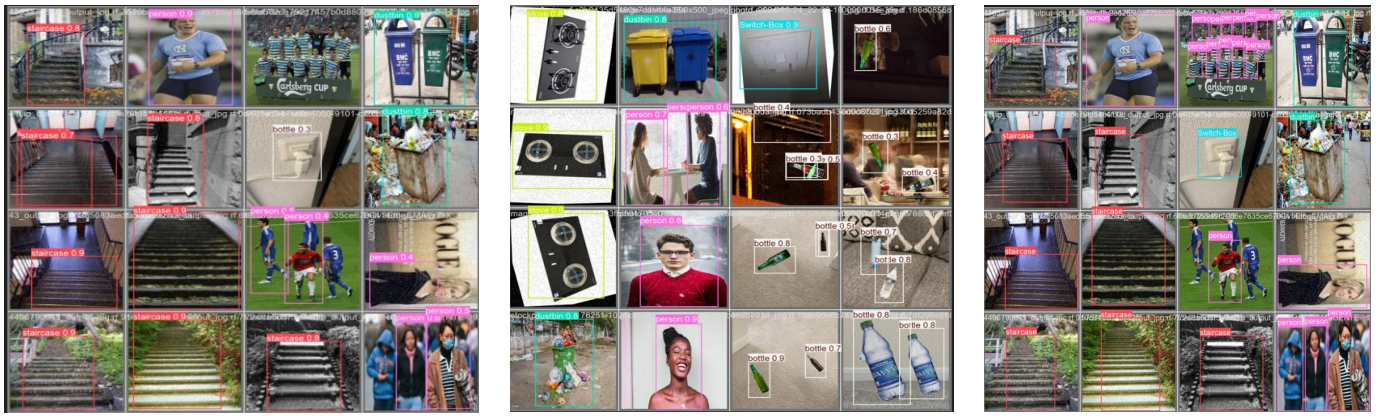

Figure-3detected objects using YOLOv8n model

As seen in Figure 3, the YOLOv8n model, known for its lightweight and efficient architecture, was trained to detect ten specific objects: Suitcase-Luggage, Switch-Box, Bottle, Dustbin, Mirror, Person, Staircase, Stove, Toilet, and Toothbrush. The training process involved using a dataset curated and annotated via Roboflow, ensuring a diverse and comprehensive representation of these objects across various environments and conditions.

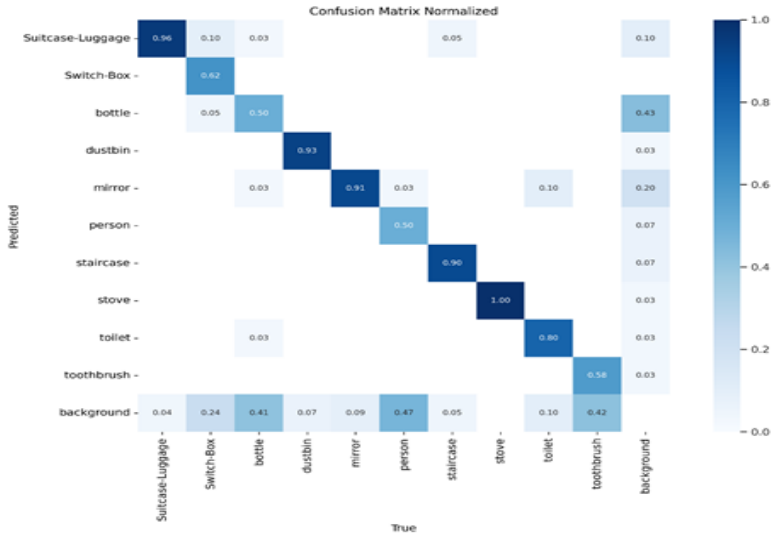

Figure-4 Confusion matrix

The provided figure-4 is a normalized confusion matrix for a machine learning model tasked with classifying images into ten categories. The confusion matrix is normalized, meaning that the values are proportions rather than raw counts, with each row summing up to 1.0. When analyzing it, the Suitcase-Luggage achieves 0.96 of TPR, Switch-Box achieves 0.62 of TPR, Bottle has 0.50 of TPR, dustbin has 0.93 of TPR, mirror has 0.91 of TPR, person with 0.50 of TPR, staircase with 0.90 of TPR, stove with 1.00 of TPR, toilet with 0.80 of TPR, toothbrush with 0.58 of TPR. Here, significant confusion with background misclassification is between 40%-45%

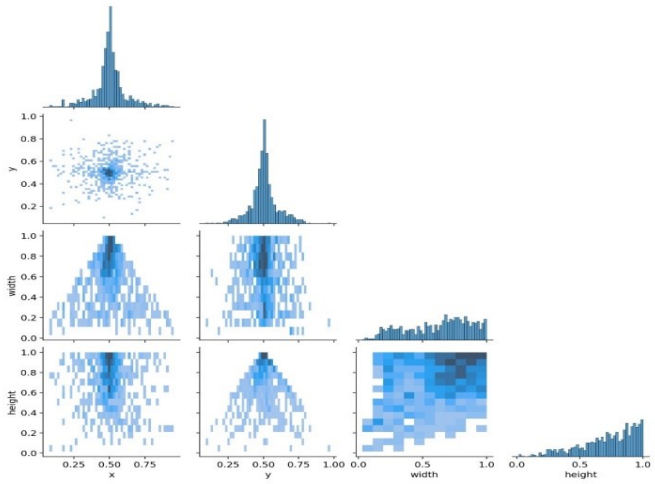

Figure-5 Correlogram analysis

The above figure-5 is a correlogram, specifically a pair plot, displaying the relationships between four variables related to object detection bounding boxes: x, y, width, and height. These variables typically represent the coordinates of the center of the bounding boxes and their respective dimensions within the normalized image space.

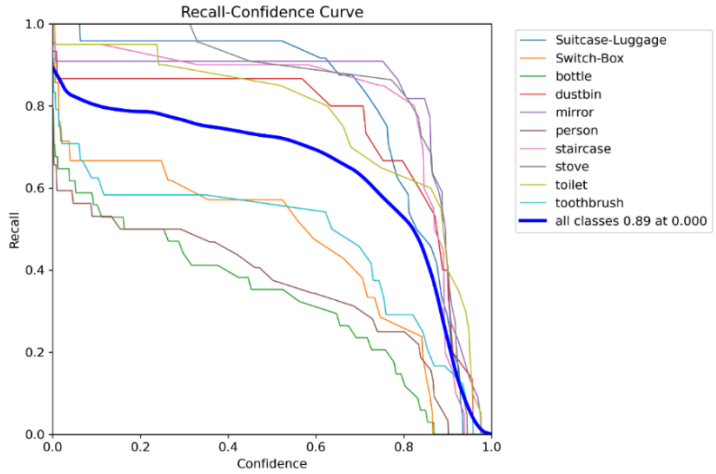

Figure-6 Recall-Confidence Curve

The provided figure-6 is a Recall-Confidence curve for a YOLOv8n object detection. When analysing it, Suitcase-Luggage maintains high recall up to a confidence of around 0.5, after which it starts to drop**.** Switch-Box shows a rapid decline in recall as the confidence threshold increases, indicating sensitivity to confidence. Bottle has steady decline in recall with increasing confidence, suggesting moderate detection performance. Dustbin and mirror maintains high recall for a broad range of confidence values, indicating reliable detection. Person Shows a sharp decline in recall

with increasing confidence, indicating challenges in detection. Staircase achieves high recall across a broad range of confidence levels. Stove has very consistent recall, dropping only slightly even at higher confidence levels. Toilet has moderate decline in recall with increasing confidence, showing relatively stable detection. Toothbrush has Significant drop in recall as confidence increases, indicating detection challenges.

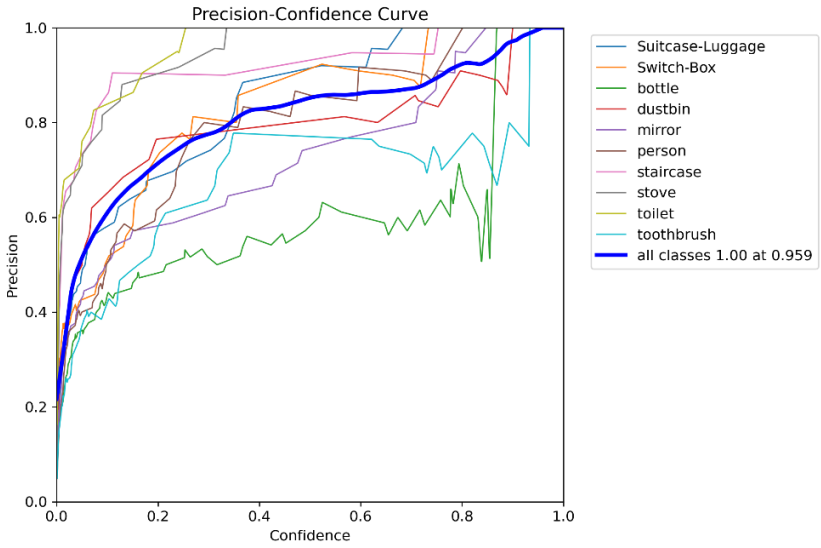

Figure-7 Precision-Confidence Curve

From the above figure-7, the curves that stay higher on the plot indicate better precision at various confidence thresholds. For example, a curve that remains high (close to 1.0) as the confidence increases indicates that the model is very precise for that class. The overall performance can be assessed by looking at the thick blue line. A higher and more consistent curve suggests that the model is performing well across different classes.

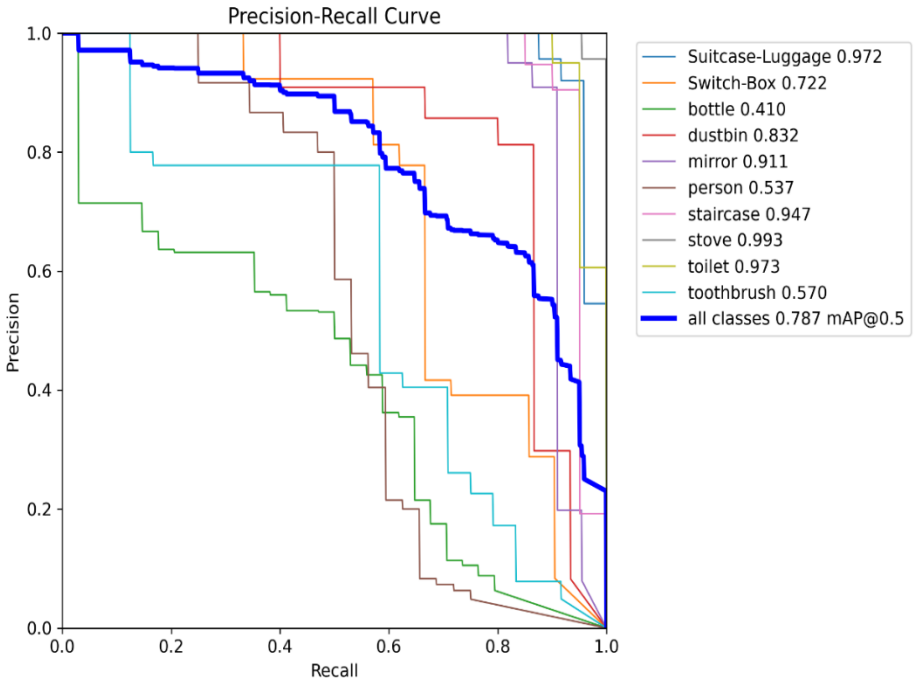

Figure-8 Precision-Recall (PR) curve

The figure-8 displays a multiple curves, with each curve representing the PR curve for a distinct class. Every curve represents the balance between accuracy and recall for a certain class. The Suitcase-Luggage model has achieved a high mAP score, which indicates excellent performance in terms of both accuracy and recall specifically for this category. Switch-Box has a moderate mAP score, suggesting a balance between precision and recall. Bottle has a lower mAP score, indicating poorer performance in distinguishing this class. Dustbin class performs relatively well. Mirror class has a high mAP score. Person and staircase has a high mAP score wherin Person has moderate performance. Stove has an excellent mAP score, indicating near-perfect classification. Toilet also performs exceptionally well wherin Toothbrush class has moderate performance. Classes with higher mAP values indicate better performance, meaning the model is more accurate in predicting the presence of these classes.

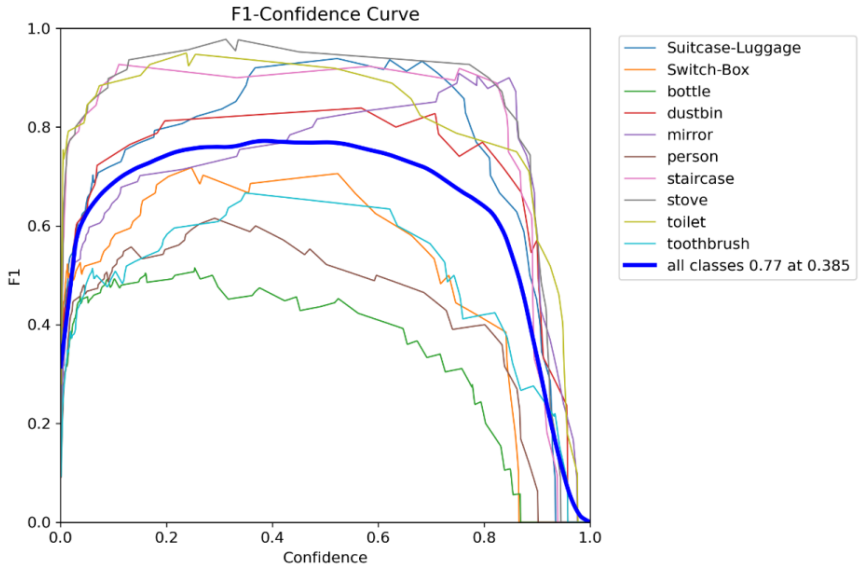

Figure-9 F1-Confidence Curve

The F1 score, shown in figure-9, is calculated as the harmonic mean of accuracy and recall. It serves as a unified measure that strikes a balance between the two. The x-axis of the Recall-Confidence curve reflects the confidence threshold, which is the estimated probability by the model that an item belongs to a certain class.

Table 2. Comparative analysis of proposed and Existing methods

Model	Accuracy (%)	Precision (%)	Recall (%)
[14]	-	80	68.2
[15]	73.26	72.84	73.47
YOLOv8n [proposed]	78.3	82.9	74.2

Conclusion

In this work, we have introduced a novel object detection model that is both lightweight and highly accurate. The model is built on the YOLOv8 architecture and aims to tackle the substantial difficulties faced by those with vision impairment in their everyday lives. This proposed model achieved superior performance compared to the original YOLOv8 model in terms of accuracy, recall, and mean average precision (mAP) by lowering the number of parameters. This experimental findings reveal that the Improved-YOLOv8n model surpassed previous object identification models in terms of both accuracy and inference time. Moreover, due to its reduced dimensions, decreased memory consumption, and enhanced inference speed, the developed model exhibited greater efficiency. This makes it a more feasible choice for mobile devices and real-time applications, particularly in situations when resources are limited. In order to improve its detection performance, the model needs undergo more training using a larger number of similar images.

Furthermore, the proposed object identification model is specifically designed to be implemented as a mobile application. Its purpose is not only to identify things in photos, but also to do so in real-time inside videos.

Reference

1. Jonas, J. B., Cheung, C. M. G., & Panda-Jonas, S. (2017). Updates on the epidemiology of age-related macular degeneration. *Asia-Pacific Journal of Ophthalmology*, *6*(6), 493-497.
2. Roopa, G. M., Prakash, C., & Pradeep, N. (2023). Computer Vision-Based Assistive Technology for Blind and Visually Impaired People: A Deep Learning Approach. *Computer Assistive Technologies for Physically and Cognitively Challenged Users*, 48.
3. Alajarmeh, N. (2021). Non-visual access to mobile devices: A survey of touchscreen accessibility for users who are visually impaired. *Displays*, *70*, 102081.
4. Szeliski, R. (2022). *Computer vision: algorithms and applications*. Springer Nature.
5. R. Kumar and S. Meher, "A Novel method for visually impaired using object recognition," 2015 Int. Conf. Commun. Signal Process. ICCSP 2015, pp. 772– 776, 2015.
6. N. A. Ismail, N. G. Yohgamalar, and M. S. Salam, "Gesture design for visually impaired people on mobile platforms: A systematic literature review," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8, pp. 1282– 1287, 2019
7. Masita, K. L., Hasan, A. N., & Shongwe, T. (2020, August). Deep learning in object detection: A review. In *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)* (pp. 1-11). IEEE.
8. Dang, B., Ma, D., Li, S., Dong, X., Zang, H., & Ding, R. (2024). Enhancing kitchen independence: Deep learning-based object detection for visually impaired assistance. *Academic Journal of Science and Technology*, *9*(2), 180-184.
9. Sajini, S., & Pushpa, B. (2024). A Binary Object Detection Pattern Model to Assist the Visually Impaired in detecting Normal and Camouflaged Faces. *Engineering, Technology & Applied Science Research*, *14*(1), 12716-12721.
10. Arifando, R., Eto, S., & Wada, C. (2023). Improved YOLOv5-based lightweight object detection algorithm for people with visual impairment to detect buses. *Applied Sciences*, *13*(9), 5802.
11. Nagarajan, A., & Gopinath, M. P. (2023). Hybrid optimization-enabled deep learning for indoor object detection and distance estimation to assist visually impaired persons. *Advances in Engineering Software*, *176*, 103362.
12. Masud, U., Saeed, T., Malaikah, H. M., Islam, F. U., & Abbas, G. (2022). Smart assistive system for visually impaired people obstruction avoidance through object detection and classification. *IEEE access*, *10*, 13428-13441.
13. Ashiq, F., Asif, M., Ahmad, M. B., Zafar, S., Masood, K., Mahmood, T., ... & Lee, I. H. (2022). CNN-based object recognition and tracking system to assist visually impaired people. *IEEE access*, *10*, 14819-14834.
14. He, C., & Saha, P. (2023). Investigating YOLO models towards outdoor obstacle detection for visually impaired people. *arXiv preprint arXiv:2312.07571*.
15. Sharma, T., Chehri, A., Fofana, I., Jadhav, S., Khare, S., Debaque, B., ... & Arya, D. (2024). Deep Learning-Based Object Detection and Classification for Autonomous Vehicles in Different Weather Scenarios of Quebec, Canada. *IEEE Access*.
16. https://universe.roboflow.com/blind-people-object-detection-klpbg/blind-people-object-detection-iuhho