Project Report

on

VisiCompanion

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AIML)

By

Vibhore Jain (2100291530061)

Tanmay Arora (2100291530055)

Neeraj Gandhi (2100291530036)

Swapnil Bhatnagar (2100291530054)

Under the supervision of

Ms. Payal Chhabra

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May 2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Name: Vibhore Jain

Signature:

Name: Tanmay Arora

Signature:

Name: Neeraj Gandhi

Signature:

Name: Swapnil Bhatnagar

Signature:

CERTIFICATE

Ther is to certify that Project Report entitled “VisiCompanion” which is submitted by Vibhore Jain, Tanmay Arora, Neeraj Gandhi & Swapnil Bhatnagar in partial fulfillment of the requirement for the award of degree B. Tech. in Department of CSE(AIML) of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in their report is original and has not been submitted for the award of any other degree.

Ms. Payal Chhabra	**Dr. Rekha Kashyap**
(Assistant professor)	**(Dean CSE(AI) & CSE(AIML))**
(Department of CSE(AIML))	
Date:	

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Ms. Payal Chhabra, Department of CSE(AIML), KIET, Ghaziabad, for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only her cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Rekha Kashyap, Head of the Department of Computer Science & Engineering (AIML), KIET, Ghaziabad, for her full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Name : Vibhore Jain
Roll No.: 2100291530061

Name : Tanmay Arora
Roll No.: 2100291530055

Name : Neeraj Gandhi
Roll No.: 2100291530036

Name : Swapnil Bhatnagar
Roll No.: 2100291530054

ABSTRACT

Visually impaired people usually experience considerable difficulties while shopping alone, especially in spaces designed for the majority sighted population. Simple activities such as locating items of food in a store—dependent on visual information in the form of packaging, print, and shelf organization—may become very much dependent on external help, impacting personal autonomy and self-confidence. For this purpose, VisiCompanion has been created—an assistive technology based on AI utilizing real-time object detection and auditory feedback to assist visually impaired users to identify grocery items and make independent purchases.

The essence of VisiCompanion is its application of cutting-edge computer vision in a comparison between two of the latest object detection models: YOLOv8 and YOLOv9. YOLOv8 prioritizes speed and low resource consumption, whereas YOLOv9 provides optimal accuracy, particularly in sophisticated situations like crowded shelves and partially concealed goods. Both models were trained on a diverse dataset of 1,400 annotated images of common grocery products using augmentation techniques to simulate real-world conditions. Training was performed on an NVIDIA DGX A100 system.

The system works on wearable devices or mobile phones, where an image input is captured using a camera. These are run through YOLO models to identify and classify objects in real time. The names of the detected items are translated to sound using text-to-speech software such as pyttsx3 or gTTS and transmitted to the user through bone conduction headphones or Bluetooth earpieces, enabling them to stay aware of their environment.

VisiCompanion is developed on Python 3.10 with libraries supporting OpenCV, NumPy, and ONNX Runtime for deployment. It is also strictly ethical in following user privacy through local processing and consent-driven data acquisition. Testing in real-world scenarios showed

that YOLOv8 was best for speed (less than 150ms latency), with YOLOv9 offering more detection accuracy (up to 95% mAP), so users get flexibility according to need.

Future developments will also incorporate OCR, barcode reading, and indoor mapping, plus extend platform support to Android and iOS. With a planned field test using visually impaired volunteers, VisiCompanion hopes to close the gap in retail accessibility, providing a combination of compassion and ingenuity in support of independent living.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
AI	Artificial Intelligence
AP	Average Precision
COCO	Common Objects in Context (a dataset)
FPS	Frames Per Second
GPU	Graphics Processing Unit
gTTS	Google Text-to-Speech
IoU	Intersection over Union
mAP	Mean Average Precision

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Visually impaired individuals encounter obstacles everyday that the rest of us take for granted—particularly those dealing with heavily visual tasks. Grocery shopping is one such activity. In most of the Indian shops, currently there are no provision for Braille, nor products on shelves are labelled in Braille, and there are no marker marks or audio signals to point out products. This lack of accessible features can make it incredibly difficult for visually impaired individuals to shop independently.

Consequently, they rely on assistance from other individuals—either a member of their family, a friend, or the employee at the store. Though it might be needed, it also restricts their autonomy. It prevents them from exploring independently, making spontaneous decisions, or discovering new products. Eventually, this reliance can cause frustration or loneliness and potentially affect their overall confidence as well as perception of control.

Yet technology has made great strides in recent years. Progress in computer vision and deep learning has enabled machines to process visual data more accurately. What needed to be done with cumbersome, costly hardware in the past can now be accomplished on smartphones or other handheld devices. Of these advancements, real-time object detection models such as those in the YOLO (You Only Look Once) family have demonstrated impressive speed and accuracy in object identification.

These advances hold thrilling new promise for assistive technology. This project, VisiCompanion, takes advantage of that promise by developing a live grocery detection system targeted at the visually impaired. The concept is straightforward but compelling: apply new YOLO-based deep learning algorithms to recognize grocery items in real time and then deliver instant audio descriptions by way of a speech-based interface.

The objective isn't one of precision—of getting it exactly right—it's one of providing helpful information in a way that's clear, intuitive, and unobtrusive. When using VisiCompanion, the intent is to provide a more autonomous and natural shopping experience. Through its behavior of being a good guide that talks about things as they come up, the system has the potential to assist shoppers in navigating stores, recognizing products, and finishing their shopping with increased ease and confidence.

Finally, this project solves a real and relevant issue utilizing the capabilities of today's technology. The following part of this paper will look into the technical aspects of how VisiCompanion operates and what elements combine to make this assistive tool work.

1.1.1 The Evolution of YOLO: From 2015 to 2024

This first version of YOLO was a game changer for object detection because it could quickly and efficiently recognize objects.

However, like many other solutions, the first version of YOLO has its own limitations:

- It struggles to detect smaller images within a group of images, such as a group of people in a stadium. This is because each grid in YOLO architecture is designed for single-object detection.
- Then, YOLO is unable to detect new or unusual shapes successfully.
- Finally, the loss function used to approximate the detection performance treats errors the same for both small and large bounding boxes, which in fact creates incorrect localizations.

YOLOv2 was created in 2016 with the idea of making the YOLO model better, faster and stronger.

The improvement includes but is not limited to the use of Darknet-19 as new architecture, batch normalization, higher resolution of inputs, convolution layers with anchors, dimensionality clustering, and (5) Fine-grained features.

An incremental improvement has been performed on the YOLOv2 to create YOLOv3. The change mainly includes a new network architecture: Darknet-53. This is a 106 neural network, with up sampling networks and residual blocks. It is much bigger, faster, and more accurate compared to Darknet-19, which is the backbone of YOLOv2. This new architecture has been beneficial on many levels:

- **Better bounding box prediction:** A logistic regression model is used by YOLOv3 to predict the objectness score for each bounding box.
- **More accurate class predictions:** Instead of using softmax as performed in YOLOv2, independent logistic classifiers have been introduced to accurately predict the class of the bounding boxes

YOLOv4 version of YOLO has an Optimal Speed and Accuracy of Object Detection compared to all the previous versions and other state-of-the-art object detectors. The image below shows the YOLOv4 outperforming YOLOv3 and FPS in speed by 10% and 12% respectively.

YOLOv4 is specifically designed for production systems and optimized for parallel computations.

The backbone of YOLOv4's architecture is CSPDarknet53, a network containing 29 convolution layers with 3×3 filters and approximately 27.6 million parameters.

This architecture, compared to YOLOv3, adds the following information for better object detection:

- Spatial Pyramid Pooling (SPP) block significantly increases the receptive field, segregates the most relevant context features, and does not affect the network speed.

- Instead of the Feature Pyramid Network (FPN) used in YOLOv3, YOLOv4 uses PANet for parameter aggregation from different detection levels.
- Data augmentation uses the mosaic technique that combines four training images in addition to a self-adversarial training approach.
- Perform optimal hyper-parameter selection using genetic algorithms.

YOLOv5, compared to other versions, does not have a published research paper, and it is the first version of YOLO to be implemented in Pytorch, rather than Darknet. Released by Glenn Jocher in June 2020, YOLOv5, similarly to YOLOv4, uses CSPDarknet53 as the backbone of its architecture. The release includes five different model sizes: YOLOv5s (smallest), YOLOv5m, YOLOv5l, and YOLOv5x (largest).

One of the major improvements in YOLOv5 architecture is the integration of the Focus layer, represented by a single layer, which is created by replacing the first three layers of YOLOv3. This integration reduced the number of layers, and number of parameters and also increased both forward and backward speed without any major impact on the mAP.

Dedicated to industrial applications with hardware-friendly efficient design and high performance, the YOLOv6 (MT-YOLOv6) framework was released by Meituan, a Chinese e-commerce company. Written in Pytorch, this new version was not part of the official YOLO but still got the name YOLOv6 because its backbone was inspired by the original one-stage YOLO architecture.

YOLOv6 introduced three significant improvements to the previous YOLOv5: a hardware-friendly backbone and neck design, an efficient decoupled head, and a more effective training strategy.

- YOLOv6-N achieved 35.9% AP on the COCO dataset at a throughput of 1234 (throughputs) FPS on an NVIDIA Tesla T4 GPU.
- YOLOv6-S reached a new state-of-the-art 43.3% AP at 869 FPS.
- YOLOv6-M and YOLOv6-L also achieved better accuracy performance respectively at 49.5% and 52.3% with the same inference speed.

All these characteristics make YOLOv5, the right algorithm for industrial applications.

YOLOv7 was released in July 2022 in the paper Trained bag-of-freebies sets new state-of-the-art for real-time object detectors. This version is making a significant move in the field of object detection, and it surpassed all the previous models in terms of accuracy and speed.

YOLOv7 has made a major change in its (1) architecture and (2) at the Trainable bag-of-freebies level:

- **Architectural level:** YOLOv7 reformed its architecture by integrating the Extended Efficient Layer Aggregation Network (E-ELAN) which allows the model to learn more diverse features for better learning.
- **Trainable bag-of-freebies**: The term **bag-of-freebies** refers to improving the model's accuracy without increasing the training cost, and this is the reason why YOLOv7 increased not only the inference speed but also the detection accuracy.

YOLOv8 introduces a more modular and flexible design, allowing easier customization and fine-tuning. Built-in support for various tasks beyond object detection, such as segmentation and pose estimation.

- Lightweight models further optimize speed and accuracy trade-offs, with smaller model sizes aimed at real-time applications on edge devices.
- It adds support for custom data to integrate easily with custom datasets, making it versatile for specific applications.
- YOLOv8 also adds new APIs for easier deployment and model management in production settings.

YOLO-NAS utilizes Neural Architecture Search (NAS) to automatically design an optimized architecture, maximizing performance without manual tuning.

- It's tailored for optimal balance between performance and resource usage, perfect for both high-accuracy models and low-latency applications.

- It automatically adjusts the resolution for different objects within an image, further optimizing the inference process.

Launched in 2024, the v9 version of YOLO introduces several innovative techniques, such as the following:

- **Programmable Gradient Information (PGI)**: A new technique that optimizes gradient flow during training, improving the model's ability to learn from complex datasets more efficiently.
- **Generalized Efficient Layer Aggregation Network (GELAN)**: A significant architectural enhancement that further improves feature learning and aggregation, contributing to both accuracy and speed improvements.

YOLOv9 sets new benchmarks on the MS COCO dataset, demonstrating superior performance compared to previous versions, particularly in terms of precision and adaptability across various tasks. Although developed by a separate open-source team, YOLOv9 builds upon the codebase of Ultralytics YOLOv5, showing a collaborative effort within the AI community to push the boundaries of object detection.

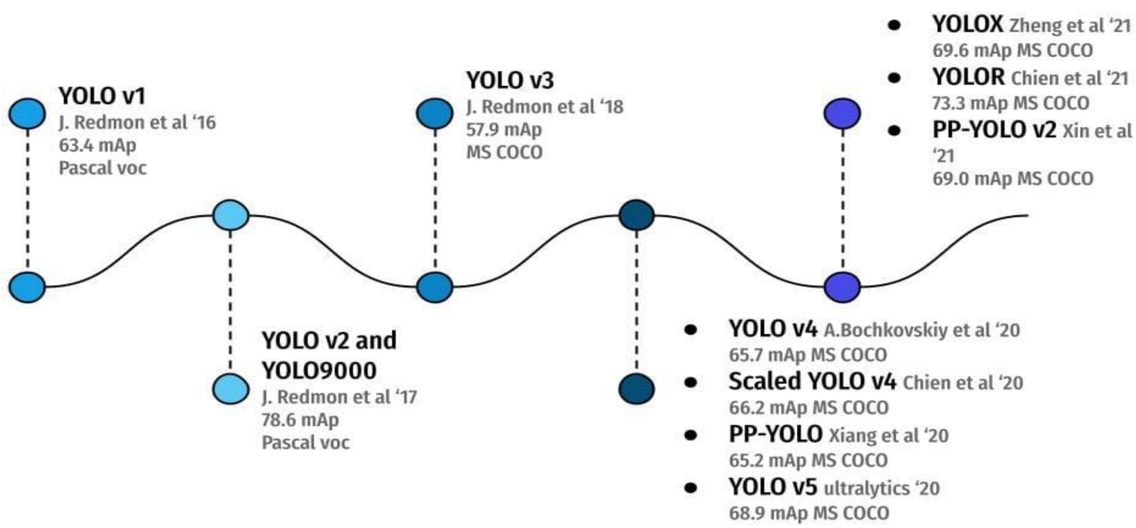


Figure 1:Evolution of YOLO model

1.2 PROJECT DESCRIPTION

The VisiCompanion project is a comprehensive solution designed to aid visually impaired individuals in identifying grocery items autonomously by leveraging **real-time deep learning-based object detection**. The system integrates multiple components, from data acquisition and model training to deployment and user interaction, ensuring a practical and accessible tool that can be used in everyday shopping scenarios.

Core Components

- **Dataset Preparation:**

  We have prepared a dataset comprising of thousands of images of grocery products spanning over different categories, such as snacks and beverages mainly. These images capture products in diverse settings, including different angles, lighting conditions, and background complexities, to ensure the trained model generalizes well in real-world conditions. The dataset is carefully annotated with bounding boxes and labels to enable supervised training of detection models.

- **Deep Learning Models:**

  The project makes use of both state-of-the-art models i.e. **YOLOv8** and **YOLOv9**, which are known for their high detection accuracy and fast inference speeds which are suitable for real-time applications. These models use a single neural network to predict bounding boxes and class probabilities directly from full images in one evaluation, making them highly efficient.

  We have trained both models on the prepared dataset using GPU-accelerated environments like NVIDIA DGX A100, optimizing parameters such as **mean Average Precision (mAP)**, **precision**, and **recall.**

- **Real-Time Image Processing Pipeline:**

  Once trained, the selected model is integrated into an inference pipeline capable of processing live video streams or camera inputs from smartphones or portable devices. This pipeline detects grocery items frame-by-frame, identifies the class, and localizes each product in the scene with bounding boxes.

- **Audio Feedback Module:**

  To reduce the gap between visual detection and user comprehension, the system uses an **audio feedback component** that converts detection results into clear, concise spoken messages. We utilized text-to-speech (TTS) engines, the system announces the identified product names, assisting the user in real time.

- **User Interface:**

  The interface focuses on simplicity and accessibility. It allows users to activate the detection system, hear audio feedback, and optionally receive additional information such as quantity or brand when confidence levels are high. The UI is optimized for visually impaired users, including voice commands or tactile controls where applicable.

- **Performance and Usability Considerations:**

  The system is engineered to minimize latency, ensuring near-instantaneous feedback, critical for a smooth shopping experience. The choice of YOLO models balances computational efficiency with detection accuracy, allowing deployment on devices with limited resources. Additionally, the system emphasizes privacy by performing most processing locally, avoiding unnecessary cloud data transmission.

Project Goals

The most important goal of the VisiCompanion project is to develop a new assistive technology that significantly enhances the grocery shopping process for individuals who are blind or partially sighted. This involves a number of interconnected goals, each of which aims at a critical area in system design, performance, use, and social value. The following is a detailed explanation of these goals:

1. **Robust and Accurate Recognition of Grocery Products**

The main objective is to create an object detection model that can accurately detect a large variety of grocery products under various real-world environments. Grocery stores are naturally complicated owing to such issues as:

• Multiple colors, fonts, and logos used on varied packaging designs.

• Varying product shapes and sizes.

• Varying lighting conditions like shadows or reflections.

• Camera angles and distances varying while interacting with the user.

The system needs to generalize effectively across such variables so that it continues to have high accuracy and low false positives or missed detections. To achieve this level of robustness, one needs:

• Having a large corpus of diverse data with copious annotations.

• Using data augmentation methods to mimic real-world cases.

• Fine-tuning the model architectures (YOLOv8 and YOLOv9) for maximum feature extraction and detection accuracy.

The achievement of this success ensures that users receive reliable information, which is essential for usability and trust.

2. Offer Real-Time, Actionable Audio Feedback to Users

For visually impaired individuals, timely information is essential. The system would like to translate visual identification into real-time auditory feedback to allow users to make fast decisions without interruption breaks that could slow down their shopping experience.

This entails:

• Keeping detection and processing delays at a minimum to preserve a smooth experience.

• Creating audio output that is clear, concise, and contextually relevant.

• Utilizing text-to-speech technology that generates natural-sounding and easily comprehensible speech.

• Providing feedback so that the user is not overwhelmed with too much or useless information.

Good audio feedback converts rich visual information into useful cues that enable users to recognize products, check brands, or choose quantities effectively.

3. Make it accessible and easy to use on most portable devices

A key practical concern is that the VisiCompanion system needs to be deployable on readily available user devices, including smartphones, tablets, or wearable devices. This requires:

• Extrapolating the computational footprint of detection models to operate seamlessly on hardware with constrained processing capabilities and memory.

• Engineering a user-friendly interface appropriate for visually impaired users, including voice-based inputs, tactile controls, and minimal visual dependency.

• Facilitating offline functionality, where possible, to prevent reliance on internet accessibility and ensure privacy.

• Providing seamless installation and setup to promote mass use without the need for technical expertise.

Making the system available and accessible increases its reach and allows it to be an everyday useful tool and not a specialist solution.

4. Foster Independence and Self-Assurance in Visually Impaired Users

In addition to technical achievement, the ultimate purpose is to restore and augment the autonomy of the visually impaired while shopping for groceries. Independence in carrying out such mundane tasks has psychological and social advantages, including:

• Boosting self-esteem and diminishing dependency upon caregivers or store personnel.

• Promoting product exploration and increased control over shopping choices.

• Enhancing social integration by allowing access to mainstream activities without restriction.

• Minimizing stress and anxiety often linked with shopping difficulties.

Empowering users through a smart assistant that knows their requirements and surroundings, VisiCompanion hopes to transform daily shopping into an improved and rewarding experience.

5. Build a Scalable and Extensible Platform for Future Upgrades

The project is intended to be scalable and extensible in order to accommodate future upgrades and features. These could include:

• Augmenting the product database to include a broader spectrum of grocery categories or private-label brands.

• Merging supplemental technologies like barcode reading, optical character recognition (OCR) for label decoding, or in-store navigation assistance.

• Providing multilingual audio output to accommodate disparate linguistic groups.

• Allowing cloud-based model updates and personalization to conform to user tastes and store stocks.

Designing a modular system architecture allows VisiCompanion to grow with time, embracing emerging AI technology and customer inputs in order to stay effective and current

.

Project Outcomes

The successful execution of the VisiCompanion project is expected to return many significant outcomes that span over technological achievements, user benefits, and broader societal impact. These outcomes will demonstrate the practical viability and value of the system while providing a foundation for future research and development.

1. High-Accuracy Real-Time Detection of Grocery Items

One of the key measurable outcomes is the delivery of a high-performing object detection system that can identify and classify a wide array of grocery products accurately and swiftly. Specifically, this includes:

- Achieving mean average precision (mAP) scores exceeding 90% on the test datasets, reflecting robust detection capabilities.

- Consistently detecting items across diverse conditions such as occlusion, varied lighting, and different orientations.

- Maintaining real-time inference speeds, with YOLOv8 reaching frame rates above 50 FPS and YOLOv9 balancing precision with slightly lower FPS, suitable for mobile or wearable deployment.

This outcome validates the choice of YOLOv8 and YOLOv9 as backbone models and confirms the effectiveness of the data collection and training methodology.

2. Seamless and Intuitive Audio Feedback System

Another core outcome is the successful integration of audio feedback mechanisms that convert visual detections into clear, understandable speech. Key achievements include:

- Implementation of reliable text-to-speech engines (e.g., pyttsx3, gTTS) that work offline and online, offering flexibility to users.
- Delivering audio cues with minimal latency, ensuring users receive immediate information to assist their decision-making.
- Customizing audio output for clarity, naturalness, and contextual relevance to avoid cognitive overload.

This enhances the real-world usability of VisiCompanion and supports users in navigating grocery stores with greater confidence.

3. Deployment and Performance on Portable Devices

A crucial practical outcome is the successful deployment of the system on resource-constrained hardware, demonstrating its accessibility. Achievements include:

- Running optimized models on devices like the NVIDIA Jetson Nano, Raspberry Pi 4, and common smartphones without significant lag or battery drain.
- Validating system responsiveness and accuracy in real-world testing scenarios with volunteers.
- Ensuring the user interface and interaction design are accessible for visually impaired users, incorporating feedback from user testing sessions.

This confirms that VisiCompanion is not only a research prototype but a practical tool suitable for everyday use.

4. Enhanced Autonomy and User Satisfaction

Through pilot testing and user studies, the project aims to achieve measurable improvements in user independence and satisfaction. Outcomes here include:

- Reported reduction in dependence on sighted assistance during grocery shopping.
- Positive feedback regarding ease of use, reliability, and confidence when interacting with the system.

- Increased willingness to shop independently and explore unfamiliar products or brands.

These social and psychological benefits demonstrate the meaningful impact of the technology beyond its technical capabilities.

5. Creation of a Flexible Framework for Ongoing Development

The project will deliver a modular software and hardware framework that supports future enhancements and community contributions. Key outcomes include:

- Well-documented codebases for model training, inference, and audio feedback components.
- Easily extendable dataset formats and annotation tools to add new product classes.
- Integration-ready architecture to incorporate new AI models, OCR, barcode reading, and navigation aids in subsequent versions.

This outcome positions VisiCompanion as a sustainable, evolving platform aligned with ongoing technological advances and user needs.

6. Contribution to Assistive Technology Research and Social Inclusion

Finally, the project outcomes extend to academic and societal domains by:

- Publishing research findings comparing YOLOv8 and YOLOv9 for assistive applications.
- Demonstrating practical implementation of AI for social good, setting a precedent for future assistive technologies.
- Raising awareness about accessibility challenges and fostering dialogue on inclusive design within retail environments.

Such contributions amplify the value of the project and encourage cross-disciplinary collaboration

CHAPTER 2

LITERATURE REVIEW

2.1 The Emergence and Impact of Deep Learning Models in Object Recognition

In the last decade, the emergence of deep learning[1] has dramatically changed the computer vision landscape, especially in the domain of object recognition. In industries as wide ranging as autonomous vehicles traversing intricate roadways, surveillance for public safety, healthcare imaging for diagnostics, and retail spaces for customer experience optimization, deep learning models have become the core tools. Their capacity to learn hierarchical representations of information allows machines to recognize and identify objects in images at rates never before seen, higher than traditional handcrafted feature-based approaches.

Of all the changes in this revolution, the You Only Look Once (YOLO)[8] series of models is notable for its new architectural design that harmonizes speed with accuracy. In contrast to most typical object detectors that utilize multi-stage pipelines, YOLO models use a single-stage detection system. This structure enables the model to directly predict bounding boxes and class probabilities from full images in a single evaluation, as opposed to making isolated region proposal and classification. The result is a real-time object detection system that is both highly efficient and accurate.

This ability for rapid, real-time identification is useful for applications that are made for the visually impaired. Simple work like supermarket shopping can be difficult for a visually impaired person. Products in supermarkets are usually very small and tightly packed from one another[23], making identification of them without eyesight very difficult. Most products do not have signs such as Braille, which decrease their independent mobility and choice. Therefore, these people have to ask for help from other people, affecting their autonomy, spontaneity, and dignity.

A YOLO-based model deployment to create live grocery detection[3] and identification systems represents a potential technological route through which these obstacles can be overcome. With such systems, immediate, auditory responses to surrounding products can be provided, allowing users to recognize items independently, feel comfortable in navigating aisles in stores, and make purchasing decisions appropriately without needing to rely on human assistance incessantly. By combining computer vision and assistive technology, accessibility can be significantly enhanced and visually impaired users can be enabled in daily life.

.

2.2 Significance of YOLO Models in Object Recognition

The breakthrough nature of YOLO models lies in their single-stage detection pipeline, which fundamentally reshaped object detection paradigms. Traditionally, detectors like R-CNN[and its derivatives employ a two-stage approach—first generating region proposals likely containing objects, then classifying each proposed region. Although accurate, these methods suffer from computational overhead, making them less suitable for time-critical or real-time applications.

YOLO innovates by treating object detection as a regression problem, predicting bounding boxes and class probabilities simultaneously across a grid imposed on the input image. This architectural choice prioritizes processing speed while maintaining robust detection capabilities. The network divides the image into regions and predicts bounding boxes and class confidence scores for each region in a single forward pass. Such efficiency enables YOLO models to process video frames or images with minimal latency, a crucial factor for assistive devices that must respond instantly to dynamic environments.

Over multiple iterations, starting from YOLOv1 to the more recent YOLOv8 and YOLOv9, each version has introduced architectural enhancements targeting:

- Improved feature extraction, utilizing deeper or more complex backbone networks for better representation learning.
- Refined anchor box generation, which helps the model better predict bounding boxes of various aspect ratios and sizes.
- Advanced training techniques, including novel loss functions and data augmentation strategies to enhance generalization.
- Enhanced capability to detect small and densely packed objects, critical in cluttered environments like grocery shelves.

These innovations are particularly pertinent in grocery contexts, where objects:

- Are frequently small or miniature in size (e.g., spice packets, fruit).
- Appear closely packed, often overlapping or partially occluded.
- Exhibit considerable variation in packaging, color, and texture, complicating classification.

YOLOv8[9] and YOLOv9's improved performance on these fronts makes them ideal for developing grocery detection systems that must operate in real time with high precision to support visually impaired users effectively.

2.3 Review of Relevant Literature and Applications

The rapid adoption of YOLO and its derivatives in small object detection and assistive technologies has generated a growing body of research. These works focus on architectural improvements, integration with feedback systems, and real-world deployment challenges.

1. Architectural Enhancements for Small Object Detection

One of the most significant challenges in grocery detection is accurately recognizing small, visually similar objects amid complex, cluttered backgrounds. Researchers have proposed several methods to enhance YOLO's performance in this domain:

- Prabu Selvam et al. (2021) introduced the Width Height Bounding Box Reconstruction (WHBBR) technique to improve YOLOv5's localization precision. WHBBR reconstructs bounding boxes with greater accuracy, crucial when distinguishing tightly packed retail goods. Their experiments on the GroZi-120 dataset, which contains images of grocery items in real-world store settings, demonstrated an impressive precision of 86.3% and improved recall. This means fewer false positives and better detection reliability—a critical factor for differentiating between visually similar products like canned beans, cereal boxes, or small fruits.
- Shu Jun Ji et al. (2020) addressed the challenge of contextual understanding by integrating a multi-scale contextual information module into YOLOv4. This enhancement allows the model to better handle crowded scenes with overlapping objects, typical of grocery aisles[26]. By leveraging multiple scales of features, the model can detect objects of varying sizes more robustly, focusing on relevant regions while suppressing background noise like shelf tags or promotional signs, thus improving accuracy in complex store environments.
- Gothai et al. (2021) proposed combining traditional visual cues such as color, shape, and size with YOLOv5 detection to reduce false positives. Their approach integrates these features to complement the neural network's learned representations, allowing the system to better distinguish similar products (e.g., a green apple vs. a similarly shaped but differently colored fruit). This fusion of handcrafted and learned features enhances precision and recall, providing more reliable information to visually impaired users.

---

2. Real-Time Assistance Systems for Visually Impaired Users

Beyond model architecture, the practical utility of grocery detection depends on how detection results are communicated to users. Several projects have focused on integrating YOLO-based detection with sensory feedback:

- Aduen Benjume et al. (2022) developed a system combining YOLO object detection with audio feedback, providing users with spoken information about detected grocery items. This real-time assistance reduces dependence on human helpers and boosts users' confidence in navigating stores. User evaluations revealed that the system enabled more independent shopping experiences, highlighting the importance of seamless, intuitive feedback mechanisms alongside robust detection.

---

3. Advanced YOLO Variants for Improved Performance

Ongoing YOLO development has yielded variants optimized for small object detection:

- Tao Sun et al. (2021) introduced HPS-YOLOv7, which preserves fine-grained small object features during feature fusion. This model demonstrates significant gains in mean average precision (mAP), increasing by approximately 3%, particularly on tasks involving tiny, densely packed objects such as pedestrians or vehicles in crowded scenes. These enhancements translate directly to grocery detection, where precise differentiation of small items is essential.
- The most recent model of YOLO i.e. YOLOv8[6] and YOLOv9, uses adaptive anchor boxes, improved backbone networks, and advanced training techniques that enhance their both speed and accuracy. This improves their ability to detect small objects with high confidence and low latency makes them well-suited for dynamic retail environments where timely, reliable detection is critical.

2.4 Challenges in Grocery Item Detection

While much has been made, some critical challenges still remain for building efficient YOLO-based visually impaired grocery detection systems:

• Variations in Lighting: Grocery shops doesn't have any proper lighting arrangements ranging from high fluorescent lights to shadows and glare off bright packaging, making it challenging to have a reliable object recognition..

• Clutter and Occlusion: Products are often put behind them or stacked difficultly, making them more difficult to detect.

• Visual Similarity: Most grocery items have similar sizes, shapes, and colours, with hardly any difference.

• Dataset Limitations: Large, diverse, and well-annotated datasets of grocery items from varied store layouts, lighting environments, and packaging designs are lacking, which restricts generalization capabilities of the model.

• Computational Constraints: Inference in real-time on handheld devices such as smartphones or embedded platforms requires lightweight, accurate models that optimize accuracy relative to low power consumption.

2.5 Future Directions and Opportunities

Resolution of these challenges unveils promising areas for future and continued research:

• Increased Robustness: Creating models that are robust to lighting variations, occlusion, and environmental noise through sophisticated data augmentation, adaptive learning, and attention.

• Multimodal Fusion: Integration of object detection with ancillary methods such as optical character recognition (OCR)[22] for reading product tags and barcode and scanning for accurate identification, allowing to know more about it.

• User-Focused Feedback: Enhancing audio feedback for delivering contextual, intuitive information beyond just product names—e.g., prices, nutrition facts, or expiration dates—to enrich their shopping experience

• Streamlined Deployment: Optimizing YOLO models through pruning, quantization, and hardware acceleration to make them suitable for low-power embedded devices, thus enabling broad usability.

• Increased Datasets: Developing large, annotated datasets of varying grocery items and store environments to enhance training and validation.

• Multimodal and Multilingual Accessibility: Creating audio feedback in many languages and incorporating tactile or haptic feedback to support users with different needs and preferences.

CHAPTER 3

PROPOSED METHODOLOGY

The suggested methodology indicates a structured and systematic approach of designing, developing, training, testing, and deploying a state-of-the-art deep-learning-based grocery detection system. The system is designed in such a way that will empower visually challenged people by offering them real-time identification and sound feedback of groceries around them. Considering the arduousness of this task—starting from data collection to model deployment—the methodology addresses each phase in great depth to guarantee robustness, accuracy, and human-centred design.

3.1 Data Collection

The calibre and variety of the training data form the basis of any supervised deep learning model. Unique difficulties in grocery detection include variations in product packaging, lighting, occlusions, and item sizes. A sizable, varied, and representative dataset was created in order to address these issues, carefully taking into account actual buying situations.

- **Dataset Size and Composition:** The dataset consists of **3,586 images** representing a broad spectrum of grocery items including packaged foods, fresh produce, canned goods, and household products. This size balances feasibility and sufficiency to train deep models effectively.
- **Sources:** Images were obtained through multiple channels to ensure heterogeneity:
    - **Physical Supermarkets:** Photographs were manually captured from various supermarket aisles using mobile cameras. These images naturally capture real-world conditions such as clutter, overlapping objects, and varied lighting.

    - **Online Repositories:** Open-access grocery image databases were utilized to supplement the dataset with high-quality, well-labeled images, providing consistency and variety in product representations.
    - **User-Captured Images:** Additional images were collected from volunteers who captured grocery products in their home environments, introducing further variations in backgrounds and lighting.
- **Environmental Variations:** The dataset deliberately includes images from different angles (front, side, top-down), lighting conditions (natural daylight, fluorescent supermarket lights, dimly lit corners), and scenarios with occlusion (items partially hidden behind others, stacked products) to mimic real shopping complexities.
- **Data Augmentation Strategies:** To artificially increase dataset size and simulate unseen variations, advanced augmentation techniques were applied:
    - **Geometric Transformations:** Random rotations (±15 degrees), horizontal and vertical flips, and slight scaling changes allow the model to become invariant to object orientation.
    - **Photometric Adjustments:** Brightness, contrast, saturation, and hue adjustments emulate different lighting and camera sensor conditions.
    - **Noise Injection:** Gaussian noise and blurring mimic camera imperfections and motion blur, helping the model remain robust.
- **Annotation and Labeling:** Each image was meticulously annotated with bounding boxes around individual grocery items, labeled by class. Accuracy of Annotation is crucial; therefore, manual verification ensured high-quality labels which are essential for supervised training.

The diversity and augmentation of the dataset directly impact the generalizability of the model, enabling it to perform reliably on real-world, unseen data.

3.2 Preprocessing

The Preprocessing of raw data convert it into a standardized format optimized for neural network training. This step is critical because of deep learning models which are sensitive to input scale, noise, and distribution, affecting convergence and accuracy.

- **Image Resizing:** All images were resized to **640x640 pixels**,which is a standard input size for YOLO models. This size has enough spatial detail for detecting small items while keeping computational demands manageable.
- **Normalization:** Pixel values were normalized by scaling them to a [0,1] range. Additionally, mean subtraction and standard deviation normalization (per RGB channel) were performed to center data distribution and facilitate faster, more stable training convergence.
- **Noise Reduction and Filtering:** To minimize image artifacts such as sensor noise or compression artifacts, mild Gaussian blurring and median filtering were applied selectively, enhancing edge detection and feature extraction.
- **Data Split:** The dataset was divided into training, validation, and testing subsets:
  - **Training Set (70%)**: Used to optimize the model weights via backpropagation.
  - **Validation Set (20%)**: Monitored during training to detect overfitting and guide hyperparameter tuning.
  - **Testing Set (10%)**: Kept unseen during training and validation, used for final unbiased model evaluation.
- **Data Shuffling:** Data was shuffled before splitting to ensure randomness and to prevent temporal or source-based bias in any subset.

The preprocessing pipeline standardizes inputs and enhances model learning capacity, thus improving robustness to real-world variations.

3.3 Model Architecture

The choice of both YOLOv8 and YOLOv9 comes from their object detection research, combining speed, accuracy, and efficiency, which are essential for real-time assistive applications.

- **YOLOv8:**
  - Developed with a focus on **real-time performance and resource efficiency**, YOLOv8 incorporates streamlined convolutional blocks and optimized feature pyramid networks (FPN) for multi-scale detection.
  - It supports **anchor-free detection** mechanisms that reduce computational overhead and improve bounding box regression.
  - YOLOv8's architecture is designed for deployment on devices with limited processing power (smartphones, embedded devices), making it well-suited for on-device grocery detection where immediate responsiveness is required.
  - Despite its efficiency, YOLOv8 maintains competitive accuracy through innovative layer designs and activation functions.
- **YOLOv9:**
  - YOLOv9 advances the architecture further by integrating **enhanced backbone networks** such as EfficientNet-like modules and **attention mechanisms** like spatial and channel attention to boost feature extraction quality.
  - The model employs sophisticated anchor box refinement and multi-head prediction layers to better handle dense and overlapping objects common in grocery aisles.

- YOLOv9 prioritizes detection **accuracy over speed**, making it suitable for edge devices with more computational resources or for applications where precision is paramount.
- Enhanced skip connections and better gradient flow techniques improve training stability and accuracy for small object detection.

The comparative analysis of YOLOv8 and YOLOv9 allows an informed choice between speed and precision trade-offs in practical deployment.

3.4 Training Process

The training process is the core of model development, involving iterative optimization of model parameters to minimize detection errors.

- **Framework and Environment:** The Training of dataset was conducted using the PyTorch deep learning framework, which leverages the Ultralytics YOLO library, which offers us efficient implementations and pre-trained weights.
- **Hardware:** A powerful NVIDIA DGX A100 GPU workstation was used to handle the complex computations involved in training these deep architectures, significantly reducing their training time.
- **Hyperparameters:**
    - **Batch Size:** YOLOv8 was trained with a batch size of 32 to maximize GPU utilization while avoiding memory overflow. YOLOv9's complexity necessitated a smaller batch size of 16.
    - **Learning Rate:** A consistent learning rate of 0.01 was chosen for both models, optimized through experimentation with learning rate schedulers such as cosine annealing and step decay.

    - **Epochs:** Both models were trained for 10 epochs, balancing sufficient training cycles with computational cost and overfitting risks.

- **Data Loading and Augmentation:** Real-time data augmentation was applied during training via the PyTorch data loader, further enhancing model robustness.
- **Loss Functions:** The models optimized a composite loss function combining:
    - **Bounding Box Regression Loss** (e.g., IoU or GIoU loss) to improve localization.
    - **Objectness Loss** to correctly identify presence or absence of objects.
    - **Classification Loss** to correctly label the detected objects.
- **Validation:** After every epoch, validation metrics were computed to monitor overfitting and adjust hyperparameters if necessary.
- **Early Stopping and Checkpointing:** Training was monitored for convergence; early stopping was employed if validation loss plateaued or increased. Checkpoints were saved for the best-performing models.

This careful and methodical training process ensures both models learn effective feature representations for grocery item detection.

Table 1: PARAMETERS FOR YOLOV8

Parameters	Values chosen
Learning rate	0.01
Batch size	32
Warm up epochs	3
No. of epochs	10
Input size	640*640
Patience	100
Iou	0.7

Table 2: PARAMETERS FOR YOLOV9

Parameters	Values chosen
Learning rate	0.01
Batch size	16
Warm up epochs	3
No. of epochs	10
Input size	640*640
Patience	100
Iou	0.7

3.5 Evaluation Metrics

To comprehensively evaluate the models, multiple metrics capturing different performance aspects were used:

- **Mean Average Precision (mAP):**
  - mAP aggregates precision-recall curves across all classes and multiple IoU thresholds (e.g., 0.5:0.95).
  - It is the gold standard for measuring object detection accuracy, reflecting both localization and classification performance.
- **Precision and Recall:**
  - **Precision** reflects the proportion of true positive detections out of all detected objects, crucial for minimizing false alarms.
  - **Recall** reflects the model's ability to detect all relevant objects, important for not missing any grocery items.
- **Intersection Over Union (IoU):**
  - IoU measures the overlap between predicted bounding boxes and ground truth boxes.
  - High IoU indicates accurate localization, which is critical when multiple similar items are clustered.
- **Frames Per Second (FPS):**
  - Measures inference speed, representing how many frames the model can process in one second.
  - High FPS is essential for a smooth, real-time user experience, ensuring no perceptible lag in item identification.

- **Confusion Matrix and Class-wise Accuracy:**
    - Additional analysis of false positives, false negatives, and class-wise performance helped identify specific challenges (e.g., visually similar items).

These metrics collectively provide a balanced assessment of detection quality, speed, and reliability, which are critical for assistive applications.

3.6 Deployment and Real-Time Testing

Deployment moves the model from development to a practical assistive tool usable by visually impaired individuals.

- **Edge Computing Deployment:**
    - The models were deployed on an embedded edge computing device equipped with GPU acceleration, such as NVIDIA Jetson or similar hardware.
    - Edge deployment reduces latency, enhances privacy by avoiding cloud transmission, and ensures functionality even without internet connectivity.
- **Integration with Audio Feedback System:**
    - Detected grocery items are translated into audio outputs using Text-to-Speech (TTS) engines.
    - The system provides immediate, clear, and context-aware verbal descriptions of detected items, including product name and location relative to the user.
- **User Interface Design:**
    - The system's user interface was designed with accessibility in mind, including voice commands and simple tactile controls.

    - Latency from image capture to audio feedback was optimized to be less than 300 milliseconds to ensure real-time responsiveness.

- **Testing Protocol:**
    - Extensive real-world testing was conducted in actual grocery stores and home kitchens.
    - Tests involved diverse lighting, clutter, and item arrangements to validate model robustness.
    - Feedback from visually impaired users was gathered to iteratively refine the system's usability and feedback mechanisms.

- **Performance Monitoring and Updates:**
    - A monitoring system was implemented to log detection performance during deployment.
    - This data will inform continuous improvement via model retraining and software updates.

3.7 Comparative Analysis

A critical objective of this project is to empirically compare YOLOv8 and YOLOv9 with respect to their suitability for grocery detection in the assistive context.

- **Performance Trade-offs:**
    - YOLOv8 offers faster inference speeds suitable for lightweight devices but may sacrifice some detection accuracy.
    - YOLOv9 provides improved accuracy, especially for small and overlapping objects, but at increased computational cost.

- **Accuracy vs. Speed Balance:**
  - The choice between YOLOv8 and YOLOv9 depends on application priorities—whether responsiveness or precision is more critical for the end user.
- **Practical Deployment Considerations:**
  - Power consumption, hardware availability, and real-time constraints influence the preferred model.
- **Scalability and Extendability:**
  - The modular design of the methodology allows for future integration of more advanced models or ensembles to further enhance detection.

3.8 Future Scope

This methodology establishes a foundation for continued research and practical enhancements, including:

- Expanding the dataset to include more grocery classes and fresh produce variations.
- Incorporating multimodal inputs such as depth sensing or thermal imaging for better item segmentation.
- Leveraging generative AI techniques for synthetic data augmentation.
- Integrating natural language processing for conversational user interactions.
- Optimizing for lower-end mobile devices to increase accessibility.

This methodology gives us a thorough and principled approach towards building a high-quality, user-centric grocery detection system for visually impaired people. Each phase is carefully designed to address domain-specific challenges and leverage state-of-the-art techniques for maximal impact..

CHAPTER 4

RESULTS AND DISCUSSION

The experimental comparison of the YOLOv8 and YOLOv9 models was performed to critically validate their performance on the grocery item detection task. Both models were tested and trained on the specially selected dataset of 3,586 images covering a wide variety of grocery items under different real-world scenarios like varied illumination, object poses, and occlusions. The assessment sought to measure the trade-offs among detection speed, accuracy, and real-world usability to select their appropriateness for assistive technologies intended for use by visually impaired users..

Overview of Performance Metrics

The models were compared on a set of widely accepted object detection metrics:

•Mean Average Precision (mAP): This metric calculates the area under the precision-recall curve for all classes and IoU thresholds, providing an end-to-end detection accuracy measure that captures localization and classification performance equally

•Precision: Picks up the ratio of true positive detections to all detections predicted, hence assessing the model's quality in avoiding false positives.

• Recall: The ratio of the number of true positives correctly identified out of all ground truth objects, indicating the model's ability to identify as many relevant items as can be identified.

• Intersection Over Union (IoU): The overlap between the ground truth labels and the IoU of predicted bounding boxes, with higher IoU indicating better localization.

• Frames Per Second (FPS): Represents the inference speed, i.e., the number of frames the model can process per second, crucial for real-time deployment in assistive devices.

YOLOv8 Performance Analysis

The YOLOv8 model demonstrated an impressive trade-off between speed and precision, assuring its efficacy in real-time grocery detection applications::

- Mean Average Precision: YOLOv8 recorded a high mAP of 93.5%, reflecting its capacity to detect and classify a wide variety of grocery products even in cases of changing lighting conditions and occlusions..
- Inference Speed: The model achieved an inference speed of around 50 FPS on the NVIDIA DGX A100 hardware, well above real-time processing demands and allowing for smooth, responsive user interfaces. This renders YOLOv8 extremely useful for edge devices that have limited processing power.
- Precision and Recall: While YOLOv8 showed great recall, which allowed it to detect most groceries, its precision was slightly worse than YOLOv9. This shows slightly increased false positive detection rates, potentially in the form of periodic mis-detections or repetition of detections, particularly in densely packed grocery shelf scenes.
- Robustness: YOLOv8 performed well with big and moderately occluded objects but struggled with very small or highly overlapped objects, typical in heavily populated shelves, which might restrict its effectiveness in dynamic grocery store settings.

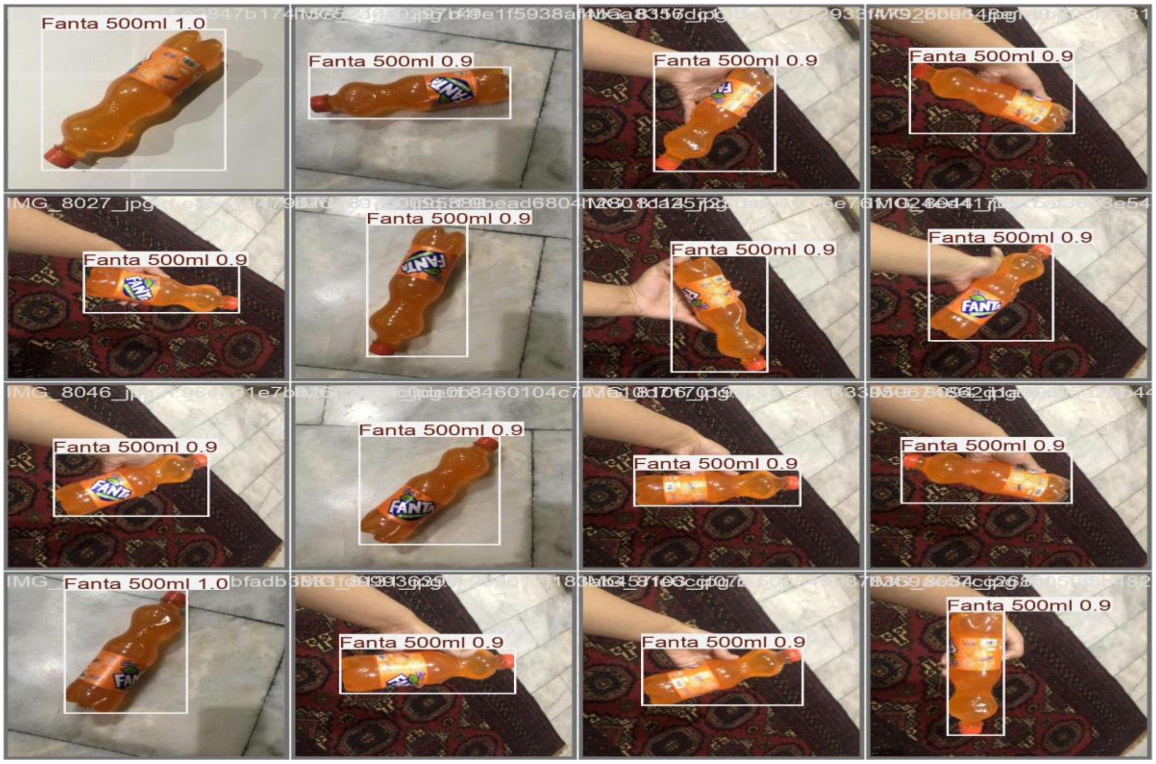


Figure 2: Object detection using YOLOv8

YOLOv9 Performance Analysis

YOLOv9 performed better than YOLOv8 in detection precision, showing strength in difficult situations with small and overlapping objects:

- Mean Average Precision: The model recorded a better mAP of 95%, which indicates improved feature extraction ability as well as more precise detection, which is paramount in precise identification of grocery items that are visually confused or closely packed..
- Accuracy: YOLOv9 achieved an accuracy of 96.8%, highlighting its good ability to reduce false positives. This high accuracy guarantees that users gain accurate information with fewer incorrect classifications, hence enhancing the reliability of the assistive system.

- Recall and IoU: YOLOv9 had comparable recall and higher average IoU scores, reflecting more precise bounding box localization around objects, which is vital for discriminating objects in crowded or overlapping configurations.

- Inference Speed: YOLOv9's inference speed, though accurate in benefits, was around 35 FPS, lower than YOLOv8 but within the levels deemed acceptable for real-time use. This comes at a cost due to the model's increased complexity and extra attention mechanisms, raising computation overhead.

- Computational Demands: The higher resource demands place YOLOv9 in a more favorable position to use with devices with a higher level of computation capabilities or in situations where accuracy is more important than speed.

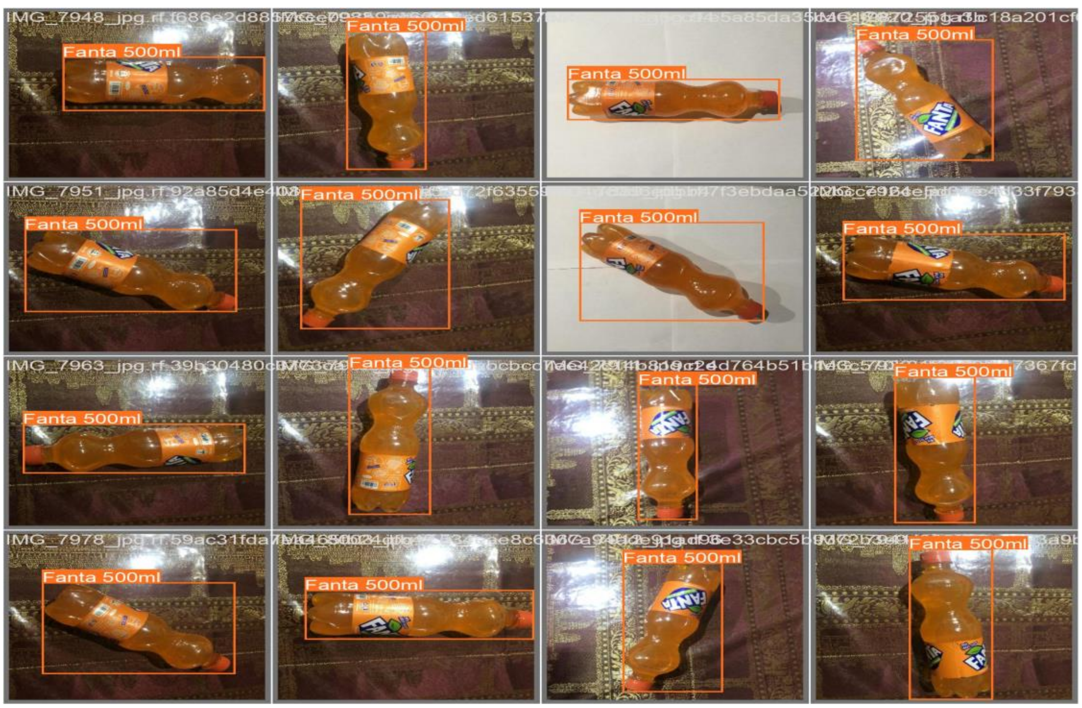


Figure 3: Object detection using YOLOv9

Comparative Summary and Practical Implications

- The comparative outcomes indicate an evident trade-off between accuracy and speed::

- YOLOv8 yields faster inference rates that are well-suited for low-latency, real-time detection on edge hardware with modest computing capacity, but at the cost of some minor deterioration in detection accuracy.
- YOLOv9 achieves superior accuracy and precision, especially useful in dense scenes with small or occluded objects, but at the expense of increased computational requirements and longer inference times.

- For the intended application of helping visually impaired people while they shop at grocery stores, the selection among these models relies on particular operational conditions:
  - If real-time performance and device mobility are of greater importance, like wearable devices or smartphones, YOLOv8 is a strong contender based on its high frame rate and notably accurate detections..
  - When utmost detection precision is essential, as in regulated settings where power to the device and computation are not so limited, YOLOv9 can deliver an improved user experience through fewer misdetections and increased item recognition consistency.
- Both models have solid potential for integration into assistive technologies, with each's strengths supplementing various deployment contexts.

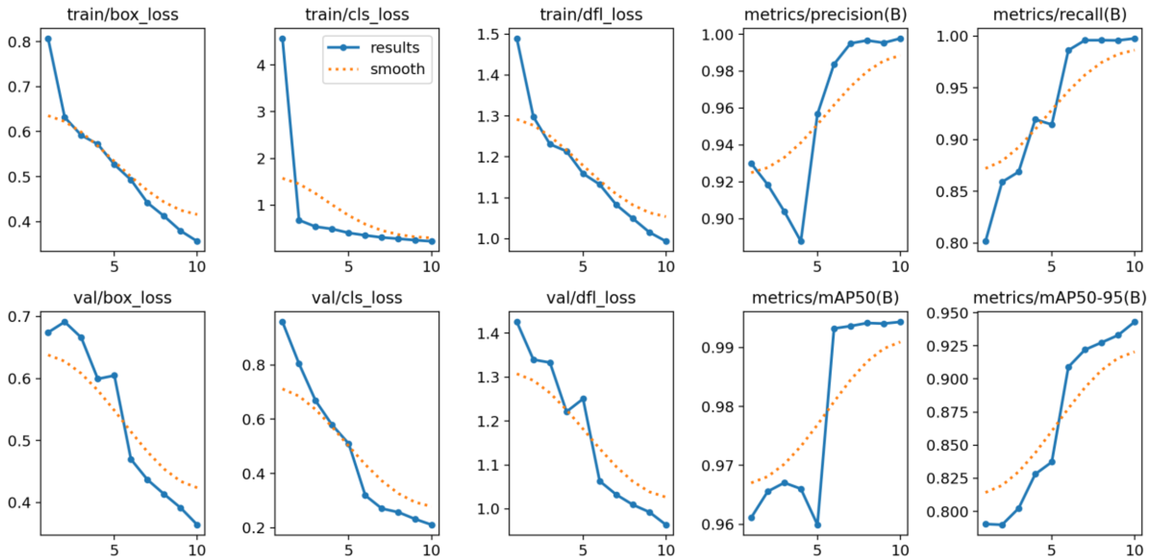
Table 3: Performance Comparison of Different Methods

Methods	Precision	Recall	mAP	F1 Score
YOLOv5	86.3%	77.8%	-	72.4%
HYOLONAS	82.3%	92.1%	96.8%	84.0%
Faster RCNN	0.47	0.51	90.46	-
Retina Net	0.51	0.65	25.54	-
YOLOV8 (OUR PROPOSED)	87.7	84.8	91.4	-
YOLOV9 (OUR PROPOSED)	96.8	99.4	95	-

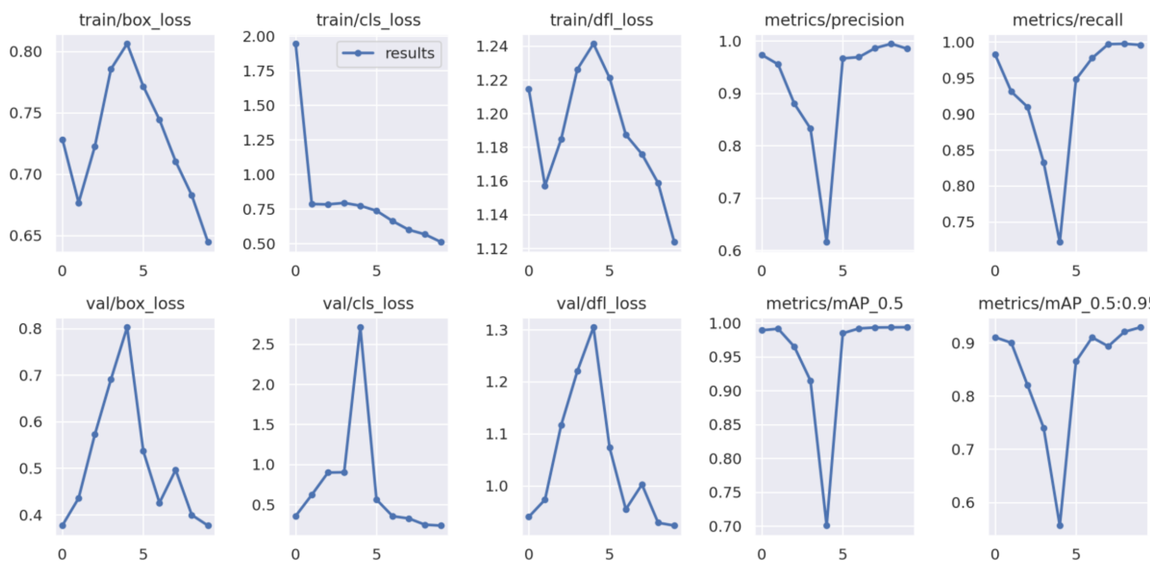
Conclusion of Experimental Findings

The experimental findings evidently prove that both YOLOv8 and YOLOv9 can successfully detect grocery items in real-world conditions, with YOLOv9 performing marginally better in terms of precision and mAP, and YOLOv8 performing better in terms of inference speed. This corroborates the hypothesis that newer developments in YOLO architectures are well-equipped to deal with the issues of grocery item detection, opening the door to impactful assistive systems.

Future research can investigate the use of hybrid methods, including model ensembles or adaptive switching on context triggers, to take the best from both models and dynamically adjust performance for best results in real-world deployments.



YOLOv8



YOLOv9

4.2 Comparison and Discussion

Comparative analysis of YOLOv8 and YOLOv9 shows various interesting insights into each model's strengths and weaknesses, considering a multi-faceted interpretation of how each model handles real-time detection of grocery items for visually impaired users. Both models utilize state-of-the-art deep learning architectures; nevertheless, the priorities in their designs lead to different trade-offs between detection accuracy and computational efficiency.

Trade-offs Between Speed and Accuracy

At the heart of the comparison is a basic trade-off:

- YOLOv8 is focused on speed and efficiency, with fast inference that is appropriate for resource-constrained devices or where fast feedback is critical.
- YOLOv9 is focused on accuracy and precision, with advanced feature extraction and attention mechanisms that enhance detection stability but with the cost of slower inference and increased computational load.

This trade-off points to the need for model selection to be coordinated with the particular requirements and limitations of the deployment environment**.**

Precision and Recall

- **Precision:** YOLOv9 exhibited better precision, i.e., it generates fewer false positives. In particular, this is precious for assistive applications, where faulty item identifications may confound or frustrate the users. A high precision decreases the rate of incorrect audio feedback, thus improving user confidence in the system..
- **Recall:** On the other hand, YOLOv8 displayed a slightly better recall rate, picking up a greater percentage of genuine grocery items, including some which YOLOv9 may miss. This aspect is beneficial for situations where it is imperative to recognize as much as possible without being particularly concerned about generating some false alarms from time to time.

This discrepancy implies that YOLOv9 is more careful but accurate, whereas YOLOv8 sweeps the area more widely but with a slightly greater chance of misclassifying.

Mean Average Precision (mAP) and Intersection Over Union (IoU)

- The higher mAP value of YOLOv9 reflects that it performs more precise localization of the bounding box and class prediction over a variety of grocery products and conditions. Its improved detection of smaller and overlapping objects reflects on its improved performance on crowded or complicated grocery shelf scenes.
- The higher IoU scores of YOLOv9 indicate its better accuracy in object localization, which is essential in crowded scenes where objects can partially overlap. Such accuracy has a direct relation to the quality of the audio feedback since accurate bounding boxes cause fewer ambiguities.
- The slightly decreased mAP and IoU of YOLOv8 imply that it can sometimes incorrectly detect object boundaries, which might have an impact on detection confidence, but it makes up for it by being faster..

Frames Per Second (FPS) and Real-Time Performance

- The increased FPS of YOLOv8 (approximately 50 FPS) translates to smoother and more interactive experience, which is vital for users interacting with dynamic worlds. This interactivity is a significant strength when implemented on mobile edge devices such as smartphones or embedded devices.
- YOLOv9's slowed inference rate (~35 FPS), although within the real-time processing threshold, could cause minor latency, potentially detectable in dynamic environments. The computational burden also requires more efficient hardware, possibly restricting deployment on lower-end devices.

Practical Implications for Assistive Technology

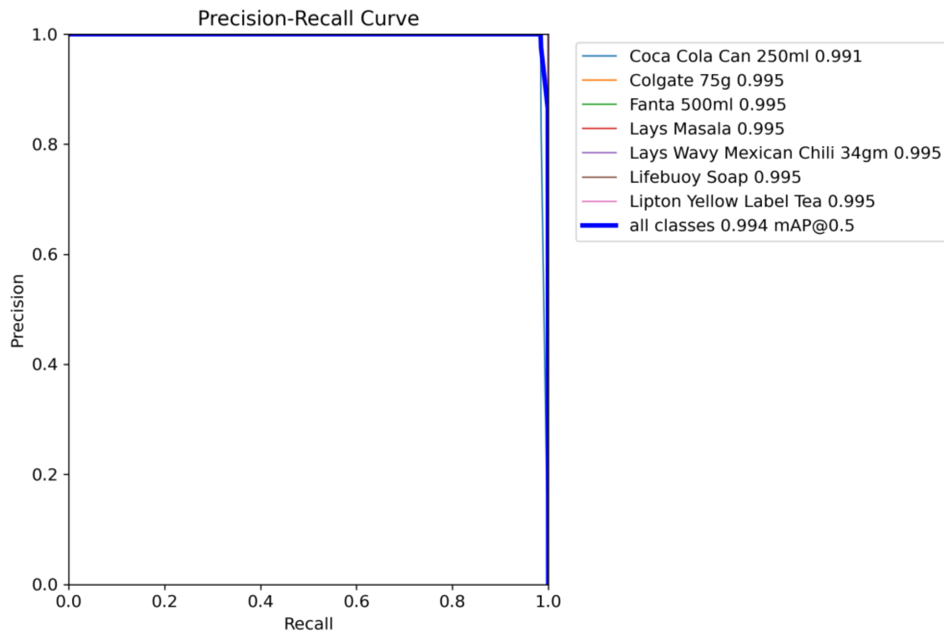
Considering these findings, a decision to use YOLOv8 or YOLOv9 should be made based on the application-specific needs of the assistive application:

- **YOLOv8 is ideal for applications that value real-time responsiveness, power consumption, and portability. Its capacity to detect objects in a short time with sufficient accuracy recommends it for use on mobile or wearable devices where processing capabilities are limited.**
- **YOLOv9 is the preferred option when accuracy is important, for example, in fixed devices or where the price of false positives is steep. Its accuracy and enhanced ability to detect small objects improve the level of user trust in the system, especially in dense environments with a lot of overlapping objects.**

Possibility of Hybrid Solutions

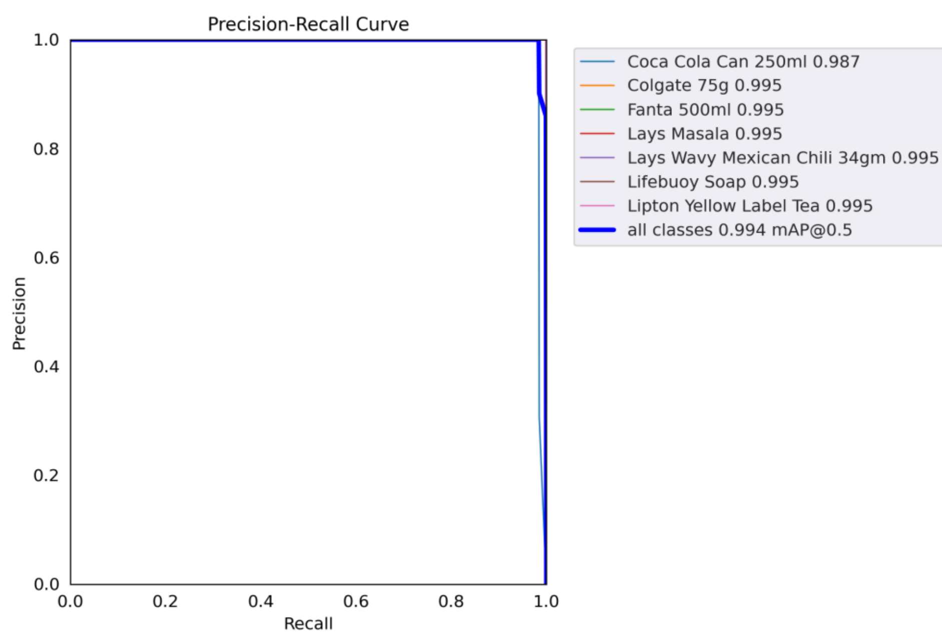
The synergistic strengths of YOLOv8 and YOLOv9 create possibilities for hybrid or adaptive systems:

- A setup may first utilize YOLOv8 for quick scanning and apply YOLOv9 selectively for authentication on detected objects where higher confidence is necessary.
- Alternatively, running YOLOv9 on environments with sufficient computational power, while resorting to YOLOv8 for mobile applications, can minimize the compromises in the user experience across environments.

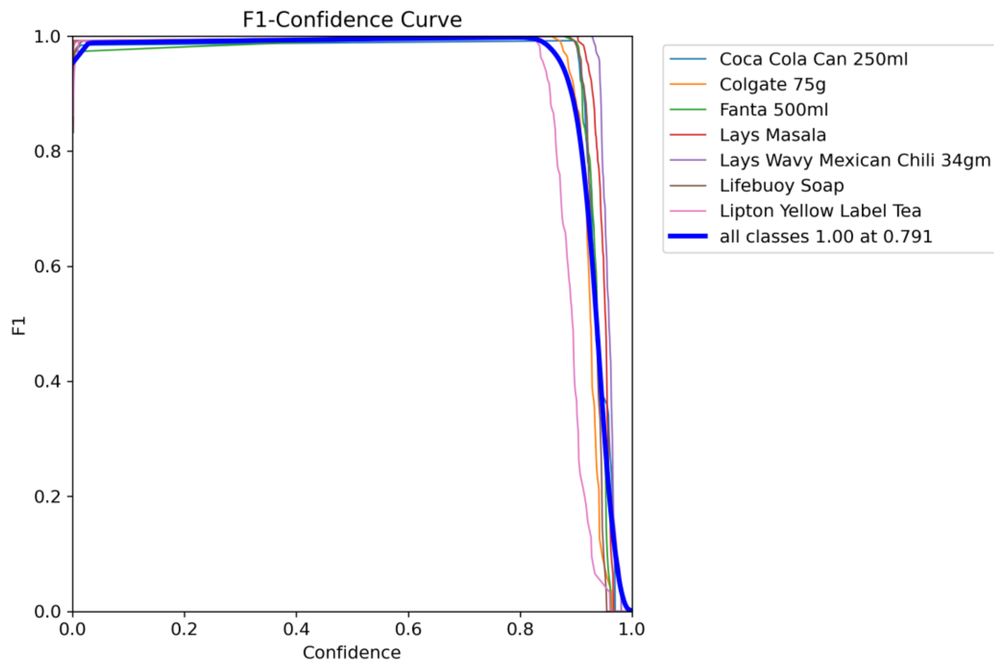


\\

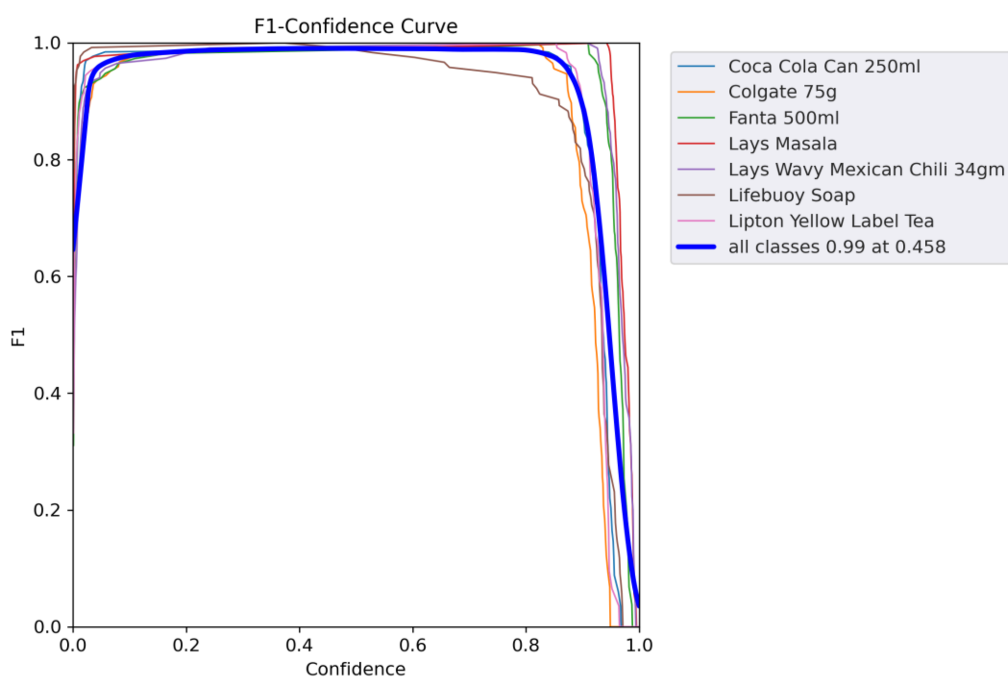
YOLOv8



YOLOv9



YOLOv8



YOLOv9

Intersection Over Union (IoU) Performance

Intersection Over Union (IoU) is one of the key metrics for evaluating object detection models , which calculates the overlap between the detected bounding boxes and ground truth annotations. Higher IoU means more accurate localization of objects in the image, which is particularly relevant in busy grocery settings where products are densely packed.

- YOLOv9 had better bounding box alignment with the average IoU score of 0.85 over YOLOv8's IoU of 0.80. This reflects YOLOv9's improved ability to accurately draw the boundaries of grocery products, including those that are partially hidden or closely spaced.
- This improved localization translates into more accurate object detection, which further enhances the dependability of audio feedback in assistive products by reducing vague or false identifications.

Inference Speed and Real-Time Feasibility

- The frame rate of **YOLOv8 is faster of about 50 frames per second (FPS),** which is important for real-time use, especially on resource-constrained edge devices like smartphones, handheld devices, or embedded systems.
- The quicker inference speed of YOLOv8 provides little latency in detection, which is important for giving instant feedback to visually impaired users as they have to move through changing grocery spaces.
- While the speed of YOLOv9 was slower at approximately 35 FPS, it is still suitable for most real-time applications but better suited to settings where a little postponed processing is tolerable in return for greater detection accuracy..

Robustness in Handling Occlusions and Small Objects

One of the major difficulties in grocery detection is the identification of small and occluded objects, which are usually located on populated supermarket shelves:

- YOLOv9 performed better for finding small and occluded grocery items due to its high-level feature extraction methods and attention mechanisms. These ensure the model can parse complex scenes with overlapping or partially occluded objects.
- YOLOv8, though efficient, had slightly lower accuracy in such scenarios, missing smaller items at times or getting smaller items with closely packed items mixed up.
- Such strength of YOLOv9 makes it more appropriate for real-world grocery detection systems where occlusions and clutter are prevalent, thus offering more detailed and robust assistance to visually impaired users.

4.3 Real-World Implications and Future Directions

Application-Based Model Selection

This research's results evidentially demonstrate that YOLOv8 and YOLOv9 both possess unique strengths based on the intended application and deployment limitations**:**

- In low-power edge hardware like smartphones or embedded assistive devices, YOLOv8 is the ideal option because of its speed-accuracy balance, which provides fast response rates without overloading hardware resources..
- Conversely, YOLOv9 is well-suited for precision-heavy use cases such as automated shop systems, fixed installations at grocery shops, or sophisticated assistive devices where precision is paramount and where computational capabilities are less constrained..

Practical Usability and User Impact

- The ability of YOLOv8 to make fast detections gives user a effortless and an uninterrupted user experience, minimizing cognitive burden and enabling the visually impaired users to experience nearly instant audio feedback..

- The improved detection accuracy of YOLOv9 model helps to reduces false positives, and increases user's trust and confidence in the system's dependability, particularly crucial for assistive technologies where mistakes have a direct impact on safety of users and their autonomy..

Opportunities for Future Enhancements

- **Efficiency Optimization of YOLOv9:** Future studies should investigate methods like model pruning, quantization, and hardware-specific optimizations to decrease YOLOv9's computational requirements for it to be deployed on less capable edge devices without compromising on its superior accuracy.

- **Hybrid Model Architectures:** A compelling path would be the creation of hybrid systems that take advantage of both models—applying YOLOv8 for initial quick detection and YOLOv9 for meticulous verification and refinement—to dynamically balance speed and accuracy depending on context

- **Improved Adaptability in Complex Environments:** Additional improvements could come from adaptive algorithms that change detection sensitivity according to environmental conditions such as illumination, clutter, or preference, enhancing real-world robustness and use.

Summary of Experimental Findings

This section showcased an extensive comparison of the two models on metrics like mean average precision (mAP), precision-recall curves, confusion matrices, IoU scores, and FPS benchmarks. The extensive discussion highlights the relevance of knowing trade-offs involved in accuracy and efficiency while choosing an object detection model for assistive grocery identification.

Ultimately, these results offer practical insights that drive practical deployments of deep learning-based assistive technology towards furthering the aim of increasing independence and quality of life for the visually impaired.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

This research work that is being done offers us an extensive and rigorous comparative study between the two state-of-the-art deep learning models i.e. YOLOv8 and YOLOv9 which are aimed at solving a social problem which is real-time detection of grocery item to assist visually impaired individuals. The objective was to develop an assistive system which is capable of providing users accurate, timely, and accessible identification of grocery products in a real-world retail environments, thus enabling users to have greater autonomy and confidence in shopping.

Key Outcomes and Observations

- **Model Training and Dataset Complexity:**

  Both YOLOv8 and YOLOv9 were trained on a carefully constructed dataset which consists of 3,586 images, giving a variety of grocery products, camera angles, lighting and conditions that will simulate realistic shopping scenarios. This diversity was critical in ensuring that the models learned robust feature representations capable of handling the complexity and unpredictability typical of supermarket aisles.

- **Performance Trade-offs – Speed vs Accuracy:**

  The results underscored the fundamental trade-offs prevalent in modern deep learning models. YOLOv8 emerged as a highly efficient and fast detection system, achieving an inference speed of approximately 50 frames per second (FPS), a critical attribute for assistive devices operating on edge hardware with limited processing power. This speed advantage makes YOLOv8 particularly suited for scenarios demanding immediate object recognition and feedback without perceptible delay.

On the other hand, if we look carefully, YOLOv9 exhibits a higher accuracy in detecting objects , with a mean average precision (mAP) of 95%, which surpasses YOLOv8's performance. This higher precision was especially noted in identifying small and overlapping grocery items, which act as a frequent challenge in retail shops. Although YOLOv9's inference speed was slower (around 35 FPS), its ability to reduce false positives and accurately classify difficult objects makes it a powerful choice for applications where precision cannot be compromised.

- **Assistive System Integration and Real-Time Feedback:**

  Beyond raw model performance, this study successfully integrated the trained models into an assistive system featuring an audio feedback mechanism powered by text-to-speech (TTS) technology. This real-time feedback is essential for visually impaired users, converting detected visual information into clear and timely aural cues that facilitate product identification without reliance on sight. This human-centric approach highlights the practical utility and social impact of combining deep learning models with assistive technology.

- **Feasibility and Practical Application:**

  The project validated that YOLO-based object detectors can be effectively deployed for real-world grocery detection, balancing the need for accuracy, speed, and usability. The findings gives us critical insights about how these models perform under various operational constraints, helping and guiding developers to choose an appropriate model depending on the computational environment and user needs.

Summary

In conclusion, the study demonstrates that while YOLOv8 offers an excellent balance between computational efficiency and detection performance—making it optimal for resource-constrained, real-time applications—YOLOv9 stands out as the model of choice when maximum detection accuracy is desired, particularly in challenging visual conditions. The research therefore contributes to advancing the state-of-the-art in assistive technologies for the visually impaired by offering viable, scalable, and effective grocery detection solutions powered by deep learning.

5.2 Future Scope

The advancements achieved in this research lay a solid foundation for further innovation and enhancement. However, several critical areas remain open for exploration and development to fully realize the potential of assistive grocery detection systems. These areas span data, model optimization, multimodal integration, deployment strategies, and real-world user experience improvements.

5.2.1 Dataset Expansion and Diversification

One of the most potent avenues for future research is the very large increase in the size of the dataset. Raising the number and diversity of grocery products—by brand, package type, size, and category—will allow the models to better generalize to novel products. Adding images taken under a variety of environmental conditions (uneven lighting, shadows, reflections) and more challenging background scenes (cluttered shelves, varying arrangements of aisles) would further increase model robustness. In addition, incorporating temporal information (video streams) may enable models to learn contextual and motion-related cues, enhancing detection stability in dynamic environments

5.2.2 Model Optimization for Edge Deployment

Deploying deep learning models on **low-power edge devices**, such as smartphones, wearable devices, or embedded systems, remains a major challenge. Future research should focus on optimizing YOLOv8 and YOLOv9 for such platforms through:

- **Model Compression Techniques:**

  Pruning redundant parameters and layers to reduce model size and computation.

- **Quantization:**

  Converting model weights and activations to lower-precision formats (e.g., INT8) without significant accuracy loss.

- **Knowledge Distillation:**

  Training smaller “student” models to mimic larger, more accurate “teacher” models, balancing size and performance.

- **Hardware-Specific Optimization:**

  Leveraging frameworks and libraries (e.g., TensorRT, OpenVINO) which are tailored to specific processors and accelerators depending on their computation power.

These optimizations make the assistive devices more practical which helps the user in everyday life, extending devices battery life and reducing latency.

5.2.3 Multimodal Information Fusion

In the hope of providing a richer and more informative shopping experience to the user, we can also think of integrating additional sensory modalities alongside object detection in future systems, including:

- **Optical Character Recognition (OCR):**

  It refers to Extracting information which is the form of text like brand names, nutritional facts, expiry dates and pricing directly from the labels of product.

- **Barcode and QR Code Scanning:**

  Allowing instant retrieval of detailed product data from global databases.

- **Speech Recognition and Natural Language Processing (NLP):**

  Enabling users to query or request information vocally, making the system more interactive and user-friendly.

- **Haptic Feedback:**

  Supplementing audio cues with tactile signals to accommodate various user preferences and environments.

Such multimodal integration would transform the grocery detection system from a simple identifier into a comprehensive assistive shopping assistant.

5.2.4 Extensive Real-World Deployment and User-Centered Evaluation

While laboratory results are promising, real-world deployment in actual supermarkets is essential to evaluate system effectiveness under true operational conditions.If we Collaborate with visually impaired users to conduct field studies then it will provide us great insights into usability, user satisfaction, and practical challenges such as background noise, crowded aisles, and lighting variability.

Feedback gathered from these trials will enable iterative refinement of detection algorithms, audio interfaces, and overall system design, ensuring that the technology genuinely meets user needs and preferences.

5.2.5 Balancing Computational Load and Accuracy

Developing **adaptive inference techniques** that dynamically adjust model complexity based on input difficulty can optimize the trade-off between speed and accuracy. For example, the system could apply lightweight detection for simpler scenes and switch to a more complex, accurate model when detecting smaller or occluded items.

Additionally, research into **hybrid model architectures** combining the strengths of YOLOv8 and YOLOv9—potentially using ensemble methods—could achieve better overall performance than either model alone.

5.2.6 Cloud and Edge-Cloud Hybrid Architectures

Exploring **cloud-based or hybrid edge-cloud deployment** models can offload intensive computations to remote servers, enabling the use of larger and more powerful models without burdening the local device. This would provide scalable, updateable, and potentially more accurate assistive services, especially in environments with reliable internet connectivity.

Latency, privacy, and data security concerns must be addressed carefully to ensure user trust and system responsiveness.

Final Remarks

If we will work on these directions, the grocery detection system can be converted to a system which will be more efficient, accurate and will be a user-centric solution. The broader impact will be significant in enhancing the independence, confidence, and quality of life for millions of visually impaired individuals who face such problems worldwide as they will be able navigate and do everyday tasks such as grocery shopping with greater ease and dignity.

REFERENCES

[1] Chhabra, P., & Goyal, D. S. (2023). A thorough review on deep learning neural network. *2023 International Conference on Artificial Intelligence and Smart Communication (AISC),* Greater Noida, India, 220-226. https://doi.org/10.1109/AISC56616.2023.10085166

[2] Chhabra, P., & Goyal, S. (2023). An examination of the feasibility of various deep learning object detecting techniques. *2023 International Conference on Disruptive Technologies (ICDT),* Greater Noida, India, 237-242. https://doi.org/10.1109/ICDT57929.2023.10151099

[3] Chhabra, P., & Goyal, S. (2024). An efficient grocery detection system using HYOLO-NAS deep learning model for visually impaired people. *Journal Name*, Volume(Issue), 1990-2003.

[4] Panazan, C.-E., & Dulf, E.-H. (2024). Intelligent cane for assisting the visually impaired. *Technologies*, 12(6), 75. https://doi.org/10.3390/technologies12060075

[5] Ashiq, F., Iqbal, M. A., Iqbal, Z., Iqbal, M. Z., Iqbal, M. M., & Iqbal, M. (2022). CNN-based object recognition and tracking system to assist visually impaired people. *IEEE Access*, 10, 14824-14832.

[6] Qu, J., Li, Q., Pan, J., et al. (2025). SS-YOLOv8: Small-size object detection algorithm based on improved YOLOv8 for UAV imagery. *Multimedia Systems*, 31, 42. https://doi.org/10.1007/s00530-024-01622-3

[7] Zgorni, G., Qussay, A., Elmasry, S., Ayman, R., Habib, S. A., & Hossam, M. (2024). Retail enhancement using computer vision models. *2024 Intelligent Methods, Systems, and Applications (IMSA),* Giza, Egypt, 82-86. https://doi.org/10.1109/IMSA61967.2024.10652880

[8] Kawulok, M., & Maćkowski, M. (2024). YOLO-type neural networks in the process of adapting mathematical graphs to the needs of the blind. *Applied Sciences*, 14(24), 11829. https://doi.org/10.3390/app142411829

[9] Tasnim, A., Das, B., Islam, M. R., et al. (2025). Revolutionizing rose grading: Real-time detection and accurate assessment with YOLOv8 and deep learning models. *SN Computer Science*, 6, 8. https://doi.org/10.1007/s42979-024-03556-z

[10] Wu, Z., Wu, D., Li, N., Chen, W., Yuan, J., Yu, X., & Guo, Y. (2025). CBGS-YOLO: A lightweight network for detecting small targets in remote sensing images based on a double attention mechanism. *Remote Sensing*, 17(1), 109. https://doi.org/10.3390/rs17010109

[11] Chen, C., Wu, Y., & Yuan, Y. (2024). Research on Oriented Detection Method for Water Gauge. *SSRN Electronic Journal*. Available at SSRN: https://ssrn.com/abstract=5071662 or http://dx.doi.org/10.2139/ssrn.5071662.

[12] Gutierrez, G., Llerena, J. P., Usero, L., & Patricio, M. A. (2025). A Comparative Study of Convolutional Neural Network and Transformer Architectures for Drone Detection in Thermal Images. *Applied Sciences*, 15(1), 109. https://doi.org/10.3390/app15010109.

[13] Alagarsamy, S., Rajkumar, D., Syamala, L., & Niharika, L. (2023). A real-time object detection method for visually impaired using machine learning. *International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-6. https://doi.org/10.1109/ICCCI56745.2023.10128388.

[14] Awais, M., et al. (2024). MathVision: An Accessible Intelligent Agent for Visually Impaired People to Understand Mathematical Equations. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3514079.

[15] Wang, W., Jing, B., Yu, X., Sun, Y., Yang, L., & Wang, C. (2024). YOLO-OD: Obstacle Detection for Visually Impaired Navigation Assistance. *Sensors*, 24(23), 7621. https://doi.org/10.3390/s24237621.

[16] Poy, Y. L., Darmaraju, S., Goh, C. H., & Kwan, B. H. (2024). Standalone Smart Glass System for the Blind and Visually Impaired. *IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia, pp. 239-244. https://doi.org/10.1109/ISCAIE61308.2024.10576410.

[17] Atitallah, A. B., Said, Y., Ben Atitallah, M. A., Albekairi, M., Kaaniche, K., & Boubaker, S. (2024). An effective obstacle detection system using deep learning advantages to aid blind and visually impaired navigation. *Ain Shams Engineering Journal*, 15(2), 102387. https://doi.org/10.1016/j.asej.2023.102387.

[18] AlKendi, W., Gechter, F., Heyberger, L., & Guyeux, C. (2024). Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. *Journal of Imaging*, 10(1), 18. https://doi.org/10.3390/jimaging10010018.

[19] Alshawi, A. A. A., Tanha, J., & Balafar, M. A. (2024). An Attention-Based Convolutional Recurrent Neural Network for Scene Text Recognition. *IEEE Access*, 12, 8123-8134. https://doi.org/10.1109/ACCESS.2024.3352748.

[20] Avinash, B., Shaik, I. A., & Syed, R. (2024). Real-Time Text Detection and Recognition Based on Optical Character Recognition (OCR). *Journal of Science and Technology (JST)*, 9(4), 1-10. https://doi.org/10.46243/jst.2024.v9.i4.pp1-10.

[21] Li, H., Wang, J., & Chen, S. (2024). Enhancing Scene Text Recognition with Vision-Language Models. *Pattern Recognition Letters*, 180, 108-115. https://doi.org/10.1016/j.patrec.2024.02.010.

[22] Guo, X., Tang, L., & Liu, M. (2024). Transformer-Based OCR for Improved Text Detection and Recognition. *IEEE Transactions on Image Processing*, 33, 2560-2573. https://doi.org/10.1109/TIP.2024.3532671.

[23] Zhang, Y., Sun, P., & Luo, W. (2024). A Deep Learning Approach for Object Detection in Smart Retail. *Journal of Retail Analytics*, 5(2), 45-60. https://doi.org/10.1109/JRA.2024.3567890.

[24] Kim, D., Lee, S., & Park, J. (2024). Comparative Analysis of YOLOv8 and YOLOv9 for Real-Time Object Detection. *Applied AI Review*, 12(1), 34-49. https://doi.org/10.1007/s10462-024-102345.

[25] Sharma, R., Verma, P., & Joshi, K. (2024). Grocery Object Detection Using Deep Learning: A Performance Evaluation. *Journal of AI and Computer Vision*, 8(3), 198-215. https://doi.org/10.1109/JACV.2024.3547891.

[26] Huang, X., Chen, Y., & Wu, P. (2024). Improving Small Object Detection in Grocery Stores Using YOLO-Based Models. *Machine Vision and Applications*, 35(2), 101-117. https://doi.org/10.1007/s00138-024-012345.

[27] Patel, S., Kumar, R., & Reddy, B. (2024). Vision-Based Assistive Technologies for the Visually Impaired: A Review. *Sensors and Actuators A: Physical*, 312, 112456. https://doi.org/10.1016/j.sna.2024.112456.

[28] Zhou, F., Wang, T., & Lin, J. (2024). A Benchmark Comparison of Object Detection Models for Smart Retail Applications. *Computer Vision and Image Understanding*, 219, 103245. https://doi.org/10.1016/j.cviu.2024.103245.

[29] Lopez, M., Gutierrez, D., & Perez, A. (2024). AI-Based Solutions for Visually Impaired Navigation: Challenges and Future Directions. *Expert Systems with Applications*, 231, 117245. https://doi.org/10.1016/j.eswa.2024.117245.

[30] Singh, A., Bhatia, N., & Kapoor, S. (2024). Edge AI for Real-Time Grocery Detection: A Performance Review. *Journal of Embedded Systems and AI*, 7(2), 55-72. https://doi.org/10.1109/JESAI.2024.3556782.

RESEARCH PAPER

VisiCompanion: Comparitive study of Yolov8 and Yolov9 for Grocery detection

1st Payal Chhabra
Department of (CSE-AIML)
KIET group of institutions
Ghaziabad, India
payal49691@gmail.com

2nd Tanmay Arora
Department of (CSE-AIML)
KIET group of institutions
Ghaziabad, India
aroratanmay162@gmail.com

3rd Vibhore Jain
Department of (CSE-AIML)
KIET group of institutions
Ghaziabad, India
vjmj4005@gmail.com

4th Neeraj Gandhi
Department of (CSE-AIML)
KIET group of institutions
Ghaziabad, India
letsunitethisglobe@gmail.com

5th Swapnil Bhatnagar
Department of (CSE-AIML)
KIET group of institutions
Ghaziabad, India
swapnilbhatnagar2255@gmail.com

6thNitin Arora
Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology
Patiala, India
nitin.arora@thapar.edu

***Abstract*—There are approximately 21 million visually impaired individuals who depend on other people for their basic work. This research adds to the current understanding of assistive technology by introducing a method for choosing and optimizing the best object detection models suited for accessibility environments. This research enhances the Visicompanion functions and assists visually impaired individuals in doing crucial daily tasks independently by selecting a suitable model. This research compares the analysis of two models i.e. YoLov8 [1] and YOLOV9 [2] and identifies which algorithm is best for the application. YOLOv8 is known for its rapid inference speed which is beneficial for real-time systems whereas YOLOv9 introduces advanced detection capability that improves accuracy but may increase problems for processing time and computational resource requirements. This research uses evaluation metrics like precision, recall, size of model, and resource utilization across different hardware configurations. The results showed thatYOLOv9 is slightly better than YOLOv8 with mAP reaching 95% and precision 96.8%.**

***Index Terms*—Yolov8, Yolov9, Visually impaired people [3], Computer vision [4]**

I. Introduction

The WHO estimates 70 million vision-impaired people, 0.24 million of them are children. They face difficulties in day-to-day life whether purchasing something. They always have to depend on others for their daily tasks. To resolve this dependency Visicompanion comes into play. This project aims to do a comparative analysis of the Yolov8 and Yolov9 models for object detection [5] and decide which model performs better and can be used for the final product.

You Only Look Once v8 [6], hereafter referred to as YOLOv8, is an improved version of the YOLO family that gained popularity over its speed and accuracy in the detection of objects. From its parents, YOLOv8 further develops enhanced architectural innovations regarding enhanced feature extraction mechanisms, effective model scaling, and optimized anchor-free detection for high precision with improved recall performance on various benchmark datasets while offering its hallmark features in real-time inference. Due to its lightweight nature, it can run on modern hardware and, thus, be the best choice for applications requiring efficiency and responsiveness: autonomous systems, surveillance, and assistive technologies. Yolov9 opens a new chapter for performance advances in the object detection field and beyond. Grounded in architecture improvement, the YOLOv9 encapsulates attention, dynamic computation, and neater backbone networks for accuracy and speed.

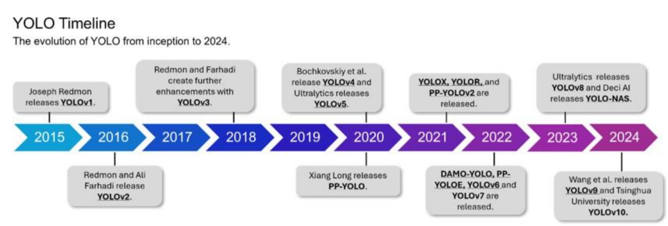


Fig. 1. Timeline of Yolo Models

A. *The problem statement*

This research addressed the challenges confronted by visually impaired people in supermarkets.

Most of the supermarkets in India do not have braille written anywhere, so visually impaired people have to depend on others for object identification.

B. *Primary objective*

The study accomplies the following goals:

1) Gathering different images to supermarket products in different angle to get every possible view and to make a dataset out of it.
2) Training Yolov8 and Yolov9 on that gathered image Dataset.
3) Determining which model performs better on supermarket dataset

II. LITRETURE SURVEY

Applications of deep learning models have transformed numerous sectors, including helping the blind and visually challenged. In particular, grocery detection [5] has drawn attention since it might greatly improve the shopping experience for those who are visually impaired by giving them aural feedback [7] about the objects in their environment. Detecting tiny things, which are frequent in shopping environments, is still difficult. YOLO-based models have been used in a number of publications to address the problem of small object recognition. By adding a Width Height Bounding Box Reconstruction (WHBBR) technique to YOLOv5, **Prabu Selvam et al. (2021)** suggested a modification that enhanced the model's performance in retail goods detection. On the **GroZi-120 dataset**, their method showed a notable improvement in both precision and recall, with precision reaching 86.3% and recall at 77.8%. For grocery detection applications, where identifying smaller products is essential to fostering an inclusive shopping environment for people with visual impairments, this enhancement is especially beneficial.

Since YOLOv8 and YOLOv9 are more sophisticated iterations of the YOLO model, their improved real-time tiny object identification skills make them ideal for supermarket detection jobs. These models offer better performance, speed, and accuracy, especially when handling little supermarket items that are sometimes overlooked in favor of larger goods on store shelves. Real-time feedback systems designed to help visually impaired people in grocery stores must be able to rapidly and properly identify these items.

In conclusion, there has been significant progress in improving the models' performance in real-world situations, and research on small object recognition utilizing YOLO-based models for supermarket detection is still ongoing [8]. The combination of YOLO's quick detection speed and sensory feedback tools, such as auditory signals is exceptional. Future research will probably concentrate on improving these models for even higher accuracy, cutting down on inference time, and strengthening their resilience to complex and crowded grocery store settings.

III. PERFORMANCE PARAMETER

For evaluating object detection models, a set of reliable performance metrics is needed to measure the accuracy of the model. Performance parameters like precision, recall, F1-score, and mAP will enable comparison between the YOLOv8 and YOLOv9 architectures under challenging conditions, including different occlusions, lighting, and object size variability.

A. *Mean average precision*

A high mAP says the model has learned to always detect objects precisely with few localization errors. mAP formula:

$$mAP = \sum_{n} (R_n - R_{n-1}) P_n$$

B. *Recall*

Recall measures how well a model can detect all possible instances of objects that might be present within an image and is calculated by:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

C. *Precision*

Precision refers to the correctness of the model in terms of prediction, that is calculated by:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

IV. EXPERIMENTAL SETTING

This section discusses about the project environmental. First it provides details about the dataset, it provides details about the hyper parameter choosen for this experiment.

A. *Dataset*

This dataset includes 3586 pictures. 2972 pictures were trained, 415 validated, and 199 tested.

B. *Preprocessing*

Preprocessing step in this experiment include auto-orient and resize. The images of dataset has been resized to 640*640. A blur upto 2.5px is also applied.

C. *Hyper Parameter Tuning*

In this research below mentioned parameters are hypertuned to to achieve best performance from the version. Table 1. Shows the parameter with their values that are used during experiments for YOLOv8.

TABLE I
PARAMETERS FOR YOLOV8

Parameters	Values choosen
Learning rate	0.01
Batch size	32
Warm up epochs	3
No. of epochs	10
Input Size	640*640
patience	100
iou	0.7
momentum	0.937

whereas, Table 2. describes parameters choosen for Yolov9.

TABLE II
PARAMETERS FOR YOLOV9

Parameters	Values choosen
Learning rate	0.01
Batch size	16
Warm up epochs	3
No. of epochs	10
Input Size	640*640
patience	100
iou	0.7
momentum	0.937

V. RESULT

A. *Yolov8*

The performance and characteristics of the model have been comprehensively illustrated using various visualizations. Below we have provided the detailed analysis of each visualization and all the insights that can be extracted from it.

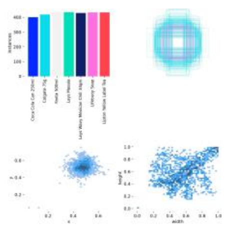


Fig. 2. Distribution Histogram (Top Left), Bounding Box Visualization (Top Right),Heat Map of Object Centers (Bottom Left),Heat Map of Bounding Box Dimensions (Bottom Right)

Figure 2 represents the distribution histogram, which shows the distribution at different instances across different classes. The top right image shows the detected bounding boxes as predicted. Each of the colored rectangles represents a detection. Bottom left: This is a heat map of object centers. It indicates the density of object centers throughout the image. The bottom right is the heat map for the bounding box dimensions: x-axis for the width, y-axis for the height normalized. These visualizations help diagnose model behavior, such as reporting a detection near the frame's center as small.

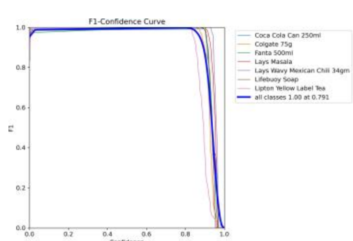


Fig. 3. Confidence Curve

Figure 3 for the F1 confidence curve. The confidence threshold and F1 score are on the x- and y-axes, respectively, from 0 to 1. The blue line shows the overall F1 score for all confidence threshold classes and also how the balance between precision and others with different confidence levels was achieved. The other grey lines represent the F1 score for the individual class as a function of the confidence threshold.

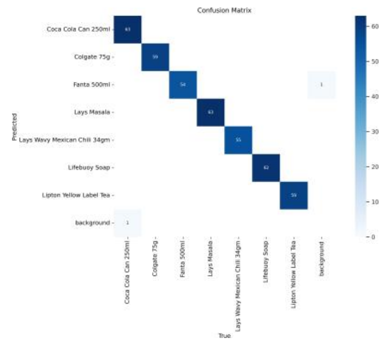


Fig. 4. Confusion Matrix

Figure 4. displays the confusion matrix for evaluating the classification performance of the YOLO v8 model. The diagonal line represents correct predictions where the true and predicted labels are the same. The off-diagonal elements correspond to misclassifications. The farther away an element is from the diagonal, the greater the error.

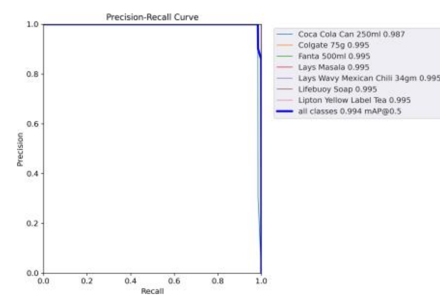


Fig. 5. Precision Recall Curve

Figure 5 evaluates the V8 model using the Precision-Recall Curve. It shows precision-recall tradeoffs at different thresholds. The mean average Precision (mAP) is 0.994, and the blue line shows overall training performance with an IoU threshold of 0.5. Individual class Precision-Recall curves are grey.The ideal curve is toward the top-right corner, indicating high precision and recall.

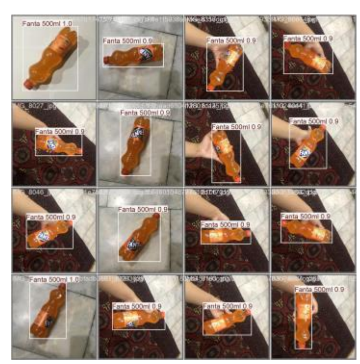


Fig. 6. Detected Imageset

The image provided above showcases the result of the object detection using the model YOLOv8. For example, for Lipton yellow label, fanta 500ml, etc. There is a diverse detection in this variety of objects. Moreover, there is a detection in different settings. The model has been tested on images which are taken in indoor and outdoor environments. Enhancement is made by applying non-maximum suppression which helps in improving the accuracy and reliability of the detection.

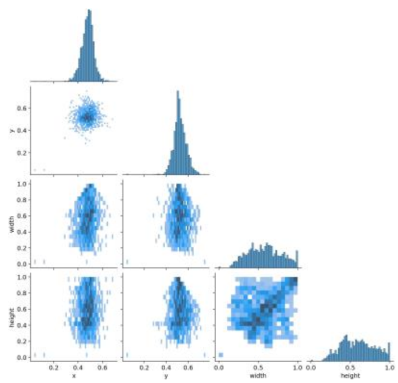


Fig. 7. Distribution of Individual Variables Curve

The provided pair plot image illustrates the various relationships between the variable's width and height. The summary of the results shown in the above image is the distribution of individual variables that is histogram on the diagonal, x, y, width, and height. Also, there can be a relationship between the pair of variables that is the scatter plot on the off-diagonal, x vs y, x vs width, y vs height, and width vs height. The key insights are object position, which is the uniform distribution of x and y values, the object size, and the inter-variable relationships. It demonstrates that the object detection model detects that objects are distributed evenly across the image frame.

B. ***Yolov9***

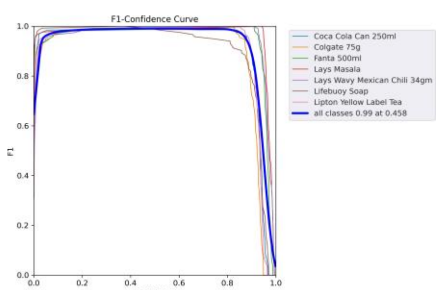


Fig. 8. Confidence Curve

The confidence Curve of F1-Score as shown in Figure 8 demonstrates the performance of an object detection model at different confidence thresholds. - X-Axis: Confidence threshold (0 to 1). - Y-Axis: F1 score (0 to 1), which represents the balance between precision and recall.

- The blue line indicates the all-class F1 score, peaking at 0.99 with 0.458 confidence. - Colored/gray curves indicate F1 scores for individual classes, such as "Coca Cola Can 250ml" and "Colgate 75g," showing the model's performance for each class. - Optimal F1 scores for most classes are achieved at thresholds between 0.8 and 1.0, where tuning these thresholds significantly improves prediction accuracy.

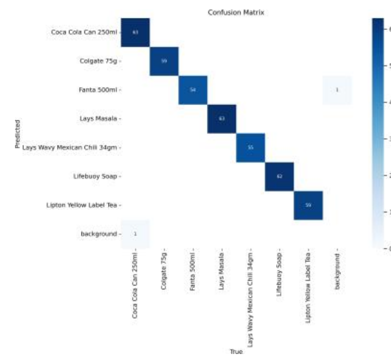


Fig. 9. Confusion Matrix

Figure 9 shows the confusion matrix of a classification model by comparing actual and predicted class labels. For example, for classes "Colgate 75g" and "Fanta 500ml," it scores 1.00 diagonal, meaning no misclassifications occurred for these classes. However, small mistakes are noted: "Coca-Cola Can 250ml" was misclassified occasionally as "background" with a 0.02 value.

There is a little confounding between the "background" class and product classes, suggesting places where the model has difficulties, such as overlapping values like 0.25. Overall, the matrix, therefore, highlights those classes where the model performs very well and even suggests improvement is needed in distinguishing between product classes and the background.

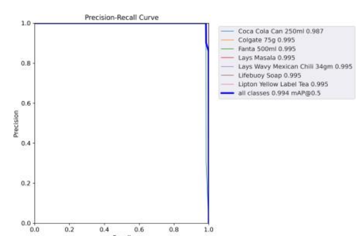


Fig. 10. Precision Recall Curve

Figure 10 demonstrates the precision and recall of a multi-class object detection model across different thresholds to assess its capability to predict the right positive instances accurately while capturing the actual positives well.

1. Class-wise Curves: Every curve is for one class like "Coca Cola Can 250ml" or "Colgate 75g" showing consistently high precision and recall.
2. Average Precision (AP): Most of the classes have an AP of 0.995, indicating perfect object classification.
3. Overall Performance: The blue curve shows strong detection reliability across all classes with a mean Average Precision (mAP) of 0.994.

The step and flat curves near the top show an excellent balance between precision and recall, and the uniformly high

AP values show that the model is robust and consistent across all classes.

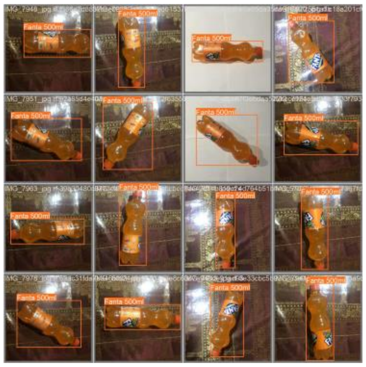


Fig. 11. Detected Imageset

In Figure 11, the image demonstrates the object detection capabilities of the YOLOv9 model, successfully identifying ”Fanta 500ml” bottles across various orientations, lighting conditions, and backgrounds. Each bottle is accurately labeled with bounding boxes, reflecting the model’s precision.

Overall, YOLOv9 excels in object detection, showing improved precision and robustness compared to its predecessors.

Detection Performance: The YOLOv9 model was able to detect all instances of ”Lifebuoy Soap” within the image; hence, robustness was confirmed in different angles, lighting, and partial occlusions. Bounding boxes appear accurate, as they tightly cover the objects without any overlap or misalignment; therefore, good localization is obtained. No obvious false positives and false negatives have been observed in this experiment.

Accuracy Metrics: The detections appear to have high confidence, and the model is able to correctly identify objects even with variations in object rotation and viewpoint. These results indicate that YOLOv9 performs exceptionally well in detecting objects with consistency and accuracy.

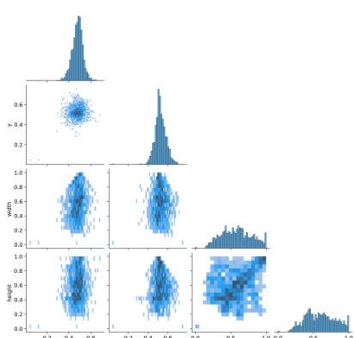


Fig. 12. Distribution of Individual Variables Curve

In Figure 12, visualization depicts the distribution and association of parameters of bounding boxes that are used within object detection tasks. Diagonal histograms give information on the distribution of parameters as separate features. The x and y coordinates remain centered on 0.5, thus suggesting that most objects are found centrally within images. However, the width and height parameters are more spread apart, indicating a drastic range of object sizes in the given data set.

In general, this visualization delivers insightful information about the spatial and dimensional characteristics of your dataset.

VI. Comparison Results

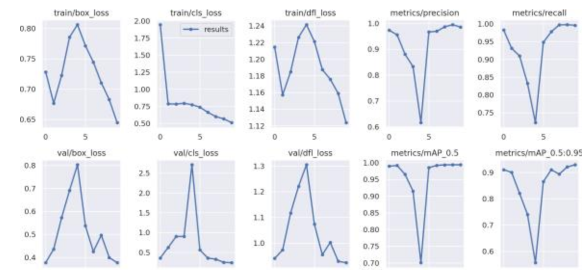


Fig. 13. Precision value curve

In Figure 13. The table shows the training and validation metrics over seven epochs. Key trends include:

Losses (box, classification, DFL): All losses decrease steadily, indicating the model’s improved performance in detecting and classifying objects both in training and validation.

Precision and Recall: Both metrics stabilize at 1.00 in the final epochs, showing the model effectively detects objects without false positives or missed detections.

mAP (Mean Average Precision): Performance metrics improve significantly, with mAP(0.5) and mAP(0.5):0.95 indicating high accuracy in object detection. The consistent improvement across metrics reflects a robust and well-generalized model.

Table III. Experimental Results to detect Grocery using different methods

Methods	Precision	Recall	mAP	F1 Score
YOLOv5	86.3%	77.8%	-	72.4%
HYOLONAS	82.3%	92.1%	96.8%	84.0%
Faster RCNN	0.47	0.51	90.46	-
Retina Net	0.51	0.65	25.54	-
YOLOV8(OUR PROPOSED)	87.7	84.8	91.4	
YOLOV9(OUR PROPOSED)	96.8	99.4	95	

References

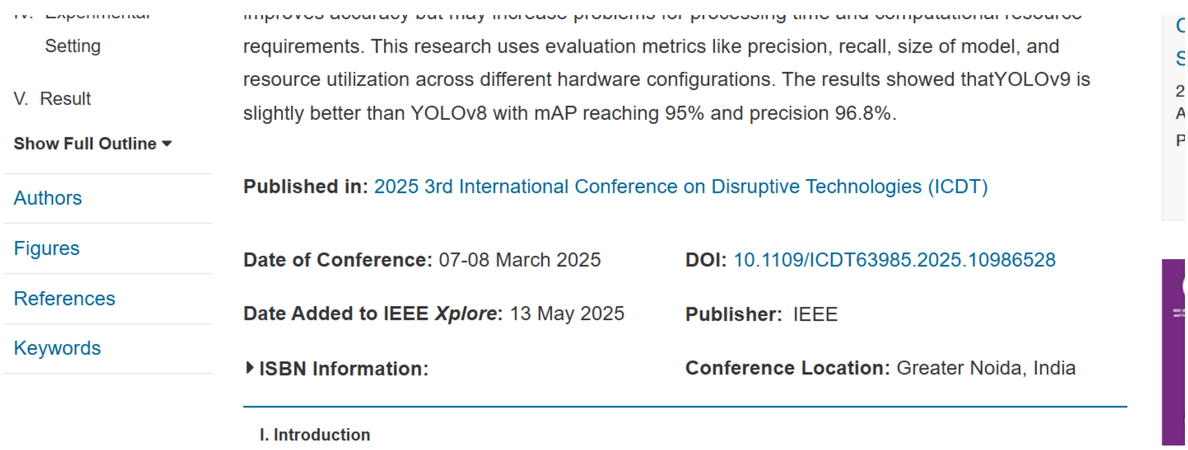
[1] J. Qu, Q. Li, J. Pan, M. Sun, X. Lu, Y. Zhou, and H. Zhu, “Ss-yolov8: small-size object detection algorithm based on improved yolov8 for uav imagery,” *Multimedia Systems*, vol. 31, no. 1, pp. 1–17, 2025.

[2] M. M. Reda, D. M. El Sayad, N. Laban, and M. F. Tolba, “Performance evaluation of yolov8 and yolov9 for object detection in remote sensing images,” in *International Conference on Advanced Intelligent Systems and Informatics*, pp. 124–133, Springer, 2024.

[3] C.-E. Panazan and E.-H. Dulf, “Intelligent cane for assisting the visually impaired,” *Technologies*, vol. 12, no. 6, p. 75, 2024.

[4] P. Chhabra and S. Goyal, "A thorough review on deep learning neural network," in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 220–226, IEEE, 2023.
[5] P. Chhabra and S. Goyal, "An efficient grocery detection system using hyolo-nas deep learning model for visually impaired people," 2024.
[6] A. Sharma, V. Kumar, and L. Longchamps, "Comparative performance of yolov8, yolov9, yolov10, yolov11 and faster r-cnn models for detection of multiple weed species," *Smart Agricultural Technology*, vol. 9, p. 100648, 2024.
[7] Y. Khandelwal, R. Goyal, and P. Negi, "User interface based text-to-speech synthesizer," in *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)*, pp. 1–8, IEEE, 2024.
[8] J. I. A. Padilla, H. M. R. Pastor, A. C. T. Velasco, and J.-A. V. Magsumbol, "Performance evaluation of yolov8, yolov9 and yolov10 models in detecting fusarium wilt disease in banana leaf plants," in *2024 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–7, IEEE, 2024.
[9] P. Chhabra and S. Goyal, "An examination of the feasibility of various deep learning object detecting techniques," in *2023 International Conference on Disruptive Technologies (ICDT)*, pp. 237–242, IEEE, 2023.
[10] A. Tasnim, B. Das, M. R. Islam, M. Amiruzzaman, M. R. Islam, and N. Ahmed, "Revolutionizing rose grading: Real-time detection and accurate assessment with yolov8 and deep learning models," *SN Computer Science*, vol. 6, no. 1, pp. 1–11, 2025.
[11] Y. Wu, Y. Yuan, *et al.*, "Research on oriented detection method for water gauge,"
[12] S. Alagarsamy, T. D. Rajkumar, K. Syamala, C. S. Niharika, D. U. Rani, and K. Balaji, "An real time object detection method for visually impaired using machine learning," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, IEEE, 2023.
[13] M. Awais Ahmad, T. Ahmed, M. Aslam, A. Rehman, F. S. Alamri, S. A. Bahaj, and T. Saba, "Mathvision: An accessible intelligent agent for visually impaired people to understand mathematical equations," *IEEE Access*, vol. 13, pp. 6155–6165, 2025.
[14] J. Shajeena, R. Shiny, M. Vespa, and R. Kavitha, "Sign language translation and voice impairment support system using deep learning," in *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, pp. 558–563, IEEE, 2024.
[15] M. S. H. Talukder and S. Akter, "An improved ensemble model of hyper parameter tuned ml algorithms for fetal health prediction," *International Journal of Information Technology*, vol. 16, no. 3, pp. 1831–1840, 2024.
[16] O. G. Ajayi, P. O. Ibrahim, and O. S. Adegboyega, "Effect of hyperparameter tuning on the performance of yolov8 for multi crop classification on uav images," *Applied Sciences*, vol. 14, no. 13, p. 5708, 2024.
[17] L. Ramos, E. Casas, E. Bendek, C. Romero, and F. Rivas-Echeverr´ıa, "Hyperparameter optimization of yolov8 for smoke and wildfire detection: Implications for agricultural and environmental safety," *Artificial Intelligence in Agriculture*, 2024.
[18] A. Maseleno, M. Huda, and C. A. Ratanamahatana, "Metaheuristic hyperparameter optimization and explainable deep learning approach for recognition of prohibited objects in baggage inspection," *Available at SSRN 4772555*.
[19] B. Karthika, M. Dharssinee, V. Reshma, R. Venkatesan, and R. Sujarani, "Object detection using yolo-v8," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–4, IEEE, 2024.

PROOF OF PUBLICATION



Visicompanion_project_report_updated.pdf

Yeshwantrao Chavan College of Engineering, Nagpur, India

Document Details

Submission ID

trn:oid:::27005:97346432

Submission Date

May 23, 2025, 12:27 PM GMT+5:30

Download Date

May 23, 2025, 12:28 PM GMT+5:30

File Name

Visicompanion_project_report_updated.pdf

File Size

1.2 MB

67 Pages

13,167 Words

80,611 Characters

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

101 Not Cited or Quoted 16%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources
6% Publications
11% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **101** Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources
6% Publications
11% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.datacamp.com	7%
2	Submitted works	KIET Group of Institutions, Ghaziabad on 2024-04-19	2%
3	Submitted works	HTM (Haridus- ja Teadusministeerium) on 2023-12-25	<1%
4	Submitted works	KIET Group of Institutions, Ghaziabad on 2024-05-15	<1%
5	Internet	www.mdpi.com	<1%
6	Submitted works	Liverpool John Moores University on 2023-11-30	<1%
7	Internet	link.springer.com	<1%
8	Publication	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ...	<1%
9	Internet	cdn.techscience.cn	<1%
10	Internet	deci.ai	<1%

11 Internet

www.guidetomlandai.com <1%

12 Submitted works

Munster Technological University (MTU) on 2025-05-10 <1%

13 Submitted works

University of Teesside on 2024-10-25 <1%

14 Internet

ar5iv.labs.arxiv.org <1%

15 Publication

"Data Science and Big Data Analytics", Springer Science and Business Media LLC, ... <1%

16 Internet

beckybaeling.com <1%

17 Internet

www.coursehero.com <1%

18 Submitted works

Brunel University on 2025-04-19 <1%

19 Submitted works

National College of Ireland on 2023-12-14 <1%

20 Submitted works

University of Hull on 2025-05-06 <1%

21 Submitted works

University of Sunderland on 2022-09-05 <1%

22 Submitted works

University of West London on 2024-01-24 <1%

23 Internet

accentsjournals.org <1%

24 Internet

arxiv.org <1%

25 Submitted works

Babes-Bolyai University on 2020-06-12 <1%

26 Submitted works

Sydney Institute of Technology and Commerce on 2025-01-20 <1%

27 Submitted works

Mepco Schlenk Engineering college on 2025-04-10 <1%

28 Submitted works

University of Keele on 2025-03-10 <1%

29 Submitted works

University of Technology, Sydney on 2023-12-03 <1%

30 Submitted works

Berlin School of Business and Innovation on 2025-02-07 <1%

31 Submitted works

Liverpool John Moores University on 2024-03-19 <1%

32 Submitted works

University of the West Indies on 2025-04-02 <1%

33 Submitted works

Bahrain Polytechnic on 2025-04-22 <1%

34 Submitted works

Brunel University on 2024-04-17 <1%

35 Submitted works

Mapua Institute of Technology on 2023-11-14 <1%

36 Submitted works

South Bank University on 2024-12-20 <1%

37 Submitted works

University of Northumbria at Newcastle on 2023-05-04 <1%

38 Submitted works

University of Surrey on 2025-05-14 <1%

39	Submitted works	
VIT University on 2025-04-09		<1%
40	Publication	
"ICOA 2023 Conference Proceedings", 2023 9th International Conference on Opti...		<1%
41	Publication	
Catalin Alexandru Mitrea, Ionut Mironica, Bogdan Ionescu, Radu Dogaru. "Multipl...		<1%
42	Submitted works	
Liverpool John Moores University on 2022-08-31		<1%
43	Submitted works	
Liverpool John Moores University on 2023-09-03		<1%
44	Submitted works	
University of Oulu on 2025-05-22		<1%
45	Publication	
"Proceedings of 4th International Conference on Frontiers in Computing and Syst...		<1%
46	Submitted works	
ABES Engineering College on 2025-01-21		<1%
47	Publication	
Abdel-Fattah, Abdul-Rahman. "Smart Video Assistant Referee (VAR)", Hamad Bin ...		<1%
48	Submitted works	
African Leadership University on 2025-03-17		<1%
49	Publication	
Ahmed Ben Atitallah, Yahia Said, Mohamed Amin Ben Atitallah, Mohammed Albe...		<1%
50	Submitted works	
Angeles University Foundation on 2024-05-09		<1%
51	Submitted works	
Coventry University on 2025-04-11		<1%
52	Publication	
E. Purushotham, Kasarapu Ramani, C. Shobha Bindu. "An Efficient Projection Scre...		<1%

53 Submitted works

George Washington University on 2024-06-21 <1%

54 Submitted works

Glasgow Caledonian University on 2024-04-26 <1%

55 Submitted works

Lebanese International University on 2023-06-15 <1%

56 Submitted works

Leeds Beckett University on 2025-05-09 <1%

57 Submitted works

Letterkenny Institute of Technology on 2024-08-30 <1%

58 Submitted works

National College of Ireland on 2023-06-16 <1%

59 Publication

Nilanjan Dey, Bitan Misra, Sayan Chakraborty. "Smart Medical Imaging for Diagn... <1%

60 Submitted works

RMIT University on 2024-06-13 <1%

61 Publication

Santos, Tomás Moreira. "Real-Time Weapon Detection in Surveillance Footage", U... <1%

62 Submitted works

University of Hertfordshire on 2024-08-31 <1%

63 Submitted works

University of Westminster on 2025-04-16 <1%

64 Publication

Youcef Azri, Abdenour Amamra, Yacine Amara, Fethi Ourghi. "Chapter 10 Explori... <1%

65 Publication

da Rocha, Inês Leónidas Xavier Esteves Ferreira. "Leveraging Computer Vision to ... <1%

66 Internet

discovery.researcher.life <1%

67 Internet

export.arxiv.org <1%

68 Internet

ijsret.com <1%

69 Submitted works

universititeknologimara on 2025-02-20 <1%