OBJECT DETECTION FOR RETAIL PRODUCT RECOGNITION

Nakul Pannoy

A Thesis in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate Studies
Thai-Nichi Institute of Technology
Academic Year 2024

Thesis Topic	Object Detection for Retail Product Recognition
By	Nakul Pannoy
Field of Study	Information Technology
Thesis Advisor	Dr. Sarayut Nonsiri

The Graduate Studies of Thai-Nichi Institute of Technology has been approved and accepted as partial fulfillment of the requirement for the Master's Degree

……………………………………….Vice President for Academic Affairs
(Assoc. Prof. Dr. Warakorn Srichavengsup)
Month……… Date……… Year…….…….

Thesis Committees

……..………………………………………….Chairman
(Assoc. Prof. Dr. Annop Munsakul)

……..…………………………………………….Committee
(Dr. Pramuk Boonsieng)

……..………………………………………….Committee
(Acting Sub Lt. Dr. Pichitchai Kamin)

……..…………………………………………….Advisor
(Dr. Sarayut Nonsiri)

NAKUL PANNOY : OBJECT DETECTION FOR RETAIL PRODUCT RECOGNITION. ADVISOR : DR . SARAYUT NONSIRI, 68 PP.

The retail sector is a vital driver of economic expansion. The retail industry must embrace technological advancements to augment productivity, streamline operations, and minimize human errors to uphold its crucial economic role. During the COVID-19 pandemic, the purchasing power volumes globally grew from February 2020 to April 2021, and the retail sector gained 35 percent in market capitalization. This growth underscores a compelling research opportunity in the retail industry.

Consequently, artificial intelligence (AI) has become a focus of significant interest and has widely adopted technology, which includes computer vision to recognize and detect retail products. This study explores the efficacy of YOLO (You Only Look Once) in retail product recognition. The research compares different YOLO versions and evaluates their ability to identify on-shelf grocery items. In this research, YOLOv8 is applied and divided into five subcategories: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8l (large), and YOLOv8x (extra-large). The models are assessed utilizing Grozi-120 and SKU110K datasets.

The evaluation metrics are precision, recall, mAP50, and mAP50-95. The result demonstrates that YOLOv8x yields the best overall performance across both datasets. Remarkably, a mAP50 score of 92.6 percent is achieved on the SKU110K.

In conclusion, the outcomes indicate that the YOLOv8 model works well for retail product detection. The model efficacy shows potential for effective retail product recognition, contributing to further technological advancement in the industry.

Graduate Studies Student’s Signature……….…………….

Field of Study Information Technology Advisor’s signature……..……..………….

Academic Year 2024

Acknowledgement

First and foremost, I would like to express my deepest gratitude to Dr. Sarayut Nonsiri, my advisor, whose invaluable guidance and advice were instrumental in completing this dissertation. Without his assistance and support, this research would have not be achieved.

Furthermore, I would also like to express my sincere appreciation to my research committee members, Assoc. Prof. Dr. Annop Monsakul, Dr. Pramuk Boonsieng, and Acting Sub Lt. Dr. Pichitchai Kamin. Their valuable suggestions and solid encouragement helped shape this dissertation into its final form.

Especially a special thank you to my family, all my professors, my UNICEF colleges, and supervisor; thank you from the bottom of my heart. Their positive attitudes and motivations are significant to me in completing this paper. Your constant support reminded me that with hard work and determination, anything is possible.

Lastly, I would like to thank my friends and the Thai-Nichi Institute of Technology (TNI), which is a part of my thesis achievement. This institute gives me the invaluable opportunity to expand my knowledge and explore new horizons beyond my original field of study. It has been an incredible journey, and I am so grateful for everything I have gained.

Nakul Pannoy

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Figures Pages

List of Figures (Continued)

Chapter 1
Introduction

1.1 Statement of the problem

The retail industry is a crucial driver of economic growth. In retail sector, the technology adoption is imperative for sustained success, as it enhances traditional processes, improves productivity, minimizes individual errors, and ensures continued economic vitality. The advancement in technologies is powering substantial innovation in the retail segment. Meanwhile, the change has rapidly risen during the Covid-19 crisis. Amidst the pandemic, the spending volumes globally grew from February 2020 to April 2021, and the retail sector gained 35% in market capitalization [1]. Furthermore, adopting technology in the retail business escalates customer attention and expectations. A recent study [2] revealed that convenience, level of modernization, and in-store digitization technologies are the considerable factors that affect customers' store preferences and selections. Moreover, the expectation of global spending on artificial intelligence (AI) would dramatically increase up to $12 billion in 2023 from $3.6 billion in 2019, or over300% increase [3].

Therefore, to maintain competitiveness, identifying a product on a retail store shelf is an essential activity. This is a common individual skill, but it is prone to errors. Relying on human resources will consistently result in cost inefficiency and time consumption [4]. Nowadays, barcodes are an ordinarily adopted technique in retail business for cost reduction, time savings, and minimizing human mistakes [3], [4].

The emergence of a new practice in retail business has been influenced by changes in consumer behavior and the development of technology-driven management strategies. This emerging retail model leverages knowledge development, negotiating power, and responsiveness to customers' preferences regarding convenience and product variety. As a result, it significantly impacts the decline of smaller and medium-sized retail establishments. The evolution of the modern retail industry has led to a new retail business cultural framework aimed at meeting customer demands and delivering unique customer experiences.

The following categories can be used to group retailers:

1. Department stores: They serve as one-stop stores for books, cosmetics, clothing, and other items. People enjoy shopping there because they make it simple to discover what you need.

2. Shopping malls: Although they house a wide variety of businesses, theyare not well-organized. Depending on which retailers lease space there, Malls maybe found outside or indoors.

3. Supermarket: Mostly carries food and common household items. Prices are normally affordable.

4. Hypermarket: A hybrid of a department shop and a supermarket.

5. Specialty Store: These establishments deliver to a particular niche, such as cosmetics or athletic apparel.

6. Convenience Store: These are little neighborhood shops that sell common goods for fast shopping.

7. Discount Store: These establishments prioritize low costs over high-end merchandise. They concentrate on volume and sell a lot of goods at once.

8. Mini-marts: Often located in nearby neighborhoods or stations. Mini-marts are similar to convenience stores but usually smaller, focusing on every essential goods.

9. Grocery Store: These are simple shops that mostly offer food and necessities for the home. Usually, these are little, family-run companies that cater to thecommunity.

Convenience shops, such as Lawson 108, Family Mart, and 7-Eleven, are among the greatest instances of how technology is used in retail. These 24-hour businesses were mostly focused on minor goods, daily necessities, and quick-access products. They were usually located in communal areas. They broadly use barcode system to minimize human mistakes, save time, and cut expenses, however, there are some limitations to this approach. For example, problems with barcode printing may result in unreadable codes even when the product is stocked, which might mean lost sales opportunities. Furthermore, erratic barcode positioning on goods and the constrained scanning range of barcode readers can impede transaction speed and detract from the general shopping experience of customers. Barcodes on uneven surfaces provide additional challenges. They are more complex to scan and more likely to cause errors.

Advanced technology is playing an increasingly vital role in present retail, especially when it comes to managing products on store shelves. Current technology significantly influences the management of products on retail shelves. Instead of barcodes, many convenience stores use RFID (Radio Frequency Identification) technology. This solution technique can reduce store costs and increase revenue. Procter & Gamble and Wal-Mart applied the RFID technology. The purposes are to lower inventory by 70%, service by 99%, and administrative expenditures by reorganizing their supply chain. Moreover, RFID generates more than $1.3 billion in supply chain income, while the misplaced or lost products cut profits by 20% [5].

RFID technology consists of two main components:

1. Tag or Transponder: This component is affixed to products and records product information.

2. Reader or Interrogator: This device communicates with the tag using radio waves, reading and writing information.

On the contrary, RFID technology faces certain challenges. In scenarios with numerous transactions, tags can experience collisions, leading to potential issues. When multiple tags are nearby and respond simultaneously, tag collisions can occur. Additionally, interference may arise when signals from two or more different reader devices overlap. Consequently, these challenges can result in reduced accuracy and unresponsiveness in the RFID system [3], [4].

Artificial intelligence (AI) [6] and machine learning (ML) [7] are today technology to enhance product management efficiency on shelves. This advancement is particularly beneficial in inventory management. Computer vision optimizes inventory control through real-time shelf analysis, identifying stock issues, and forecasting needs. This method automates inventory tracking, preventing overstocking and shortage of inventory and maintaining organized shelves. Poor inventory planning is substantial sources of financial loss. Consequently, many companies, especially those with multiple branches and regional distribution centers, are prioritizing this aspect. With the immediate growth of the retail sector, the demand for AI has noticeably increased. Investment in AI services has intensely increased [3]. The future of retail innovation may well be dominated by AI, employing more sophisticated methodologies and advanced technologies. Technology like computer vision, facial recognition, and AIdrive the operations in stores that don't have checkout counters. Conversational robotscreate digital maps and offer services in relevant places. Chatbots can improve customerservice, and voice commands will enhance the overall shopping experience. For instance, Amazon Go [8] is a cashier-less store powered by AI, developing the shoppingexperience. Meanwhile, the growing popularity of Alexa has contributed significantly to the improvement of the overall shopping process. Moreover, Fashion AI, driven bythe innovative capabilities of Alibaba, presents an attractive mix-and-match fashionrecommendation system tailored to suit the unique lifestyles of individual customers. Additionally, Ask-eBay simplifies the search for over 60 million products in its catalog.

Shopper Sense exemplifies the effective use of AI in retail product and inventorymanagement. Their AI technology leverages sales history and market trends to forecastand create purchasing plans that minimize overstock risks. This approach increases stock management efficiency. By analyzing sales data, customer preferences, and various factors affecting purchases through machine learning, Shopper Sense predictsmarket demands and formulates tailored purchasing orders for different occasions, suchas festivals or promotional events. This strategy mitigates the risk of excess inventoryand optimizes sales during periods of anticipated higher demand. Furthermore, AI'scapacity to refine predictions and planning processes allows stores to adapt theirinventory strategies to align with current and projected market trends more effectively.

In Croatia, Konzum Smart [9] is the first unmanned grocery store with over 1,700 consumer products, including beverages, snacks, deli items, and daily essentials. It utilizes AI and ML technology from AiFi, featuring 150 motion-detecting cameras to track customer movements and product details. The system will then calculate the price and deduct it from their bank account when customers leave the store. These methods speedthe transaction, eliminates the need to queue for payment, does not need to interact withthe cashier, and creates a positive customer experience.

Smart retail stores [10] provide an unstaffed environment entirely enabled by various-store smart technologies that raise the customers' shopping experience. Smart retail stores provide competitive prices by reducing labor costs and deploying technology to increase product sales and strengthen business operations. Moreover, it involves blending traditional shopping methods with technology, allowing retailers to communicate with target customers through electronic devices and the Internet. Customers can access product details, complete purchases, and enjoy shopping convenience anytime, anywhere, reducing the need for human intervention. Examples of smart retail solutions include people counting, audience detection, and business data intelligence. Recent smart retail innovations like self-service kiosks, and self- checkout options offer further convenience to customers. In addition, the store ambiance, convenience, use of new technology, and scale of modernization have significantly impacted store favorites.

Maintaining track of things on shelves in a computer system is crucial for identifying lost, misplaced, damaged, or unsaleable items. This tactic helps keep employees informed about any issues on the shelves. While the barcode scanning has drawbacks, including the inability to identify items, which results in inaccurate inventory counts, the product identification technologies have the potential to greatly enhance product management.

In addition, a planogram is used to arrange product displays in designated spaces. The planogram [11] is a graphic depiction or arrangement created to increase product sales. It optimizes product arrangement maps and helps maintain a competitive advantage. The planogram is essential for area mapping, inventory display, and shelf space planning in retail sales techniques. This tool is necessary for inventory management and are mainly used for designing layouts for product displays and trade fair exhibitions. It helps merchants to collect the information needed for planning, presenting, and selecting merchandising alternatives.

Moreover, the planogram provides essential full-sized photos and thorough visual blueprints for setting up trade show exhibits and visualizing products. The planogram is crucial in retail planning for inventory control, aiding retailers, especially those with brick-and-mortar stores, in optimizing their layouts for increased sales. It also facilitates retailers make the most of limited space, maximizing benefits and increasing sales. Figure 1.1 illustrates the planogram.

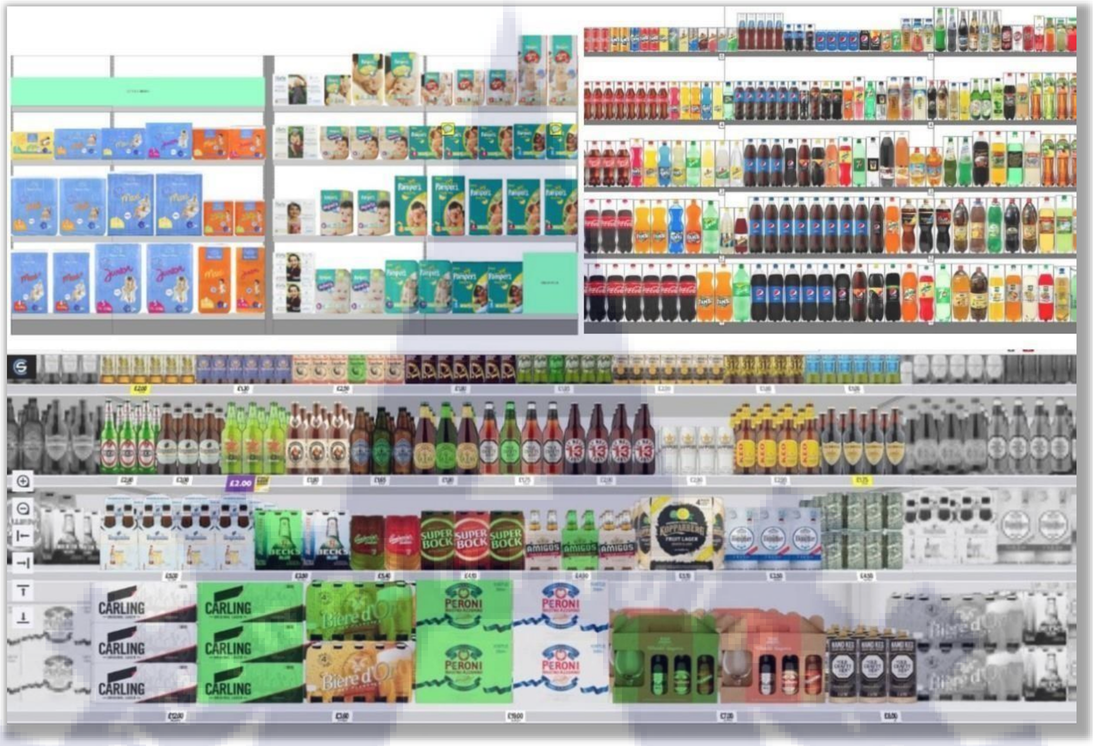


Figure 1.1: Planogram

1.2 Objective of the Study

1.2.1 To develop the product recognition model and explore its effectiveness in identifying products on retail shelves.

1.2.2 To evaluate the performance and compare it with different versions of YOLOv8 algorithms for detecting products on shelves in retail environments.

1.3 Scope

1.3.1 This research will focus on product detection on the shelf by locating the items in retail stores.

1.3.1.1 The study utilizes the following datasets:

1.3.2.1 Grozi-120

1.3.2.2 SKU110K

1.3.2.3 Freiburg Grocery

1.3.2 The object detection algorithms, YOLOv8, are applied for analysis.

1.4 Expected Benefits

1.4.1 Empty shelf detection: Recognizing vacant shelves, averting lost sales opportunities, and increasing product sales in every establishment. This entails outperforming rivals by guaranteeing that customers can obtain necessary goods at the appropriate time and location.

1.4.2 Stock tracking, audit, and survey: Manually checking store shelves takes a lot of time, causes cost inefficiency, and is prone to mistakes. Object recognition and detection technologies automate stock tracking, reduce human errors, and give businesses more accurate information.

Chapter 2
Theory and Related Work

This section reviewed previous research on object detection usingvarious fields with different techniques, focusing on the studies that utilize YOLOalgorithms of their variations.

2.1 Technology Solutions of Object Detection for Retail Products Recognition

Current developments in artificial intelligence (AI) and computer vision technologies have opened up new opportunities to identify products on shelves. Several research studies have used CCTV, object recognition, product detection, and machine learning techniques to automatically detect and monitor space on retail shelves. These solutions provide real-time shelf analysis, precise inventory management, and restocking of the shelves.

Intense rivalry in retail focuses on meeting consumer requirements through services, product quality, technology, price, and other factors. Recent technological advancements, such as barcode scanning and RFID, have been applied to improve inventory control, reduce business costs, save time, minimize human mistakes, and make self-checkout possible. Random barcode positioning could potentially disrupt the shopping process, leading to slower transactions. Likewise, RFID technology may experience issues due to radio wave interference.

As a result, computer vision is now a highlight in the retail industry. These solutions have several advantages in terms of performance; however, some limitations need to be considered.

Advantages:

1. Accuracy Increase: The technology solution analyzes photos or videos taken by in-store cameras using sophisticated computer vision algorithms. These systems can detect the shelf's availability. Moreover, manual auditing of retail shelves is time-consuming and error-prone. Object detection technologies automate stock tracking, providing retailers with more reliable data.

2. Real-time monitoring: Technology-based solutions provide real-time monitoring of shelf readiness. Retailers can instantly identify empty shelves and

restock products by continuously analyzing image data. Real-time tracking helps maintain shelf availability, increase customer satisfaction, and reduce lost sales opportunities.

3. Efficient inventory management: Technological solutions automate the process of monitoring and managing inventory levels. By observing real-time stock status, retailers can optimize their inventory management practices and reduce the problem of overstocking or insufficient stock. These results are improved supply chain efficiency, reduced moving costs, and better use of shelf space.

4. Data-driven insights: Data gathered by technology solutions can provide valuable insights into customer purchasing patterns, product demand, and shelf performance. Retailers can leverage this data to make informed decisions about restocking, assortment planning, and optimizing store layouts. These insights help retailers optimize operations, improve product availability, and increase the shopping experience.

5. Cost savings: Even though initial operating costs may be involved, technological solutions can lead to long-term cost savings. By reducing stock and improving inventory management, retailers can reduce cost revenue, avoid unnecessary emergency restocking, and increase efficiency in labor use. These cost savings increase profits and operational efficiency.

6. Planogram compliance of products on shelves: Technology solutions ensure that the planogram aligns precisely with the design. It is a notification system for misplaced products.

7. Assistance for visually impaired individuals: Visually impaired individuals may struggle to identify packaged foods. Artificial intelligence, by reading labels and text aloud, enables independent shopping.

Limitations:

1. Infrastructure and operating costs: A certain amount of infrastructure,such as cameras, sensors, and data processing systems, is needed to implement a technical solution for product detection. Retailers must make the required network infrastructure, software, and hardware investments. There can be significant up-front expenses for this.

2. Technical Challenges: Image quality, lighting conditions, occlusion, and product packaging formats are some challenges. Ensuring detection accuracy across different store environments and product types can be a complex issue requiring refinement of algorithms and models from time to time.

3. Integration complexity: Integrating technical solutions into existingstorage systems and processes can be challenging. Merchants need to ensure integrationwith inventory management systems, point-of-sale systems, and other operational software. Seamless integration may require technical expertise and coordination with Information Technology (IT) teams or third-party solution providers.

4. Maintenance and Support: Technology solutions require regular maintenance and updates to ensure optimum performance. Retailers need to allocate resources for system maintenance, data management, and technical support to resolve anypossible issues. This ongoing maintenance, however, might increase the operating costs.

5. Privacy and data security: The technological solution collects and processes image data from in-store cameras. Venders must prioritize data privacy and security to protect customer data and comply with the law. Implementing data protection and ensuring secure storage and transmission of data are important considerations.

It is vital for retailers to carefully evaluate the benefits and limitations of technological solutions for product detection. Various factors need to be considered, such as cost, scalability, compatibility with existing systems, and the unique needs of retail operations.

2.2 Theory

2.2.1 Artificial Intelligence (AI)

Artificial intelligence, or AI, refers to techniques enabling computers to mimic human thinking. This involves training computers with data and algorithms to learn, reason, and predict potential outcomes like humans do. AI can perform complex tasks that humans have previously completed. On the other hand, it gets less human involvement and reduces human error. Five concepts of artificial intelligence:

- Perception: AI can recognize objects through visual and auditory that perform as sensors, such as cameras, microphones, and other data input devices. The information is imported for processing, analysis, and prediction.

- Representation and reasoning: AI can store and utilize knowledge through knowledge representation techniques. By creating decision rules based on expert input, AI systems can engage in inferential reasoning, drawing conclusions from the knowledge they possess.
- Learning: AI using machine learning algorithms from big data. It creates a model based on data inputted by humans or gathered by the machine to understand patterns and make decisions.
- Natural interaction: AI needs to learn how to understand human- to-human interaction.
- Social impact: Ethics, security, and privacy must be considered because AI can decide or take actions that can impact people.

2.2.2 Machine learning (ML)

Machine Learning is a part of AI representing a set of algorithms trained on dataset to generate models that allow machines to perform tasks that only humans could do before, such as predicting price fluctuation, analyzing data, and categorizing images. On the other hand, it focused on enabling machines to learn fromdata, understand patterns, and make decisions with less human involvement.

Any business or industry adopting the technology can widen the gap between leaders and laggards by the competitive advantage. Businesses can reduce the time required for various data analyses and labor costs.

The working principle of machine learning:

Machine learning has principles similar to those of humans, who needto learn from experience. For example, to teach a child to differentiate betweenpencils and pens, the child first needs to be educated on what pencils look like and how pens are. As a result, children can learn and differentiate between two things.

Machine learning also works similarly by entering basic datasets andcommands to allow computers to “learn” and classify objects, including people and things. To get more accurate results, developers must provide new datasets and train the system consistently. The purpose is to develop a system that can recognize patterns and make decisions at a later time.

Three techniques of machine learning:

Depending on the input data nature, machine learning can be categorized into 3 major groups.

- Supervised Learning:

It allows machines to find the answers independently after learning the dataset. As the number of trials increases, the algorithm accuracy in classifying new data improves. For example, an image of a pen is used as input data to train computers. The computers do not recognize the pen images. It needs to be trained in how to analyze feature extraction competencies. A pen has a button and writes with ink. Later, the information will be processed and categorized so that the computers can eventually distinguish between the pen and other objects.

- Unsupervised Learning:

In contrast to supervised learning, unsupervised learning learns patterns solely from unlabeled datasets and is allowed to take any actions on that data without human supervision. The method is that humans enter the dataset and determine the objectives from that dataset. It enables machines to analyze, classify, and create patterns from the dataset. For instance, a pen image is input and does not recognize that it is a pen shape. The computers will do the feature extraction and then analyze the input image's appearance. However, it cannot be classified. The clustering method is applied in this case. The computer may take the pen and group it with a highlighter pen or other stationery with a button at the end of the handle, and use ink for writing.

- Reinforcement Learning:

This is a method of learning based on the interaction between agents and the environment, learning from agents under different actions to achieve maximum results through attempts to improve decision-making systems.

On the other hand, this is how machines set certain conditions and implement these conditions through trial and error. Developers can set feedback loop goals and reward conditions such as Alpha Go. Surround the area on the chessboard with your pieces and occupy more territory than the opponent. Alpha Go will learn if the opponent makes this move and what moves it will make to achieve the required conditions. This is to engage as much space as possible on the chessboard.

Machine Learning Benefits:

Machine learning techniques can be used for benefits and purposes. For example, Google has developed Google Maps to facilitate the best routes, distances,and estimated times. Additionally, Google Translation brings the benefits of automationto work with machine learning to assist people in comprehending the meanings of words or sentences. Well-known chat programs like LINE have adopted Speech-to-texttechnology to save typing time.

Additionally, in the retail business, machine learning can provide benefits as follows:

1. Apply a recommender system to generate revenue: Entrepreneurs can use ML to recommend new products to customers based on their historical purchases. Prioritizing customer retention is crucial for sustainable success and more cost-effective than focusing solely on acquiring new ones.

2. Analyze buying history and customer behavior: Analyze the historical information on those who have left and those who have stayed and different customer behaviors. This insight helps management plan the right strategy to maintain customer retention and keep those likely to go.

3. Enhance planning and marketing activity: ML is about making predictions. Businesses can use ML to predict trends, costs, and demand to help with budgeting.

4. Cut unexpected downtime via predictive maintenance: ML identifies which equipment will likely experience failure. Management can use this insight data to plan repairs, perform preventive maintenance, and minimize costs.

5. Learn the pattern to protect against fraud: ML can learn patternsand detect fake items.

6. Increase efficiency and reduce cost: ML can decrease business expenses by decreasing human mistakes, improving inventory competence, enhancingcustomer satisfaction, and improving efficiency.

2.2.3 Deep learning (DL)

Deep learning, an area within machine learning and a subset of artificial intelligence is stimulated by the structure of the human brain. It utilizes layered neural networks to simulate human decision-making processes. DL has revolutionized the future of AI [12]. It attempts to mimic how humans think by analyzing large amounts of data to find patterns and make decisions. To achieve this process, deep learning utilizes the multi-layered structure of algorithms called neural networks to learn and understand data without needing labels or direct instructions, enabling them to classify and process complex information. The neural network model with many layers can better categorize and analyze data.

Deep learning is used in many AI efforts, such as image classification, self-driving cars, natural language and image processing, and predictive forecasting

The recent deep learning research with special emphasis on architectures, applications, and trends [13] utilizes input data to form a model that can recognize patterns or relationships between data. To improve the model's effectiveness, the model settings are adjusted by the fine-tuning hyperparameter and tested. This process will be repeated until the model's accuracy reaches targets.

To work on deep learning, the accepted open-source software, TensorFlow [14] and Keras [15] are used as deep learning toolkits. TensorFlow is a powerful application in distributed data processing, while Keras is the advancedsoftware for creating and processing the neural network model.

Deep learning is an existing technology in AI development and can be applied in many fields that need learning and classification abilities in a large amount of complex data.

2.2.4 Object Recognition and Retail Product Recognition

Object recognition is a technology that identifies objects, places, people, writing, and actions within digital images and videos. The object recognition overview can be illustrated in Figure 2.1.

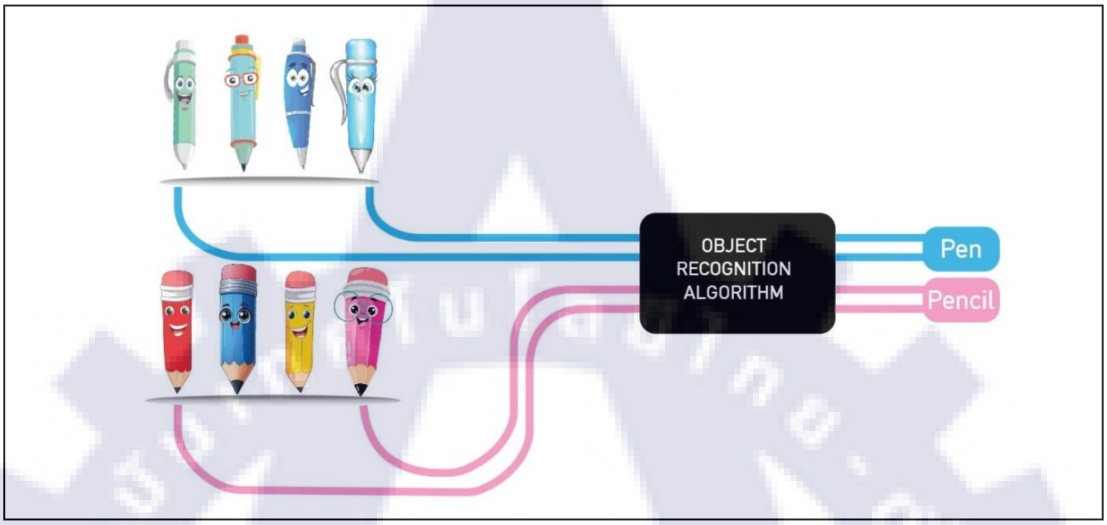


Figure 2.1: Object Recognition Overview

Retail product recognition is a technique that analyzes images of shelves in retail stores, identifying the presence of products and determining their location using bounding box coordinates, as illustrated in Figure 2.2.

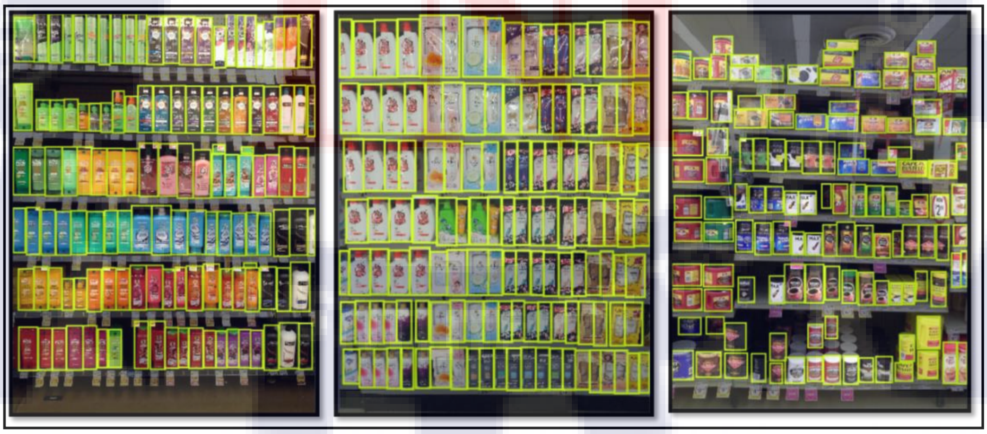


Figure 2.2: Retail Product Recognition Technique

This method is a fundamental tool in machine learning and deep learning techniques. The aim is to train machines to understand and interpret the content of images just as humans do.

2.2.5 Object Detection

Object Detection algorithm is a computer vision process that merges image classification and object localization. It is a crucial application in the domain of computer vision that identifies and locates objects within images or videos [16]. Object detection finds practical use in various everyday applications, including autonomous driving, security systems, robotics, and remote sensing target detection [17, 18]. The main objective is to detect and locate instances of objects in images or videos, highlight those objects using bounding boxes, and categorize them into their corresponding classes.

By tradition, object detection relied on HOG, SIFT, DPM, Haar, and VJ detector. Unfortunately, these methods have some limitations. For example, SIFT is incompatible with illumination changes and has a high computational cost because it is prolonged. HOG has lengthy identification times, while VJ detector [19] requires extensive training. As a result, Convolutional Neural Network (CNN) is reintroduced in conjunction with Deep Learning for object detection to solve the problems of traditional methods [20].

The classification of the object detection technique is depicted in Figure 2.3. The relationship between the object detection and object recognition methods is illustrated in Figure 2.4 while the difference between image classification, object localization, and object detection is shown in Figure 2.5.

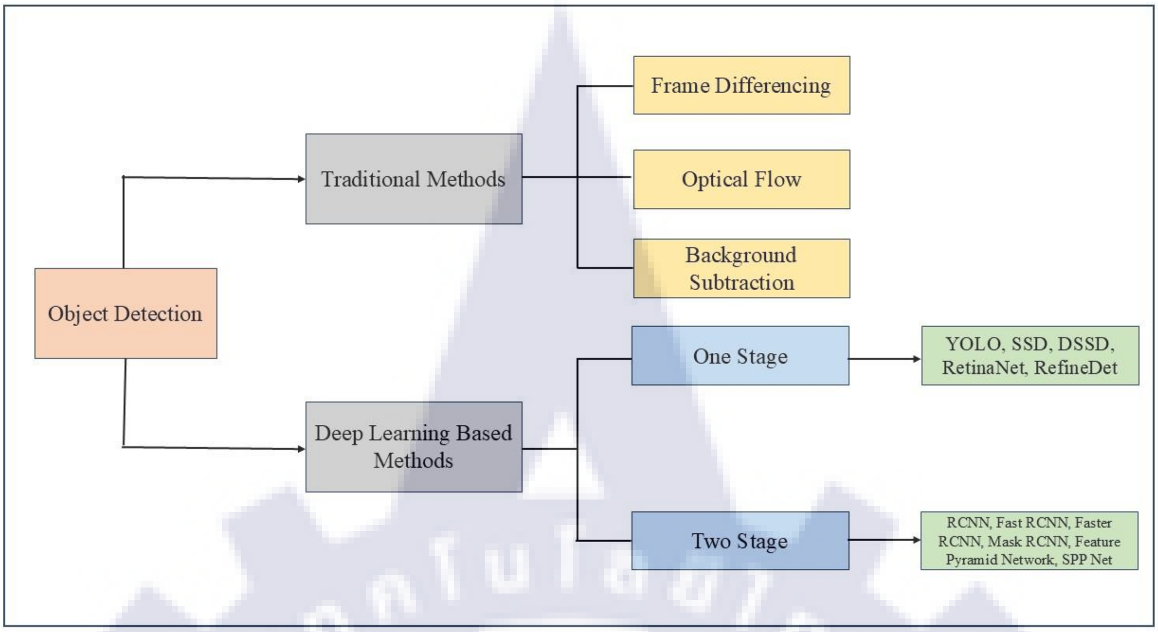

Object Detection
Traditional Methods
Deep Learning Based Methods
Frame Differencing
Optical Flow
Background Subtraction
One Stage
Two Stage
YOLO, SSD, DSSD, RetinaNet, RefineDet
RCNN, Fast RCNN, Faster RCNN, Mask RCNN, Feature Pyramid Network, SPP Net

Figure 2.3: Object Detection Approaches [20]

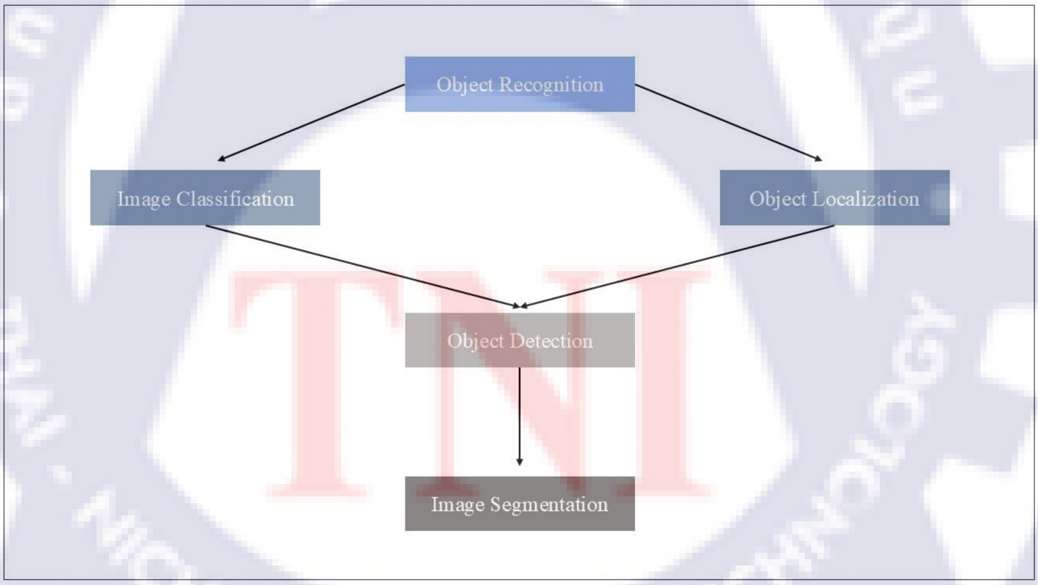


Figure 2.4: The Relationship between Object Recognition, and Object Detection

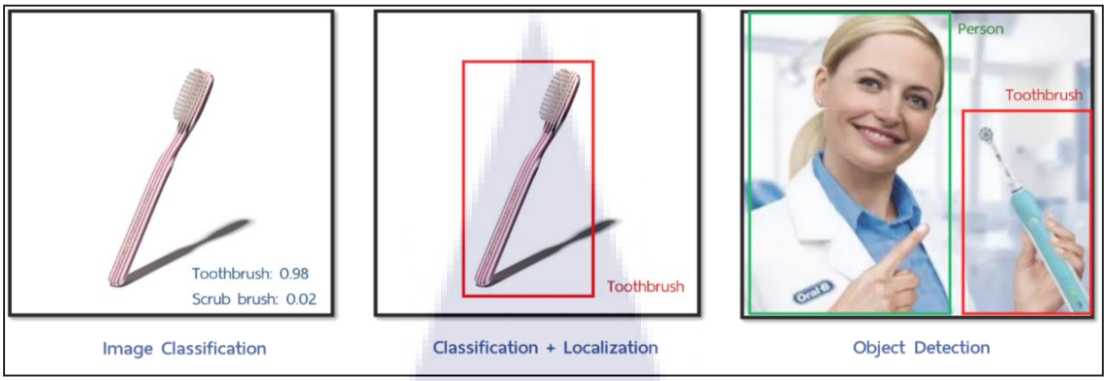


Figure 2.5: The different between image classification, object localization, and object detection technique

Object detection based on deep learning is classified into two approaches: Anchor-based and Anchor-Free. The Anchor-based approach can also be divided into Two-stage detectors, which rely on region proposals and One-Stage detectors, based on regression [20].

In Two-Stage detector architecture, the object proposals are first identified within the image. These proposals are then classified and localized in the second stage. In essence, the two-stage detector performs localization and classification of the region. Although this technique provides accurate results, it makes proposals slower and has a more complex structure [20].

One-stage object detection models skip the region proposal step and, instead, directly classify objects through a dense sampling of locations. This approach focuses on practicable spatial region proposals using a simpler architecture. The object localization and classification processes are simultaneously operated. This simplified process makes faster inference speeds compared to Two-Stage models.

The example approaches of Two-Stage and Single-Stage detector are demonstrated in Figure 2.3. The structural differences between Two-stage and Single-Stage detector structures are further illustrated in Figure 2.6.

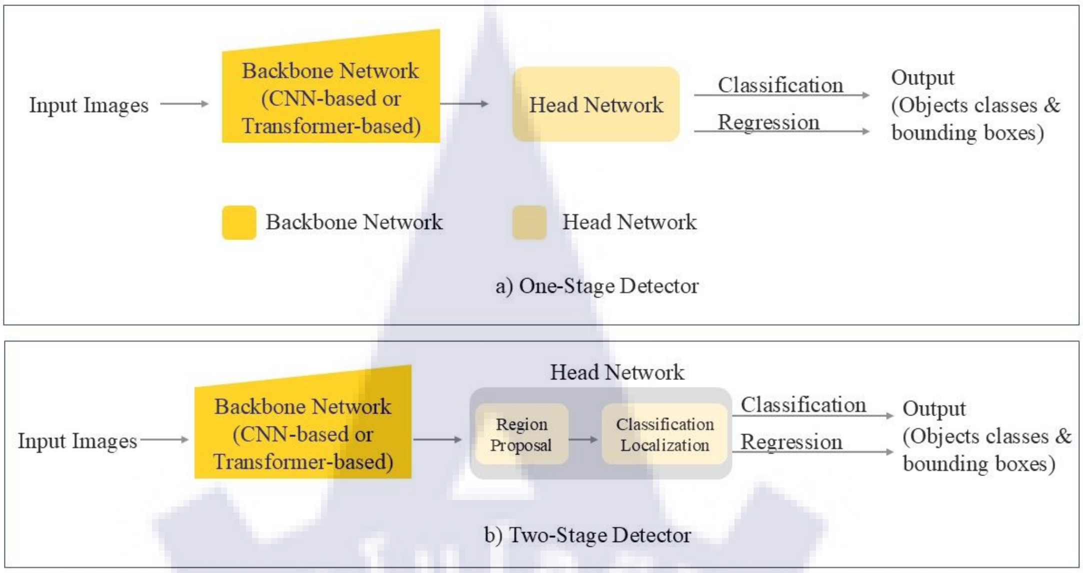

Input Images
Backbone Network (CNN-based or Transformer-based)
Head Network
Classification
Regression
Output (Objects classes & bounding boxes)
Backbone Network
Head Network
a) One-Stage Detector
Head Network
Input Images
Backbone Network (CNN-based or Transformer-based)
Region Proposal
Classification Localization
Classification
Regression
Output (Objects classes & bounding boxes)
b) Two-Stage Detector

Figure 2.6: Two-Stage and Single-Stage architecture [20]

The different between the image recognition and the object detection is the image recognition puts a label on the whole picture while the object detection finds and labels specific details within an image or video.

2.2.6 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a type of deep learning technique specializing in pattern recognition, which is one of the most used DL networks [21], [22], [23]. It learns directly from images.

CNN is formed by several layers that process and transform an input to produce an outcome. It can be trained to perform image analysis tasks, including image classification, object detection, object segmentation, and image processing. CNN's primary advantage is the ability to automatically detect substantial features without human supervision [21].

CNN simulates how humans process visual information. Like humans' eyes focusing on specific areas while still aware of the surroundings, CNN breaks down images into smaller sections. They analyze these sections for features like lines and color differences. By combining information from the focal point and the surrounding context, CNNs gain a deeper understanding of the image, much like human vision.

Figure 2.7 is depicted CNN fundamental architecture. The association between AI, ML, DL, and CNN is demonstrated in Figure 2.8.

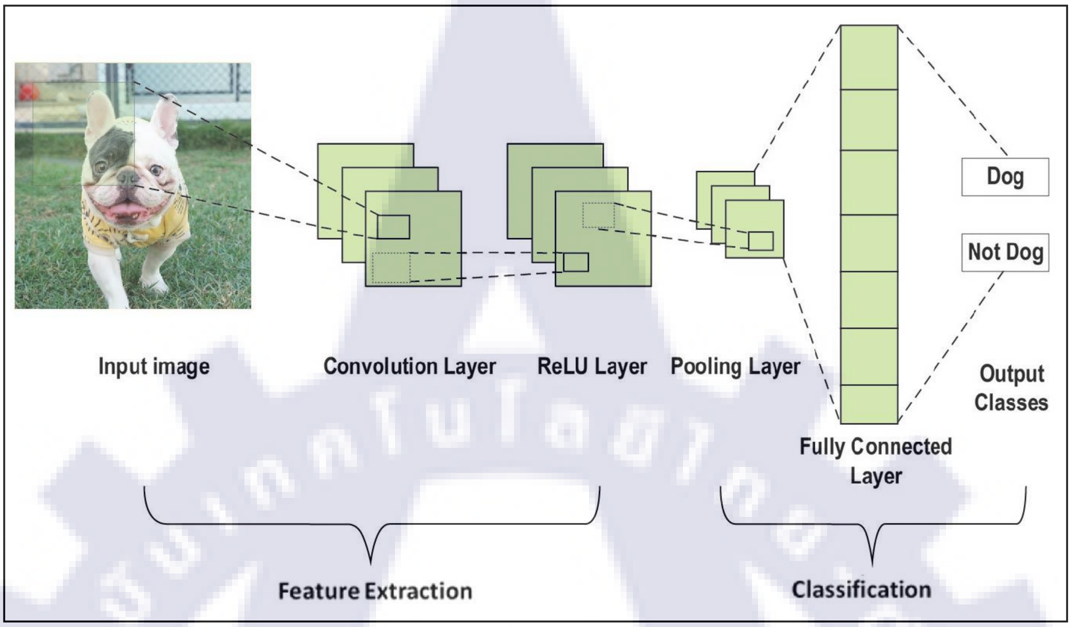


Figure 2.7: The fundamental architecture of Convolutional Neural Networks (CNNs)

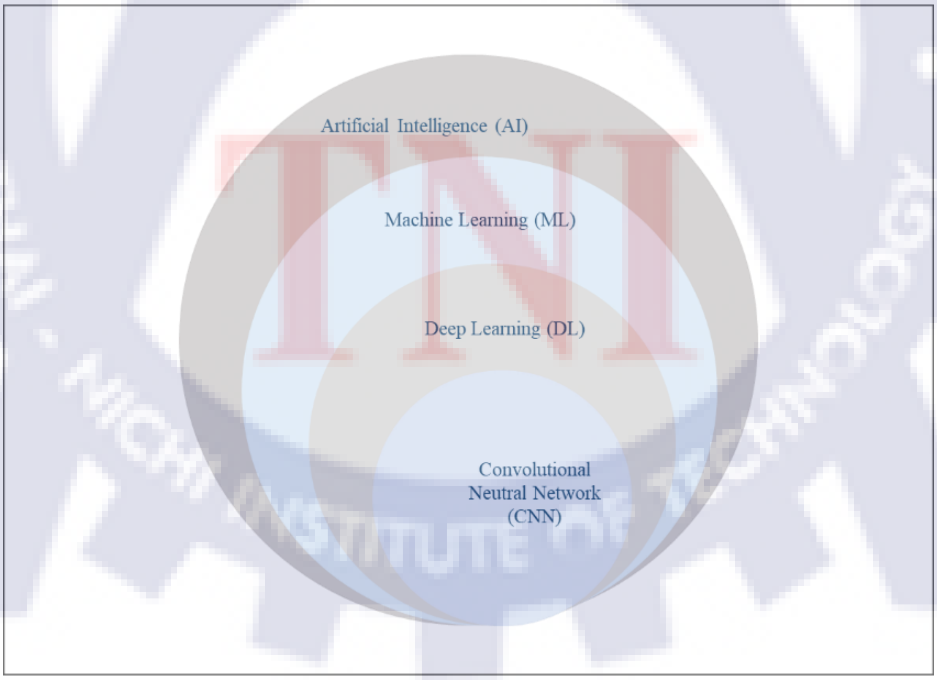

Artificial Intelligence (AI)
Machine Learning (ML)
Deep Learning (DL)
Convolutional Neutral Network (CNN)

Figure 2.8: The relationships between AI, ML, DL, and CNN

2.2.7 YOLO (You Only Look Once)

Although the R-CNNs technique tends to be more accurate, the apparent problem is speed. Thus, One-Stage Object Detection has been developed to improve the speed of object detection and solve the problem encountered by R-CNNs. The YOLO approach, introduced by authors [24] in 2016, was the first regression-based object detection technique.

The algorithm developed to enhance the speed efficiency is called YOLO. It is a unique balance of speed and accuracy, allowing the rapid and reliable identification of objects in images [25], [26]. YOLO is a cutting-edge detection system that operates in real time. YOLO algorithm uses convolutional neural networks (CNN) to identify objects. In addition, YOLO is the object detection algorithm that is categorized as One-State Object detection. It has become one of the most used of several object detection algorithms due to its accuracy and speed, which are applicable to real-time systems.

The concept is to simplify overall problems as a single problem in the form of regression analysis. The processes involve predicting the bounding box surrounding the objects and the object possibility in that frame simultaneously. This method makes the YOLO faster than other algorithms.

While YOLO performed at faster processing speeds than other techniques, the localization error was higher than that of alternative object detector methods. It also struggled to detect more than two objects of the same class within a grid cell. Moreover, it had difficulty identifying small-sized objects. YOLO’s bounding box coordinate inaccuracy was also significant [20].

2.2.8 YOLOv5

YOLOv5 [26], released in 2020 by Ultralytics, developed in Pytorch instead of Darknet [27] and integrated an algorithm called AutoAnchor. It is one of the latest and most frequently utilized versions of a widely popular deep learning neural network architecture. Renowned for its speed and accuracy, it is primarily employed for various computer vision tasks within the broader machine learning field, including object detection, image classification, and image segmentation. It is a famous object detection algorithm that detects from images, video, and live camera feeds. YOLOv5 is a lightweight and fast object detection algorithm. It achieves high accuracy in standard object detection for the COCO dataset. YOLOv5 is versatile and can be used for various object detection tasks. It is easy to use and implement, with a straightforward training process and simple configuration files. YOLOv5 is an open-source algorithm. The COCO dataset [28] contains several images with detailed object annotations. Combining the COCO dataset with YOLOv5 can lead to highly accurate object detection results.

YOLOv5's architecture consists of three main primary components.

- Backbone: This is the network core. It is a modified CSPDarknet53 which is an adaptation of the original Darknet architecture found in earlier versions.
- Neck: It is the part that bridges the backbone and the head, employing SPPF and a modified CSP-PAN structure to combine features.
- Head: This is the same as YOLOv3 head and responsible for generating the final output.

YOLOv5 leverages diverse data augmentation techniques [29] to enhance the training process and improve model performance. These practices include Mosaic Augmentation, Copy-Paste [30], Random Affine Transformations, MixUp [31], HSV Augmentation, additional techniques from Albumentations set [32], and Random Horizontal Flip.

In addition, YOLOv5 come in five different scaled versions: YOLOv5n (nano), YOLOv5s (small), YOLOv5m (medium), YOLOv5l (large), and YOLOv5x (extra large). These versions can be applied to several requirements, offering a balance between model accuracy, size, and speed for diverse scenarios.

Last but not least, YOLOv5 is an open-source object detection model dynamically maintained by Ultralytics. The latest updates notably improve the accuracy and efficiency while preserving its high speed.

2.2.9 YOLOv7

YOLOv7, improved from YOLOv4, performs with speed and accuracy effectiveness, with speed in the 5-160 fps range [26], [33]. The core competency of YOLOv7 is real-time object detection, which is a main component of modern computer vision. The real-time object detection model attempts to identify and detect the objects of interest in real-time. This model allows developers to proficiently track objects in moving frames, such as video and live CCTV input.

Real-time object detection is a step forward from traditional object detection. While the first pattern tracks objects in video files, the conventional model locates and identifies the object in the still frame, such as an image.

Unlike traditional real-time object detectors, YOLOv7 focuses on improving the training process rather than solely emphasizing architecture development. Moreover, the model introduced the trainable bag-of-freebies, modules, and methods designed to enhance precision without impacting inference speed. YOLOv7 also proposed some architectural changes, including [26]:

A) Extended Efficient Layer Aggregation Network (E-ELAN) [34]: It serves as the computational block in the YOLOv7 backbone. Additionally, it enhances the network's learning capability by shuffling and merging features from different groups without ending the original gradient path.

B) Model scaling for concatenation-based models: YOLOv7, a concatenation-based architecture, proposed a new strategy for scaling concatenation-based models. This strategy maintains the optimal structure by scaling the block's depth and width with the same factor.

YOLOv7 introduces a wide range of key features, including [35]:

a) Planned re-parameterization convolution: YOLOv7 offered a novel strategy which is applicable to layers across different networks architectures by employing a gradient propagation path concept.

b) Dynamic Label Assignment: The model proposed a new label assignment method called Coarse-to-fine lead guided label assignment, in which the lead head is responsible for generating the final output.

c) Extended and compound scaling: The YOLOv7 algorithm introduced the “extend and “compound scaling” methods for the real-time object detector.

d) Efficiency: The proposed method significantly decreases the parameters and computation of modern real-time object detectors by approximately 49% and 50%, correspondingly. This results in faster inference speed and improved detection accuracy.

2.2.10 YOLOv8

YOLOv8 [36], launched in January 2023 by Ultralytics, provided five measured versions: 1) YOLOv8n (nano), 2) YOLOv8s (small), 3) YOLOv8m (medium), 4) YOLOv8l (large), and 5) YOLOv8x (extra-large). This YOLO version offers advantages over previous versions and supports several vision tasks such as object detection, tracking, segmentation, pose estimation, and classification. It is improved in terms of accuracy in object detection, attaining better mean average precisions (mAP) and faster speeds.

Moreover, the YOLOv8 architecture has improved the network design, Backbone network, anchor-free detection head, and changes in lost function values. The object detection performance in YOLOv8, specifically with smaller objects, has been enhanced using CloU [37] and DFL [38] loss functions for bounding box loss and binary cross-entropy for classification loss.

Furthermore, YOLOv8 can also be run from the command line interface (CLI) and installed as a PIP package. Various integrations for labelling, training, and deploying are also included in this algorithm, making it a comprehensive solution for object detection tasks.

Fundamental improvement in YOLOv8:

1. CSPDarknet53 Backbone architecture: YOLOv8 applies CSPDarknet53, which combines the strengths of Darknet and CSPNet architecture. This pattern results in enhanced feature extraction capabilities, streamlined information flow between layers and increased accuracy.

2. PANET Integration: YOLOv8 integrates the Path Aggregation Network (PANet), facilitating the incorporation of information from different scales within the image.

3. Dynamic Anchor Assignment: The aim is to improve the handling of diverse object sizes and aspect ratios. By adopting the anchor box dimension during the training, the model is optimized for varying object shapes and sizes, resulting in better generalization across different datasets and scenarios.

4. Improved Training Process: The training process in YOLOv8 is significantly improved, resulting in more efficient and faster training time. The training

pipeline has been enhanced, enabling quicker model development and performance gains.

5. Model Variants: YOLOv8 proposes 5 distinct versions (YOLOv8n, YOLOv8s, YOLOv8M, YOLOv8L, and YOLOv8x). Each version is tailored to meet different needs, allowing users to choose the form that best fits their resources and purposes.

6. Compatibility and Integration: YOLOv8 is compatible with TensorFlow and PyTorch, the well-known deep learning frameworks. This compatibility ensures seamless integration into existing computer vision pipelines.

YOLOv8 modules offer the following benefits:

- Accuracy and Precision: It is capable of recognizing objects in various sizes and orientations with consistent performance across diverse datasets.
- Efficiency: YOLOv8, a single-stage object detector, is lightweight, provides inference speeds, and requires minimum computational resources for efficient object detection.
- Open-source and community-driven: YOLOv8 is open-source, fostering continuous improvement and collaboration within the developers' and researchers' communities.
- Real-time detection: The model can process images and detect objects in milliseconds, making it suitable for real-time applications.
- Flexibility in deployment: It supports various platforms such as GPUs, CPUs, and edge devices. This flexibility enables deployment on different hardware configurations and makes it practical for operation in diverse environments.
- Robustness to Occlusions: The model can infer the presence and location of the object even when partially hidden, enhancing its performance in real-world scenarios with potential occlusions.

In this research, YOLOv8, the state-of-the-art algorithm, is selected as the experimented algorithm, and the results will be compared with different YOLOv8 versions to determine the best performance.

2.2.11 YOLO-NAS

Released in May 2023 by Deci AI, YOLO-NAS is a state-of-the-art object detection mode that is designed to detect small objects, improve localization accuracy, and increase the performance-per-compute ratio [26].

Many researchers have been experimenting with integrating deep learning and backdrop modeling techniques to enhance the identification of abandoned objects. This includes the combination of deep learning models with conventional background removal. Moreover, it covers the technique of utilizing deep learning background modeling methods that can adjust to different lighting and scene conditions. Background modeling is an essential technique in computer vision and image processing to distinguish foreground items or specific areas from a static background in an image sequence or video stream. The primary purpose of background modeling is to model static or gradual changes that can be subtracted from the current frame to separate dynamic objects.

YOLO-NAS innovation includes the following:

1) Quantization-aware modules [39] called QSP and QCI combine re-parameterization techniques for 8-bit quantization, reducing the model size and computational requirements without sacrificing performance.

2) YOLO-NAS provides the optimal balance between accuracy and latency, providing high-performance object detection with minimal processing time.

3) This algorithm offers three versions, small, medium, and large, with and without quantization, providing flexibility for different resource constraints and accuracy requirements. The model has integrated quantization-aware blocks and selective quantization to support optimized performance.

4) YOLO-NAS is compatible with pre-training on renowned datasets like COCO, Roboflow 100, and Object365 improving the suitability for object detection tasks across various environments [40].

5) Optimized by AutoNAC, YOLO-NAS has been streamlined for Python integration via the Ultralytics Python package.

6) A pre-training methodology leverages automatically labelled data, self-distillation, and massive datasets.

2.2.12 MS COCO

The object detection dataset required the use of various object types and locations to verify the detection competence. For the experiment's convenience, the used dataset must include various object types and locations. MS COCO is the comprehensive dataset.

The Microsoft Common Objects in Context (COCO) dataset is a benchmark dataset for evaluating the performance of computer vision models. It gathers various kinds of object images and includes over 330,000 pictures. More than 200,000 images are labeled. The dataset consists of varied formats. The COCO dataset is divided into 80 categories, and the median image ratio is 640x480. Additionally, out of 330,000 total images, 250,000 feature images of people, indicating the location of individuals and objects in the scenes [41].

2.2.13 PASCAL Visual Object Classes (VOC)

PASCAL VOC dataset 2012 (PASCAL VOC 2012) is an ordinary standard dataset for object detection. It consists of 20 object categories, including vehicles, homes, animals, and so on: airplanes, bicycles, boats, cars, motorcycles, bottles, chairs, dining tables, food, sofa, television, birds, cats, dogs, sheep, and people. Each image in this dataset contains pixel-level segmentation annotations, bounding box annotations, and object-level annotations. According to the Roboflow website, there are eight versions of PASCAL VOC 12. Different versions provide different image numbers and purposes. For instance, PASCAL VOC 2012 v9 contains 208,772 in total and is divided into three categories: training set: 205,350 images, validation set: 3,422 images, and test set: 0 images [42].

2.3 Related Work

Most renowned convenience stores, such as 7-Eleven, still rely on their employees to verify that products are available on the shelves and in stock. Every time a gap is discovered, the store staff must replenish the shelves. Figure 2.9 illustrates current products on shelves in a convenience store.

Figure 2.9: The current product on shelf in a retail store

There are several studies using YOLO techniques in various related industries. For example, complex procedures are involved in producing consumable commodities in the food sector. To improve accuracy and speed in defect identification, F. K. Konstantinidis et al. [43] explored how to automatically identify dairy products in a production line in real time to automate quality-related standards and packaging procedures to increase accuracy and speed in defect detection. YOLOv5 was the chosen framework for their investigation. The experiment was initiated by creating 110 different images. Eighty-eight images were used for training and another 22 for validation. The experimental results revealed no significant disparity between the two algorithms (YOLOv5 and Mask R-CNN), achieving an impressive accuracy rate of 99%.

Likewise, the use of recycled materials in the food packaging industry has grown during the past several years. However, only a few recycling methods can be used, such as barcodes, which are sometimes worn out, leading to problems. Sree Chandan Kamireddi's research [44] focused on object detection models to create a dataset containing images of Tetra packs with visible recycle logos. Four deep learning models (Faster-RCNN [45], YOLO, SSD [46], and Mask R-CNN [47]) were investigated which performed best. The dataset was sourced from nearby supermarkets and amplified using https://roboflow.com/. YOLOv5 came out on top in speed and overall performance compared to the other models.

Besides the food packaging industry, there was an idea to detect a real-time packaging defection system based on deep learning techniques. T. Vu et al. [48]

developed an approach to detect defects in the packaging in real-time with the YOLO algorithm using YOLOv5. The data was collected from FFmpeg, the multimedia open-source. Two hundred image files of damaged boxes and another 200 images of intact boxes were captured by a production line. The mAP score showed the model accuracy at 78.6%.

In the retail sector, many convenience stores currently do not have a well-organized structure for inventory management. This inefficiency creates added expenses for inventory control and storage. The overstock and the shrinkage reflect the inventory availability. This is comparable to studies by P. H. Toranzo et al [49] that used YOLOv5 to identify, measure, and confirm bottles and canned goods. When the various models were compared and the algorithm's effectiveness was assessed, the findings showed that YOLOv5 produced the most accurate identification, counting, and verification of canned products and bottled goods in supermarkets.

In addition to the inventory control issue, recent research, including work by H. Kumar [50], has focused on developing real-time item detection systems to identify empty shelves. A collection of 28,000 images that included 80 popular grocery items was used in this study. The results confirmed the YOLOv8 algorithm's efficiency in recognizing unoccupied shelf spaces, demonstrating that it could identify more than 50% of the empty shelf area

Furthermore, another concern is that it is hard to identify inferior categories separately. Products can sometimes only differ slightly in appearance, whichmakes it difficult for humans to distinguish between them. J. Abyasa et al. [51] researched the brand recognition of grocery products using YOLOv8, focusing on the inter-class similarities. The dataset consisted of 3,095 images developed with 20grocery product classes. The dataset was divided into three sets: training (70%), validation (20%), and testing (10%). The results showed that YOLOv8 competed well in recognizing products even when met with inter-class similarities. Subsequently, the outcomes showed that this approach could efficiently automate checkout procedures insmart retail stores. The comparative analysis of prior studies is shown in Table 2.1.

Nevertheless, previous studies have not explored the performance of YOLOv8 versions for the Grozi-120 dataset. YOLOv8 are advanced models providing high speed and reliability, while Grozi-120, a task-adoption one-shot learning dataset, is a database of 120 grocery products. It is developed to address the challenge of training object recognition and localization models on data that differs in quality from the data used for testing.

Finally, previous studies have certain constraints. They struggle in extremely bright or dark lighting conditions, and obstructed views can hinder their ability to detect empty shelves accurately. In addition, factors such as poor image quality, inappropriate camera angles, and low or blurry resolution can impair the precision of detecting empty shelves. Besides, the standard model evaluation datasets like PASCAL VOC [52], which includes only 20 object classes, and MS COCO, with its 80 object categories, are insufficient for directly applying current object detection algorithms to retail product recognition. Figure 2.10 illustrates a comparison of the VOC2012 and COCO dataset results using three algorithms: Faster R-CNN, SSD, and YOLOv2 [53]. The comparison reveals a noticeable decrease in detector accuracy as the number of classes increases.

This research challenge is still the interclass classification with differentmethods and algorithms. Some grocery products have only minor unlike in a particularpoint of the exterior. Moreover, dissimilar environmental factors can impact the accuracy of product recognition, including background settings, lighting conditions, occlusions, and the way products are arranged. The objective of addressing this challenge is to set benchmarks and overcome these limitations and uncertainties.

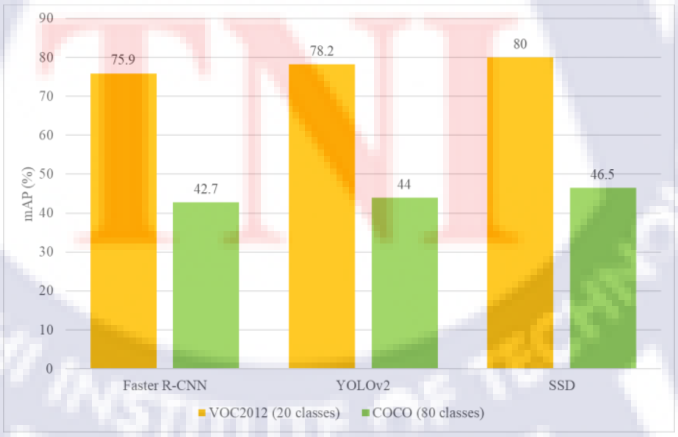


Figure 2.10: The comparison result on VOC2012 and COCO datasets with different algorithms [3]

Table 2.1: The related works comparison

No.	Ref./ Year	Objectives	Techniques	Dataset	Results
1	[27], 2023	‣ Increase accuracy and identify the speed of detecting the defect ‣ Identify dairy products within a product line in real-time to automate quality related standards and packaging procedures	‣ YOLOv5 ‣ Mask R-CNN	YogDATA (Yogurt cup recognition dataset): 110 images	No noticeable difference between 2 algorithms which is 99% of accuracy.
2	[28], 2022	‣ Create a dataset of Tetra Pak images featuring visible recycling logos. ‣ Evaluate and benchmark various algorithms.	‣ Faster-RCNN ‣ YOLOv5 ‣ SSD ‣ Mask R-CNN	Images of the logo side of the Tetra packs: 130 images from a smartphone	YOLOv5 is the best model in the comparison. (mAP = 0.771)
3	[32], 2023	‣ To introduce a real-time packaging defect detection system.	YOLOv5	Packaging defect dataset: 400 images from a production line.	mAP: 78.6%

Table 2.1: The related works comparison (Cont.)

No.	Ref./ Year	Objectives	Techniques	Dataset	Results
4	[33], 2023	‣ To detect, count, and verify the status of bottles and canned products in supermarkets. ‣ To compare the algorithms efficiency.	‣ YOLOv5 ‣ Efficient Det ‣ DET Faster R-CNN	‣ CanBo-Pe: 1,640 images ‣ CanBo-Pe +Status: 2,400 vitro images.	YOLOv5s achieved the mAP50 score of 0.935 and the mAP50- 95 score of 0.893.
5	[34], 2023	‣ To develop a model that can identify empty shelves in real-time.	‣ YOLOv8 Roboflow AutoML	‣ GroZi-3.2K: 11,585 images ‣ SKU110K: 11,762 Images ‣ WebMarket: 300 images Grocery Product: 28,000 images, 80 classes	Roboflow AutoML achieved a mAP = 49.5%
6	[35], 2023	To analyze the performance of YOLOv8 for product recognition in grocery store with the inter-class similarity problem.	YOLOv8	‣ Internally developed dataset: 3,095 images	YOLOv8 was able to recognize products with inter-class similarity problems, achieving an mAP50-95 score of 0.81.

Chapter 3
Methodology

The methodology section describes the dataset, data processing, workflow, algorithm, and assessment. The methodology processes are summarized in Figure 3.1.

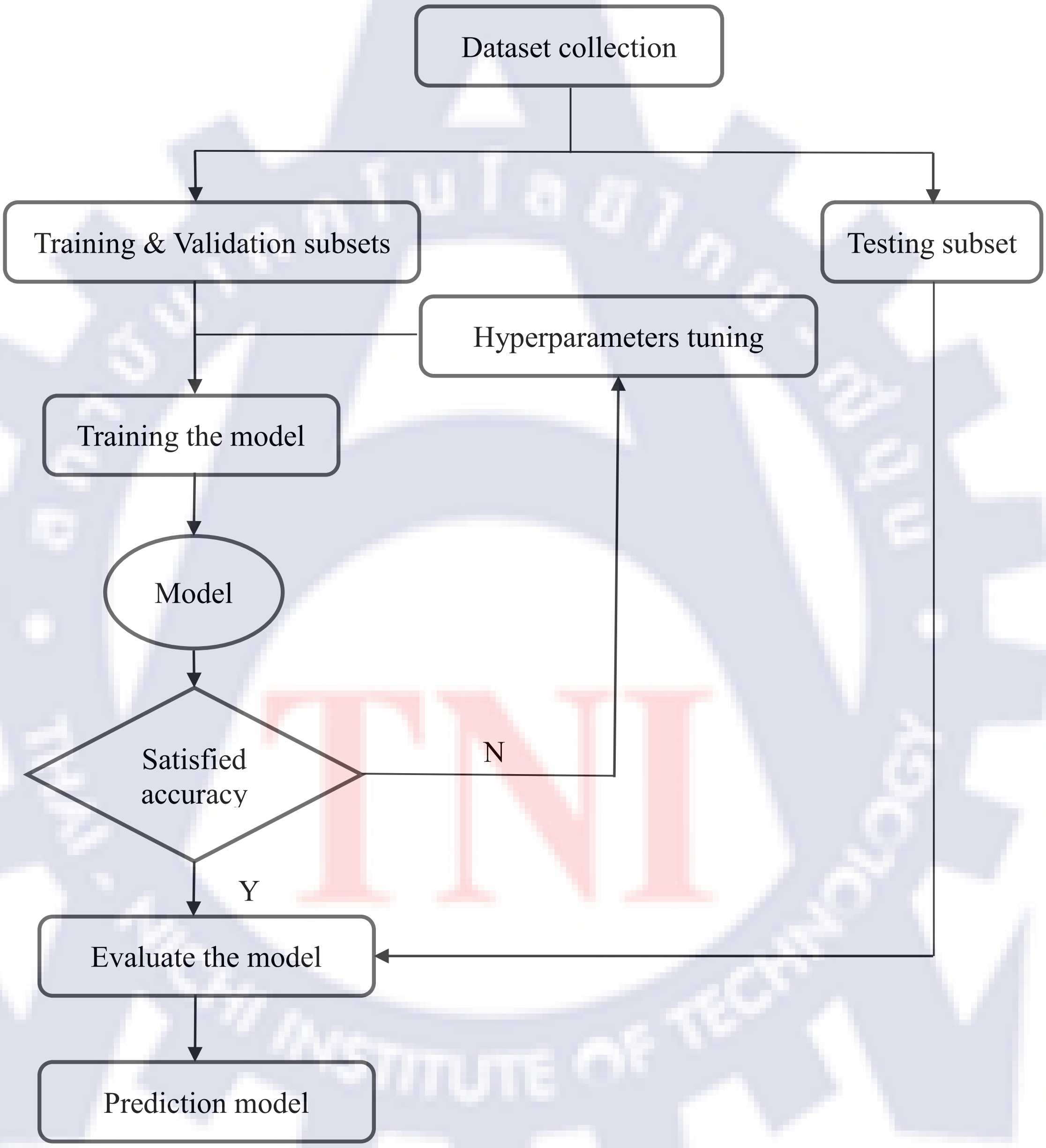


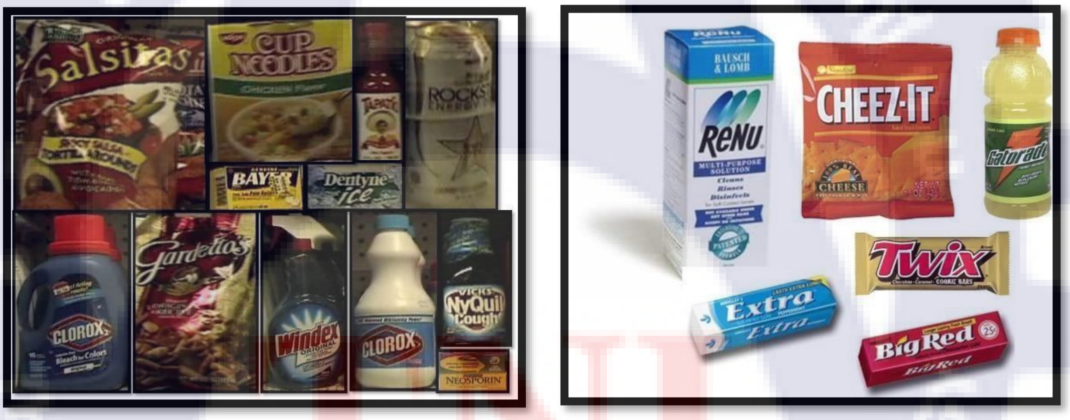
Figure 3.1: Object detection for Retail Product Recognition Workflow

3.1 Dataset Collection

In this paper, the experiment uses the public standard dataset to assist in testing models and comparing the results. This study features three datasets (Grozi-120, SKU110K, and Freiburg Groceries), concisely summarized in Table 3.1. All datasets are retrieved from the Roboflow website.

GroZi-120 [54]: This dataset contains 120 grocery classes with variations in color, size, and shape. Each item has two types of images: high-quality studio images (in vitro) and natural environment images (in situ). The in situ images include 4,973 annotated in situ test images and 29 test videos showcasing the products on various shelves. The on-shelf in situ images are low resolution.

The dataset comprises 10,958 images, divided into 7,502 training images, 2,323 validation images, and 1,133 test images. Figure 3.2 illustrates the sample images of in situ and in vitro photos in the Grozi-120 dataset.



in situ

in vitro

Figure 3.2: The example of in situ and in vitro in Grozi-120 dataset images

SKU110K [55], [56]: This dataset has bounding box annotations for objects found on store shelves, and it consists of a single class named ‘object.’ SKU110K contains 9,998 densely packed supermarket shelf images with over 1.7 million annotated bounding boxes and SKU category labels captured in densely packed shelf images from thousands of supermarkets worldwide. Every picture is resized to a resolution of one megapixel. There are several scales, lighting circumstances, camera angles, and noise levels. The example photos of SKU110K are shown in Figure 3.3.

The SKU110K dataset is split into 6,998 images for training purposes, 2,000 images for validation, and 1,000 images set aside for testing.

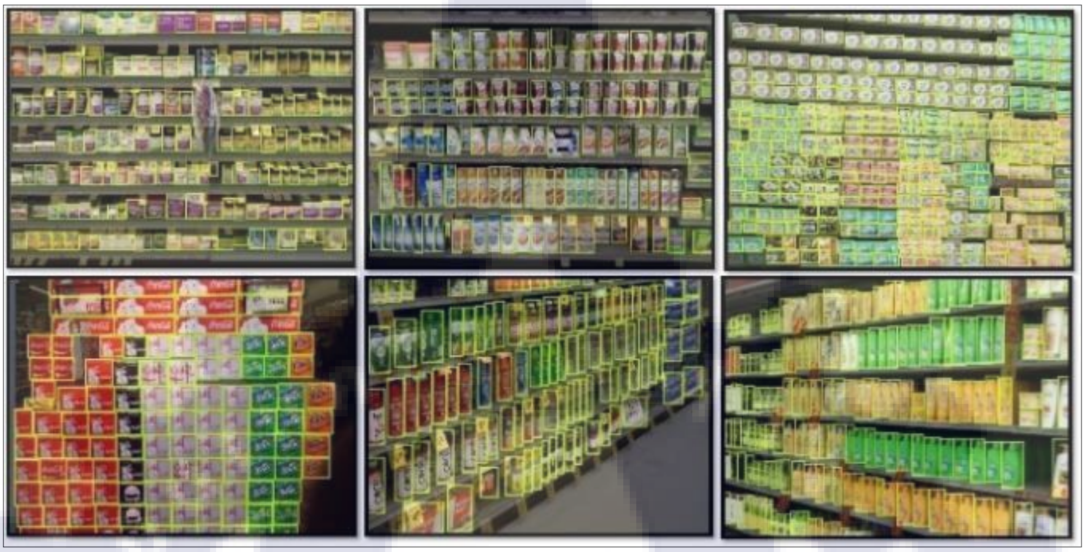


Figure 3.3: The example of SKU110K Dataset images

Freiburg Groceries [57], [58]: The Freiburg Groceries dataset contains 4,933 images covering 25 different classes of grocery products, with 97 to 370 images per class. The images are captured using four different smartphone cameras in several locations, including stores, offices, and apartments in Freiburg, Germany. This dataset is available for allowed download. Images with various aspect ratios were padded to squares. The dataset is divided into 3,946 training images, 493 validation images, and 494 test set images. Figure 3.4 shows example pictures from the Freiburg Groceries dataset.

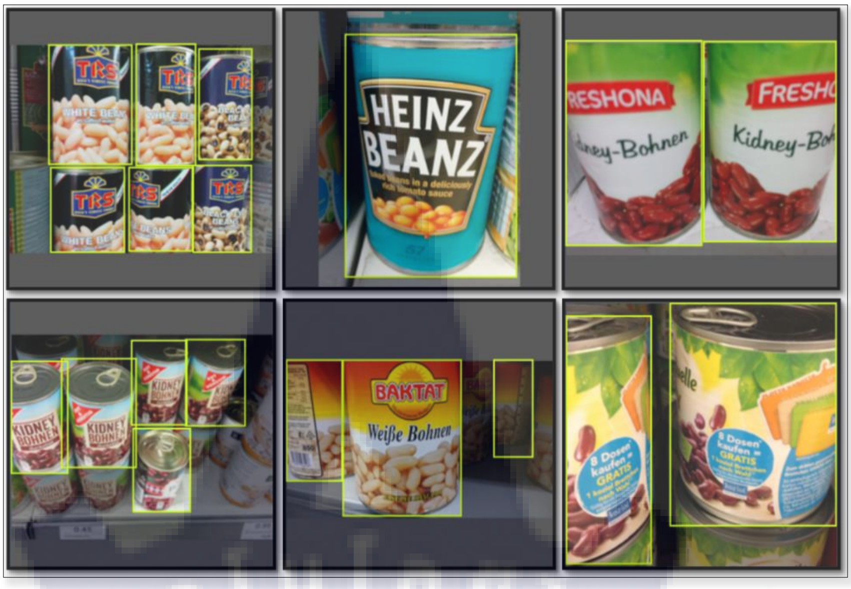


Figure 3.4: The example of Freiburg Groceries Dataset images

Table 3.1: Data Splitting Descriptions

Dataset	Training	Validation	Testing	Total
GroZi-120	7,502	2,323	1,133	10,958
SKU110K	6,998	2,000	1,000	9,998
Freiburg Grocery	3,946	493	494	4,933

3.2 Training, Validation, and Testing subsets

To create a product recognition model, the data is divided into three distinct datasets: a training set, a validation set, and a test set. The training set represents the data that trains the model. During each epoch, the model is repeatedly trained on this data in the training set.

The validation set is a set of data, separated from the training, used to validate the model during the training. This validation process provides the information used to adjust the hyperparameter settings. The particular configurations used are detailed in Table 3.2.

Table 3.2: Hyperparameters Configuration

Hyperparameters	
Parameters	**Value**
Optimizer	Adam
Color	RGB
Input Image size	640*640
Epochs	150
Patience	75
Batch size	16

The model will classify each input from the validation set during the training. The classification will be executed based only on what it is learned about. One reason is that the model needs the validation set to assure that the model does not overfit the data in the training set.

Overfitting is the condition that the model becomes good at classifying the data in the training set, but it cannot generalize and make accurate classifications on data that it is not trained. If the training results are excellent, but the results on the validation data are lagging. Then, the model is likely overfitting.

The test set is data used to evaluate the model's performance after it has been trained. It is separated from both the training set and the validation set as the unseen data.

In the scope of this research, the dataset collections are partitioned into a train set, a validation set, and a test set in a 70:20:10 ratio. Figure 3.5 demonstrates the dataset splitting.

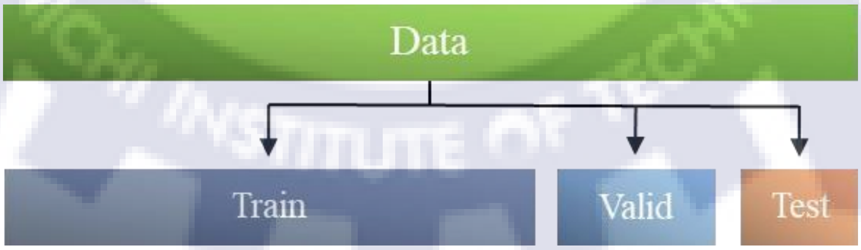


Figure 3.5: Dataset Splitting

3.3 Transfer learning

Transfer learning is a machine learning approach which reduces training time for the Deep learning model. A pre-trained model is reused as the beginning point of a new work. This is practical when the new work is similar to the first work. When the pre- trained is trained, it will give a result. When it is reused and trained for the second task, it provides a different result. Furthermore, transfer learning avoids overfitting since the model already has learned data features that are useful in the new work. Transfer learning aims to increase a target learner's performance by applying other relevant data. Unlike traditional machine learning techniques, the training and testing data come from the same feature space and distributions [59]. Transfer learning workflow is illustrated in Figure 3.6.

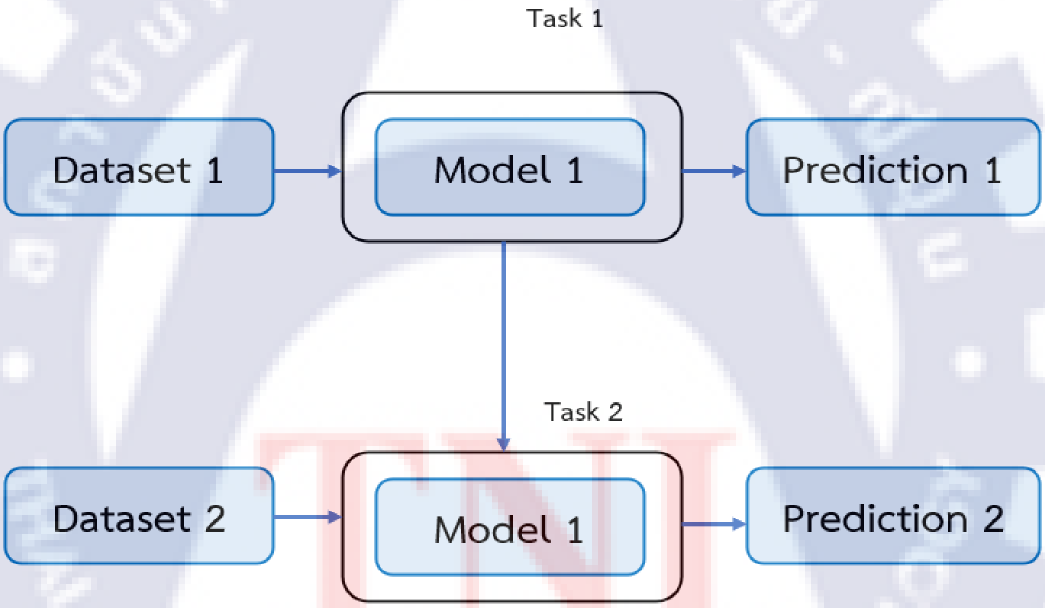


Figure 3.6: Transfer learning workflow

3.4 YOLO Models Training

Object detection [60] is an essential field in computer vision that identifies objects and locates them in images or videos. YOLO, one of the best object detection algorithms, is the single-stage neural network that instantaneously predicts object bounding boxes and class probabilities. It has been developed until now, YOLOv8.

YOLOv8 model can detect objects quickly and reliably. Thus, in this experiment, YOLOv8 is selected and compared with different YOLOv8 versions to determine the best performance.

The images are divided into smaller pictures for better small object detection using ReLU [41] activation. YOLOv8 deploys the modified CSPDarknet53 backbone [20] as YOLOv5. However, the CSPLayer, used in YOLOv5, is substituted by the C2f module. Besides, it utilizes PAN-FPN at the neck process after the extraction by Darknet53.

YOLOv8 also adopts a new anchor-free model with a decoupled head [20] that allows the pattern to reduce the number of box predictions and parameters to speed up the training. Moreover, it uses a new feature pyramid network architecture that improves feature extraction, better recognizes objects of different sizes, and improves accuracy. Thus, it gains more competence and can widely be applied in CPU and GPU.

In this research experiment, the sharing configurations are 150 epochs, a patience value of 75, and a batch size of 16 as shown in Table 3.2.

3.5 Evaluation Metrics

Evaluation metrics are the machine learning model performance assessment. They offer quantitative measures for selecting models and tuning hyperparameters. To evaluate the effectiveness of detection metrics, Intersection over union (IoU), accuracy, precision, recall, F1-score, and mean average precision (mAP) were employed as benchmarks in the evaluation of product recognition. The experiments were conducted on the tested dataset based on the confusion matrix. The equations were presented respectively.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \qquad (1)$$

IoU is used to assess the object localization accuracy, measuring the overlap between the ground truth and predicted bounding boxes. IoU scores can be interpreted as follows:

- IoU = 0: False Negative: This unique case means the model completely missed an object it should have detected.
- IoU >= 0.5: True Positive: The predicted bounding box significantly overlaps with the ground truth box, indicating the object has been correctly detected and identified.
- IoU < 0.5: False Positive: This indicates the model has drawn a bounding box, but the object within the box is misclassified.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2}$$

Accuracy evaluates how many of the models predict the results correctly. The aim is to make the fewest number of mistakes. TN stands for true negatives. TP is for true positives. FN and FP denote a false negative and a false positive, respectively. The sum of TP and TN indicates the correct prediction in total, whereas the total of executed prediction is the sum of TP, TN, FP, and FN. This dataset is vulnerable due to its imbalance. For example, if only 1% of the data set is positive labels, and the guess everything is negative, the achievement is 99% accurate.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

Precision quantifies how well the ML model correctly predicts the target labels. It describes that when an object exists, the bounding box created by the model is precise in proportion to the precision value. Precision can detect potential accuracy manipulation by evaluating the reliability of positive predictions. The system aims to optimize these metrics to make as few mistakes as possible when predicting the positive labels.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

Recall describes how accurately the ML model identifies true positives from all the actual positive samples within the dataset. It means the ability of the model to accurately identify a portion of the actual ground truths according to recall value and classify them to a particular class. The goal of this system is to find and to try every positive label there is to be found.

$$\text{F1-score} = 2x\,\frac{(\text{Precision x Recall})}{\text{Precision + Recall}} \tag{5}$$

The F1-score is subsequently assessed through the balance between precision and recall. Both factors, precision, and recall, must be optimized concurrently to maximize the F1-Score's results. The score indicates how good the quality of the predictions is and how completely the model has predicted the labels from the dataset.

$$\text{mAP} = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}} \text{AP}_\text{i} \tag{6}$$

The mAP describes the model's accuracy. The high mAP means the model has fewer false positives (FP) and fewer false negative (FN) rate. The higher the mAP, the better the model precision, and the higher the recall. This indicates the model is trustworthy in identifying objects with a high degree of accuracy while avoiding incorrect detections. Improvement of data quality and algorithm optimization can enhance the mAP outcomes.

Chapter 4
Experiment and Results

This section describes the experiment of YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x with three datasets, GroZi-120, SKU110K, and Freiburg datasets. Each model is trained, evaluated, and then compared to find the best performance.

4.1 Experimental Environment

This experiment operates on YOLO version 8 in conjunction with Python version 3.10.12. The underlying framework is provided by Torch version 2.1.0, with CUDA version 12.1 enabling hardware acceleration on an NVIDIA A100-SXM4-40GB GPU (with 40514 MB of VRAM).

The operating system runs on Linux Ubuntu 22.04.3 LTS (Jammy Jellyfish) within the Google Colab Pro+ environment. For computational resources, the system is equipped with an Intel (R) Xeon (R) CPU @ 2.00GHz (8-core) with 51 GB of RAM and 202 GB of data storage capacity.

The experimental environment to train the models is summarized in the Table 4.1 below:

Table 4.1: Experimental Environment

Cloud-hosted	Google Colap Pro+
Algorithm	YOLO Version 8
CPU	Intel (R) Xeon (R) CPU @ 2.00GHz (8-core)
GPU	NVIDIA A100-SXM4-40GB GPU with 40514 MB of VRAM
RAM	51 GB
Data Storage	202 GB
Operating System	Linux Ubuntu 22.04.2 LTS (Jammy Jellyfish)
Software Environment	Python 3.10.12, Torch 2.1.0, CUDA 12.1

4.2 Experimental Results

This study evaluates the performance of YOLOv8, a leading object detection framework. The analysis uses six essential metrics: IoU, Precision, Recall, F1-score, mAP50, and mAP50-95 to provide the aspects of the model's capabilities. To ensure robust results, the choice of 150 epochs and a patience setting of 75 are executed.

Table 4.2 demonstrates a detailed breakdown of YOLOv8's performance metrics on the Grozi-120 dataset. YOLOv8s showed the highest precision (29.1%), while YOLOv8x provided the best overall performance in Recall (29.5%), mAP50 (30.6%), and mAP50-95 (30.1%).

Figure 4.1 illustrates the example images displaying the YOLOv8x model's performance when tested on the Grozi-120 dataset. In each image, the first number represents the class ID, whereas the second number indicates the confidence score given by the model to the identified object.

Table 4.2: Experiment Results with GROZI-120 Dataset

Grozi-120				
Model	***Precision***	***Recall***	***mAP50***	***mAP50-95***
YOLOv8n	0.270	0.266	0.293	0.288
YOLOv8s	**0.291**	0.268	0.301	0.296
YOLOv8m	0.288	0.286	0.303	0.298
YOLOv8l	0.284	0.285	0.302	0.297
YOLOv8x	0.252	**0.295**	**0.306**	**0.301**

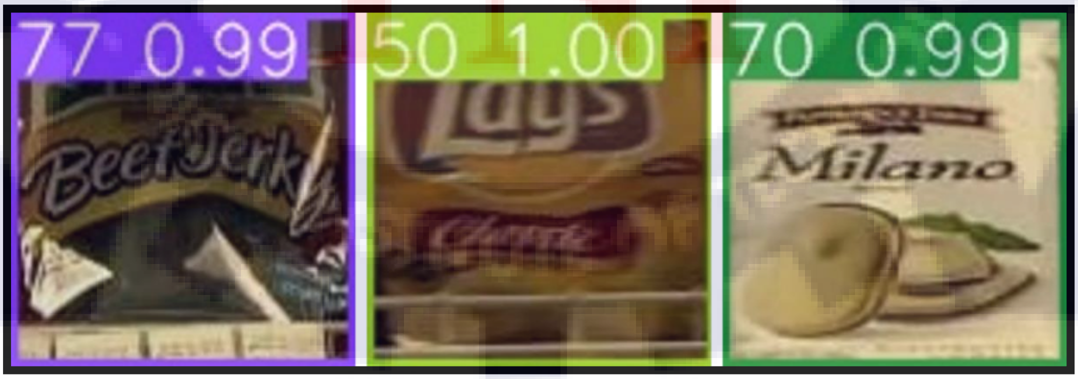


Figure 4.1: The Example of YOLOv8x Working with Grozi-120 Test Set

Table 4.3 illustrates the YOLOv8 performance for each metric using the SKU110K dataset. This analysis underscores YOLOv8x's dominance, with it achieving the highest scores in Recall (87.5%), mAP50 (92.6%), and mAP50-95 (59.6%). Interestingly, YOLOv8l demonstrates the best precision (29.1%); however, the precision result is not significantly different from that of other YOLOv8 versions.

YOLOv8x still consistently performs well across metrics, which establishes its position as a robust choice for object detection tasks.

Figure 4.2 displays the example of SKU110K image results testing with YOLOv8x.

Table 4.3: Experiment Results with SKU110K Dataset

SKU110K				
Model	***Precision***	***Recall***	***mAP50***	***mAP50-95***
YOLOv8n	0.896	0.833	0.901	0.564
YOLOv8s	0.902	0.854	0.915	0.581
YOLOv8m	0.905	0.866	0.922	0.590
YOLOv8l	**0.907**	0.871	0.925	0.595
YOLOv8x	0.906	**0.875**	**0.926**	**0.596**

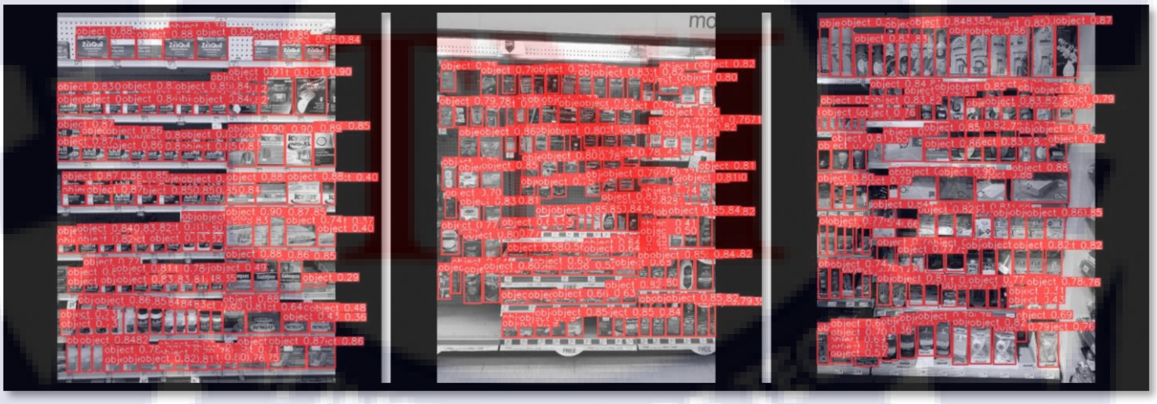


Figure 4.2: The Example of YOLOv8x Performance with SKU110K Test Set

Table 4.4 presents the performance of YOLOv8 on the Freiburg Groceries dataset. Among the different YOLOv8 variations, YOLOv8m was the most precise, achieving a score of 88.7%, while YOLOv8s demonstrated the highest recall score (82.1%). When considering overall performance, YOLOv8x obtained the top spot with the best mAP50 (89.7%) and mAP50-95 (77.5%) scores.

Figure 4.3 shows examples of the YOLOv8x performance with the Freiburg Grocery test set represented in images. The numerical value following the object class in each image indicates the confidence level of the detection.

Table 4.4: Experiment Results with Freiburg Groceries Dataset

Freiburg Groceries				
Model	***Precision***	***Recall***	***mAP50***	***mAP50-95***
YOLOv8n	0.818	0.777	0.847	0.711
YOLOv8s	0.844	**0.821**	0.881	0.742
YOLOv8m	**0.887**	0.81	0.894	0.769
YOLOv8l	0.864	0.812	0.895	0.774
YOLOv8x	0.873	0.8	**0.897**	**0.775**

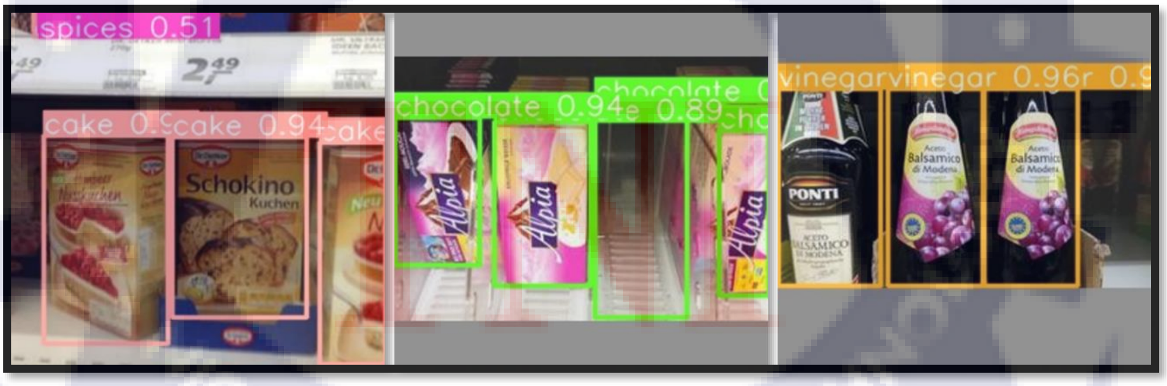


Figure 4.3: The Example of YOLOv8x Performance with Freiburg Test Set

To evaluate the model, the mAP is applied as a primary performance metric in the experiment. It is calculated by generating a precision-recall curve for each class, using varying Intersection over Union (IoU) thresholds. The average precision (AP) is determined from this curve. The threshold represents the IoU used to evaluate detected objects in object detection. After calculating the AP for each class in the dataset, the mAP is finally computed.

4.3 Performance Analysis

The experiment shows that YOLOv8x gets the best scores in overall performance when evaluated with those three datasets (Grozi-120, SKU110K, and Freiburg Groceries). The bounding box is precise in proportion to the precision value, with a high detection rate of 90.6%. Additionally, over 80% of relevant elements can be detected. The model's accuracy varies with the IoU threshold; for example, YOLOv8x shows an accuracy of 92.6% at 0.5 and 59.6% at thresholds between 0.5 and 0.95 in the SKU110K dataset. Tables 4.2, 4.3, and 4.4 demonstrate a correlation between model size and accuracy; larger models perform better. However, this increased accuracy comes with the trade-off of slower execution and larger storage needs.

The breakdown of the YOLOv8x model's performance is explained below, accompanied by an analysis and interpretation:

4.3.1 F1-Confidence Curve

The F1-Confidence curve is a tool used in machine learning to evaluate a classification model performance, especially YOLOv8x in this case. The curve shows how the F1-score varies with different confidence levels.

Grozi-120: The F1-score of 0.28 indicates a balanced trade-off between precision and recall. As Figure 4.4 shows, this optimal F1-score is reached at a confidence threshold of 0.893. In other words, the model achieves the best balance between recall and precision when only considering predictions with a confidence level of 0.893 or higher.

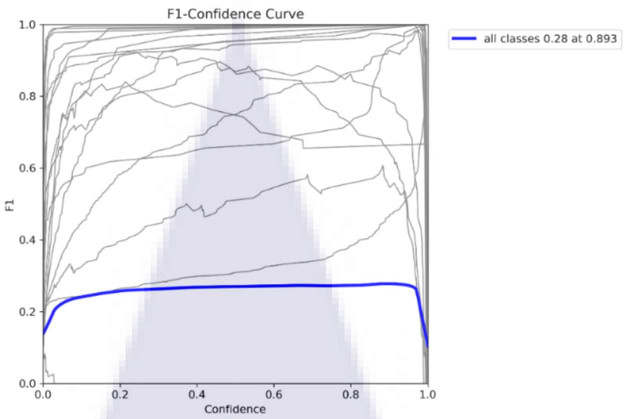


Figure 4.4: F1-Confidence Curve: YOLOv8x Working on Grozi-120 Dataset

SKU110K: The model demonstrates a well-balanced trade-off between precision and recall, reaching an impressive F1-score of 0.89 at a threshold of 0.351. This suggests that the model attains the top balance between recall and precision by only considering predictions with a confidence level of 0.351 or higher. This indicates the good performance of identifying both positive and negative cases. Figure 4.5 demonstrates the correlation results.

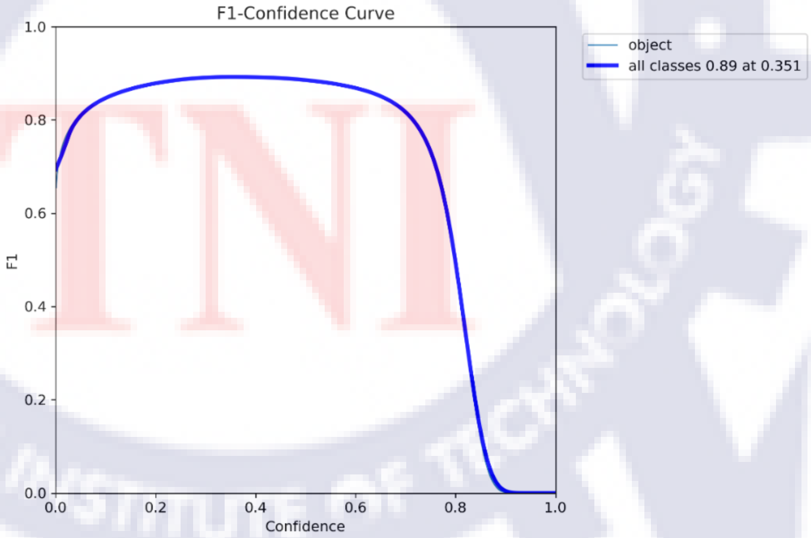


Figure 4.5: F1-Confidence Curve: YOLOv8x Running on SKU110K Dataset

Freiburg Groceries: Figure 4.6 illustrates the YOLO8x implementation on the Freiburg Groceries dataset. The optimal balance between precision and recall, measured by the F1-score, is attained at a threshold of 0.490, resulting in an F1-score of 0.86. The model effectively balances precision and recall, performing well across varying confidence levels.

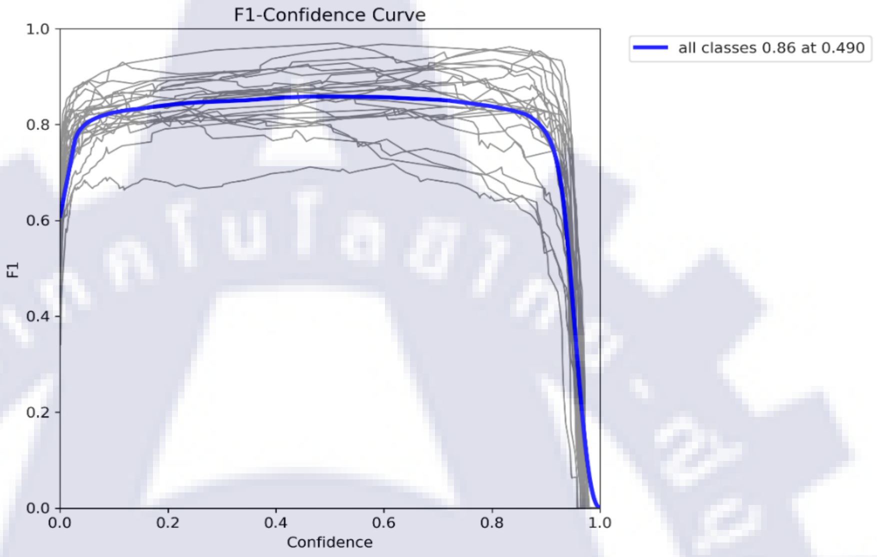


Figure 4.6: F1-Confidence Curve: YOLOv8x Act on Freiburg Groceries Dataset

4.3.2 Precision-Confidence Curve

The Precision-Confidence Curve serves as a mechanism in machine learning to evaluate the effectiveness of object detection models. It shows how the model's precision varies with the different confidence thresholds.

Grozi-120: Figure 4.7 illustrates the precision curve and reveals a perfect precision score of 1.00 at a threshold of 0.34. This indicates the model achieves 100% accuracy in identifying relevant instances for all classes at this confidence level. Although the confidence is low, the model shows high precision in detecting objects. This means that even at the highest confidence level, there is still a significant part of false positives.

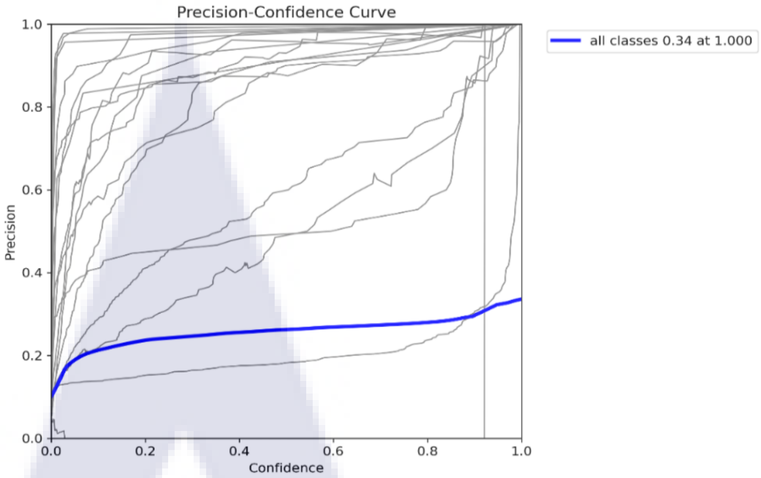


Figure 4.7: Precision-Confidence Curve: YOLOv8x Working on Grozi-120 Dataset

SKU110K: At a threshold of 0.943, the precision curve demonstrates perfect precision (1.00), meaning the model correctly identifies all relevant instances with an accuracy of 100% for all classes. When the model predicts the existence of an object, it is typically correct. Figure 4.8 shows the Precision curve of the YOLOv8x performance on the SKU110K dataset.

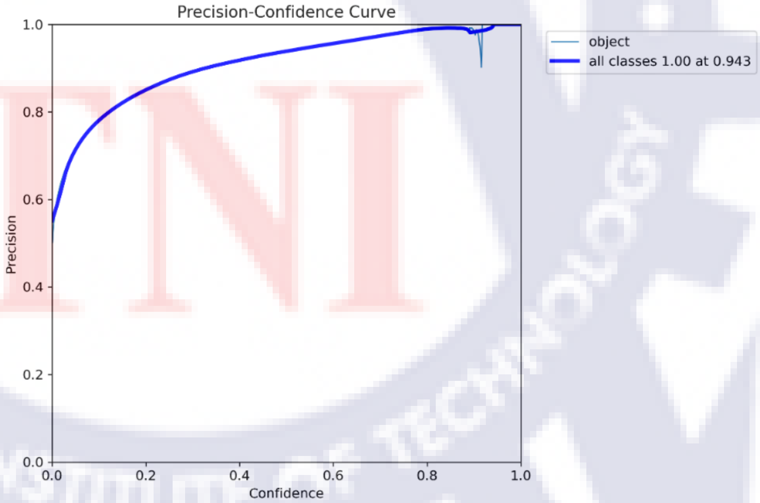


Figure 4.8: Precision-Confidence: YOLOv8x Performance on SKU110K Dataset

Freiburg Groceries: Figure 4.9 illustrates the precision curve for YOLOv8x on the Freiburg Groceries dataset. It indicates the perfect precision (1.00) at a threshold of 0.991. This means that at the confidence level of 0.991, the curve reaches a precision of 1.0. YOLOv8x accurately identifies all relevant instances across all classes with 100% precision.

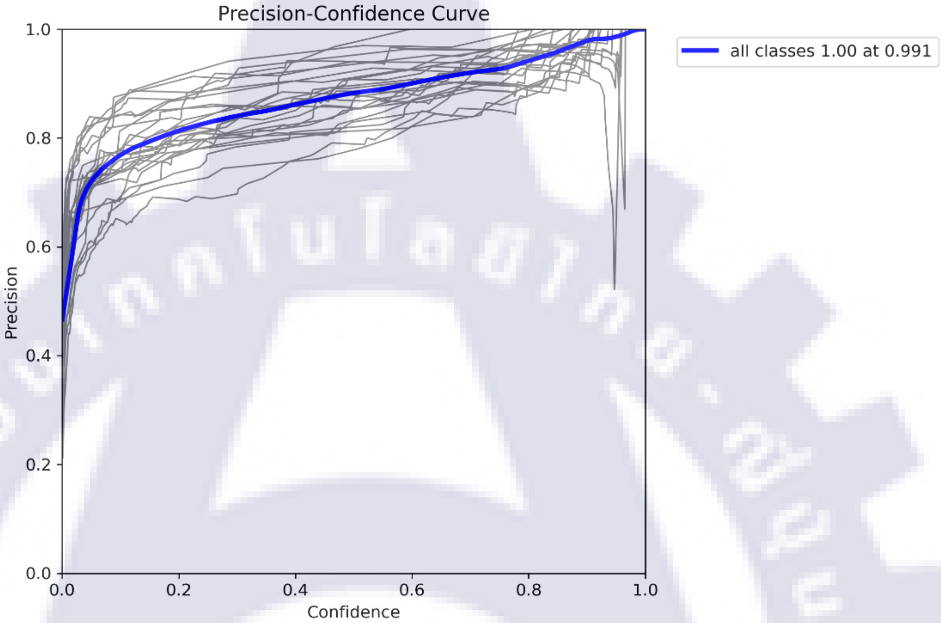


Figure 4.9: Precision-Confidence: YOLOv8x Act on Freiburg Groceries Dataset

4.3.3 Recall-Confidence Curve

The Recall-Confidence Curve, applied in machine learning, measures the performance of object detection models. It illustrates the relationship between recall and confidence scores.

Grozi-120: The Recall curve indicates a recall value of 0.33 (33%) at the threshold of 0.000. This means that at the lowest confidence threshold (0.000), the YOLOv8x model can successfully identify 33% of the actual positive instances across all classes. Figure 4.10. illustrates that as the confidence threshold increases, the recall value decreases. This shows that there is a trade-off between recall and confidence.

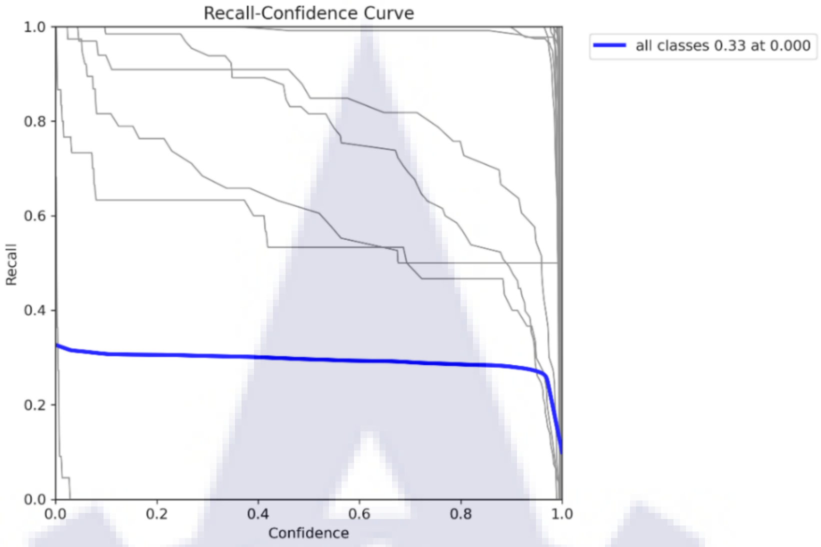


Figure 4.10: Recall-Confidence Curve: YOLOv8x Performance on Grozi-120 Dataset

SKU110K: The Recall curve shows a recall value of 0.94 (94%) at the threshold of 0.000. This suggests that at the lowest confidence threshold (0.000), the YOLOv8x model can successfully identify 94% of the actual positive instances across all classes. Figure 4.11 presents the Recall-Confidence Curve pointing to the performance of YOLOv8x on the SKU110K dataset.

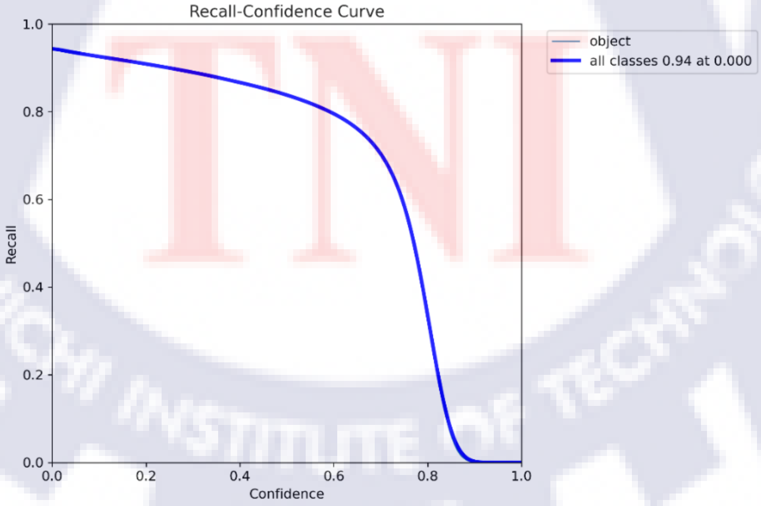


Figure 4.11: Recall-Confidence Curve: YOLOv8x Performance on SKU110K Dataset

Freiburg Groceries: The Recall curve shows that the model can successfully identify 96% of actual positive objects across all classes at the threshold of 0.000. This suggests that the model can find 96% of all ground truth objects even with the lowest confidence threshold. Figure 4.12 presents the Recall-Confidence Curve demonstrating the YOLOv8x model's performance on the Freiburg Groceries dataset.

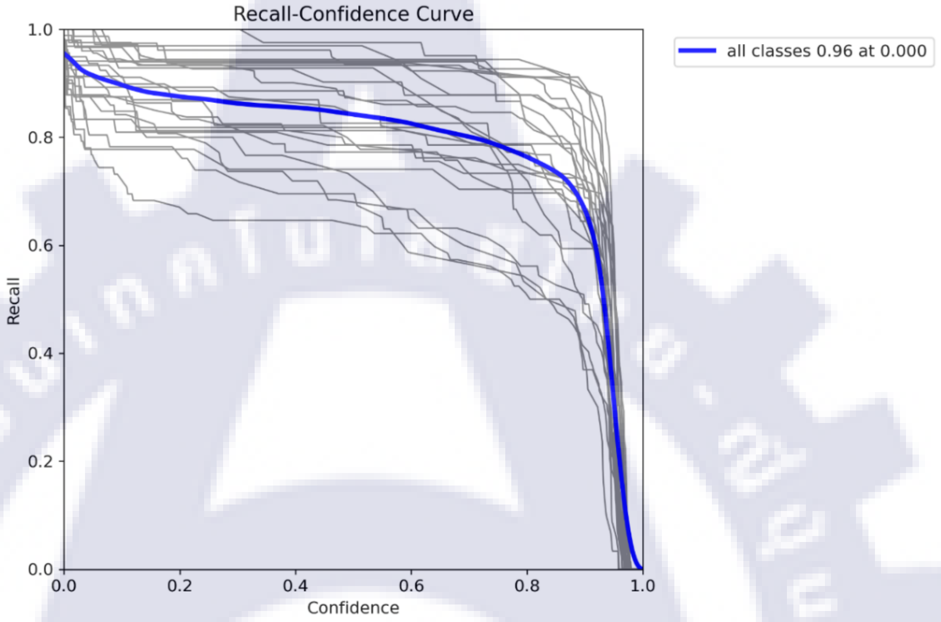


Figure 4.12: Recall-Confidence Curve: YOLOv8x Performance on SKU110K Dataset

4.4.4 Precision-Recall Curve

The precision-recall (PR) curve demonstrates the balance of a model between precision (the ability of the model to identify only correct objects) and recall (the ability of the model to find objects that exist.) for a given object detection model.

Grozi-120: The mAP@0.5 value of 0.308 can be interpreted as the model having a mean average precision of 30.8% at a confidence threshold of 0.5.

The model starts with high precision (close to 1.0, near 100%) when it is highly confident in its predictions. It is very accurate. However, this high confidence threshold at low recall values means it misses many objects in the dataset. Figure 4.13 demonstrates the PR curve showing the model's performance on the Grozi-120 dataset.

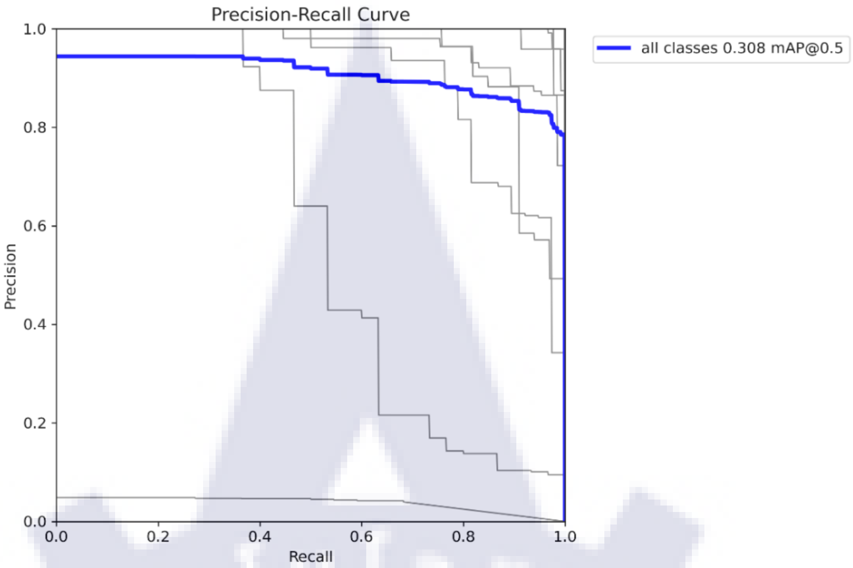


Figure 4.13: Precision-Recall Curve: YOLOv8x Performance on Grozi-120 Dataset

SKU110K: The Precision-Recall (PR) curve has an average precision (AP) value of 0.929 at the IoU threshold of 0.5, which is considerably high. This AP value indicates that the model performs well in detecting the object class overall. Moreover, it achieves a good balance between precision and recall. The PR curve of the model performance on the SKU110K dataset is demonstrated in Figure 4.14.

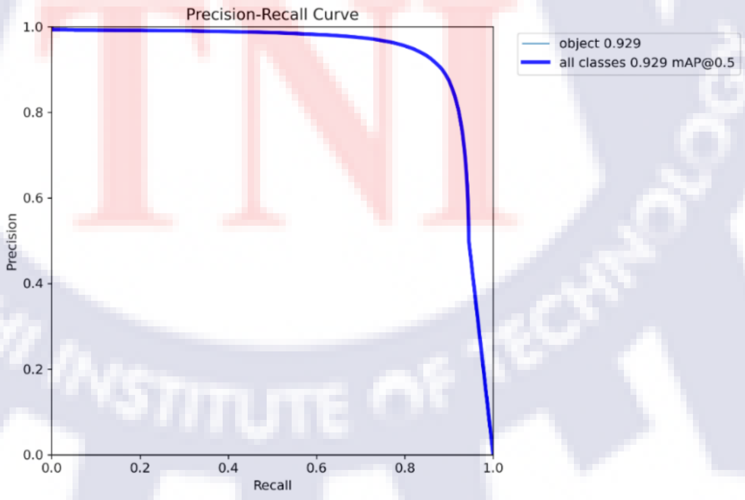


Figure 4.14: Precision-Recall Curve: YOLOv8x Performance on SKU110K Dataset

Freiburg Groceries: The mean average precision (mAP@0.5) of 0.922 indicates good performance. The model effectively identifies positive objects correctly with a balance of precision and recall when the intersection over union (IoU) threshold is set to 0.5. Figure 4.15 demonstrates the YOLOv8x performance on the Freiburg Groceries dataset using a precision-recall (PR) curve.

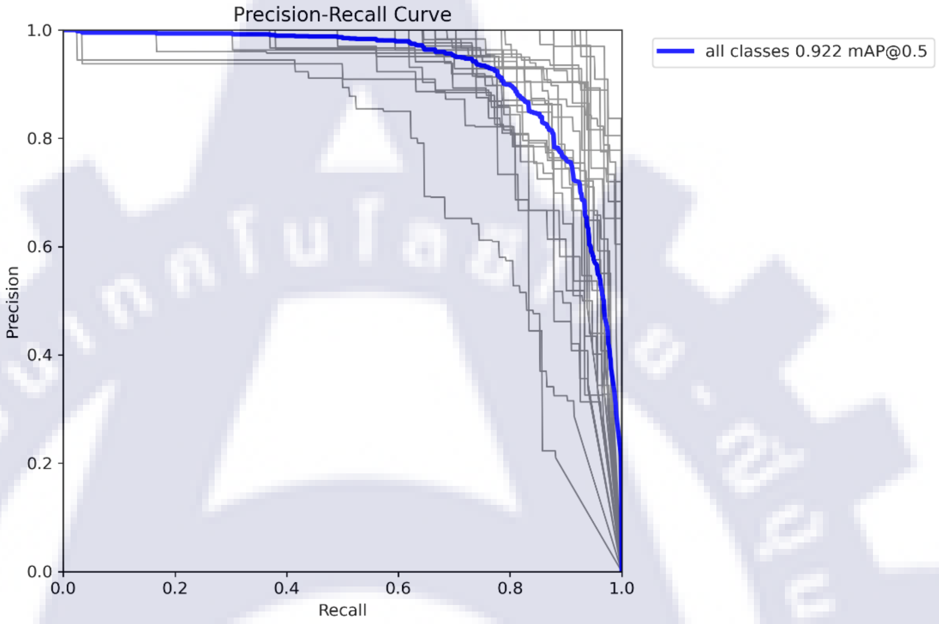


Figure 4.15: Precision-Recall Curve: YOLOv8x Act on Freiburg Groceries Dataset

4.4.5 Confusion Metrix

A Confusion Metrix is a tool that is also employed to evaluate the model's performance by comparing the actual values against the predicted values [61].

Grozi-120: The matrix shows a main diagonal line of darker blue squares, indicating that the model frequently gets correct predictions. It is also good at identifying many of the classes.

Furthermore, the darker blue square where "True 3" intersects "Predicted 3" shows that the model accurately identifies objects of class 3.

However, there are a few areas where misclassifications are more frequent than others. For example, the model incorrectly identifies items belonging to class 63 as belonging to class 66 or 69. Figure 4.16 illustrates the Confusion Metrix chart describing YOLOv8x performance on the Grozi-120 dataset.

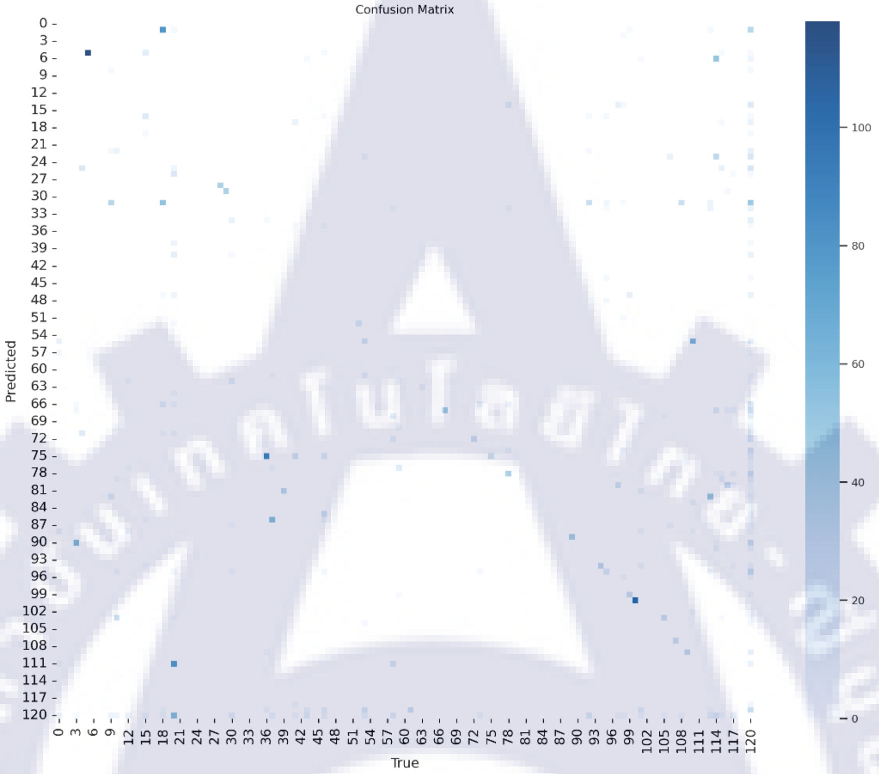


Figure 4.16: Confusion Metrix: YOLOv8x Performance on Grozi-120 Dataset

SKU110K: There are two categories: "object" and "background". The x-axis shows the true categories, while the y-axis represents the model's prediction.

In the top right corner, the model incorrectly predicted "background" (negative class) as an "object" (positive class)." 35,752 instances were classified as "object" but were actually "background".

This error type is called false positive (FP). It is misclassification, where the model predicts the existence of an object when it actually does not exist.

Furthermore, the model accurately predicted "object" (positive class) in the top left cell of the matrix, and the true class is also "object." 272,151 instances were rightly identified. True Positive (TP) represents the successful detection of an "object".

Besides, 26,742 instances were classified as "background" but were actually "objects" in the bottom left box. The mistaken prediction is a False negative (FN), which represents the failure to detect an "object".

For the True Negatives, this value is not directly visible, but it would be instances correctly classified as "background".

The model has a promising performance in identifying actual "objects" (272,151 correct classifications). However, there is a noticeable number of false positives (35,752), which indicates that there are errors in the background areas of the objects. The number of false negatives (26,742) is also significant, meaning the model sometimes misses objects and labels them as background.

Figure 4.17 describes the Confusion Metrix of YOLOv8x on the SKU110K dataset.

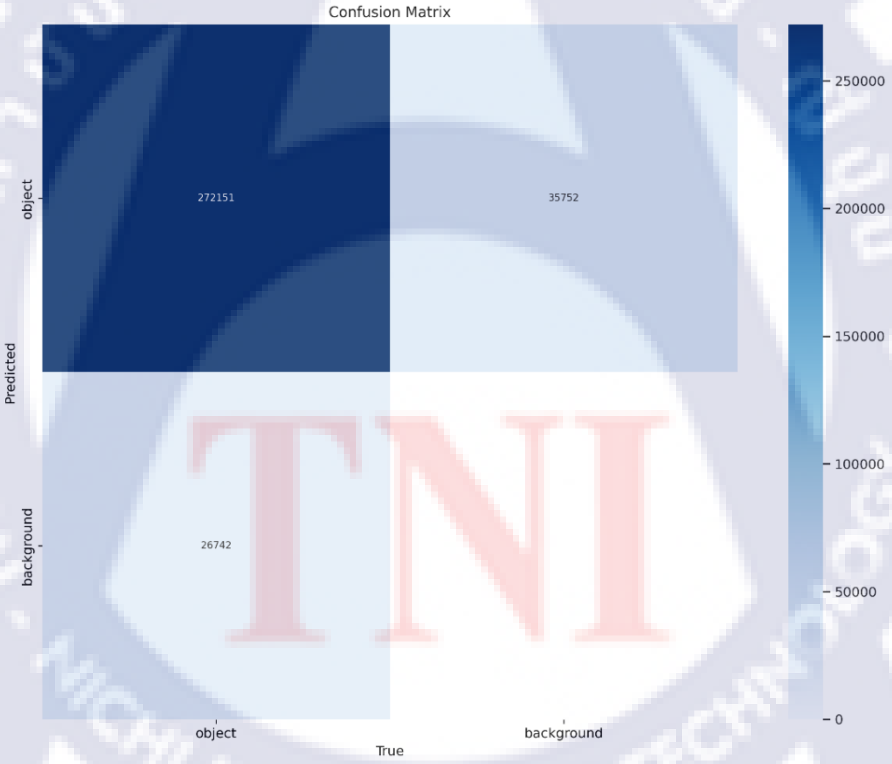


Figure 4.17: Confusion Metrix: YOLOv8x Performance on SKU100K Dataset

Freiburg Groceries: The Confusion Metrix chart demonstrates a solid line of dark blue squares going diagonally, meaning the model frequently makes correct predictions. It is good at detecting the items on which it has been trained. For example, the darkest blue square where “True beans” meets “Predicted beans” indicates the model precisely recognizes beans.

However, some lighter blue squares from this line show that the model sometimes confuses one item for another. For example, there is distinguished confusion between “chocolate” and “coffee,” noticed by the lighter blue box where labels cross.

Figure 4.18 presents the Confusion Metrix results from YOLOv8x on the Freiburg Groceries dataset.

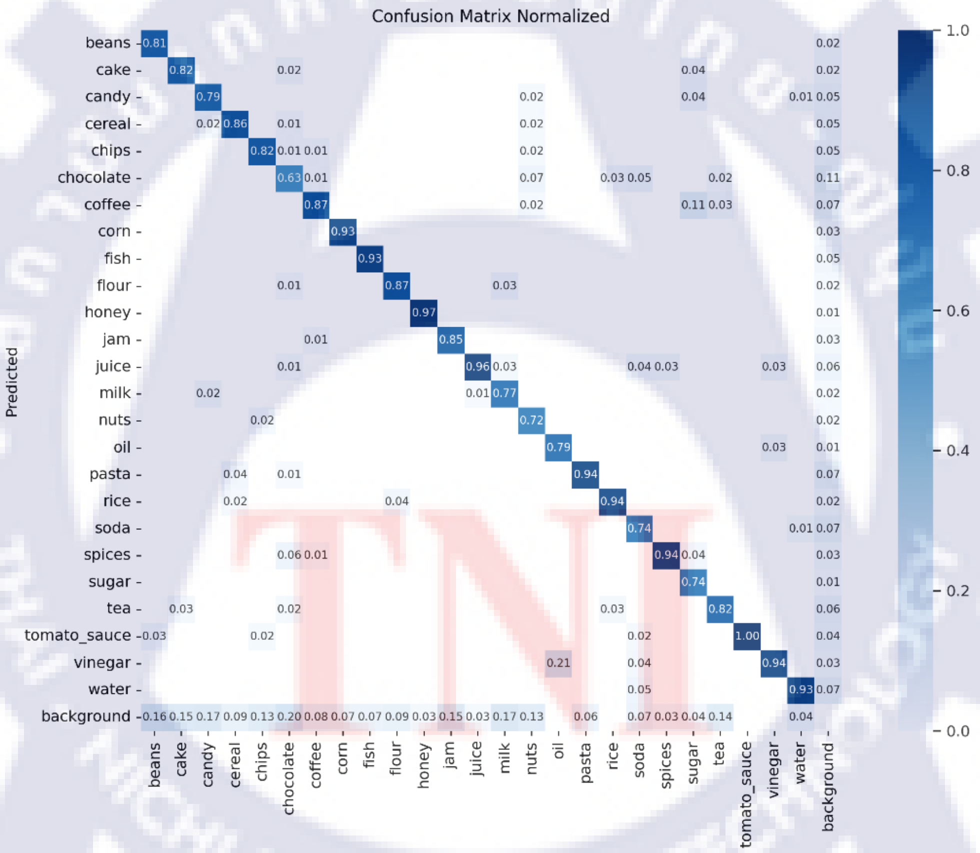


Figure 4.18: Confusion Metrix: YOLOv8x Act on Freiburg Groceries Dataset

Chapter 5
Conclusion and Future Works

5.1 Key Findings

The research found that YOLOv8 algorithms are practical tools for identifying and classifying retail products on shelves in retail stores. This is important for many tasks, such as inventory management, planogram compliance, reducing human errors, preventing sales loss, enhancing the shopping experience, and maintaining customer satisfaction.

This research paper applied the different versions of YOLOv8 algorithms (nano, small, medium, large, and extra-large) for Grozi-120, SKU110K, and Freiburg Groceries datasets. YOLOv8 experiment has been carried out through several stages of the data science process, including data collection, data splitting, transfer learning, data training, data evaluation, hyperparameter fine-tuning, and data evaluation. The evaluation metrics are IoU, precision, recall, F1-score, mAP50, and mAP50-95. YOLOv8 proves capable of detecting retail products on shelves, achieving notable accuracy scores. Performance varies based on dataset complexity: 92.6% on the SKU110K, 89.7% on the Freiburg Groceries, and 30.6% on the more challenging Grozi-120 dataset. Based on the results, YOLOv8x achieved the best overall performance. However, Grozi-120 performed poorly due to its large dataset and many classes, which impacted the model and led to lower mAP50 and mAP50-95 scores. The low performance reveals that the varied classes and many different product types remain challenging for object detection algorithms.

The PR curve indicates that SKU110K and Freiburg Groceries dataset show a good performance with YOLOv8x, achieving the values of 0.929 and 0.922 mAP@0.5, respectively. This suggests the model has a high average precision across all classes at a recall threshold of 0.5. The model is effective in accurately detecting positive objects. In contrast, YOLOv8x performance on Grozi-120 is less impressive, with a mAP@50 value of 0.308. This suggests that while presenting high confidence in its predictions, the model tends to miss many objects in the dataset at lower recall rates.

For the confusion metrics, the test results with the SKU110K dataset indicate that although the model can detect instances well (0.91), it fails to predict the objects and background correctly. This leads to many mistaken classifications. As a result, the model needs to be developed to reduce false positives (1.00). Besides, the metrics also show good performance when tested with Freiburg Groceries and Grozi-120 datasets, even though it needs some improvement and further fine-tuning to reduce the misclassifications.

Additionally, this study deployed the YOLOv8 algorithm to challenge the inter-class similarity, a common issue in recognizing visually similar retail products distinguished by color or element variations in particular areas. The results revealed that the predicted bounding box does not accurately cover the actual ground truth bounding box, leading to the conclusion that the model failed to classify the objects correctly, causing the prediction to be incorrect. One influencing factor to this misclassification was the low resolution and clarity of the test images. The small and unclear images hindered the model's ability to identify and classify the products accurately.

In simpler terms, datasets with fewer classes, such as the Freiburg Groceries, the 25-class dataset, showed more robust performance and yielded higher accuracy. YOLOv8 can effectively handle moderate complexity. Meanwhile, SKU110K, the single-class dataset, achieved the highest whole accuracy and mAP scores. YOLOv8's performance showed its effectiveness in handling focused product categories. This demonstrates an inverse relationship between the number of dataset classes and model accuracy: the accuracy decreases as the dataset complexity increases (more classes). Moreover, the evaluation indicates a positive correlation between model size and accuracy, with larger models consistently performing better. However, the large model size must trade-off with more extensive storage and slower execution.

Finally, according to the experimental results, YOLOv8x has the ability to be applied in a real-world object detection algorithm for retail product recognition and to be beneficial to the retail industry.

5.2 Contributions and Future Works

For the contribution of this paper, the research completes the literature gap by providing an inclusive analysis and the experiment exploration of the performance of different YOLOv8 versions across three public retail product datasets, Grozi-120, SKU110K, and Freiburg Groceries datasets, since no previous research papers have examined.

Future research should expand on this work by experimenting with more diverse datasets. Moreover, different YOLO model comparisons could also generate valuable insights. Combining YOLOv8 with other object detection algorithms or techniques should be conducted to get more results for the analysis. The real-world implementation should be planned for a further step to identify the practical applications and the limitations of algorithms in retail settings.

References

References

[1] R. Mostaghel et al., “Digitalization driven retail business model innovation: Evaluation of past and avenues for future research trends,” *Journal of Business Research*, vol. 146, no. C, pp.134–145, July 2022.

[2] S. Sunetra et al., “Factors that influence retail store preference and impact of in-store digitization,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 4, pp. 8715–8722, November 2019.

[3] Y. Wei et al., “Deep learning for retail product recognition: Challenges and techniques,” *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, pp. 1-23, November 2020.

[4] B. Santra and D. P. Mukherjee, “A comprehensive survey on computer vision based approaches for automatic identification of products in retail store,” *Image and Vision Computing*, vol. 86, no. C, pp. 45-63, June 2019.

[5] J. Flores et al., “RFID technology in storage management: A bibliometric study on efficiencyand cost reduction in the retail sector,” *International Journal of Electronics and Communication Engineering*, vol. 10, no. 10, pp. 1–13, October 2023.

[6] I.H. Sarker, “AI-based modeling: Techniques, applications and research issues towards automation, intelligent and Smart Systems,” *SN Computer Science*, vol. 3, no. 2, pp. 1-20, February 2022.

[7] I.H. Sarker, “Machine learning: Algorithms, real-world applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, pp. 1-21, March 2021.

[8] J. Suk et al., “Consumers’ perceived benefits and costs for amazon go based on social media data usingtext mining,” in *24th International Conference on Human-Computer Interaction (HCI International 2022)*, Berlin, Germany, June 26-July 1, 2022, pp. 221–236.

[9] R. Brečić et al., “Local foodsales and point of sale priming: Evidence from a supermarket field experiment,”*European Journal of Marketing*, vol. 55, no. 13, pp. 41–62, April 2021.

[10] J. Chen and C. Yu-Wei, "How smart technology empowers consumers in smart retail stores? the perspective of technology readiness and situational factors," *Electronic Markets*, vol. 33, no. 1, pp. 1-24, April 2023.

[11] J. Laitala and L. Ruotsalainen, "Computer vision based planogram compliance evaluation," *SSRN Electronic Journal*, vol. 13, no. 18, pp. 10145, September 2023.

[12] X. Wang et al., "Recent advances in deep learning," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, pp. 747-750, April 2020.

[13] S. Sengupta et al., "A review of deep learning with special emphasis on architectures, applications and recent trends," *Knowledge-Based Systems*, vol. 194, no. 105596, pp. 1-78, April 2020.

[14] A. Agrawal et al., "Tensorflow eager: A multi-state, python-embedded DSL for machine learning," *Proceedings of the ACM on Programming Languages (PACMPL)*, Athens, Greece, October 20-25, 2019, pp. 1-12.

[15] K. Sakthivel et al., "Traffic sign recognition system using CNN and Keras," *International Journal of Health Sciences*, vol. 6, no. S3, pp. 4986–4994, June 2022.

[16] T. Diwan et al., "Object detection using yolo: Challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, August 2022.

[17] J. Kaur and W. Singh, "Tools, techniques, datasets and application areas for object detection in an image: A review." *Multimedia Tools and Applications*, vol. 81, no. 27, pp. 38297–38351, April 2022.

[18] L.C. Jiao et al., "New generation deep learning for video object detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3195–3215, August 2022.

[19] S.S.A. Zaidi et al., "A Survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, no. 11, pp. 103514, June 2022.

[20] B. More and S. Bhosale, "A comprehensive survey on object detection using Deep Learning," *Revue d'Intelligence Artificielle*, vol. 37, no. 2, pp. 407–414, April 2023.

[21] L. Alzubaidi et al., “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, pp. 1-74, March 2021.

[22] G. Yao et al., “A review of convolutional-neural-network-based action recognition,” *Pattern Recognition Letters*, vol. 118, no. 1, pp. 14–22, February 2019.

[23] A. Dhillon and G.K. Verma, “Convolutional neural network: A review of models, methodologies and applications to object detection,” *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, December 2019.

[24] J. Redmon et al., “You only look once: Unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 26–July 1, 2016, pp. 779–788.

[25] H. Liu et al., “Object detection and recognition system based on computer vision analysis,” *Journal of Physics: Conference Series*, vol. 1976, no. 1, pp. 012024, July 2021.

[26] J.R. Terven and D.M. C.-Esparza, “A comprehensive review of YOLO: From YOLOv1 and beyond,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680-1716, November 2023.

[27] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019),* New York, USA, December 8, 2019, pp. 8026-8037.

[28] T.Y. Lin et al., “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*, Zurich, Switzerland, September 6-12, 2014, pp. 740–755.

[29] Ambitious-Octopus et al., *“Ultralytics YOLOv5 Architecture,”* Ultralytics YOLO [Online]. Available: https://docs.ultralytics.com/yolov5/tutorials/ architecture_description/ [Accessed: August 5, 2024].

[30] G. Ghiasi et al., “Simple copy-paste is a strong data augmentation method for instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, June 19–25, 2021, pp. 2918–2928.

[31] H. Zhang et al., “Mixup: Beyond empirical risk minimization,” in *The International Conference on Learning Representations (ICLR)*, Vancouver, British Columbia, Canada, April 30–May 3, 2018, pp. 1–13.

[32] A. Buslaev, et al., “Albumentations: Fast and flexible image augmentations,” *The Information Journal*, vol. 11, no. 2, pp. 125, February 2020.

[33] C. Wang et al., “Yolov7: Trainable bag-of- freebies sets new state-of-the-art for real-time object detectors,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*Vancouver, British Columbia, Canada, June 18–22, 2023, pp. 7464–7475.

[34] C.Y. Wang et al., “Designing network design strategies through gradient path analysis,” *Journal of Information Science and Engineering*, vol. 39, no. 4, pp. 975-995, July 2023.

[35] G. Jocher and Sergiuwaxmann, “*YOLOv7: Trainable Bag-of-Freebies,*” Ultralytics YOLO [Online]. Available: https://docs.ultralytics.com/models/yolov7/#how-do-i-install-and-run-yolov7-for-a-custom-object-detection-project [Accessed: October 5, 2024].

[36] G. Jocher et al., “*YOLO by Ultralytics,*” Ultralytics [Online]. Available: https://github.com/ultralytics/ [Accessed: March 13, 2024].

[37] Z. Zhang et al., “Distance-iou loss: Faster and better learning for bounding box regression,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, February 7-12, 2020, pp. 12993-13000.

[38] X. Li et al., “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, New York, USA, December 6, 2020, pp. 21002-21012.

[39] X. Chu et al., “Make repVGG greater again: A quantization-aware approach,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, New Orleans, Louisiana, USA, June 19-24, 2022, pp. 252-253.

[40] G. Jocher, “*YOLO-NAS,*” Ultralytics YOLO [Online]. Available: https://docs.ultralytics.com/models/yolo-nas/ [Accessed August: 6, 2024].

[41] Roboflow Inc., "*COCO Dataset Computer Vision Project*," Roboflow Universe [Online]. Available: https://universe.roboflow.com/microsoft/coco [Accessed: February 29, 2024].

[42] Roboflow Inc., "*PASCAL VOC 2012 Image Dataset*," Roboflow Universe [Online]. Available: https://universe.roboflow.com/jacob-solawetz/pascal-voc-2012/dataset/9 [Accessed: February 29, 2024].

[43] F.K. Konstantinidis et al., "Automating dairy production lineswith the yoghurt cups recognition and detection process in the industry 4.0 era," *Procedia Computer Science*, vol. 217, no.9, pp. 918–927, January 2023.

[44] S.C. Kamireddi, *"Comparison of Object Detection Models – to Detect Recycle Logos on Tetra Pack*," B.Sc. Thesis (Computer Science), Blekinge Institution Techonology, Karlskrona, Sweden, 2022.

[45] T. Mahendrakar, "Performance study of yolov5 and faster R-CNN for autonomous navigation around non-cooperative targets," *2022 IEEEAerospace Conference (AERO)*, Montana, USA, March 5-12, 2022, pp. 1339.

[46] D. Li and J. Huang, "A learning-based approach towards automated tuning of SSD configurations," *ACM Transactions on Storage (TOS)*, vol. 18, no. 3, pp. 1-35, September 2022.

[47] K. He et al., "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 22-29, 2017, pp. 2961-2969.

[48] T. Vu et al., "A Yolo-based real-time packaging defect detection system," *Procedia Computer Science*, vol. 217, no. 1, pp. 886–894, January 2023.

[49] P. H. Toranzo et al., "Detection and verification of the status of products using Yolov5," *Proceedings of the 20th International Conference on Smart Business Technologies*, Bangkok, Thailand, July 13-15, 2023, pp. 83-93.

[50] H. Kumar, *"Computer Vison (Ai) Based Retailer Shelves Monitoring System to Notify Empty Shelves,"* M.S. Thesis (Information Technology), International Institute of InformationTechnology Bangalore, Bengaluru, Kamataka, 2023.

[51] J. Abyasa et al., "Yolov8 for product brand recognition as inter-class similarities," *2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, Nanjing, China, August 25-27, 2023, pp. 1-6.

[52] S. Prakash, P. Shah, and A. Agrawal, “Exploiting CNNs for semantic segmentation with pascal VOC,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, Louisiana, USA, June 19-14, 2022, pp. 1-24.

[53] R. Li and J. Yang, “Improved yolov2 object detection model,” in *2018 6th International Conference on Multimedia Computing and Systems(ICMCS)*, Rabat, Morocco, May 10-12, 2018, pp. 1–6.

[54] Roboflow Inc., “*Grozi120 Computer Vision Project,*” Roboflow Universe [Online]. Available: https://universe.roboflow.com/retvispublic-dutr5/grozi120 [Accessed: February 25, 2024].

[55] Roboflow Inc., “*SKU 110 Computer Vision Project,*” Roboflow Universe [Online]. Available: https://universe.roboflow.com/jacobs-workspace/sku-110k [Accessed: February 25, 2024].

[56] Ultralytics, “*SKU-110K,*” SKU-110K - Ultralytics YOLOv8 [Online]. Available: docs.ultralytics.com/datasets/detect/sku-110k [Accessed: March 12, 2024].

[57] Roboflow Inc., “*Freiburg-Groceries Image Dataset,*” Roboflow Universe [Online]. Available: https://universe.roboflow.com/michael-ringer/freiburg-groceries/dataset/10 [Accessed: May 5, 2024].

[58] P. Jund et al., “The freiburg groceries dataset,” in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, Santiago, Chile, December 11-13, 2016, pp. 1-7.

[59] A. Farahani et al., “A concise review of transfer learning,” in *2020 International Conference on Computational Science and Computational Intelligence (CSCI),* Las Vegas, NV, USA, December 16-18, 2020, pp. 344-349.

[60] T. Diwan et al., “Object detection using yolo: Challenges, architectural successors, datasets and applications,” *MultimediaTools and Applications*, vol. 82, no. 6, pp. 9243–9275, August 2022.

[61] S. Visa et al., “Confusion matrix-based feature selection,” *Expert Systems with Applications*, vol. 39, no. 3, pp. 120-127, January 2011.

Biography

Name	Mr. Nakul Pannoy
Date of Birth	November, 21, 1976
Address	199/119 The Alcove Thonglor 10, Sukhumvit 63, Klongton nua, Wattana, Bangkok 10110 E-mail: NakulDotGov@gmail.com

Educational Background

2024	Master of Science in Information Technology Thai-Nichi Institute of Technology
2004	Master of Science in Marketing Thammasat University
1998	Bachelor of Arts in Economics Thammasat University

Working Experiences

2024-Present	ICT Assistant UNICEF, Thailand
2004-2024	Computer Management Assistant (LAN) Embassy of the United States of America, Bangkok
1999-2004	Lead Expeditor and Shipment Assistant Embassy of the United States of America, Bangkok
1998-1999	Internal Auditor Nestlé Food Co., Ltd, Bangkok