MediCaption: Integrating YOLO-Driven Computer Vision and NLP for Advanced Pharmaceutical Package Recognition and Annotation

1st Aarthi Lakshmipathy
1189423
Department of Computer Science
Lakehead University

2nd Madhurima Vardhineedi
1192227
Department of Computer Science
Lakehead University

3rd Venkata Ramana Patnaik Sekharamahanthi
1187427
Department of Computer Science
Lakehead University

4th Devanshi Dineshbhai Patel
1194465
Department of Computer Science
Lakehead University

5th Saurav Saini
1185835
Department of Computer Science
Lakehead University

Abstract

In the realm of medicinal safety and comprehension, this project presents a groundbreaking system integrating advanced image analysis, text recognition, and database management technologies. ”MediCaption” integrates cutting-edge computer vision and deep learning technologies, capitalizing on YOLO for real-time analysis of pharmaceutical packaging. The system harmonizes precise object detection with sophisticated natural language processing (NLP) techniques to generate comprehensive image captions, thereby refining diagnostic precision and patient care. We propose a solution that entails meticulous dataset preparation, image standardization, and iterative model fine-tuning with YOLOv7 using non-traditional evaluation metrics. Text extraction ensures the delivery of user-friendly and regulation-compliant medicinal information. The establishment of an extensive prescription database facilitates a thorough examination of drug interactions, offering personalized summaries for healthcare professionals to enhance patient care. The project addresses accessibility challenges and pioneers transformative applications of computer vision in pharmaceutical quality control.

Index Terms

Medication Safety, Image Analysis, Text Recognition, Database Management, YOLOv7, Optical Character Recognition, Natural Language Processing, Personalized Healthcare, Drug Interaction Analysis.

I. Introduction

The ascendancy of computer vision technology is progressively assuming a pivotal role across diverse industries, notably in the realms of pharmaceuticals and healthcare. This surge in prominence is concurrent with the escalating utilization of visual data. Image processing emerges as a versatile tool with multifaceted applications, extending from optimizing pharmaceutical administration to enhancing the quality of patient care [13]. The imperative for cutting-edge technology is impelled by the healthcare sector’s relentless pursuit of precision in treatment modalities and the dynamically evolving nature of the pharmaceutical industry. The pressing need for highly skilled management of drug-related information, encompassing elements like inventory oversight, adherence to regulations, and improved accessibility for those with visual impairments, accentuates the gravity of the circumstance [2]. The precise identification of pharmaceuticals assumes paramount importance, delineating the delicate boundary between life and death. A thorough comprehension of drug identification and understanding establishes the foundational framework of this intricate ecosystem. This underscores the necessity for an intelligent system seamlessly integrating natural language processing with computer vision for the meticulous annotation and identification of medical images [23]. The emergence of computer vision, conjoined with sophisticated deep learning techniques, has revealed unparalleled prospects to address these imperative requirements, ushering in a trans-formative shift in essential research initiatives. The convergence of deep learning and computer vision pioneers inventive methodologies, culminating in groundbreaking and indispensable contributions to meet the demands of the contemporary healthcare landscape [1]. This initiative stands at the vanguard of scholarly exploration, delineating the convergence of computer vision, deep learning, image captioning and healthcare. It aims to conceive a sophisticated system that leverages YOLO, an efficacious object detection algorithm, to meticulously scrutinize both the front and back sides of pharmaceutical packaging. One significant strength of YOLO (You Only Look Once) is its unparalleled efficiency in real-time object detection [18].

YOLO processes images in a single pass, dividing them into a grid and simultaneously predicting bounding boxes and class probabilities for objects within each grid cell [18]. When compared to other techniques, YOLO's unique approach results in remarkably fast and accurate identification of objects, making YOLO a preferred choice for applications demanding timely and precise recognition. The incorporation of YOLO in the implementation of MediCaption is crucial due to its exceptional capability to accurately identify intricate details on medical packages with high precision and accuracy. By proficiently detecting unfamiliar text elements and a myriad of diverse labels, YOLO guarantees precise recognition, establishing the fundamental infrastructure for exact labeling integration within our system. Moreover, YOLO when applied to image captioning, seamlessly combines object detection and advanced natural language processing (NLP) techniques [14]. Leveraging YOLO's real-time object detection capabilities facilitates swift and comprehensive scrutiny of medical images, enabling the creation of detailed annotations. Annotation serves as the initial step in linking images to captions, providing detailed labels and descriptions for identified objects. Captioning then utilizes these annotations, transforming them into human-readable text to convey a comprehensive and contextually rich narrative associated with the image. These annotations form the fundamental basis for our image captioning approach, enhancing the complexity of medical analysis and bolstering decision-making procedures. The significance of captioning in medical imaging is substantial, as it provides contextual details and elaborate descriptions of visual data [4]. In developing our intelligent system, we are also using OCR/NLP for text extraction and apply caption pre-processing for user-friendly interpretation by ensure regulatory compliance, and we also integrate image and text analysis for a comprehensive solution. This aids in thorough documentation, precise communication among healthcare professionals, and improves accessibility for individuals with visual impairments. The precision of captions further streamlines the analysis, interpretation, and retrieval of medical images, ultimately enhancing diagnostic accuracy and patient care [4].

In the subsequent sections, we present a comprehensive research strategy designed to address the aforementioned inquiries. The document is structured as follows: Section 2 elucidates the problem statements along with the associated research questions, while Section 3 outlines the objectives of our project, Section 4 provides the essential technological background adopted in this project, Section 5 delves into pertinent previous research, Section 6 outlines the dataset created for annotating medical images and performing captioning, while Section 7 discusses prospective research avenues to train YOLO on the New Annotated Medical Imaging Dataset and generate coherent captions. Anticipated results are detailed in Section 8, and Section 9 delineates the current progress along with planned milestones.

II. Problem Statement and Research Questions

The inception of ”MediCaption” is driven by critical considerations in contemporary healthcare, where the oversight and accessibility of pharmaceuticals significantly impact patient well-being, regulatory compliance, and operational efficiency. Historically, the labor-intensive process of identifying and labeling prescription packaging, encompassing intricate details like drug names, dosages, and expiry dates, has been prone to errors, a challenge exacerbated by the growth of the pharmaceutical sector and the diversity in drug packaging. Concurrently, constructing the intelligent system ”MediCaption” that leverages YOLO object detection for precise medical image identification and annotation poses multifaceted challenges. Developing a proficient machine learning model within MediCaption capable of comprehending medical intricacies and generating clinically relevant captions stands as a pressing hurdle. Ensuring a flexible framework that manages complexities, size fluctuations, and maintains consistent, accurate object detection and captioning adds substantial complexity to framework advancement. Moreover, achieving a balance between performance and resource-intensive YOLO characteristics within MediCaption demands innovative strategies to optimize computational efficiency while upholding precision in medical image labeling and captioning.

In response to the healthcare sector's significant challenge of unclear medication labels, the ”MediCaption” project integrates computer vision and natural language processing. The research questions arises in this study are:

1) How does integrating YOLO enhance annotation accuracy and speed in MediCaption, and which techniques can optimize its real-time medical image processing efficiency?
2) Which training approaches are most effective for enabling MediCaption, equipped with YOLO, to generate precise descriptions of medical images, especially given their varying complexity?
3) In what ways can MediCaption improve the recognition of text on complex medicine labels, and what are the primary obstacles and health-related implications in understanding these labels?
4) How is MediCaption being designed to be accessible to all users, including those with limited tech knowledge, and how do Annotated Medical Image Datasets play a role in enhancing caption generation?
5) What additional benefits does YOLO bring to the use of Annotated Medical Imaging Datasets, and what are the optimal techniques for training YOLO for effective caption generation in medical contexts?

We are interrogating these inquiries and formulating an accessible and user-centric solution within the scope of this research endeavor. The ultimate goal of ”MediCaption” is to empower individuals with comprehensive medication information, fostering safer and more informed healthcare decisions. It seeks to explore the hurdles individuals face in interpreting medication labels, develop robust methods for text detection and interpretation on packaging, and create a universally accessible, user-friendly solution.

III. Objectives

1) **Enhanced Image Processing and Machine Learning Integration:** ”MediCaption” utilizes the YOLO v7 architecture [5] for advanced image identification and annotation, targeting error-prone areas in prescription packaging such as drug names, dosages, and expiry dates. This integration is aimed at reducing medication management errors. Concurrently, the project develops a proficient machine learning model capable of generating clinically relevant captions for medical images. This model addresses the complexities inherent in medical imagery, thereby enhancing the accuracy and relevance of information conveyed to users.
2) **Optimization and Accessibility in Medical Image Processing:** The project focuses on optimizing computational efficiency while balancing the performance and resource-intensive nature of YOLO. Techniques to enhance annotation accuracy and processing speed are explored to ensure real-time efficiency in medical image analysis. ”MediCaption” is designed to be accessible to a wide range of users, including those with limited technical knowledge. This is achieved through the utilization of annotated medical image datasets, which play a critical role in enhancing the system’s caption generation capabilities.
3) **Advanced Captioning for Comprehensive Medication Understanding:** A significant feature of ”MediCaption” is its advanced captioning capabilities, which are crucial for bridging the gap between medical professionals and laypersons. The system not only identifies and annotates medical images but also generates detailed, understandable captions. These captions include vital information such as drug names, dosages, potential side effects, and usage instructions, making it easier for all individuals to comprehend and manage their medications safely. This approach aims to empower users with comprehensive medication information, fostering safer and more informed healthcare decisions.

.

IV. Background

Our efforts build upon and expand the principles derived from YOLO and Natural Language Processing (NLP) integrated with image captioning. We offer an overview of the foundational concepts employed in our work.

The YOLO (You Only Look Once) object detection algorithm is a real-time, single-stage object detection system designed to efficiently detect and classify objects in images [27]. Developed by Joseph Redmon and Ali Farhadi in 2015, YOLO offers a unique approach by processing the entire image in one pass, making it faster than traditional two-stage detectors [18].

A. Description of how the YOLO algorithm works

1) Grid System: YOLO divides the input image into a grid of cells. The grid structure helps organize the detection process across the entire image.
2) Bounding Box Prediction: For each grid cell, YOLO predicts bounding box coordinates (x, y, width, height) relative to the cell. Each cell is responsible for predicting multiple bounding boxes.
3) Object Probability Prediction: YOLO predicts the probability of an object’s presence within each grid cell. This probability indicates whether an object is present in that particular region of the image.
4) Class Prediction: In addition to predicting bounding boxes and object probabilities, YOLO performs object classification by assigning class probabilities for each bounding box. This means YOLO can identify and label multiple objects within a single image.
5) Non-Maximum Suppression: After predicting bounding boxes and class probabilities, YOLO applies non-maximum suppression to remove redundant or overlapping bounding boxes. This stage guarantees the depiction of each object through the most assured bounding box prediction.
6) Single Pass Processing: YOLO processes the entire image in one pass, making it significantly faster than two-stage detectors. This is achieved by predicting bounding boxes and class probabilities for all objects simultaneously.
7) Loss Function: YOLO uses a specific loss function that combines localization loss, confidence loss, and class prediction loss. This loss function is optimized during training to improve the accuracy of bounding box predictions and object classifications.
8) Darknet Framework: YOLO was originally implemented using the Darknet framework, which is an open-source neural network framework. Darknet supports training YOLO on custom datasets, making it versatile for various object detection tasks.

B. Versions of YOLO

1) Yolo V1 :

- Single-Stage Detector: YOLOv1 was revolutionary for introducing a single-stage object detection system, processing the entire image in one pass.
- Grid-based Prediction: The algorithm divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell [27].

- Real-time Processing: YOLOv1 demonstrated real-time object detection, making it significantly faster than traditional two-stage detectors.
- Darknet Framework: Implemented using the Darknet framework, providing flexibility for training on custom datasets [27].

2) Yolo V2 :

- Introduction of Anchor Boxes: YOLOv2 introduced anchor boxes to improve bounding box predictions, allowing the model to better adapt to object scales and aspect ratios.
- Hierarchical Classification: YOLOv2 incorporated a hierarchical classification system, enabling the detection of a diverse range of objects across multiple categories [27].
- Joint Training for Multiple Datasets: YOLO9000, a variant of YOLOv2, demonstrated the ability to detect objects from over 9000 classes by jointly training on multiple datasets.
- Darknet-19 and Darknet-53 Architectures: YOLOv2 utilized these architectures for feature extraction, enhancing the model's ability to capture complex patterns [27].

3) Yolo V3 :

- Feature Pyramid Network (FPN): YOLOv3 incorporated FPN to capture features at multiple scales, improving the detection of objects of different sizes [26].
- YOLOv3 Architecture: The model employed a larger architecture with three detection scales [26], enhancing its ability to handle small and large objects.
- Darknet-53 Architecture: An upgraded version of the Darknet architecture, providing deeper and more expressive feature extraction [26].
- YOLO9000 Dataset: Continued support for a diverse range of objects and categories, maintaining compatibility with the YOLO9000 dataset.

4) Yolo V4 :

- CSPNet (Cross-Stage Partial Network): YOLOv4 introduced CSPNet to enhance the efficiency of information flow across stages, improving performance [16].
- Spatial Attention Mechanism: Incorporation of a spatial attention mechanism [16] for better focus on relevant regions in the input.
- YOLOv4 Architecture: A more advanced architecture with a larger number of parameters, further boosting accuracy.
- Hardware Acceleration: Optimized for hardware acceleration, allowing for faster inference on specialized hardware.

5) Yolo V5 :

- Introduction of YOLOv5: YOLOv5 was developed independently of the original YOLO series by the Ultralytics team [8].
- PyTorch Implementation: YOLOv5 adopted the PyTorch deep learning framework, simplifying development and research [9].
- Model Architecture Simplification: YOLOv5 focused on a simplified and streamlined architecture with various model sizes (e.g., YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x) while maintaining high performance [20].
- Efficient Inference: The model demonstrated efficient inference times on various platforms.

6) Yolo V6 :

- Anchor-Free Approach: YOLO-v6 adopts an anchor-free methodology, resulting in a remarkable 51 percent improvement in speed compared to anchor-based approaches [9].
- Decoupled Head Architecture: YOLO-v6 incorporates a decoupled head architecture, introducing additional layers to separate features from the final head [9]. This architectural enhancement has been empirically proven to enhance overall performance.
- Two-Loss Function: YOLO-v6 employs a two-loss function strategy. The Varifocal Loss (VFL) is utilized as the classification loss, derived from the focal loss [9]. VFL treats positive and negative samples with varying degrees of importance, contributing to a balanced learning signal. The Distribution Focal Loss (DFL), coupled with SIoU/GIoU, [9] serves as the regression loss. DFL is specifically applied for box regression in YOLO-v6 medium and large variants, considering the discretized probability distribution as a representation of the continuous distribution of box locations. This approach proves efficient, particularly when dealing with blurred ground truth box boundaries [9].

7) Yolo V7 :

- Architectural Reforms: YOLO V7 introduces significant architectural reforms to enhance detection accuracy and maintain high-speed performance. One notable addition is the implementation of E-ELAN (Extended Efficient Layer Aggregation Network) in the YOLO-v7 backbone [11].
- Trainable Bag-of-Freebies (BoF): YOLO V7 incorporates trainable Bag-of-Freebies, contributing to improvements in model robustness and performance [11]. This approach focuses on integrating pre-trained components that enhance the overall capabilities of the model.

- Accuracy vs. Speed Customization: Acknowledging the diverse requirements of different applications, YOLO V7 offers customization options for prioritizing either accuracy or speed. This flexibility allows users to fine-tune the model based on specific application needs [11].
- Compound-Scaling Mechanism: YOLO V7 employs a compound-scaling mechanism [11] that facilitates coherent scaling of both width and depth in concatenation-based networks. This ensures the maintenance of optimal network architecture while adapting to various input sizes.
- Parameter-Specific Scaling with NAS: YOLO V7 leverages Network Architecture Search (NAS) to enable parameter-specific scaling [11]. While NAS can identify the best factors for scaling, YOLO V7 introduces independent scaling factors, offering a nuanced approach to optimization.

8) Yolo V8 :

- Versatility Across Vision AI Tasks: YOLOv8 supports a comprehensive range of vision AI tasks, including detection, segmentation, pose estimation, tracking, and classification, enhancing its applicability across diverse domains [25].
- State-of-the-Art Backbone and Neck Architectures: YOLOv8 incorporates advanced backbone and neck architectures, resulting in superior feature extraction and enhanced object detection performance [25].
- Anchor-free Split Ultralytics Head: YOLOv8 adopts an anchor-free split Ultralytics head, a novel approach that improves accuracy and contributes to a more efficient detection process compared to traditional anchor-based methods [25].
- Optimized Accuracy-Speed Tradeoff: With a dedicated focus on achieving an optimal balance between accuracy and speed, YOLOv8 is well-suited for real-time object detection tasks in various application areas, ensuring efficient and effective performance.

C. Why we choose YOLO V7 and Advantages of YOLO V7

The preference for YOLOv7 stems from its incorporation of extensive architectural improvements and distinctive advantages. YOLOv7 introduces the extended efficient layer aggregation network (E-ELAN) into its backbone, a significant enhancement contributing to increased accuracy in object detection while maintaining swift detection speeds [5]. This architectural refinement ensures a more sophisticated and efficient structure, ultimately enhancing the model's overall performance.

An essential feature of YOLOv7 is its adaptability to diverse application requirements. Acknowledging the varying priorities of applications, whether they prioritize accuracy or speed, YOLOv7 leverages network architecture search (NAS) for precise parameter scaling [5]. This advanced mechanism enables the identification of optimal factors for different use cases, providing flexibility without compromising the model's structural integrity [5]. Additionally, YOLOv7 incorporates a compound-scaling mechanism, ensuring coherent scaling of both width and depth in concatenation-based networks [11]. This guarantees the preservation of optimal network architecture even when scaled for different sizes, offering a well-balanced approach to address diverse application needs. The combination of these architectural enhancements and scalability features positions YOLOv7 as a robust and versatile choice in the realm of object detection.

D. Image captioning in computer vision

Image captioning within computer vision involves the creation of cohesive textual depictions for images. This process melds computer vision with natural language processing, employing neural networks like CNNs to extract image features and RNNs or transformers for generating captions [19]. It leverages pre-trained models on extensive datasets comprising image-caption pairs, optimizing performance through transfer learning [20]. Image captioning serves diverse functions, such as assisting individuals with visual impairments, enabling content-based image retrieval, and enhancing human-computer interaction by furnishing descriptive narratives for images.

1) Image captioning with YOLO : The amalgamation of YOLO (You Only Look Once) with image captioning entails the seamless integration of real-time object detection capabilities with advanced natural language processing (NLP) techniques [19]. YOLO, as an efficient real-time object detection algorithm, assumes a pivotal role in discerning and pinpointing objects within images. Subsequently, the identified objects are fed into an NLP model, which, in turn, generates elucidative captions based on the perceived visual elements [14]. This fusion harmonizes the swift object detection prowess of YOLO with the nuanced contextual comprehension provided by NLP, culminating in the generation of precise and insightful image captions. This synthesis contributes to the development of sophisticated systems designed to comprehend and interpret visual content with exceptional efficacy [20].

2) Optical Character Recognition (OCR) Role in Image Captioning: The convergence of Optical Character Recognition (OCR) technology and the formidable object detection capabilities of YOLO (You Only Look Once) epitomizes a cutting-edge paradigm in image captioning. OCR, an avant-garde technological facet, specializes in the discernment and interpretation of textual entities intricately woven into images [6]. Within the domain of YOLO-driven image captioning, the synthesis of OCR emerges as a pivotal strategy for deciphering complex medical imagery, pharmaceutical labels, and other visually critical healthcare data. The collaborative synergy between YOLO's robust object detection prowess and OCR's proficiency in extracting textual information ensures a comprehensive understanding of image content [6]. Through precise identification and interpretation of text elements within images, OCR significantly contributes to the generation of meticulous, contextually rich captions. This integration serves as a linchpin in automating and optimizing the captioning process, ensuring the seamless infusion of relevant textual data into the overarching narrative [16]. Whether extracting nuanced medication details, dosage specifications, or other intricate textual constituents from pharmaceutical packaging, the harmonious amalgamation of YOLO and OCR augments the overall efficiency and precision of the captioning workflow [16]. This sophisticated integration not only streamlines the process but also amplifies the accuracy of caption generation for medical images, thereby cultivating a more discerning and insightful approach to healthcare analysis and decision-making.

V. Literature Review

A. 2021: Laying the Foundations

*1) **Healthcare Breakthroughs**:* The review of YOLO's transformative journey begins in 2021. Pioneers like Wu, Haiwen, et al., [28] and Han, Yun, et al. [7] harnessed neural networks for critical healthcare applications such as drug package text detection and blister package identification. Tan, Lu, et al. [26] took this further, implementing YOLOv3 for real-time pill identification. This was not just about technological abilities; it marked the beginning of AI's role in enhancing medication safety, although the challenges of dataset limitations and the need for highly specialized training data were apparent.

*2) **Smart City Infrastructure**:* In the urban environment, Laroca et al. [10] introduced YOLOv5 for automated license plate recognition. This technology symbolized AI's integration into the fabric of urban management, improving traffic monitoring and vehicle tracking. However, this technology faced hurdles in adapting to diverse vehicle types and environmental conditions.

B. 2022: Expanding Horizons

*1) **Cognitive Computing**:* In 2022, AI applications witnessed a significant expansion, moving beyond healthcare into more diverse fields. The research by Al-Malla, Muhammad Abdelhadie, et al. [3] on image captioning models was particularly groundbreaking. Their work in developing AI systems capable of enhanced context-awareness and human-like understanding represented a major leap in the field of cognitive computing. This advancement not only showcased the versatility of AI in understanding and interpreting complex visual data but also marked a pivotal moment in the journey towards more intuitive and intelligent AI systems, capable of bridging the gap between technology and human-like perception and interaction.

C. 2023: The Era of Advanced Integration and Diverse Applications

*1) **Healthcare Integration**:* The innovative work by Paglinawan, Charmaine C., et al., [16] in healthcare integration is a landmark achievement. They effectively combined YOLOv4 with Tesseract OCR to revolutionize medicine classification. This blend of object detection and optical character recognition in healthcare technology faced challenges, particularly in integrating two distinct AI systems. Despite these hurdles, such as handling diverse data formats and fine-tuning for various medical scenarios, their research has opened new possibilities for AI in medical diagnostics and patient care.

*2) **Public Safety and Transportation**:* Nandhakumar, R. G.,et al.'s [12] application of YOLOv3 for license plate detection marks a significant stride in traffic law enforcement. This utilization of AI technology demonstrated the adaptability of YOLOv3 in identifying traffic violators and enhancing road safety. The study navigated through challenges such as variable lighting conditions and plate obfuscation, which sometimes affected detection accuracy. Nonetheless, their work underscores the potential of YOLOv3 in traffic management and law enforcement, paving the way for broader applications in public safety.

D. 2023: Expanding YOLO Innovations

*1) **Performance Analysis**:* In their comparative study, Horvat, Jelečević, and Gledec [9] meticulously analyze the performance differences between YOLOv5 and YOLOv6, focusing on aspects such as detection speed, accuracy, and computational efficiency. Their research is vital for professionals needing to make informed decisions on model selection, considering factors like real-time processing capabilities, accuracy under varied conditions, and the balance between speed and precision. This comparison is particularly relevant in environments where computational resources are limited or where the trade-off between speed and accuracy is critical.

*2) **Commercial and Visual Data Analysis**:* Prexawanprasut, Takorn, et al. [17] expanded AI's commercial and industrial applications by using machine learning and OCR for product categorization based on packaging features, showcasing AI's growing utility in streamlining and enhancing efficiency in the commercial sector. El Abbadi, Nidhal Khdhair's [6] study further propelled the evolution of YOLO technology by focusing on scene text detection and recognition using YOLOv5 and MSERs, highlighting continual improvements in YOLO's accuracy and versatility. Complementing these studies, Terven, Juan, and Diana Cordova-Esparza [27] provided a comprehensive review of the YOLO technology from YOLOv1 to YOLOv8. Their analysis offered valuable insights into the development and advancements of this influential algorithm, tracing its journey and spotlighting its diverse applications across different domains. Together, these studies underscore the dynamic and evolving nature of AI and YOLO technology, emphasizing its increasing relevance and adaptability to various real-world challenges.

*3) **Image Captioning and Linguistics**:* Negi and Buch's [14] integration of YOLO models with LSTM networks marks a notable advancement in generating descriptive linguistics from images, enhancing AI's contextual understanding. Concurrently, Saroja and Brown Mary's work on merging YOLOv5 with the Xception V3 model pushes the boundaries of image captioning, resulting in more detailed, accurate, and context-rich descriptions. These innovations open avenues for diverse applications, ranging from enhancing digital accessibility to facilitating content creation and improving surveillance systems.

*4) **Traditional Medicine Application**:* Cheng, and Cai's [25] application of YOLOv8 in traditional Chinese medicine for the detection and classification of medicinal slices bridges modern AI technology with ancient practices. This study not only demonstrates YOLO's capabilities in handling complex pattern recognition but also highlights its potential in supporting the preservation and efficiency of traditional medicinal practices. Such applications underscore the adaptability of YOLO models in specialized fields, offering a blend of cultural heritage preservation with technological innovation.

*5) **Multimodal AI Applications**:* The exploration by Rinaldi, Russo, and Tommasino [19] in combining natural language processing with deep neural networks, potentially utilizing advanced YOLO models, signifies a major advancement in multimodal AI applications. This approach enhances AI's ability to interpret and narrate visual data, creating accurate and context-aware image captions. Such a development has vast implications, from revolutionizing digital media and education to aiding in the creation of assistive technologies for the visually impaired.

*6) **Biometric Systems**:* Dewi, Chen, and Christanto's [5] research on using the YOLOv7 model for hand recognition underscores its precision and reliability in biometric systems. This study demonstrates the model's effectiveness in accurately detecting fine details, essential for secure and efficient biometric authentication. The implications of this research extend beyond hand recognition, offering potential advancements in secure access control, identity verification, and sophisticated surveillance systems.

*7) **Smart Transportation Systems**:* Wang, and Wang's [11] enhancement of YOLOv7 for detecting small traffic signs plays a critical role in the safety and efficiency of autonomous driving systems. This advancement is not only crucial for the development of reliable autonomous vehicles but also impacts urban planning and traffic management by ensuring better recognition of road signs, leading to improved traffic flow and reduced accidents.

*8) **Sports Entertainment**:* Shashank and team's [22] creative use of YOLOv8 for generating audio commentary from cricket match videos showcases the model's adaptability in the entertainment sector, particularly in sports broadcasting. This application illustrates how AI can not only enhance the viewer experience by providing interactive and immersive content but also suggests potential for similar innovative uses in other areas of sports and entertainment, enhancing audience engagement and offering novel viewing experiences.

E. Advancements in AI Models for Image Processing and Captioning

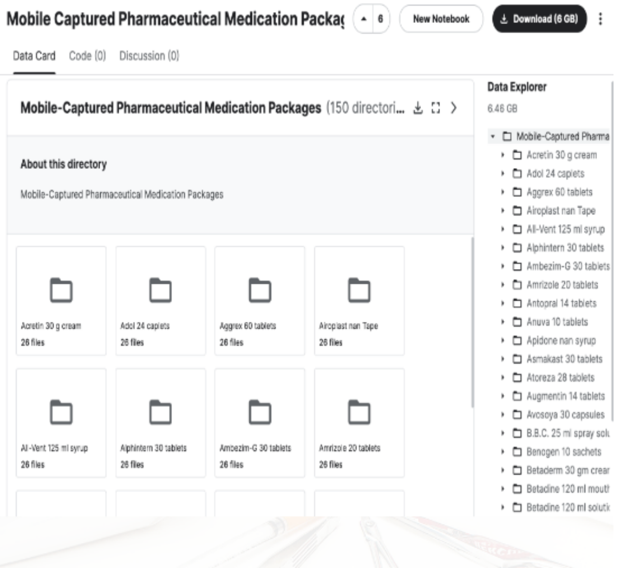
In recent years, significant strides have been made in the application and integration of advanced AI models for image processing and captioning. Shurdhaj, Elda, and Ulehla Christián [24] demonstrated the versatility of YOLO technology through its application in real-time vehicle detection within intelligent transportation systems, highlighting its potential in traffic management and smart city infrastructure. Concurrently, Saroja and Brown Mary [20] enhanced the capabilities of YOLOv5 by integrating it with the Xception V3 model, thus advancing the field of image captioning. Their work illustrated the adaptability of YOLO models for application-specific refinement. Adding to these developments, Ondeng, Ouma, and Akuon's [15] 2023 study, "A Review of Transformer-Based Approaches for Image Captioning," likely explored the synergy between advanced AI models like Transformers and image captioning techniques. This review potentially offered insights into how Transformer models, renowned in natural language processing, could amplify image captioning tasks, possibly in collaboration with YOLO models. These collective efforts mark a significant leap in the intersection of AI, computer vision, and natural language processing, shaping the future of intelligent image analysis and interpretation.

VI. Dataset

A "Drug-detection Image Dataset" is a collection of images specifically curated for training and testing machine learning models or algorithms to detect and identify drugs, medication, or pharmaceutical products in images shown in Figure 1 and 2. The key characteristics of the dataset are listed as below.

The dataset typically contains images where drugs or pharmaceutical products are present. The dataset may encompass a wide range of drug types, forms (e.g., pills, capsules, vials), and packaging styles to ensure that the model can handle diverse scenarios.The dataset include images with variations in lighting, background, orientation, and occlusions to mimic real-world conditions. The availability of labeled datasets facilitates research and the development of intelligent systems that can assist in the reliable detection and management of drugs in various healthcare settings.

We've strategically curated three distinctive datasets pivotal for our project's objectives. The primary dataset is a set of photos of pill capsules, which provides a foundation for training our model to identify medications based on subtle characteristics such as their distinct sizes, colours, and forms. To enhance this, the following dataset consists of pictures of the packaging—which



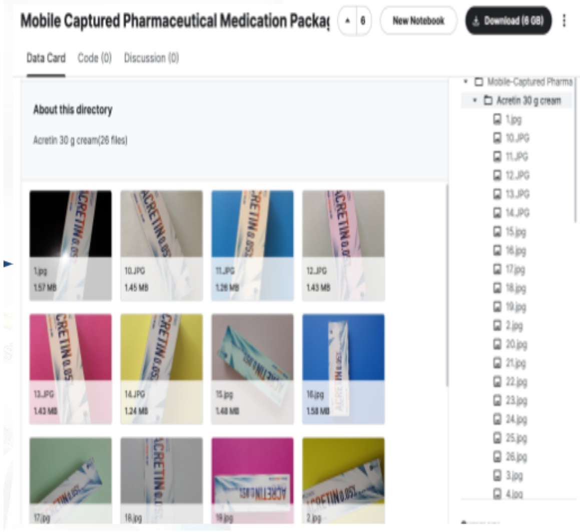
Mobile Captured Pharmaceutical Medication Packag

Mobile-Captured Pharmaceutical Medication Packages (150 directori...

About this directory

Mobile-Captured Pharmaceutical Medication Packages

Fig. 1: Dataset 1: Folders [21]



Mobile Captured Pharmaceutical Medication Packag

About this directory

Acretin 30 g cream(26 files)

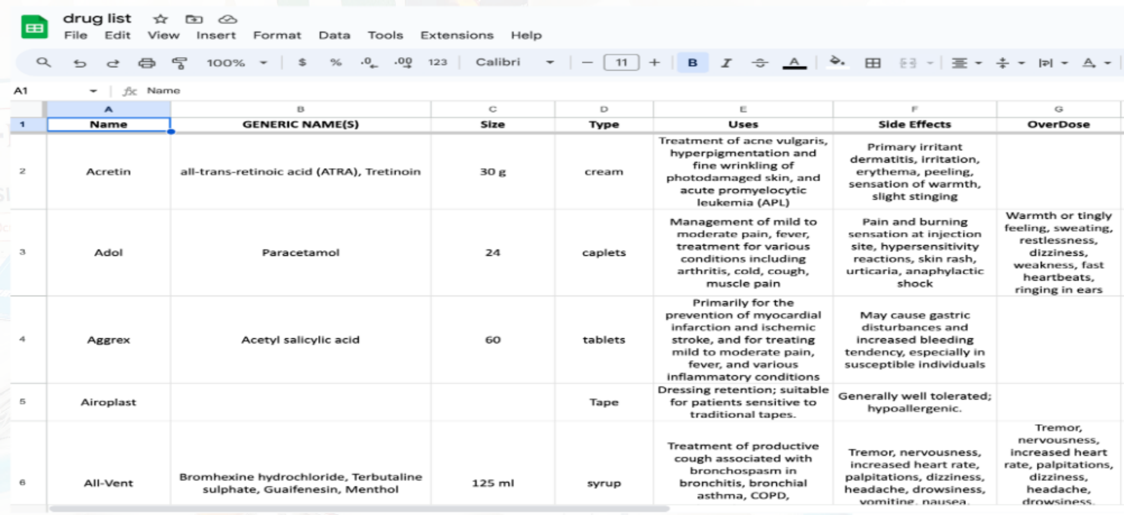
Fig. 2: Dataset 1: Drug images for extracting labels [21]

includes extractable medication labels and manufacturing information. The significance of this dataset lies in its ability to provide further contextual data. The third dataset functions as a repository for all the information related to medicines.It includes important details such as dosage, maximum limit, and recommended times of intake, all of which are arranged in an orderly manner in an Excel or CSV file.

VII. Proposed Solution

Upon obtaining these datasets, the fundamental things we undertake are to carefully prepare the photographs and captions. To maintain consistency, this entails standardising the dimensions of the images and tokenizing the descriptions in order to set the foundation for developing an extensive vocabulary. Images are carefully annotated to improve the model's comprehension. Bounding boxes are drawn around medications, and labels with names or relevant usage information are added. Using expert labelling programmes like LabelImg or RectLabel simplifies this annotation procedure and guarantees accuracy and productivity. [16]

Our plan includes training the model using selected versions of YOLO by using these different datasets in order to identify the one that best fits the particular characteristics of each dataset. After the model is trained, we employ a non-traditional evaluation technique that includes a battery of metrics (BLEU, METEOR, CIDEr, and ROUGE) to carefully evaluate the output caption quality against a pre-selected test dataset.



Name	GENERIC NAME(S)	Size	Type	Uses	Side Effects	OverDose
Acretin	all-trans-retinoic acid (ATRA), Tretinoin	30 g	cream	Treatment of acne vulgaris, hyperpigmentation and fine wrinkling of photodamaged skin, and acute promyelocytic leukemia (APL)	Primary irritant dermatitis, irritation, erythema, peeling, sensation of warmth, slight stinging	
Adol	Paracetamol	24	caplets	Management of mild to moderate pain, fever, treatment for various conditions including arthritis, cold, cough, muscle pain	Pain and burning sensation at injection site, hypersensitivity reactions, skin rash, urticaria, anaphylactic shock	Warmth or tingly feeling, sweating, restlessness, dizziness, weakness, fast heartbeats, ringing in ears
Aggrex	Acetyl salicylic acid	60	tablets	Primarily for the prevention of myocardial infarction and ischemic stroke, and for treating mild to moderate pain, fever, and various inflammatory conditions	May cause gastric disturbances and increased bleeding tendency, especially in susceptible individuals	
Airoplast			Tape	Dressing retention; suitable for patients sensitive to traditional tapes.	Generally well tolerated; hypoallergenic.	
All-Vent	Bromhexine hydrochloride, Terbutaline sulphate, Guaifenesin, Menthol	125 ml	syrup	Treatment of productive cough associated with bronchospasm in bronchitis, bronchial asthma, COPD,	Tremor, nervousness, increased heart rate, palpitations, dizziness, headache, drowsiness, vomiting, nausea,	Tremor, nervousness, increased heart rate, palpitations, dizziness, headache, drowsiness,

Fig. 3: Dataset 3: Drug list and details

We then set out to tweak and improve the model using these evaluation indicators as our markers. The key is fine-tuning, which coordinates changes in model weights, hyper-parameters, and configurations to coordinate an increase in accuracy metrics and a corresponding decrease in processing time. The process of iteration plays a crucial role in developing a model that can provide precise and perceptive captions while still functioning effectively within the constraints of real-world application requirements.

For text extraction, OCR or NLP techniques extract usage details from labels, then pre-process captions with padding, tokenization, and word embeddings. This synthesized caption aims to summarize multiple details into a human-interpretable format. The project emphasizes usability for users to interpret medicine-related information easily and complies with regulations when handling sensitive medical data. The holistic approach integrates image analysis and text recognition to deliver a comprehensive and user-friendly medicinal information system. [27]

The following step is creating an extensive database of user prescriptions, which will allow for a thorough examination of any drug interactions. This step is most crucial for our project as we generate captions and text summary for each patient individually and store them in a personalized database. Not only does this tailored database list a person's prescription drugs, but it also explores how well each drug works with others. The main objective is to develop a strong system that can provide users with comprehensive drug histories that highlight potential interactions and expected side effects. The purpose of this resource is to provide healthcare workers with personalised, educational text summaries. These summaries include a person's medication history, outline possible interactions between different medications, and provide information about anticipated adverse effects. The main goal is to provide healthcare staff with relevant data from a tailored database, enabling them to make educated decisions and enhancing patient care.

VIII. Expected Results

The project's goal is to develop a transformative system that seamlessly integrates recent advances in image analysis, text recognition, and database management to reshape the landscape of medication comprehension and safety. The system strives for precise medication identification by leveraging subtle visual cues and contextual understanding through meticulous dataset curation and extensive training using various iterations of YOLO. Adopting novel evaluation metrics ensures a comprehensive assessment of caption quality, serving as a guidepost for iterative refinements, increasing accuracy, and streamlining operational efficiency.

The incorporation of OCR and NLP techniques represents a significant step forward, allowing the extraction of intricate medication usage details to be strategically condensed into intelligible human-readable summaries. The overall objective is to create an intuitive and user-centered medical information hub. This sophisticated system is intended to empower both users and healthcare providers. It aspires to revolutionise healthcare decision-making by providing tailored medication histories, highlighting potential drug interactions, and outlining expected side effects.

This project's ideology goes beyond development to promote a safer and more informed approach to medication management. In addition to prioritising user accessibility, the system strictly adheres to data handling regulations, ensuring confidentiality and compliance at all times. Finally, this endeavour aims to bridge the gap between complex medical information and everyday comprehension, leading to an era in which informed healthcare decisions are accessible to all.

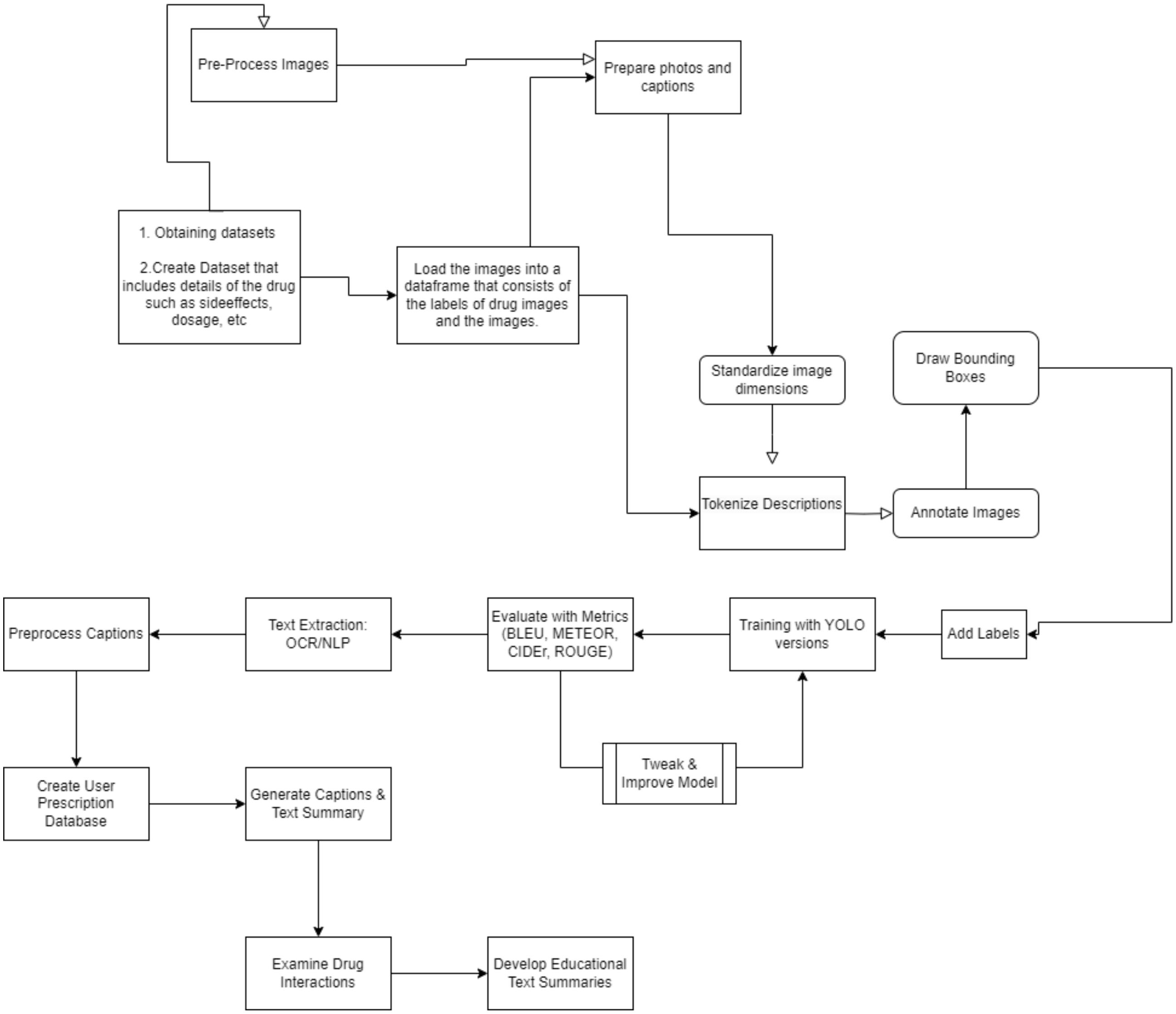

Pre-Process Images
Prepare photos and captions
1. Obtaining datasets
2.Create Dataset that includes details of the drug such as sideeffects, dosage, etc
Load the images into a dataframe that consists of the labels of drug images and the images.
Standardize image dimensions
Draw Bounding Boxes
Tokenize Descriptions
Annotate Images
Preprocess Captions
Text Extraction: OCR/NLP
Evaluate with Metrics (BLEU, METEOR, CIDEr, ROUGE)
Training with YOLO versions
Add Labels
Tweak & Improve Model
Create User Prescription Database
Generate Captions & Text Summary
Examine Drug Interactions
Develop Educational Text Summaries

Fig. 4: Project flow chart

After tracking the patient details using the healthcare history, the health care professionals who will have the access to the tailored database generated through the project using which they can monitor the drug side effects, efficiency over numberous users each with different medical conditions. This curated database serves as an invaluable tool, affording healthcare professionals the ability to navigate through a plethora of patient profiles. Through this expansive lens, they can discern patterns in drug responses, comprehensively evaluating side effects and gauging the efficacy of medications across diverse user spectra. The profound advantage lies in the granularity of insights unearthed: understanding how different medications interact with varying medical conditions and individual physiological compositions.

The database facilitates a panoramic view, where healthcare professionals can gauge the performance of specific drugs within specific cohorts, identifying correlations and trends that illuminate the nuanced landscape of drug efficiency and side effects. This wealth of information doesn't just bolster decision-making; it becomes the cornerstone for personalized care strategies. The ability to discern which medications perform optimally across different conditions or demographics empowers healthcare professionals to tailor treatments, optimizing outcomes and minimizing risks of adverse effects.

Moreover, this database operates as an evolving repository, accumulating invaluable real-world data over time. This dynamic pool of information becomes an invaluable asset, fostering continuous learning within the healthcare domain. And the best part of it is the decrease in the manual repetitive work that needs to be done while collecitng the data by an individual, as this projects helps to eliminate the third party by collecting the information from the patients automatically and tailoring the database without any human effort.

IX. Planned Milestones

A. *Progress to Date*

- OCT 3rd - OCT 24th : Topic Selection and Literature Survey
- OCT 25th - NOV 14th : Dataset Preparation and Proposal Summary
- NOV 15th - DEC 3rd : Proposal Report and Presentation

B. *Future Research Plans*

- JAN 11th, 2024 - JAN 31st, 2024 : Process the datasets and train the model and find best evaluation approach
- FEB 1st, 2024 - FEB 22nd, 2024 : Develop the front end application and train the model
- FEB 23rd, 2024 - MAR 14th, 2024 : Fine-tune the parameters and hyperparameters and process the results using Evaluation Metrics
- MAR 15th, 2024 - April 4th, 2024 : Final Project Report Submission and Presentation

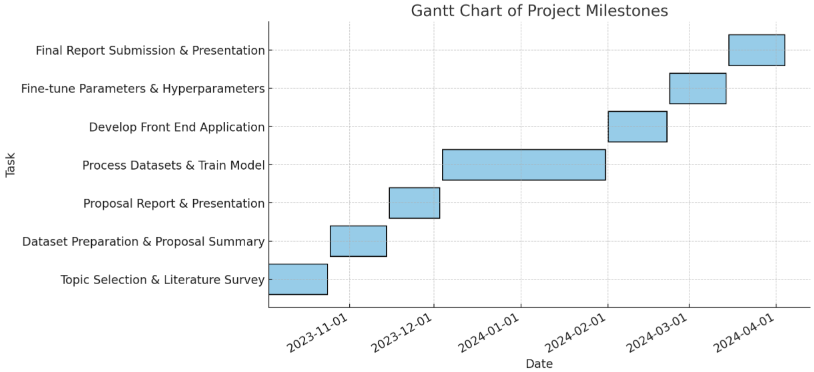


Fig. 5: Gantt chart for Project Milestones

References

[1] Ahsan Ahmad, Aftab Tariq, Hafiz Khawar Hussain, and Ahmad Yousaf Gill. 2023. Revolutionizing Healthcare: How Deep Learning is poised to Change the Landscape of Medical Diagnosis and Treatment. *Journal of Computer Networks, Architecture and High Performance Computing* 5, 2 (2023), 458–471.

[2] Amira Ahmed. 2023. Leveraging Artificial Intelligence for Advancements in the Pharmaceutical Field: A Comprehensive Review. (2023).

[3] Muhammad Abdelhadie Al-Malla, Assef Jafar, and Nada Ghneim. 2022. Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data* 9, 1 (2022), 1–16.

[4] Djamila-Romaissa Beddiar, Mourad Oussalah, and Tapio Seppänen. 2023. Automatic captioning for medical imaging (MIC): a rapid review of literature. *Artificial Intelligence Review* 56, 5 (2023), 4019–4076.

[5] Christine Dewi, Abbott Po Shun Chen, and Henoch Juli Christanto. 2023. Deep Learning for Highly Accurate Hand Recognition Based on Yolov7 Model. *Big Data and Cognitive Computing* 7, 1 (2023), 53.

[6] Nidhal Khdhair El Abbadi and others. 2023. Scene Text detection and Recognition by Using Multi-Level Features Extractions Based on You Only Once Version Five (YOLOv5) and Maximally Stable Extremal Regions (MSERs) with Optical Character Recognition (OCR). *Al-Salam Journal for Engineering and Technology* 2, 1 (2023), 13–27.

[7] Yun Han, Sheng-Luen Chung, Qiang Xiao, Jing-Syuan Wang, and Shun-Feng Su. 2021. Pharmaceutical blister package identification based on induced deep learning. *IEEE Access* 9 (2021), 101344–101356.

[8] Marko Horvat and Gordan Gledec. 2022. A comparative study of YOLOv5 models performance for image localization and classification. In *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 349–356.

[9] Marko Horvat, Ljudevit Jelečević, and Gordan Gledec. 2023. Comparative Analysis of YOLOv5 and YOLOv6 Models Performance for Object Classification on Open Infrastructure: Insights and Recommendations. In *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 317–324.

[10] Rayson Laroca, Luiz A Zanlorensi, Gabriel R Gonçalves, Eduardo Todt, William Robson Schwartz, and David Menotti. 2021. An efficient and layout-independent automatic license plate recognition system based on the YOLO detector. *IET Intelligent Transport Systems* 15, 4 (2021), 483–503.

[11] Songjiang Li, Shilong Wang, and Peng Wang. 2023. A small object detection algorithm for traffic signs based on improved YOLOv7. *Sensors* 23, 16 (2023), 7145.

[12] RG Nandhakumar, K Nirmala Devi, N Krishnamoorthy, S Shanthi, VR Pranesh, and S Nikhalyaa. 2023. Yolo Based License Plate Detection Of Triple Riders And Violators. In *2023 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–6.

[13] Shaik Naseera, GK Rajini, B Venkateswarlu, and M Priyadarisini. 2017. A review on image processing applications in medical field. *Research Journal of Pharmacy and Technology* 10, 10 (2017), 3456–3460.

[14] Pooja R Negi and Sanjay H Buch. Effective Image Captioning With YOLO and LSTM. (????).

[15] Oscar Ondeng, Heywood Ouma, and Peter Akuon. 2023. A Review of Transformer-Based Approaches for Image Captioning. *Applied Sciences* 13, 19 (2023), 11103.

[16] Charmaine C Paglinawan, Marielle Hannah M Caliolio, and Joshua B Frias. 2023. Medicine Classification Using YOLOv4 and Tesseract OCR. In *2023 15th International Conference on Computer and Automation Engineering (ICCAE)*. IEEE, 260–263.

[17] Takorn Prexawanprasut, Lalita Santiworarak, Piyaporn Nurarak, and Poom Juasiripukdee. 2023. Employing Machine Learning and an OCR Validation Technique to Identify Product Category Based on Visible Packaging Features. In *Proceedings of the 2023 6th International Conference on Machine Vision and Applications*. 114–119.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[19] Antonio M Rinaldi, Cristiano Russo, and Cristian Tommasino. 2023. Automatic image captioning combining natural language processing and deep neural networks. *Results in Engineering* 18 (2023), 101107.

[20] M SAROJA and Ani Brown Mary. 2023. Image Captioning Using Improved YOLO V5 Model and Xception V3 Model. (2023).

[21] A. Shah. 2023. Mobile Captured Pharmaceutical Medication Packages. https://www.kaggle.com/datasets/aryashah2k/mobile-captured-pharmaceutical-medication-packages/data. (2023). [Online; accessed 5-December-2023].

[22] Karnati Sai Shashank, N Praneeth Prasad, K Spoorthy Reddy, and L Sridhara Rao. 2023. Upload Cricket Match Video to Generate Audio Commentary by YOLOv8 and Transformer. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*. IEEE, 1152–1157.

[23] Hoo-Chang Shin, Le Lu, and Ronald M Summers. 2017. Natural language processing for large-scale medical image analysis using deep learning. *Deep learning for medical image analysis* (2017), 405–421.

[24] Elda Shurdhaj and Ulehla Christián. 2023. Real Time Vehicle Detection for Intelligent Transportation Systems.

[25] Yaying Su, Baolei Cheng, and Yijun Cai. 2023. Detection and Recognition of Traditional Chinese Medicine Slice Based on YOLOv8. In *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*. IEEE, 214–217.

[26] Lu Tan, Tianran Huangfu, Liyao Wu, and Wenying Chen. 2021. Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification. *BMC medical informatics and decision making* 21 (2021), 1–11.

[27] J Terven and D Cordova-Esparza. 2023. A comprehensive review of YOLO: From YOLOv1 and beyond. arXiv 2023. *arXiv preprint arXiv:2304.00501* (2023).

[28] Haiwen Wu, Ri-Gui Zhou, and Yaochong Li. 2021. A Neural Network Model for Text Detection in Chinese Drug Package Insert. *IEEE Access* 9 (2021), 39781–39791. DOI:http://dx.doi.org/10.1109/ACCESS.2021.3064564