Video-Based Human Activity Recognition of Cough Action Using Deep Learning

Mike Chene

A project report submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2025

School of Engineering, Computer & Mathematical Sciences

Abstract

This study investigates the use of deep learning for recognizing human coughing actions in video footage, with an emphasis on achieving reliable, real-time performance. By integrating human action recognition methods with advanced object detection models, the research aimed to build a system capable of accurately identifying coughing behavior under various conditions. This report also explored different training strategies and model configurations to optimize performance. Among the models tested, the proposed approach achieved the highest accuracy, reaching an F1 score of 0.926. These results suggest that the system is not only effective but also well-suited for potential applications in health monitoring and public safety, where timely and accurate detection of flu-like symptoms is essential.

Keywords: Cough action recognition, Human skeleton, Key point detection, YOLO, Transformer, Deep learning

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Mike CHENE Date: 17/06/2025

Acknowledgment

First and foremost, I am profoundly grateful to my parents for their steadfast support and encouragement over the past two years. Their generous financial assistance and constant belief in my academic journey have been the cornerstone of my ability to undertake and complete my Master's studies at Auckland University of Technology (AUT), New Zealand. Their sacrifices and dedication have been a continuous source of strength and motivation.

Secondly, I have to acknowledge the English Language Academy (ELA) for their exceptional guidance and support in helping me improve my English language skills. The training I received at ELA greatly contributed to my academic preparedness and confidence in pursuing this degree in an English-speaking environment.

I would also like to express my deepest gratitude to my supervisor Dr. Wei Qi Yan. Throughout this study, he provided not only expert academic guidance and invaluable knowledge, but also unwavering support that greatly enriched my learning experience. I firmly believe that this work would not have been possible without his mentorship and dedication.

Lastly, I am sincerely thankful to AUT for offering me the opportunity to pursue a Master's program that is not available in my home country, French Polynesia. Being part of AUT has allowed me to grow both academically and personally, and I am truly honored to have completed my studies at such a distinguished institution.

Mike Chene

Auckland, New Zealand

June 2025

Chapter 1
Introduction

This chapter lays the foundation for this study and is structured into five key sections. It begins by outlining the background and motivation behind the research. This is followed by a clear statement of the research question guiding the investigation. The subsequent sections present the main objectives of the study, highlight its contributions, and conclude with an overview of the report's structure.

1.1 Background and Motivation

Managing sick leave in companies is a recurring challenge, especially due to its unpredictable nature. By definition, sick leave allows employees to take time off work for medical reasons. However, unexpected absences can disrupt workflow, particularly in small businesses where immediate replacements are not available. Additionally, the presence of an ill employee in shared workspaces can increase the risk of further infections, exacerbating productivity losses. For example, 75.6% of adults visited a general practitioner between November 2023 and the same month in 2024 (*Annual Update of Key Results 2023/24*, 2024). Given these challenges, it becomes crucial for managers and business owners to implement proactive strategies to ensure both employee well-being and operational efficiency.

Infectious diseases such as influenza, COVID-19, and the common cold are generally believed to spread through airborne droplets released when individuals cough or sneeze. As a result, the early detection of these symptoms could play an important role in limiting the spread of illness, particularly in high-risk settings like hospitals, workplaces, and public transportation. In relation to this, artificial intelligence (AI) is increasingly being explored in the healthcare domain for its potential to analyze large volumes of medical data. This includes possible applications in drug development, clinical trial optimization, diagnostic support, patient monitoring, and personalized treatment, all of which may contribute to improved efficiency, greater accuracy, and better overall patient outcomes (Koski & Murphy, 2021; Saraswat et al., 2022; Shaheen, 2021; Talati, 2023).

Automated symptom detection is an emerging field in computer vision and artificial intelligence that focuses on identifying physical signs of illness, such as coughing and sneezing, through video-based analysis. Traditionally, symptom monitoring has relied on self-reports or medical examinations, which can be time-consuming, subjective, and impractical for large-scale public health surveillance. In many cases, individuals may experience symptoms like coughing or sneezing but overlook their potential illness, continuing daily activities while unknowingly spreading infections. However, recent advancements in deep learning and video analysis offer the potential for real-time, automated symptom detection, reducing reliance on manual monitoring and improving early disease identification.

Human Action Recognition (HAR) is a fundamental task in computer vision that involves detecting, identifying, and classifying human actions from video or sensor data. Despite challenges related to accuracy and scalability, ongoing research continues to develop more robust and efficient approaches. More specifically, HAR can be described as the task of assigning a label to a sequence of images or video frames that corresponds to a particular human action or activity (Zhang et al., 2019).

The most intuitive way to detect a cough is through sound. While traditional methods rely on handcrafted features like Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Codes (LPCs), these approaches often need expert tuning and don't always work well across different devices or environments. Deep learning offers a more flexible alternative. Models like Convolutional Neural Networks (CNNs) and Residual Neural Networks (RNNs) can learn useful patterns directly from raw audio or spectrograms, removing the need for manual feature design. Turning cough sounds into spectrograms lets these models treat the problem like image recognition, capturing subtle details in the sound. RNNs such as Long Short-Term Memories (LSTMs) also help by tracking the timing and flow of a cough (Amoh & Odame, 2016; Hamdi et al., 2022). Cough actions can also be detected by analyzing motion-based body movements captured through accelerometer signals, extracting time-domain features, and classifying them using a neural network model (Diab & Rodriguez-Villegas, 2024).

Developing a system capable of observing human activity and notice behaviors related to cold and flu symptoms can be a solution to predict potential sickness. For example, the cold and flu system can detect people showing signs of illness in a company and alarm the person in charge that the risk of the individual becoming sick or infecting others in rising. With that information, prevention measures can be taken to ensure the health or workers, and avoid any sudden diminution of productivity due to sudden sick leave. Therefore, using tools such as HAR in Closed-Circuit Television (CCTV) footage can solve the concerns of unforeseen sick people among the workforce of the enterprise.

In many workplaces, employers are required to notify employees at least 14 days before directing them to take leave. However, since cold and flu symptoms typically manifest within

1 to 3 days, using symptom detection to preemptively send employees home would be ineffective. However, real-time detection can still serve as a valuable preventive measure. By identifying symptomatic individuals early, employers or managers can implement sanitary protocols, such as encouraging mask usage, increasing ventilation, providing hand sanitizers, or temporarily adjusting seating arrangements to minimize the risk of further transmission. Therefore, the motivation of this study is to provide a tool that can be used in companies to assist managers in decision making concerning the sanitary protocol that workers have to follow when a potential sick person is detected.

1.2 Research Questions

The aim of this research is to develop a video-based detection system that can identify cold and flu symptoms, specifically coughing, using deep learning models. This study will focus on training and evaluating YOLOv12 on a relevant dataset to assess their effectiveness in recognizing the coughing action in real-time video footage. The performance of both models will be compared using standard evaluation metrics, such as accuracy. Therefore, the main research questions of this report are:

- How performant is YOLOv12 in detecting and analyzing a coughing person to achieve accurate real-time action analysis and recognition?

In order to refine this research question, we can split it again:

- How accurately can deep learning models detect coughing actions in video footage?
- What are the main challenges in video-based symptom recognition?
- How do different architectures compare in terms of accuracy, speed, and computational cost?
- How feasible is incorporating video based human activity recognition in existing surveillance systems?

Beyond model performance, this research will also examine the challenges associated with implementing human activity recognition in surveillance systems, including factors like computational efficiency, environmental variations, and privacy concerns. Additionally, an evaluation of the cost and feasibility of deploying such a system in real-world settings will be

conducted. The findings from this study aim to provide insights into the limitations of current approaches and suggest potential improvements for future research in vision-based action recognition.

1.3 Contributions

The focus of this research project is to achieve accurate recognition and analysis of coughing actions in video footage using object detection methods combined with computer vision techniques. We propose a system that leverages an advanced object detection framework alongside temporal modeling to identify and interpret human actions associated with coughing. A real-time, efficient, and highly accurate pipeline was developed to detect and analyze coughing behavior in various environments. This pipeline is built upon existing state-of-the-art models, adapted and integrated to suit the specific requirements of coughing action recognition. To improve the model's accuracy and robustness, we also built a custom video dataset annotated with coughing people and normal people for training. By the end of this project, we were able to:

- Fine-tune and adapt state-of-the-art deep learning models to accurately distinguish individuals performing coughing actions in both video and image data.
- Generate reliable image and video-level inferences using the refined models to support real-time human action recognition.
- Apply targeted optimization techniques to enhance model performance and improve classification accuracy across diverse scenarios.
- Create a dedicated dataset of coughing actions to serve as a robust training resource for deep learning-based action recognition systems.

Furthermore, all models will be evaluated and compared based on their performance metrics, and if possible computational resource requirements during training. This comparative analysis will highlight the model best suited for practical deployment, while also identifying the one with greater potential for future research and advanced applications.

1.4 Objectives of This Report

In this report, we outline a research methodology that integrates human action recognition with YOLOv12, which aims to identify and people coughing in video footage. Additionally, we offer a comprehensive review of the contemporary literature surrounding HAR, Human Pose Estimation (HPE), CNNs modelling techniques.

Subsequently, we introduce a novel HAR system which discerns a specific human action by detecting the action as an object from image data and translating this data into linguistic information suitable for processing by the models.

Therefore, the specific objectives of this report are twofold: Firstly, to create an accurate dataset of coughing action in order to train deep learning models. Secondly, to utilize this dataset to perform HAR in videos and images, consequently generating a model capable of detecting a specific human action in real-life scenarios. The development environments utilized for this endeavor are Python through the use of Google Colab (*Colab.Google*, n.d.).

To assess the effectiveness of the proposed methodology, a comparative analysis is conducted between CNNs architectures and the Transformer model, with the objective of highlighting the distinct advantages and limitations of each approach.

1.5 Structure of This Report

The description of this report is as follows:

- Chapter 2 provides a comprehensive overview of HAR beginning with its foundational concepts and key datasets, and progressing to the role of deep learning in enhancing system performance. It concludes with a focused discussion on the YOLO family of models, emphasizing their real-time detection capabilities and potential in advancing HAR applications.
- Chapter 3 introduces the architecture of YOLOv12, offering insight into their design principles and their differences to better understand how it can perform at multiple levels. This chapter also outlines the research methodology and experimental setup, concluding with a comparison of the expected outcomes.
- Chapter 4 details the results from the experiments proposed in the methodology. It includes the process of data collection and the subsequent analysis, with a focus on visual evaluation of the results and metrics.
- Chapter 5 provides a comprehensive analysis of the experimental findings and summarizes the key outcomes of the study. In addition, it critically evaluates the advantages and limitations of the method used.
- Chapter 6 explores possible future research directions and strategies for improving the current approach. This final chapter aims to identify areas for further development and suggest ways to enhance the system's performance in future applications.

Chapter 2 Literature Review

The focus of this report is on pose capturing based on dynamic motion for deep learning, this chapter will introduce a plenty of traditional methods and the relevant knowledge of deep learning.

2.1 Introduction

This literature review begins by exploring HAR providing an overview of its definition, key advantages, current limitations, and widely used datasets in the field. Following this, we examine how AI has been applied to HAR tasks, highlighting their strengths in spatial feature extraction and their role in advancing recognition accuracy. We then shift our focus to the YOLO framework, discussing its evolution, real-time performance capabilities, and relevance to HAR applications. Finally, we conclude by reflecting on the key findings and identifying gaps or directions for future research.

2.2 HAR

As the need for systems that can accurately and efficiently interpret human behavior continues to grow, HAR has evolved from early handcrafted approaches to deep learning methods that can learn patterns directly from raw data. This subsection explores how HAR has progressed, the challenges it still faces, and its relevance to tasks like detecting cough actions in visual data.

HAR can play a pivotal role in modern computer vision, with a broad spectrum of applications spanning surveillance, healthcare, human-computer interaction, robotics, and sports analytics. An earlier review provides a thorough and methodologically robust synthesis of research in its domain, adhering to established guidelines and covering a broad temporal and thematic scope (Aggarwal & Ryoo, 2011). It can serve as a foundational reference for researchers interested in the intersection of sports sciences, human activity recognition, and related technological applications. Furthermore, HAR has progressed overtime, going from early handcrafted feature-based techniques to sophisticated deep learning frameworks that harness convolutional and recurrent neural networks to effectively capture both spatial structures and temporal dynamics (Dwivedi et al., 2024).

In team sports, HAR can be used to automatically identify and analyze player actions and

interactions during games, enabling applications such as performance analysis, game summarization, highlight generation, referee decision assistance, and injury prevention. It involves recognizing complex, fast-paced actions often involving multiple players and objects, requiring advanced video understanding, temporal tracking, and multi-object detection capabilities (Yin et al., 2024).

Another approach to HAR is with HPE, human poses can be extracted as 2D body landmarks using the OpenPose detector from CCTV-like videos, and these poses are processed to generate low- and high-level spatio-temporal features that capture body posture and movement dynamics. HPE could handle occlusions and missing data, enhancing robustness and accuracy in action recognition (Angelini et al., 2020). By capturing the spatial and temporal relationships of keypoints, HPE enables the identification and classification of various human actions, enhancing the understanding of complex movements in sports and physical exercise context (Badiola-Bengoa & Mendez-Zorrilla, 2021; Cao & Yan, 2024). Moreover, HPE can serve as a foundation for action recognition and enabling accurate analysis of human behavior across various applications, including healthcare and human-computer interaction (Q. Wu et al., 2020).

Despite its advancements, HAR still faces numerous limitations. The recognition pipeline generally encompasses several interconnected stages, including data acquisition, preprocessing, feature extraction, temporal modeling, classification, and evaluation. Nonetheless, each of these stages introduces potential challenges. For instance, the complexity of HAR is significantly heightened by variations in camera viewpoints, subject appearances, lighting conditions, and occlusions (Shafizadegan et al., 2024). Another issue can appear with backgrounds and camera motion. Many HAR algorithms perform well in controlled indoor environments but struggle in outdoor or uncontrolled settings due to background noise and camera movements (Kong & Fu, 2022). Therefore, background clutter and dynamic scenes introduce noise into feature extraction, degrading recognition performance, and recent works have tried to address these by using skeleton models, 3D point clouds, temporal pyramids, and

dynamic time warping (Jegham et al., 2020; Wang et al., 2020).

In conclusion, HAR refers to the automated identification and analysis of human behaviors from visual or sensor-based data, with the goal of interpreting and classifying actions across diverse contexts. This definition holds particular relevance in the present study, as it underpins the development of a coughing action recognition system designed to detect individuals exhibiting cough-like behaviors in both video footage and still images. However, previous studies have highlighted the limitations of HAR which have to be taken into account when researching and experimenting in that domain.

2.3 HAR Datasets

In the field of HAR, there exist various datasets. The HMDB51 dataset is a well-known and richly detailed resource for human action recognition, offering a total of 6,766 manually annotated video clips spread across 51 different action categories. Each action has at least 101 video examples, drawn from a wide range of real-world sources such as movies, public video archives, and YouTube. The dataset is designed to reflect the natural complexity of human motion in everyday settings, making it highly valuable for training and evaluating recognition models. The actions are grouped into five main types: basic facial expressions (like smiling or laughing), facial actions involving objects (such as drinking or eating), general body movements (like jumping or running), body-object interactions (like playing golf or riding a bike), and human-to-human interactions (such as hugging or shaking hands). Each clip includes detailed metadata on visible body parts, camera motion, viewpoint, video quality, and number of actors, allowing for flexible and fine-grained analysis. All videos are standardized to 240 pixels in height and 30 frames per second, with stabilization applied to most clips to reduce the impact of camera shake. With its diversity, realism, and rich annotations, HMDB51 stands as a strong benchmark for advancing human action recognition research (Kuehne et al., 2011).

The UCF101 Dataset is one of the largest and most widely used collections for human action recognition, containing over 13,000 video clips across 101 action categories. These clips

are taken from real YouTube videos, capturing natural variations like camera movement and busy backgrounds, which adds realism and complexity. The actions are grouped into five main categories: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, and Sports. Each video runs at 25 frames per second with a resolution of 320×240 pixels, and most clips last just over seven seconds. This makes UCF101 a rich and challenging dataset for developing and testing action recognition models in real-world conditions (Soomro et al., 2012).

The KTH dataset includes 2,391 video clips of 25 actors performing six different actions: boxing, handclapping, handwaving, jogging, running, and walking. To add diversity and challenge model performance, each action is recorded in four different background settings. This setup helps test how well recognition models can identify basic but dynamic human movements in varying visual environments. As a result, the KTH dataset has become a widely used benchmark for evaluating action recognition systems (David & Abbas, n.d.; Thi et al., 2014).

The Sneeze-Cough Dataset (BIISC) was developed to support public health research by focusing on detecting flu-like symptoms, especially sneezing and coughing. It features 960 color video clips recorded from 20 participants aged between 20 and 50, with an even mix of men and women. Alongside the targeted flu-related actions, the dataset also includes six everyday background activities like drinking, using a phone, and stretching, which add useful variety and context. All recordings were made indoors under semi-controlled lighting, with each video captured at a resolution of 480×290 pixels and a frame rate of 5 frames per second. Each clip lasts about 15 seconds, making the dataset well-suited for training models in health-related action recognition (Gupta et al., 2023; Thi et al., 2014).

The NTU RGB+D Dataset is a large and diverse resource created for human action recognition. It contains more than 56,000 video samples and millions of frames collected from 40 participants between the ages of 10 and 35. The dataset spans 60 different types of actions,

including everyday activities, person-to-person interactions, and health-related behaviors. All data was captured using the Microsoft Kinect v2 across 80 different camera viewpoints, providing a rich variety of angles. Each video sample includes RGB footage, depth maps, infrared images, and detailed 3D skeletal data showing the movement of 25 key body joints (Shahroudy et al., 2016).

2.4 Artificial Intelligence and HAR

Nowadays, machine learning can be used in many modern technologies such as voice assistants, recommendation systems, and spam filters. Among the classical machine learning models, there is the Support Vector Machine (SVM) which utilizes classification and regression therefore usable for HAR. For example, some methods can outperform several existing techniques at the time, demonstrating the effectiveness of combining local space-time descriptors with SVMs for action recognition tasks (Schuldt et al., 2004).

Deep learning can be used in video-based human action recognition by automatically learning hierarchical features from raw video frames, capturing both spatial and temporal information through architectures like Convolutional Neural Networks (CNNs), 3D CNNs, and Recurrent Neural Networks (RNNs) including LSTMs. These models extract complex motion patterns and high-level representations from video data, enabling robust recognition of various human actions without manual feature engineering (D. Wu et al., 2017). Moreover, deep learning can be used for human action recognition by employing neural network architectures, such as sequential models with convolutional layers, to automatically extract features from raw time-domain sensor data representing human movements (Sikder et al., 2021).

A notable benefit when it comes to CNNs in HAR lies in the training, and optimization of the models. When optimized and paired with efficient detection and tracking methods, CNNs become powerful tools for real-time human action recognition. This is particularly important in areas like surveillance, human-computer interaction, and robotics, where quick and accurate understanding of human behavior can make a real difference in performance and usability

(Archana & Hareesh, 2021).

CNNs can offer the advantage of automatically learning hierarchical features directly from raw input data. As a result, they eliminate the reliance on handcrafted features, which are often tailored to specific problems and may struggle to generalize across diverse scenarios. This inherent capability enables CNNs to adaptively extract discriminative spatial and temporal patterns relevant to human actions, thereby enhancing recognition accuracy in a wide range of contexts. Furthermore, by extending traditional 2D CNNs to 3D CNNs, models can capture both spatial and temporal information simultaneously. This is crucial for HAR since actions are inherently dynamic and involve motion patterns over time. 3D convolutions enable the model to learn motion information encoded in multiple adjacent frames, enhancing the understanding of temporal dynamics (Ji et al., 2013; Liang & Yan, 2024).

In summary, CNNs provide a powerful framework for HAR by enabling automatic, robust, and scalable learning of spatiotemporal features from raw video data, supporting multimodal integration, and facilitating real-time applications with high accuracy and adaptability to complex environments.

2.5 YOLO

YOLO, short for You Only Look Once, is a fast and efficient object detection model that looks at the entire image in one go to identify and locate objects. Over the years, it has gone through many versions, each improving on speed, accuracy, and design. Newer models like YOLOv8, YOLO-NAS, and YOLOv12 bring in smarter features like attention mechanisms and automatic architecture search, helping them perform even better across different tasks. Because of this balance between speed and accuracy, YOLO is now widely used in areas like self-driving cars, video surveillance, robotics, and even healthcare (Kumar Thakur & Chauhan, 2025; Sapkota et al., 2025; Terven et al., 2023; Tian et al., 2025). Another interesting usage of YOLO is in emotion detection within online classrooms where YOLO-based models can detect faces and recognize emotional states from facial expressions, facilitating real-time analysis of student engagement (Parambil et al., 2025).

In action recognition, YOLO is widely used thanks to its ability to quickly and accurately detect objects, making it well-suited for real-time tracking of people in dynamic scenes like sports or surveillance. For example, in soccer player tracking, YOLO helps identify players and objects from UAV footage, allowing for detailed analysis of movement and team behavior with strong accuracy and flexibility (Rezende et al., 2025). Another example is in security applications where YOLOv8 can be employed to detect suspicious human activities in restricted areas, offering rapid processing and high accuracy for theft detection with real-time alert (Reddy et al., 2025; Sai Mrudhun et al., 2024). Moreover, YOLO can extract spatial features from individual frames, which are then fed into an LSTM network to capture temporal dependencies and model the sequential dynamics of human movements, enabling robust action recognition even under challenging conditions like occlusions and varying illumination (Elnady & Abdelmunim, 2025). Instead of breaking the process into separate steps like older methods, YOLO uses a single neural network to handle everything at once, which makes it much faster compared to other detectors like SSD and EfficientDet, making it suitable for real-time action recognition tasks (Yilmaz & Navruz, 2025).

YOLO can also be used in HAR by incorporating HPE, which involves detecting key points in images or videos to understand body movement. By accurately identifying the positions and dynamics of multiple individuals' keypoints, YOLO is able to capture both spatial and temporal patterns that are essential for distinguishing different actions. This method leverages YOLO's well-known efficiency and precision, which in turn supports effective multi-person action recognition even in complex and rapidly changing environments (Maji et al., 2022). In addition, the model's fast detection speed and high accuracy make it well-suited for real-time applications, including deployment on platforms such as unmanned aerial vehicles for monitoring human activities in large or remote areas (Ding et al., 2024). Moreover, YOLO's accurate object detection can also work alongside Dynamic Spatial-Temporal Modeling for Skeleton-based Action Recognition (DG-STGCN) by identifying relevant objects, such as tools in a workspace. This combination can improve the clarity and precision of action recognition,

particularly in complex settings like assembly lines (Hsiao et al., 2024). In fall detection, YOLOv8 detects human presence and, combined with pose estimation, tracks body postures to distinguish falls from normal activities, enabling timely alerts (G et al., 2025). Consequently, various actions such as standing, sitting, or engaging in sports can be classified effectively, even in scenarios involving multiple subjects or partial occlusions.

In summary, YOLO has emerged as a highly efficient and adaptable framework for action recognition, offering a well-balanced combination of speed, accuracy, and flexibility. Its capability to simultaneously detect objects and human movements makes it particularly effective in complex environments such as surveillance systems, sports analytics, and healthcare monitoring. Moreover, the integration of human pose estimation and temporal modeling can further enhance YOLO's ability to capture fine-grained human actions. Nonetheless, this study will focus exclusively on employing YOLO for object detection, as this approach is expected to deliver faster performance with lower computational overhead, making it more appropriate for deployment in real-world scenarios.

2.6 Summary

To conclude, this literature review has explored the development and significance of HAR highlighting its relevance in various real-world applications such as surveillance, healthcare, and human-computer interaction. We first outlined the fundamental principles of HAR, followed by an overview of the key datasets that have enabled progress in this field. Next, we examined the role of artificial intelligence, particularly deep learning, in improving the accuracy and robustness of action recognition systems. Finally, we focused on the YOLO family of models, emphasizing their strengths in real-time object detection and their growing relevance in HAR tasks. By integrating fast and efficient detection capabilities with temporal understanding of human motion, models like YOLO represent a promising direction for future research in action recognition.

Chapter 3 Methodology

This chapter describes the methodology used to develop a deep learning-based video recognition system for detecting coughing and sneezing. The approach includes dataset selection, preprocessing steps, model architecture, training procedures, and evaluation metrics.

3.1 Dataset

Based on previous work involving the BIISC dataset, a custom dataset has been created using Roboflow. The BIISC dataset videos were filmed at a 10 frames per second rate. However, it is noticeable that the subjects in the videos are moving slowly to attenuate movement blur. Horizontal flip videos were included in the original folder, but removed to lighten workload during manual annotation using the Roboflow tool.

This new dataset focuses on cough actions which are present in the videos with COUGH in their name. After selecting all the videos related to the cough action, a video segmentation was performed to collect images for the training, testing, and validation of the trained models. To avoid the duplication of images, an image sampling was performed using a sampling ratio of 1 image per second which resulted in a raw dataset containing 1823 images.

The annotation consisted on rectangle boxing of the subjects present on the images by declaring whether they were class 1 = coughing, or class 2 = normal.

Dataset split	Percentage	Number of images
Training	66%	1203
Validation	26%	480
Testing	8%	140

Table I: Repartition of the dataset split training, validation, and testing before data augmentation.

Preprocessing:

- **Auto-Orient**: this preprocessing step is important for detecting human actions because it ensures that all input images or video frames are properly aligned, regardless of how they were captured. Videos recorded in different orientations or on various devices may appear rotated or flipped if the orientation metadata is not correctly handled. This misalignment can distort body posture and motion cues, leading the model to misinterpret actions or fail to detect them altogether. Therefore, by auto-orienting

frames before processing, the model receives consistent and correctly positioned visual information, improving the accuracy and reliability of human action recognition. Fortunately, the dataset used to create our new dataset follow strict rules of recording to ensure that all the videos were filmed the same way and that the subjects performed the actions in a set manner.

- **Resize**: This preprocessing technique was mainly used to help control computational cost and memory usage by reducing large, high-resolution inputs to a manageable scale while preserving essential features. This consistency enables more efficient training, better convergence, and more accurate recognition of human actions.

Data augmentation:

- **Brightness**: Bringing variations to the brightness of each video frame increases the diversity of lighting conditions the model is exposed to during training, which is important for recognizing actions like coughing. In real-world settings, lighting can vary significantly depending on the environment. For example, indoor company rooms versus outdoor public areas. Consequently, if the training dataset contains mostly well-lit footage, artificially altering brightness helps the model learn to recognize the same action in darker or overexposed scenes, therefore improving its generalization to new environments.
- **Cutout**: Cutout increases the robustness of the model by randomly masking out square regions of each video frame, forcing it to rely on a wider range of visual cues. This is particularly relevant when recognizing a person coughing, as parts of the body may be occluded in real-life footage. For instance, if a subject’s hand or face is blocked by another person or object, the model should still be able to detect the action based on remaining features like shoulder movement or body posture. Applying cutout during training helps the model learn to handle these kinds of partial occlusions.
- **Horizontal Flip**: Flipping each video frame horizontally helps diversify the training data and is especially useful for recognizing actions like coughing. In real-world scenarios, such as sports, a player's movements can differ noticeably based on their

dominant hand. If most examples in the dataset show right-handed individuals coughing, applying horizontal flips allows the model to also learn how the action might appear from left-handed individuals, improving its ability to generalize.

- **Noise**: Adding noise to training data can enhance the model's ability to generalize by preventing overfitting and encouraging it to learn more robust, high-level features rather than memorizing specific patterns. This technique simulates real-world variability, forcing the model to become resilient to imperfections it may encounter during inference. As a result, it improves the model's accuracy on unseen data by teaching it to focus on the essential structure of the input rather than noise-sensitive details.

In summary, the final dataset used for the coughing action recognition model was derived from a selected portion of the BIISC dataset videos, which were first converted into image frames and then annotated using Roboflow. The annotated data was subsequently divided into training, validation, and testing sets. Before training, the dataset underwent key preprocessing steps including auto-orientation and resizing to ensure uniformity and compatibility with the model. Additionally, to further enhance the model's robustness, accuracy, and generalization capabilities, multiple data augmentation methods such as brightness variation, cutout, horizontal flipping, and noise addition were applied to the training set. The new dataset used in this research project contains 4229 images in total split as shown in Table II.

Dataset split	Percentage	Number of images
Training	85%	3609
Validation	11%	480
Testing	3%	140

Table II: Repartition of the dataset across training, validation, and testing after data augmentation.

3.2 YOLOv12 Architecture

Understanding the architecture of YOLOv12 is important because it brings together years of development in a design that balances high accuracy with real-time performance. Compared to earlier versions, YOLOv12 introduces a more advanced backbone that combines dynamic convolutional layers with attention mechanisms, helping the model better capture spatial and contextual information across different scales. YOLOv12 also utilizes refined training techniques like task-aligned label assignment, which together make the model more adaptable and precise in complex scenarios. These changes set YOLOv12 apart from its predecessors, making it especially valuable for tasks like human action recognition where both speed and accuracy are essential.

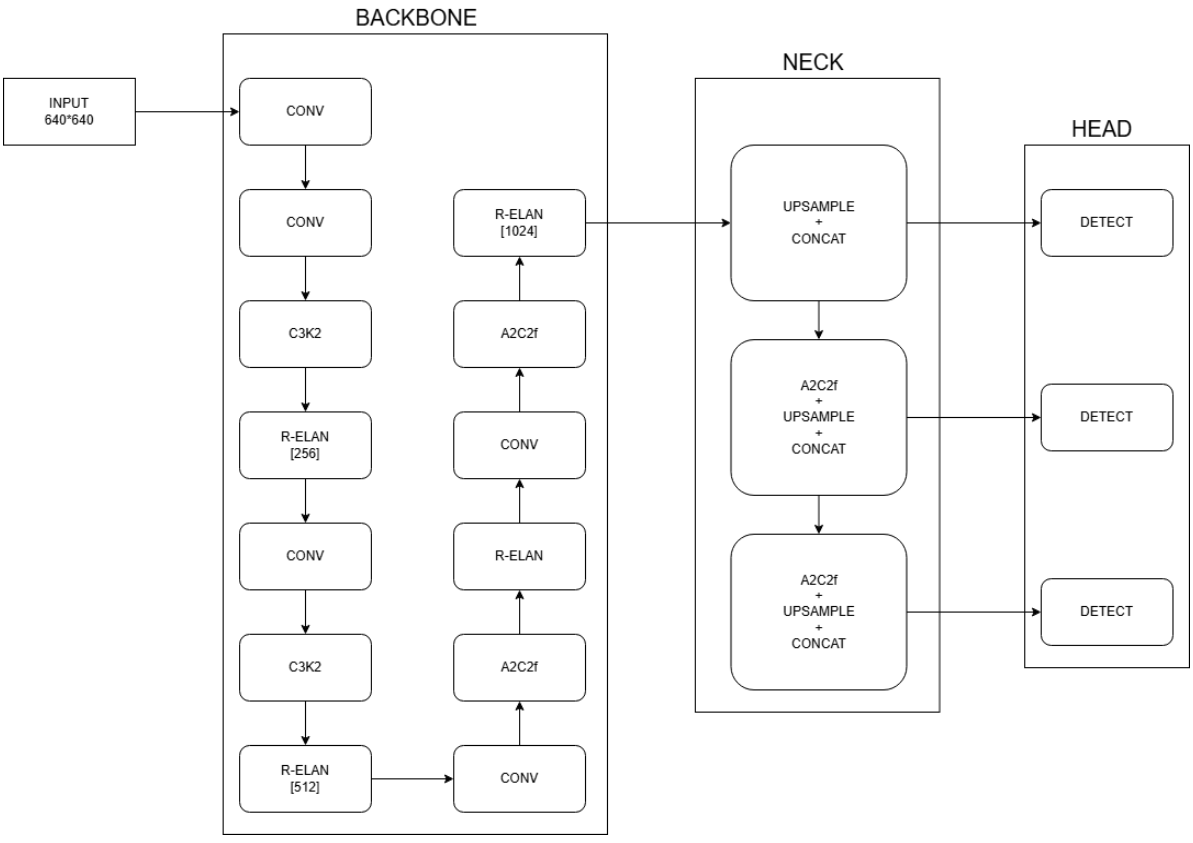

Figure I: Simplified YOLOv12 architecture

As shown in Figure I, YOLOv12 has three main parts: the backbone, the neck, and the head. Compared to its predecessors, the architecture of YOLOv12 introduces the Residual

Efficient Layer Aggregation Networks (R-ELAN) and A2C2f blocks. To summarize its functioning, the input which can be 640*640 images is sent to the backbone, which processes the data through multiple convolutional blocks, C3K2 blocks, and a A2C2f block. The backbone then sends the output to the neck that performs upsampling and concatenation before forwarding it to the head which handles the muti-layered detection. Therefore, YOLOv12 introduces an attention-driven architecture that strikes a balance between the speed of traditional CNN-based models and the enhanced representational power of attention mechanisms which marks a significant step forward, as earlier attention-based models often fell short in terms of speed, limiting their practicality for real-time tasks like detecting coughing actions in workplace video surveillance.

C3K2

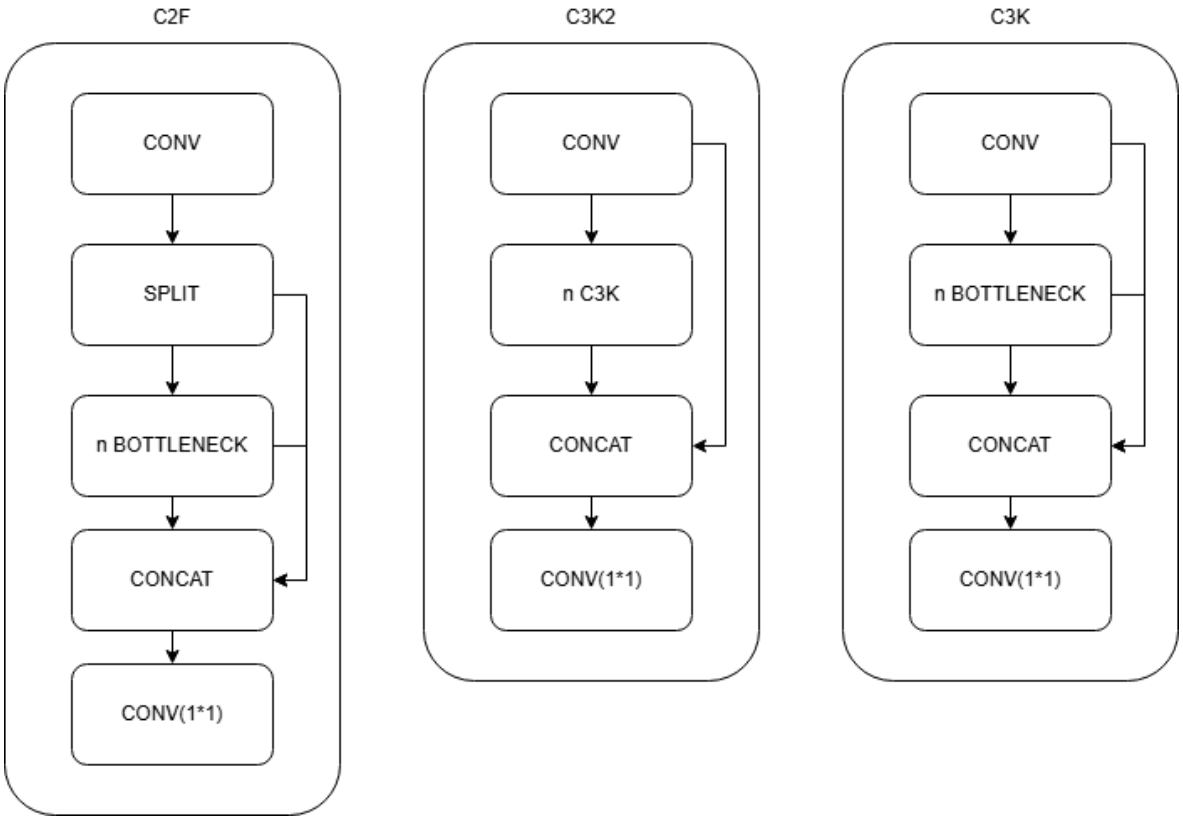

Figure II: Simplified representation of C2F, C3K2, and C3K blocks for comparison

Similarly to YOLOv11, YOLOv12 employs C3K2 blocks in its backbone, an improved version of the earlier versions bottleneck block. This block can boost feature extraction while keeping the model fast and lightweight by using several small convolutions on different parts

of the feature map instead of larger ones. These features are then merged, helping the model stay accurate with fewer parameters than older designs like YOLOv8's C2f blocks. The C3K2 block expands on the simpler C3K structure by adding extra convolution layers around it and combining their outputs to improve feature integration even further. As a result, this design can allow YOLOv12 to maintain a strong balance between speed and performance, making it a practical choice for real-time object detection tasks.

To summarize, YOLOv12 brings a big improvement to real-time object detection by combining speed, efficiency, and accuracy through several smart design choices. It introduces Area Attention and R-ELAN blocks, while also using Flash Attention to reduce memory overhead and make attention nearly as fast as CNNs. Instead of stacking three heavy blocks in the backbone like older YOLO versions, it uses just one R-ELAN blocks, making optimization easier and speeding up inference. Additionally, the new architecture of YOLOv12 replaces linear layers with convolutional ones plus batch normalization to make the most of convolution's speed and efficiency. This is especially useful for action recognition because YOLOv12 can quickly and accurately detect subtle movements, like coughing, in real time without slowing down or missing important details.

A2C2f

Much like Transformers, YOLOv12 integrates attention blocks into its backbone to enhance feature extraction. Specifically, it extends the C2f architecture by incorporating area-attention and A2 Block layers, which allow the model to operate in both attention and standard convolution modes. By doing so, it benefits from the strengths of both approaches. Attention mechanisms have transformed deep learning by enabling models to dynamically focus on the most informative parts of the input, thereby improving performance and interpretability. In the context of object recognition, this ability to prioritize salient regions within an image leads to more accurate detection and segmentation outcomes (Shen et al., 2024).

This area-based attention strategy is specifically designed to reduce the computational burden associated with traditional self-attention, which typically suffers from quadratic complexity. Instead of applying attention globally across the entire feature map, Area Attention divides it into equal-sized, non-overlapping segments either horizontally or vertically, allowing

the model to limit the scope of attention and improve efficiency. For example, a feature map with dimensions (H, W) can be partitioned into L segments of size (H/L, W) or (H, W/L). This segmentation relies on a simple reshape operation, making it a more efficient solution compared to complex partitioning methods such as Shifted Window, Criss-Cross Attention, or Axial Attention (Khanam & Hussain, 2025).

R-ELAN

R-ELAN, introduced in YOLOv12, builds on the original ELAN by tackling issues like gradient blocking and unstable optimization, especially when attention mechanisms are involved. While ELAN splits output and processes them separately before merging, this can disrupt gradient flow and lacks a direct connection from input to output. R-ELAN solves this by introducing a residual shortcut that links the input to the output with a small scaling factor, which stabilizes training in a way similar to layer scaling used in vision transformers, but without the added computational burden. Instead of splitting the input, R-ELAN first adjusts the channel dimensions using a transition layer, then processes the entire feature map through subsequent blocks and merges the results into a more efficient bottleneck. Area Attention, also referred to as the A2 Module in YOLOv12, is a novel attention mechanism designed to improve the efficiency of self-attention in computer vision tasks. It works by partitioning spatial regions of the feature map into equal-sized, non-overlapping segments either horizontally or vertically. For a feature map of dimensions (H, W), it is divided into L segments of size (H/L, W) or (H, W/L) This new structure improves feature aggregation and model stability while keeping the architecture fast and lightweight (Alif & Hussain, 2025; Tian et al., 2025). Consequently, R-ELAN can improve YOLOv12's ability to detect coughing actions by enhancing feature extraction stability and efficiency, which can enable the model to capture subtle motion cues more accurately in real time.

Conclusion

In conclusion, YOLOv12 marks a major advancement in real-time object detection by combining the speed and efficiency of CNN-based designs with the improved feature

representation of attention mechanisms. Its refined architecture, centered around R-ELAN and A2C2f blocks, enables faster and more accurate detection, making it especially suitable for practical tasks such as identifying coughing actions in workplace surveillance footage.

3.3 Evaluation Metrics

To evaluate and compare the performance of the models used in this study, we will rely on key metrics such as precision in eq (2), recall in eq (3), and the F1 score in eq (4), as they offer complementary perspectives on detection quality. Precision reflects the proportion of correctly identified positive cases out of all instances the model predicted as positive, while recall measures the model's ability to correctly identify all actual positive cases. Since both metrics are important, especially when dealing with imbalanced data, the F1 score combines them into a single value by calculating their harmonic mean. In addition, analyzing the F1 curve across different confidence thresholds allows us to observe how each model performs under various decision boundaries. By using this combination of metrics, we aim to ensure a fair and well-rounded comparison that highlights both the accuracy and reliability of each model in practical scenarios.

$$P = \frac{TP}{TP+FP} \tag{2}$$

$$R = \frac{TP}{TP+FN} \tag{3}$$

P = Precision

R = Recall

TP = True Positives

FP = False Positives

FN = False Negatives

$$F1 = 2\frac{P*R}{P+R} \quad (4)$$

F1 = F1 score

3.4 Fine-Tuning

This subsection explores the fine-tuning process of YOLOv12 to determine the most suitable batch size and the right optimizer for detecting coughing actions in both images and video footage. Since these hyperparameters play a key role in shaping how the model learns from the data, it is important to find a balance that promotes both stability and generalization. Similarly, too few epochs may result in underfitting, where the model fails to capture key features of the coughing action, while too many can lead to overfitting and reduced performance on unseen footage. Through systematic experimentation and observation, this fine-tuning process helps adapt YOLOv12 to the specific characteristics of cough recognition in real-world visual contexts.

Batch size

Batch size	Precision	Recall	MAP@50	MAP@50-95
4	0.946	0.839	0.913	0.78
8	0.911	0.915	0.962	0.831
16	0.917	0.833	0.912	0.789
32	0.949	0.852	0.921	0.792

Table III: Performance comparison of YOLOv12 from validation with batch size variations performed on 30 epochs

Choosing the right batch size is crucial, as a value that is too small can result in noisy gradients and unstable learning, while a batch size that is too large may smooth out meaningful variations in the data. As shown in Table III, the model achieves its best performance with a batch size of eight, yielding higher scores in precision, recall, mean average precision at 0.5, and mean average precision across the 0.5 to 0.95 range. At the same time, it is worth noting

that a batch size of 32 also delivers strong results and trains significantly faster due to its larger size. For this reason, both batch sizes will be used in subsequent training and evaluation to balance performance and efficiency.

Optimizer

Optimizer	Precision	Recall	MAP@50	MAP@50-95
ADAM	0.883	0.858	0.91	0.761
SGD	0.909	0.904	0.939	0.809
AdamW	0.905	0.906	0.937	0.771
NAdam	0.925	0.894	0.945	0.789
RAdam	0.856	0.874	0.903	0.762

Table IV: Performance comparison of YOLOv12 from validation with different optimizers performed on 30 epochs with a batch size of 32

In YOLO, an optimizer is a training component that updates the model's weights based on the loss function to improve prediction accuracy over time. Based on Table IV, Nadam showed a better precision and mean average precision at 0.5 than the other optimizers, but worse results for recall, mean average precision from 0.5 to 0.95. Stochastic Gradient Descent (SGB) achieve higher mean average precision from 0.5 to 0.95, while maintaining competitive values in precision, recall, and mean average precision at 0.5. SGB offers computational efficiency by using small subsets of data (mini-batches) for updates, often leading to faster convergence and better generalization (Ketkar, 2017). Consequently, the optimizer used for training the model will be SGD.

Chapter 4 Results

The main content of this chapter is to collect video data and demonstrate the experimental results. In the end, in this chapter, we also discuss the limitations of this project.

4.1 F1 Score to Confidence Results from Training

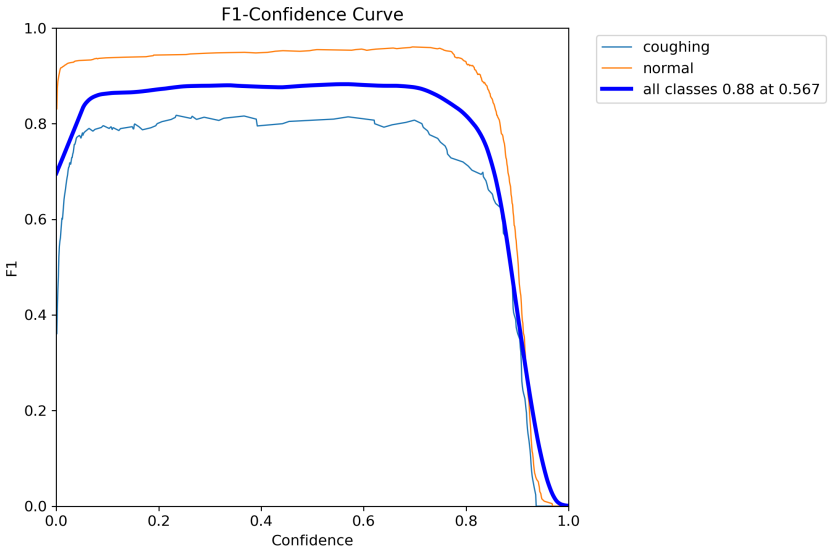

Figure III: F1 to confidence graph for batch size 8 performed on 100 epochs

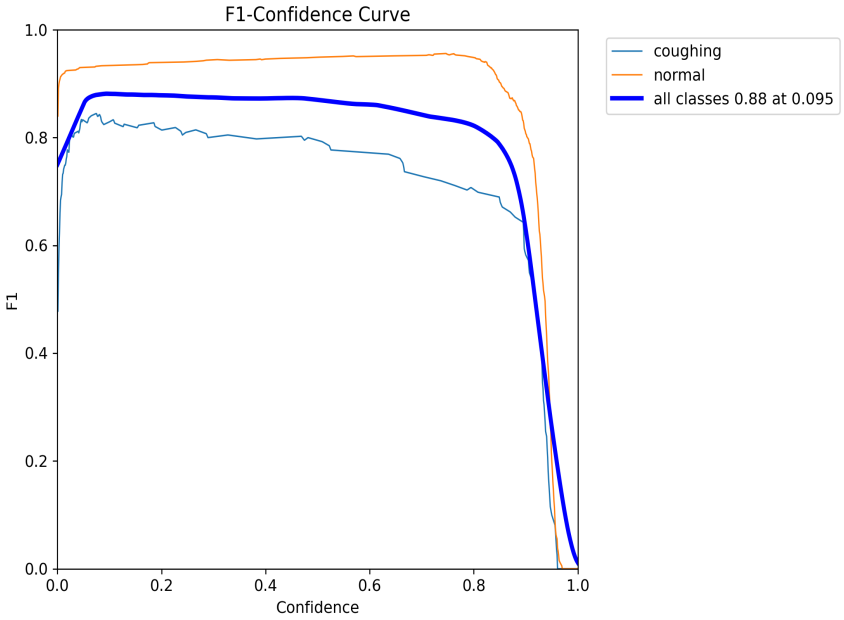

Figure IV: F1 to confidence graph for batch size 32 performed on 100 epochs

As shown in Figure III, the F1 score remains relatively stable across confidence thresholds from 0.1 to 0.8, whereas in Figure IV, it declines steadily over the same range. Notably, both figures reveal a consistent downward trend in the F1 score for coughing recognition as confidence increases, while the F1 score for normal recognition follows a similar pattern in

both graphs. These findings suggest that although both models perform similarly in recognizing normal actions, the model trained with a batch size of 8 demonstrates greater stability in detecting coughing actions during training.

4.2 Overall Results from Training on 100 Epochs

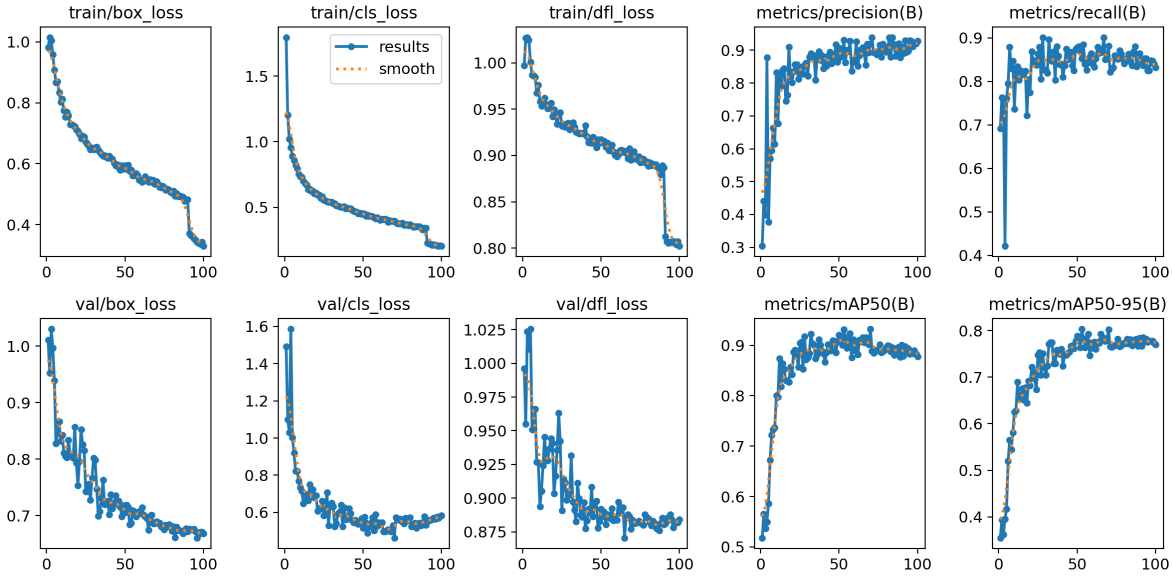

Figure V: Results from training YOLOv12 with batch size 8 on 100 epochs

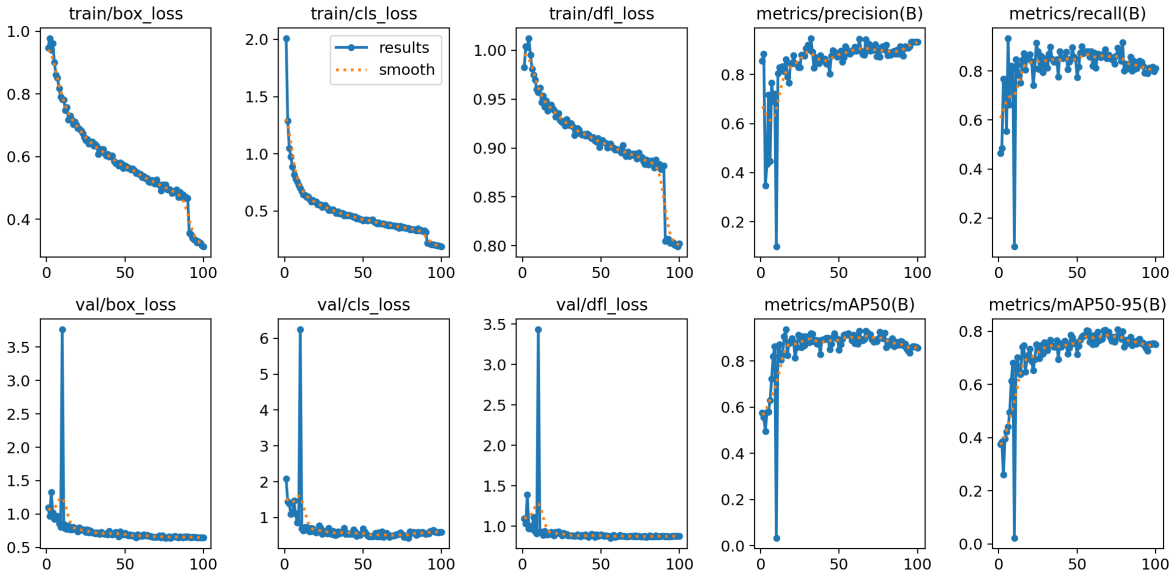

Figure VI: Results from training YOLOv12 with batch size 32 on 100 epochs

Figures V and VI illustrate the model's training behavior over 100 epochs. Although the boxing, classification, and distribution focal losses (DFL) follow similar trends during training for both configurations, a noticeable difference appears during validation: all three losses start significantly lower with the smaller batch size. This suggests that training with a batch size of 8 may lead to better generalization for cough action detection. In contrast, the lower validation performance with a batch size of 32 could indicate a higher risk of overfitting.

Another thing we can notice is that both models are improving their precision the more epochs they are trained on. However, the recall, mean average precision at 0.5 and mean average precision from 0.5 to 0.95 all start to decrease after 50 epochs. Similarly to the phenomenon observed in validation losses, this pattern suggests a case of overfitting after 50 epochs.

4.3 Results from Validation on Image Set

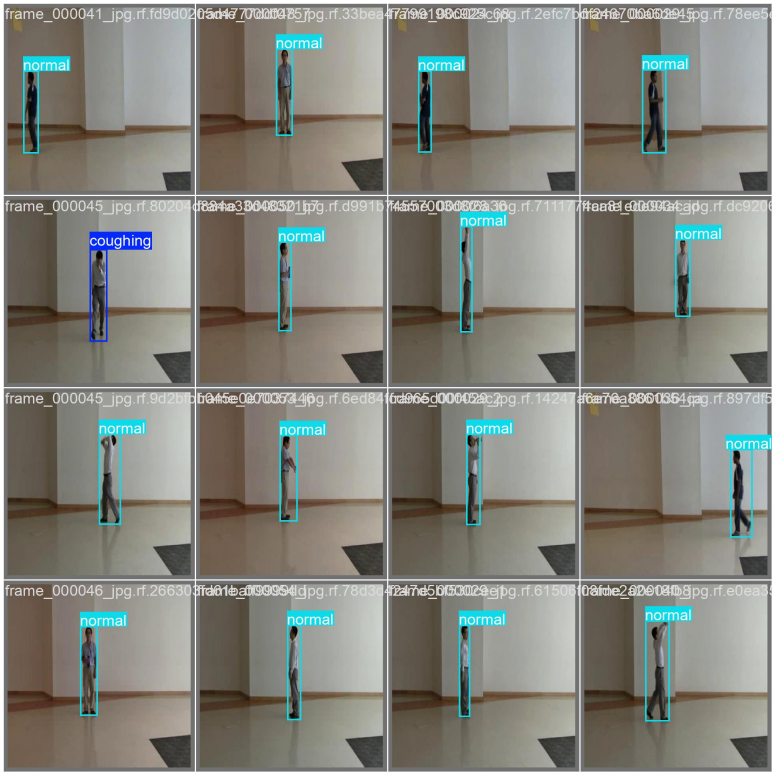

Figure VII: Validation batch with correct object labels

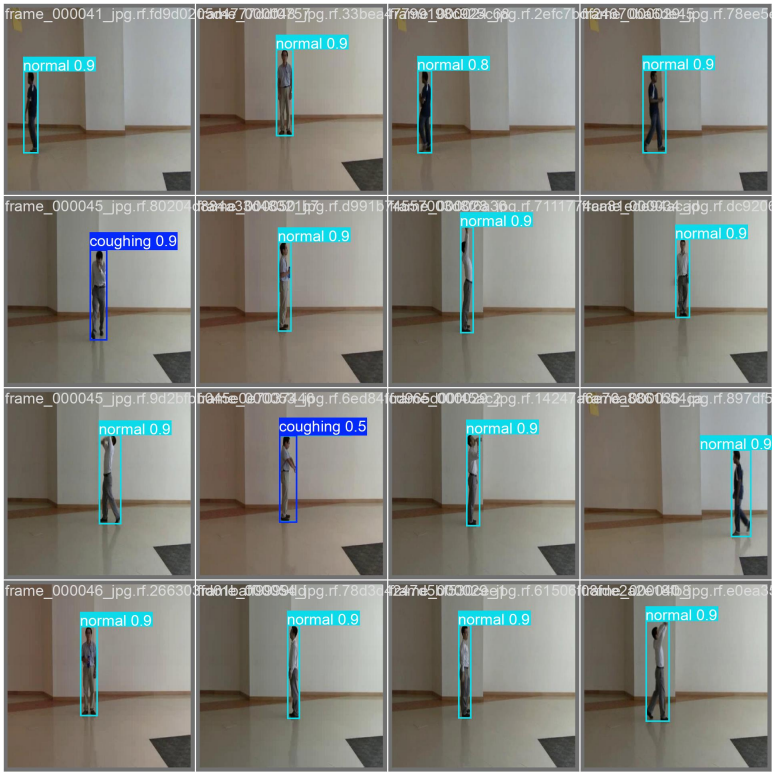

Figure VIII: Prediction of model trained with batch size of 8 on validation batch

Figures VII, VIII, IX, and X demonstrate that the model is capable of identifying coughing actions in validation images, while also correctly recognizing instances where individuals are behaving normally without any visible symptoms. This indicates that the model has learned to distinguish between symptomatic and non-symptomatic actions to a certain degree. However, both versions of the model still struggle with false positives, often misclassifying normal behavior as coughing. This suggests that the model may be overly sensitive to certain movements or poses, potentially due to similarities in body posture or motion between normal and coughing actions. Additionally, these results also highlight a need for more refined features or better-balanced training data to improve the model's discrimination ability.

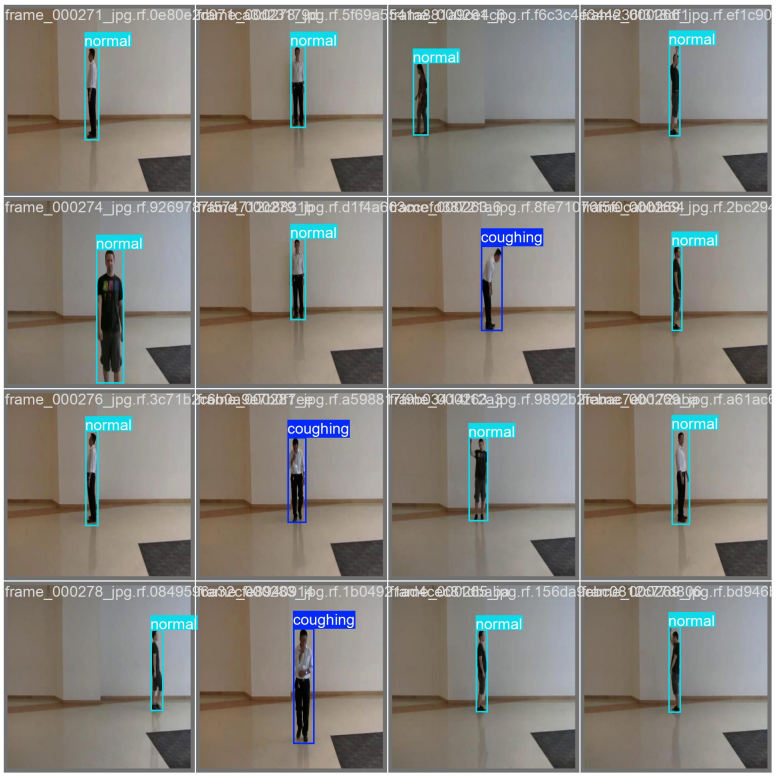

Figure IX: Validation batch with correct object labels

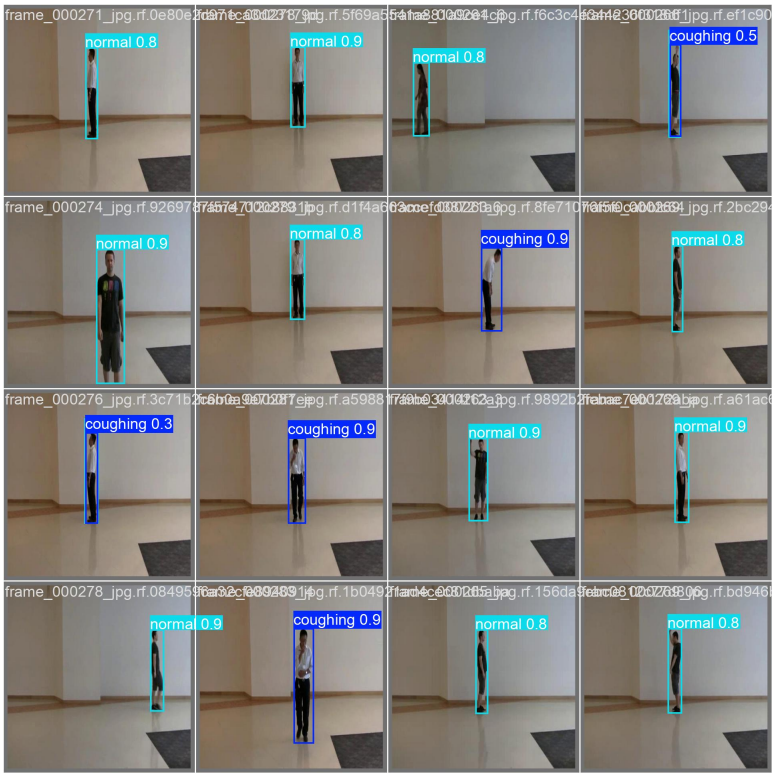

Figure X: Prediction of model trained with batch size of 32 on validation batch

Another reason the model might mistake normal actions for coughing is the poor quality of some images in the dataset. If the images are blurry, dark, or low in resolution, the model may struggle to see the small details that show someone is actually coughing, like a hand near the mouth or a change in facial expression. When those details are missing or unclear, the model might rely too much on general body posture, which can look similar in many actions. This can lead to mistakes, especially if parts of the body are blocked or if the image is noisy. Therefore, improving the image quality and using basic enhancements like sharpening or brightening could help the model make more accurate predictions.

4.4 Video Inference

To test the model's performance in a real-world scenario, we applied it to video inference. A short video was recorded in which the subject first behaved normally and then performed a coughing action. This allowed us to evaluate how well the model could detect the transition between normal behavior and visible symptoms in a continuous video stream, rather than isolated images.

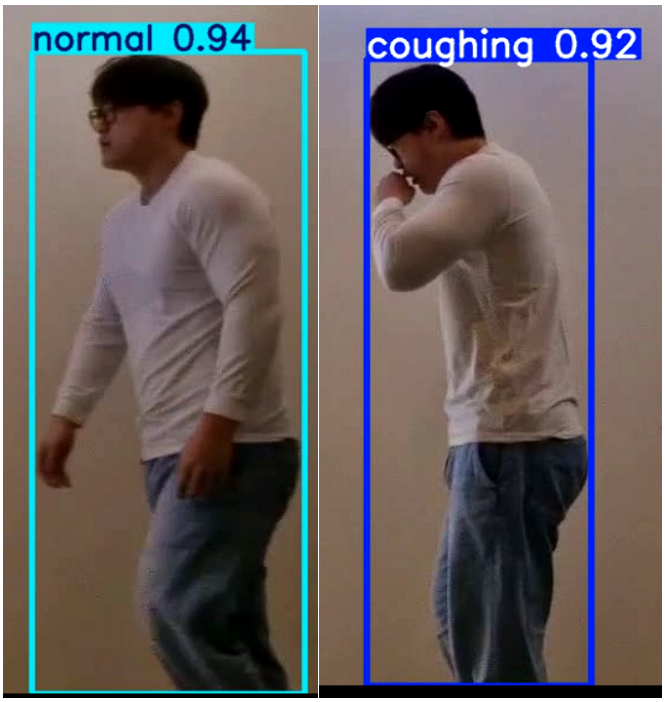

Figure XI: Results obtained from the application of the trained model on a test video.

Figure XI shows that the model is able to detect both normal behavior and coughing actions in video frames that were not part of the training set. This demonstrates the model's ability to generalize beyond the data it was trained on and accurately recognize coughing as a distinct human action in real-world footage. By successfully identifying and distinguishing coughing from normal behavior in previously unseen video, the model proves its potential for real-time action recognition, where each frame is treated as an object detection task. Consequently, these results highlight the effectiveness of the model in capturing subtle actions and applying them consistently across continuous video input.

4.5 Comparison with Other Models

In this subsection, we compare the performance of YOLOv12 with YOLOv8, YOLOv11, and RF-DETR, a real-time transformer-base architecture, by evaluating key detection metrics, including precision, recall, mean average precision at 0.5, and mean average precision from 0.5 to 0.95. These metrics can provide a clear view of YOLOv12's trained model to detect coughing actions accurately and consistently, allowing us to assess its strengths and limitations in real-time cough action recognition in video footage and images.

Model	Precision	Recall	MAP@50	MAP@50-95	F1
YOLOv12	0.943	0.910	0.941	0.814	0.926
YOLOv11	0.925	0.903	0.924	0.798	0.914
YOLOv8	0.890	0.875	0.910	0.766	0.882
RF-DETR	0.893	0.850	0.940	0.804	0.871

Table V: Performance comparison of 4 different models on the dataset based on validation metrics

The results in Table V show that YOLOv12 performs the best among the four models tested, with an F1 score of 0.926. It reaches the highest scores in all metrics, including a precision of 0.943 and a recall of 0.910. These values mean that the model is both accurate when it makes predictions and consistent in finding most of the coughing actions. It also leads in mean average precisions, which shows it can correctly localize actions across different levels of overlap.

Overall, these results suggest that YOLOv12 is highly reliable for recognizing coughing behavior in our dataset.

Just behind YOLOv12 is YOLOv11, which also gives strong results with a F1 score of 0.914. With a precision of 0.925 and a recall of 0.903, it is only slightly lower than YOLOv12. Its mean average precision scores are also close, especially the mean average precision at 0.5 (0.924). This shows that YOLOv11 is still quite effective at detecting and localizing actions, though not quite as consistent as YOLOv12. The small gap between the two models likely reflects improvements in YOLOv12's architecture that make it more refined and stable.

YOLOv8, on the other hand, shows the weakest performance overall with a F1 score of 0.882. It has the lowest precision (0.890) and recall (0.875), along with the lowest mean average precision scores. This means it makes more mistakes when predicting and also misses more true coughing actions. While YOLOv8 is known for being lightweight and fast, these results suggest that it may not be the best choice for tasks that need higher accuracy, like action recognition related to health.

Finally, RF-DETR gives mixed results, with its mean average precision scores that can be convincing, and the lowest F1 score overall (0.871). Its recall is the lowest at 0.850, which means it tends to miss more actual coughing cases. This suggests that RF-DETR can be good at making precise predictions when it does detect something, but it doesn't always catch everything. It might be useful in combination with other models, but on its own, it may not be ideal if the goal is to catch every possible sign of coughing.

Chapter 5
Analysis and Discussions

This chapter presents an analysis and comparison of the experimental results, examining how the outcomes vary under different conditions. It also discusses the potential reasons of the results obtained from the experiment.

5.1 Analysis

In summary, we combined human action recognition and deep learning to detect the coughing actions in videos. Firstly, an object detection dataset was created based of a bigger dataset containing videos of various human actions performed by a group of subjects. Then the model was fine-tuned on two different aspects: batch size and optimizer, to figure out a reliable, yet fast way to train our model. Secondly, two versions of the model were created and compared based on the metrics obtained from the trainings and the validations. The more promising model was then used for video inference on new data. The model was capable of recognizing the subject coughing in the video with good accuracy, which highlights its potential for future usage. Thirdly, after comparison with other models, YOLOv12 proved to be the most effective model for recognizing coughing actions in videos because it combined high accuracy with reliable detection across different situations. It consistently identified coughing events while avoiding many false positives, which shows that its predictions were both precise and thorough. Compared to the other models, YOLOv12 was better at locating the action within the frame and adapting to variations in how coughing appears. This strong performance highlights its potential for real-world applications, where recognizing subtle human actions, like coughing, needs to be both dependable and accurate.

5.2 Discussions and Limitations

In this study, we trained and fine-tuned the YOLOv12 model to detect coughing actions in both videos and images, relying on a dataset composed of annotated frames that reflect real-world human behavior. Since coughing is often a brief and subtle action that occurs in complex environments with multiple people, varying lighting, and background clutter, the training process was carefully designed to improve YOLOv12's ability to handle such challenges. As a result, the model showed notable improvements in accuracy, precision, and mean average precision when compared to earlier YOLO versions. Furthermore, the video inference results clearly demonstrate the model's ability to perform well in realistic scenarios, reinforcing its practical value. Thanks to its efficient architecture, fast detection speed, and robustness under

difficult conditions, YOLOv12 presents a highly competitive approach for human action recognition, particularly in applications that require real-time analysis and cost-effective deployment.

We initially considered using skeletal joints and human pose estimation to evaluate human posture in this system, but it was ultimately deemed unnecessary for our application. Instead of relying on pose comparison or joint completion assessments, we focused on a more direct recognition approach using object detection, which can be faster and less computationally heavy because it predicts bounding boxes around whole objects rather than estimating multiple precise keypoints for each body part, which requires more detailed processing and higher resolution features. By combining YOLO models with visual input, we developed an efficient real-time system that analyzes a person's behavior without the need for explicit keypoint extraction. This method offers a streamlined and effective solution while also showcasing how YOLOv12 can handle visual tasks by learning patterns directly from image data, without depending on skeletal modeling.

Although the proposed HAR) system demonstrates encouraging performance, it continues to face several limitations that hinder its effectiveness in real-world scenarios. The system remains vulnerable to environmental variations, as conditions such as inadequate lighting, background clutter, and occlusions can substantially degrade its accuracy. Furthermore, it encounters challenges in generalizing across different users and contexts, which restricts its ability to deliver consistent results beyond controlled environments. Actions that are visually similar or happen quickly are especially hard to distinguish, which further affects reliability. Real-time deployment also presents difficulties, since running these models efficiently often requires substantial computational resources. Moreover, building large and accurately labeled datasets remains a time-consuming and expensive task, adding to the overall development burden. Privacy concerns must also be considered, particularly in settings like healthcare or public spaces, and the lack of transparency in deep learning models makes it harder to explain or justify their decisions. Taken together, these limitations highlight the gap between current capabilities and the demands of real-world use.

Chapter 6 Conclusion and Future Work

This chapter outlines the purpose and approach of the project, while also recommending future research paths in light of the experimental findings, observed limitations, and possible improvements his chapter, we will summarize the subject and method of this project and propose new research direction according to the result and insufficiency of the experiment as well as the future work.

6.1 Conclusion

This study begins by reviewing existing research in human action recognition, starting with foundational work and gradually moving toward more recent advances in deep learning models such as CNNs, while also addressing the integration of human pose estimation. Following this, the research methodology introduces and contrasts the architecture of YOLOv12 with its predecessors, providing insights into its design principles and how these differences may influence performance. The experimental setup is described to ensure clarity in how the models were developed and tested. The results are then presented, including the process of data collection, the evaluation metrics used, and a visual analysis of the outputs. Finally, the study concludes by analyzing the findings in depth, drawing attention to the effectiveness of the proposed approach, while also discussing its limitations and suggesting areas for future improvement.

In conclusion, the aim of this study was to explore the potential of deep learning for recognizing human coughing actions in videos, with a focus on real-time performance and reliability. By combining human action recognition techniques with state-of-the-art object detection models, the goal was to develop a system capable of accurately identifying coughing behavior in various conditions. Moreover, this research also aimed to evaluate different training strategies and model configurations to find the most effective setup. The results have shown that the model performed better than some other models used for comparison, with an overall F1 score of 0.926. As a final point, the study highlights how such a system could contribute to health monitoring and public safety by providing a fast and dependable way to detect coughing actions in real-world environments.

6.2 Future Work

In the future, this work could be extended by training the model to recognize a broader range of human actions, such as sneezing, yawning, or speaking. Expanding the dataset with more diverse video samples, featuring different people, environments, and lighting

conditions, would help improve the model's generalization and adaptability to real-world situations. This would not only make the system more robust but also more applicable in various health and public safety contexts.

Moreover, integrating temporal modeling techniques like LSTM or Temporal Convolutional Networks could enhance the system's ability to understand motion over time, making it better at distinguishing between actions that appear visually similar in individual frames. In addition, future efforts could focus on optimizing the model for real-time performance on low-resource devices, which would support practical deployment in surveillance or healthcare settings. Combining video analysis with other data sources, such as audio or sensor input, might also improve recognition in cases where visual information is limited, while raising important questions around privacy and responsible use that should be carefully addressed.

Finally, this whole study could be done again but using newer state-of-the-art methods, especially Transformers which are a big area of focus nowadays. For example, Transformer can replace recurrent and convolutional layers with multi-head self-attention mechanisms, enabling superior parallelization, faster training, and state-of-the-art performance in tasks like machine translation (Vaswani et al., 2017). Unlike CNNs, which have limited receptive fields, Transformers can model relationships between all patches in a video sequence globally. This is crucial for understanding actions that involve interactions across distant spatial or temporal region (Shaikh et al., 2024). Additionally, Transformers like Swin can generate hierarchical feature maps, enabling multi-scale modeling of actions—useful for recognizing actions at varying speeds or scales (Liu, Lin, et al., 2021; Liu, Ning, et al., 2021).

References

Aggarwal, J. K., & Ryoo, M. S. (2011). Human Activity Analysis: A review. *ACM Computing Surveys*, *43*(3), 1–43. https://doi.org/10.1145/1922649.1922653

Alif, M. A. R., & Hussain, M. (2025). *YOLOv12: A Breakdown of the Key Architectural Features* (No. arXiv:2502.14740). arXiv. https://doi.org/10.48550/arXiv.2502.14740

Amoh, J., & Odame, K. (2016). Deep Neural Networks for Identifying Cough Sounds. *IEEE Transactions on Biomedical Circuits and Systems*, *10*(5), 1003–1011. https://doi.org/10.1109/TBCAS.2016.2598794

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. ACM Transactions on Multimedia Computing, Communications and Applications.

Angelini, F., Fu, Z., Long, Y., Shao, L., & Naqvi, S. M. (2020). 2D Pose-Based Real-Time Human Action Recognition With Occlusion-Handling. *IEEE Transactions on Multimedia*, *22*(6), 1433–1446. https://doi.org/10.1109/TMM.2019.2944745

Annual Update of Key Results 2023/24: New Zealand Health Survey | Ministry of Health NZ. (2024, November 18). https://www.health.govt.nz/publications/annual-update-of-key-results-202324-new-zealand-health-survey

Archana, N., & Hareesh, K. (2021). Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks. *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 173–177. https://doi.org/10.1109/ACCESS51619.2021.9563316

Badiola-Bengoa, A., & Mendez-Zorrilla, A. (2021). A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise. *Sensors*, *21*(18), Article 18. https://doi.org/10.3390/s21185996

Bibbò, L., & Vellasco, M. M. B. R. (2023). Human Activity Recognition (HAR) in Healthcare. *Applied Sciences*, *13*(24), Article 24. https://doi.org/10.3390/app132413009

Cao, X., & Yan, W. Q. (2024). Pose Estimation for Swimmers in Video Surveillance. *Multimedia Tools and Applications*, *83*(9), 26565–26580. https://doi.org/10.1007/s11042-023-16618-w

Cao, X. (2022) Pose Estimation of Swimmers from Digital Images Using Deep Learning. Master's Thesis, Auckland University of Technology.

Cao, Y, Yan, W. (2024) Lips reading using deep learning. Optimization, Machine Learning, and Fuzzy Logic: Theory, Algorithms, and Applications (Chapter 17). IGI Global.

Chen, Z., Yan, W. (2024) Real-time pose recognition for billiard player using deep learning. Deep Learning, Reinforcement Learning and the Rise of Intelligent Systems, pp.188-208, Chapter 10, IGI Global.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

David, Z. S., & Abbas, A. H. (n.d.). Human Action Recognition Using Interest Point Detector with k-th Dataset.

Diab, M. S., & Rodriguez-Villegas, E. (2024). A TinyML Motion-Based Embedded Cough Detection System. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. https://doi.org/10.1109/EMBC53108.2024.10782961

Ding, J., Niu, S., Nie, Z., & Zhu, W. (2024). Research on Human Posture Estimation Algorithm Based on YOLO-Pose. *Sensors*, *24*(10), Article 10. https://doi.org/10.3390/s24103036

Dong, K., Yan, W. (2024) Player performance analysis in table tennis through human action recognition. Computers, 13(12), 332.

Dwivedi, A., Shuaib, M., Joshi, A., Diwakar, M., Singh, P., & Mishra, A. K. (2024). Deep Learning Enabled Human Action Recognition. *International Conference on Advancement in Electronics & Communication Engineering (AECE)*, 513–518. https://doi.org/10.1109/AECE62803.2024.10911226

Elnady, M., & Abdelmunim, H. E. (2025). A Novel YOLO LSTM Approach for Enhanced Human Action Recognition in Video Sequences. *Scientific Reports*, *15*(1), 17036. https://doi.org/10.1038/s41598-025-01898-z

Gupta, H., Imran, J., & Sharma, C. (2023). Flu-Net: Two-stream Deep Heterogeneous Network to Detect Flu like Symptoms from Videos Using Grey Wolf Optimization Algorithm. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7733–7745. https://doi.org/10.1007/s12652-023-04585-x

Hamdi, S., Oussalah, M., Moussaoui, A., & Saidi, M. (2022). Attention-based hybrid CNN-LSTM and Spectral Data Augmentation for COVID-19 Diagnosis from Cough Sound. *Journal of Intelligent Information Systems*, *59*(2), 367–389. https://doi.org/10.1007/s10844-022-00707-7

Herrera, A., Beck, A., Bell, D., Miller, P., Wu, Q., Yan, W. (2008) Behavior analysis and prediction in image sequences using rough sets. International Machine Vision and Image Processing Conference (pp.71-76)

Huan, Y., Yan, W. (2025) Semaphore recognition using deep learning. Electronics 14 (2), 286

Hsiao, T.-S., Hou, H.-Y., Lin, P. T., Chang, C.-Y., Chen, Y.-Y., & Yang, C.-L. (2024). Integrating YOLO and DG-STGCN Systems for Enhanced Human Action Recognition. *International Conference on Advanced Robotics and Intelligent Systems (ARIS)*, 1–5. https://doi.org/10.1109/ARIS62416.2024.10679957

Jegham, I., Ben Khalifa, A., Alouani, I., & Mahjoub, M. A. (2020). Vision-Based Human Action Recognition: An Overview and Real World Challenges. *Forensic Science International: Digital Investigation*, *32*, 200901. https://doi.org/10.1016/j.fsidi.2019.200901

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

https://doi.org/10.1109/TPAMI.2012.59

Ketkar, N. (2017). Stochastic Gradient Descent. In *Deep Learning with Python* (pp. 113–132). Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2766-4_8

Khanam, R., & Hussain, M. (2025). *A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions* (No. arXiv:2504.11995). arXiv. https://doi.org/10.48550/arXiv.2504.11995

Kong, Y., & Fu, Y. (2022). Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, *130*(5), 1366–1401. https://doi.org/10.1007/s11263-022-01594-9

Koski, E., & Murphy, J. (2021). AI in Healthcare. In M. Honey, C. Ronquillo, T.-T. Lee, & L. Westbrooke (Eds.), *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/SHTI210726

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A Large Video Database for Human Motion Recognition. *International Conference on Computer Vision*, 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543

Kumar Thakur, M., & Chauhan, S. (2025). Brain Tumor Detection and Classifcation Using YOLO Algorithms. *International Conference on Inventive Computation Technologies (ICICT)*, 116–123. https://doi.org/10.1109/ICICT64420.2025.11004729

Liang, C., Lu, J., Yan, W. (2022) Human action recognition from digital videos based on deep learning. ACM ICCCV.

Liang, C., Yan, W. (2024) Human action recognition based on YOLOv7. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.126-145, Chapter 6, IGI Global.

Liu, C., Yan, W. (2020) Gait recognition using deep learning. Handbook of Research on Multimedia Cyber Security (pp.214-226)

Liu, Y., Nand, P., Hossain, A., Nguyen, M., Yan, W. (2023) Sign language recognition from digital videos using feature pyramid network with detection Transformer. Multimedia Tools and Applications.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* (No. arXiv:2103.14030). arXiv. https://doi.org/10.48550/arXiv.2103.14030

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2021). *Video Swin Transformer*. arXiv. https://doi.org/10.48550/arXiv.2106.13230

Lu, J., Shen, J., Yan, W., Boris, B. (2017) An empirical study for human behaviors analysis. International Journal of Digital Crime and Forensics 9 (3), 11-17.

Lu, J., Nguyen, M., Yan, W. (2018) Pedestrian detection using deep learning. IEEE AVSS.

Lu, J., Nguyen, M., Yan, W. (2020) Human behavior recognition using deep learning. International Conference on Image and Vision Computing New Zealand.

Lu, J., Nguyen, M., Yan, W. (2020) Comparative evaluations of human behavior recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 176-189.

Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. International Symposium on Geometry and Vision.

Luo, Z., Nguyen, M., Yan, W. (2022) Kayak and sailboat detection based on the improved YOLO with Transformer. ACM ICCCV.

Luo, Z. (2022) Sailboat and Kayak Detection Using Deep Learning Methods. Master's Thesis, Auckland University of Technology, New Zealand.

Luo, Z., Nguyen, M., Yan, W. (2021) Sailboat detection based on automated search attention mechanism and deep learning models. International Conference on Image and Vision Computing New Zealand.

Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2636–2645. https://doi.org/10.1109/CVPRW56347.2022.00297

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multimedia Tools and Applications, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

Parambil, M. M. A., Bouktif, S., Gochoo, M., & Alnajjar, F. (2025). Comparing Emotion Detection Methods in Online Classrooms: YOLO Models, Multimodal LLM, and Human Baseline. *2025 IEEE Global Engineering Education Conference (EDUCON)*, 1–7. https://doi.org/10.1109/EDUCON62633.2025.11016401

Reddy, K. U. K., Shaik, F., Swathi, V., Sreevidhya, P., Yashaswini, A., & Maheswari, J. U. (2025). Design and Implementation of Theft Detection Using YOLO Based Object Detection Methodology and Gen AI for Enhanced Security Solutions. *International Conference on Inventive Computation Technologies (ICICT)*, 583–589. https://doi.org/10.1109/ICICT64420.2025.11005144

Rezende, F. dos A., Hudson, T. M., Silva, P. A. F., Alves, W. F. de O., Mendes, A. L. C., & Brandão, A. S. (2025). Soccer Player Tracking Using UAV Imagery: A Comparative Study of YOLO and Traditional Image Processing Algorithms. *2025 International Conference on Unmanned Aircraft Systems (ICUAS)*, 147–154. https://doi.org/10.1109/ICUAS65942.2025.11007880

Roboflow: Computer Vision Tools for Developers and Enterprises. (n.d.). Retrieved June 16, 2025, from https://roboflow.com

Sai Mrudhun, P., Keerthana, M., Guru Nithysh, K., Akash, T., & R, M. (2024). Unified Detection Framework

for Robbery Events: Integrating YOLOv8, Fast R-CNN, and RetinaNet with Explainable AI Validation and Real-time Android Application Integration. *IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 1–6. https://doi.org/10.1109/CVMI61877.2024.10781669

Sapkota, R., Flores-Calero, M., Qureshi, R., Badgujar, C., Nepal, U., Poulose, A., Zeno, P., Vaddevolu, U. B. P., Khan, S., Shoman, M., Yan, H., & Karkee, M. (2025). YOLO Advances to Its Genesis: A Decadal and Comprehensive Review of the You Only Look Once (YOLO) Series. *Artificial Intelligence Review*, *58*(9), 1–83. https://doi.org/10.1007/s10462-025-11253-3

Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*, *10*, 84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing Human Actions: A Local SVM Approach. *International Conference on Pattern Recognition, ICPR 2004.*, *3*, 32-36 Vol.3. https://doi.org/10.1109/ICPR.2004.1334462

Shafizadegan, F., Naghsh-Nilchi, A. R., & Shabaninia, E. (2024). Multimodal Vision-Based Human Action recognition Using Deep Learning: A Review. *Artificial Intelligence Review*, *57*(7), Article 7. https://doi.org/10.1007/s10462-024-10730-5

Shaheen, M. Y. (2021). *Applications of Artificial Intelligence (AI) in Healthcare: A review*. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019. https://doi.org/10.1109/CVPR.2016.115

Shaikh, M. B., Chai, D., Islam, S. M. S., & Akhtar, N. (2024). From CNNs to Transformers in Multimodal Human Action Recognition: A Survey. *ACM Trans. Multimedia Comput. Commun. Appl.*, *20*(8), 260:1-260:24. https://doi.org/10.1145/3664815

Shen, J., Yan, W., Miller, P., Zhou, H. (2010) Human localization in a cluttered space using multiple cameras. IEEE International Conference on Advanced Video and Signal Based Surveillance.

Shen, Z., Zhang, M., Zhao, H., Yi, S., & Li, H. (2024). *Efficient Attention: Attention with Linear Complexities* (No. arXiv:1812.01243). arXiv. https://doi.org/10.48550/arXiv.1812.01243

Sikder, N., Ahad, M. A. R., & Nahid, A.-A. (2021). Human Action Recognition Based on a Sequential Deep Learning Model. *International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 1–7. https://doi.org/10.1109/ICIEVicIVPR52578.2021.9564234

Soomro, K., Zamir, A. R., & Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. arXiv. https://doi.org/10.48550/arXiv.1212.0402

Talati, D. (2023). AI in healthcare domain. *Journal of Knowledge Learning and Science Technology ISSN:*

2959-6386 (Online), *2*(3), Article 3. https://doi.org/10.60087/jklst.vol2.n3.p262

Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, *5*(4), 1680–1716. https://doi.org/10.3390/make5040083

Thi, T. H., Wang, L., Ye, N., Zhang, J., Maurer-Stroh, S., & Cheng, L. (2014). Recognizing Flu-Like Symptoms from Videos. *BMC Bioinformatics*, *15*(1), Article 1. https://doi.org/10.1186/1471-2105-15-300

Tian, Y., Ye, Q., & Doermann, D. (2025). *YOLOv12: Attention-Centric Real-Time Object Detectors*. arXiv.Org. https://arxiv.org/abs/2502.12524v1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, L., Huynh, D. Q., & Koniusz, P. (2020). A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. *IEEE Transactions on Image Processing*, *29*, 15–28. https://doi.org/10.1109/TIP.2019.2925285

Wang, X., Yan, W. (2019) Human gait recognition based on self-adaptive hidden Markov model. IEEE/ACM Transactions on Biology and Bioinformatics.

Wang, X., Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. International Journal of Neural Systems.

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification. Springer Multimedia Tools and Applications.

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. Neural Computing and Applications 32 (11), 7275-7287.

Wang, X., Yan, W. (2022) Human identification based on gait manifold. Applied Intelligence.

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer Nature.

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Springer Nature.

Wu, D., Sharma, N., & Blumenstein, M. (2017). Recent Advances in Video-Based Human Action Recognition Using Deep Learning: A Review. *International Joint Conference on Neural Networks (IJCNN)*, 2865–2872. https://doi.org/10.1109/IJCNN.2017.7966210

Wu, Q., Xu, G., Zhang, S., Li, Y., & Wei, F. (2020). Human 3D Pose Estimation in a Lying Position by RGB-D Images for Medical Diagnosis and Rehabilitation. *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 5802–5805.

https://doi.org/10.1109/EMBC44109.2020.9176407

Yang, B., Yan, W. (2024) Real-time billiard shot stability detection based on YOLOv8. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.159-172, Chapter 8, IGI Global.

Yang, G., Nguyen, M., Yan, W., Li, X. (2025) Foul detection for table tennis serves using deep learning. Electronics 2025, 14(1), 27.

Yang, G. (2025) ChatPPG: Multi-Modal Alignment of Large Language Models for Time-Series Forecasting in Table Tennis. Master's Thesis, Auckland University of Technology, New Zealand.

Yilmaz, E. N., & Navruz, T. S. (2025). Real-Time Object Detection: A Comparative Analysis of YOLO, SSD, and EfficientDet Algorithms. *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, 1–9. https://doi.org/10.1109/ICHORA65333.2025.11017287

Yin, H., Sinnott, R. O., & Jayaputera, G. T. (2024). A Survey of Video-Based Human Action Recognition in Team Sports. *Artificial Intelligence Review*, *57*(11), 1–55. https://doi.org/10.1007/s10462-024-10934-9

Yu, Z., Yan, W. (2020) Human action recognition using deep learning methods. International Conference on Image and Vision Computing New Zealand.

Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., & Chen, D.-S. (2019). A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, *19*(5), Article 5. https://doi.org/10.3390/s19051005

Zheng, K., Yan, Q., Nand, P. (2017) Video dynamics detection using deep neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence.

Zhou, H., Nguyen, M., Yan, W. (2023) Computational analysis of table tennis matches from real-time videos using deep learning. PSIVT 2023.

Zhu, W., Peng, B., Yan, W. (2024) Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. IEEE Transactions on Multimedia, 26 (7359 - 7371).

Zhu, Y., Peng, B., Yan, W. (2022) Ski fall detection from digital images using deep learning. ACM ICCCV.

.