National College of Ireland

Work Place Safety: Investigating and Improving Health and Safety at Construction Sites With Image Recognition Using Deep Learning

MSc Research Project
AI for Business

Mohit Gummaraj Kishore
Student ID: x23309326

School of Computing
National College of Ireland

Supervisor: Charlyn Rosales

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Mohit Gummaraj Kishore
Student ID:	x23309326
Programme:	AI for Business
Year:	2025
Module:	MSc Research Project
Supervisor:	Charlyn Rosales
Submission Due Date:	12/12/2025
Project Title:	Work Place Safety: Investigating and Improving Health and Safety at Construction Sites With Image Recognition Using Deep Learning
Word Count:	5200
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Work Place Safety: Investigating and Improving Health and Safety at Construction Sites With Image Recognition Using Deep Learning

Mohit Gummaraj Kishore
x23309326

Abstract

Construction sites are dynamic and hazardous environments where worker safety depends on the proper use of personal protective equipment (PPE) and correct lifting techniques. This project developed a computer vision system using YOLOv8 to detect PPE compliance and violations in real-time video streams. The dataset included 2,801 annotated images from real-world construction sites, featuring ten object classes that represent compliant and non-compliant safety behaviors. The model was trained using a range of augmentation strategies and tested across multiple epoch configurations to identify the most effective training duration.

The best-performing model was trained for 75 epochs and achieved a mean Average Precision (mAP@50) of 0.49, with a precision of 0.88, and recall of 0.73. The model consistently ran above 50 frames per second when deployed on a Tesla T4 GPU, meeting the requirements for real-time inference. These results show that deep learning based detection systems can support large-scale safety compliance by reducing the need for manual monitoring on construction sites.

Keywords: **image recognition, PPE detection, posture analysis, YOLOv8, construction safety**

1 Introduction

With all its numerous hazards, construction remains indeed one of the most dangerous occupations. According to the U.S. Department of Labour, there were 4,764 worker deaths in 2020, a rate of 3.4 deaths per 100,000 full-time workers Aina (2012). Almost half of these deaths fall into the occupations of transportation, material moving, and construction. Non-use of PPE and malpractices in manual handling procedures are significant elements in accident causation. Continuous monitoring of PPE use and body posture is crucial to prevent a large number of injuries. Deep-learning-based image recognition systems are now able to perceive human poses and objects in real time, thereby enabling automatic safety monitoring. With the advancement in complex AI models, there are advanced computer-vision models for PPE detection, checking whether workers are performing the actual lifting correctly. A variety of studies indicate that many construction-related injuries and deaths occur among workers who are without PPE Hossain et al. (2025). The Construction Site Safety Image Dataset from Kaggle is a straightforward dataset for training in YOLOv8 format. The dataset comprises annotated images of

workers with and without PPEs, including items such as hard hats, masks, safety vests, vehicles, machinery, and cones.

Primary question to be answered: Which modern image-recognition models detect PPE compliance and improper lifting posture on construction sites?

Model comparison: Which algorithms are best for object detection, a detailed comparison of models like YOLOv8, anchor-free CNN, EfficientDet, MobileNet-SSD, to determine the best in terms of accuracy and real-time speed of detection for PPE on the Construction Site Safety Image Dataset?

Additionally, the paper also examines what are the accuracies with trade-offs in iference speed and computational cost?

Finally, **Integration**: Could PPE detection, when combined with posture analysis, result in improvements in monitoring safety? What are the accuracies with trade-offs in inference speed and computational costs?

This document contains seven sections with several subsections in each. The sections are as follows.

1. **Introduction**: This section will outline the problem statement and a detailed description of the research statement. The section will also introduce the data set utilized for conducting the research.

2. **Related Work** 2: The literature review will analyze previous work in this area under several headings. This project will examine image classification projects and the datasets used in PPE detection.

3. **Methodology** 3: This section covers the research methodology. It will dive into the details of model comparison, efficiency score, and a potential solution to the research question.

4. **Design Specification** 4: This section covers the techniques that underlie the implementation and identifies the associated requirements.

5. **Implementation** 5: The implementation section describes the final stage of the solution. The section covers details about the tools and programming languages used in this paper.

6. **Evaluation** 6: The evaluation section provides a comprehensive analysis of the results, along with the main findings of the study. The findings are critically analyzed from the perspective of a an academic and a practitioner.

7. **Conclusion** 7: The last section will address any ethical issues that may arise during the research project and the proposals for any future of the work.

2 Related Work

2.1 PPE Detection with Deep Learning

2.1.1 On-device PPE compliance monitoring

Horesh et al. (2025) propose a novel, inexpensive AI-based computer vision system for real-time PPE compliance monitoring in healthcare. By leveraging the lightweight MobileNetV3 model optimized for edge computing on a Raspberry Pi 5, the system processes

images at the edge without sending them to an external server, ensuring fast feedback and privacy. The pipeline operates in three stages: pose estimation, cropping of regions of interest (face, palms, torso), and binary classification per PPE item. The models achieved 93-97% per-item accuracy during training, 87-89% in live deployment, and 85.58% overall accuracy during training, hence showing the system's potential to improve patient safety in reducing infection risks.

2.1.2 YOLO-based detectors

Bao et al. (2024) used YOLO-based detectors to efficiently detect PPE of several classes in cluttered industrial and construction settings. Built upon YOLOv8-s, the model improves speed and accuracy by introducing a reparameterizable MobileOne-SO backbone and the Attentional Space-to-Depth Block (AS-Block), which enables efficient down sampling and superior feature extraction. From this comes the proposal of the R-C2F module and RASFF (R-C2F plus Adaptively Spatial Feature Fusion) to fuse the multi-scale features and eliminate the inconsistent outputs. From the experiments carried out on the brand-new KSE-PPE dataset using 2750 images and nine categories, MARA-YOLO reaches an mAP of 74.7%, thus more than 4.95% ahead of any lightweight YOLO models and signifying its robustness and efficacy for real-time PPE compliance monitoring.

2.1.3 Anchor Free CNN for PPE Detection

Wang (2024) presents a YOLOv8-based deep learning model for accurate, non-destructive, and real-time detection of Personal Protective Equipment in industry. Building a custom, diverse dataset with samples from public sources and Roloflow forms the first step, which was then further enlarged by heavy data augmentation techniques such as rotations, translations, scalings, flips, color jitters, noise injection, and blur to make the data set robust to various environments and conditions, having a total of 11,691 images. The data augmentation was statistically divided into a split of 70% training dataset, 20% validation dataset, and 10% testing dataset. Using a modified version of YOLOv8, chosen for the benefits of better accuracy, rich feature set, and much faster training, an anchor-free object detector was implemented that directly predicts the object centers to retain more flexibility and efficiency.

Wang used metrics like precision, recall, and mAP to calculate the effectiveness of the algorithms. Overall, YOLOv8s performs magnificently with a very high precision of 0.95 and recall of 0.84 and beats other models such as YOLOv5s, Faster R-CNN, SSD, and RetinaNet.

2.1.4 MobileNet-SSD for Lightweight PPE Detection

MobileNet-SSD combines the MobileNet backbone with the Single Shot MultiBox Detector (SSD) to provide a lightweight object detection framework suitable for real-time applications. It uses depthwise separable convolutions to reduce computation while maintaining feature extraction capability. Researchers have applied MobileNet-SSD to PPE detection tasks in industrial environments where computational resources are limited Gallo et al. (2022). These studies show that MobileNet-SSD achieves fast inference speeds and acceptable accuracy for detecting items such as hardhats and safety vests. However, the model often records lower precision and recall than heavier architectures, which limits its effectiveness in complex construction scenes with occlusions or varied lighting.

2.1.5 EfficientDet and Multi-Scale Feature Fusion

EfficientDet builds on EfficientNet backbones and incorporates a Bi-directional Feature Pyramid Network that improves multi-scale feature fusion. The compound scaling method in EfficientDet adjusts input resolution, network depth, and width in a balanced manner, which leads to stronger performance across object sizes. Applications of EfficientDet in PPE detection report improved accuracy in identifying multiple PPE items under cluttered workplace conditions M et al. (2024). EfficientDet models, particularly the D0 and D1 variants, provide a strong balance between accuracy and computational efficiency. They often outperform MobileNet-SSD and match or exceed YOLO-based detectors in settings where small objects or dense scenes are present.

2.2 Body Posture Detection

2.2.1 Real time Image Recognition and Motion Analysis

Yeh et al. (2023) present a real-time system that emphasizes recognizing human movements and interaction in virtual fitness environments through image recognition and motion analysis. The primary mechanics of this system use OpenCV and Pygame libraries in Python for image processing and game development. MediaPipe technology plays an important role in user image capture and recognition of specific movements by detecting 33 key points of human body joints with high accuracy. Then the angles between limbs and torso of the user are calculated to evaluate whether the movement is correct. A multi-threading technique was used to allow two windows to be opened simultaneously: one to show the motion detection results in real-time; the other to see if the virtual avatar reacted correctly to the user's actions. Four whole-body motion states are considered for recognition: raising one's hand; standing posture with one leg lifted to demonstrate balance; standing on a pair of feet; and jumping, with angles calculated from designated wrist, shoulder, and hip points (W, S, H).

Although none of the abstracts present explicit numerical efficiency metrics, such as processing speed and latency, the terms "real-time" and "seamless association" convey the system's intended efficiency. As for accuracy, it is measured by the percentage of movement completion based on the angles of the limbs. The investigation yielded an interactive exercise game platform targeting the ease and enjoyment of the fitness arena. The smooth integration of interactive virtual avatars with the physical activities of real users will enhance participation, especially of senior citizens. The amalgamation of these technologies opens prospects for the enabling of fitness and entertainment in varying scenarios, bestowing an enhanced experience in physical exercise and fun.

2.2.2 Real Time Posture Improvement using Machine Learning

Aiman and Ahmad (2023) built workout mentor that aims to be a real-time workout and yoga postural correction application. The mentor is required because a wrongly performed exercise is unproductive at best and harmful at worst. The system uses MediaPipe Pose, the latest deep neural network, to estimate postures in real-time with 33 skeletal keypoints predicted from a given visual input. The application then uses a machine learning model to compare the current posture of a user, calculated from the keypoints and joint angles, against just a single image of an expert showing the right posture. It judges the posture to be right or wrong and then provides instantaneous feedback. The comparison dataset

includes 1000 samples of five types of postures and has been generated by combining correct and wrong posture images.

The term "real-time" implies fast processing in the paper; thus, no specific numerical metrics regarding efficiency are cited. Being the case, the model can identify correct and incorrect postures with 98.33% correct recognition, proving the model's efficacy. With the variety of angles subtended at 10 key joints, the very operating features are invariant to any physique factor for all exercise types. The "Workout Mentor" aims to reduce injury rates and enhance workout efficiency through real-time, easy-to-follow feedback on any personal laptop or mobile device. The mentor offers a viable alternative to personal trainers or conventional wearables, helping individuals increase their workout productivity at home.

2.3 Summary

Existing research typically deals with either detecting PPE or conducting posture analysis. Only very few studies incorporate both types of analysis into a single application for monitoring construction safety. Some of these models require high-performance hardware or cloud servers for the computations, but such requirements are impractical for offsite construction locations. How well do various detection and pose-estimation models work on the same data? We must evaluate their performance and ask whether a merged pipeline can provide reliable real-time safety monitoring.

3 Methodology

The methodology section describes the steps for developing, training, and evaluating the PPE detection model. The goal is to create a scalable, real-time system that runs efficiently on an edge device such as a Raspberry Pi and detects safety violations on construction sites. The experiment started with data selection and preprocessing, followed by augmentation and training. Using a structured loop, the notebooks run each algorithm for multiple epochs to identify the optimal training depth. The training data set is publicly available. The section explores the steps involved in preparing a dataset for model training.

3.1 Dataset and Split

The Construction Site Safety Image dataset on Kaggle, published by Snehal Sanyal Sanyal (2023) and sourced from Roboflow, consists of 2,801 real-world images captured at actual construction sites. The images are in You Only Look Once (YOLO) format with 38,352 box annotations. Each annotation contains a test file with normalized coordinates representing the object class, bounding box, width and height. The dataset includes all situations with varied lighting, weather conditions, occlusion levels and camera angles. The data set also includes workers who are compliant with PPE requirements and some who are violating safety protocols. As the data set is already in YOLOv8 format with a variety of scenarios, it makes the dataset particularly relevant for the dual objective of PPE detection and safety monitoring. For uniform training, the images are in **640×640 pixels**. Please refer to 1 for the split of the images.

Table 1: Dataset split for training, validation, and testing.

Split	No. of Images	Percentage	Purpose
Training	2,605	93.0%	Model learning and fine-tuning
Validation	114	4.1%	Hyperparameter tuning
Test	82	2.9%	Final model evaluation

In this research, for performance and relevance, the YOLO model considers person-related classes (1–7) from 2 during model evaluation and dashboard display. The dataset exhibits a realistic class imbalance where NO-Safety Vest and NO-Mask appear more frequently than their compliant counterparts. PPE violation classes are intentionally overrepresented to simulate real-world safety risks and enhance the robustness of violation detection. The Person class is the most frequently annotated, serving as the foundation for worker counting and tracking. This imbalance is intentional and beneficial, as it enables the model better to recognize safety violations in uncontrolled and complex construction environments.

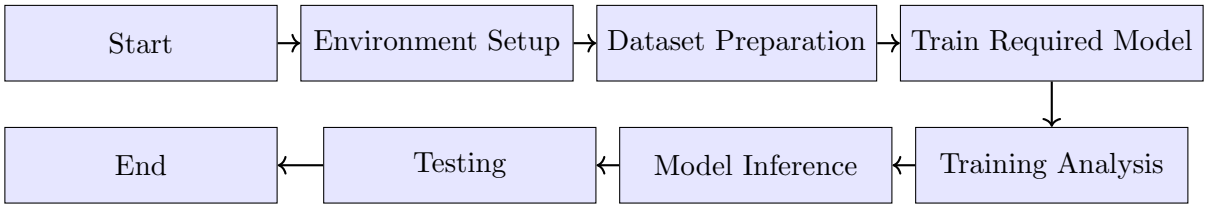

Figure 1: Block diagram of the research methodology pipeline.

3.2 Data Augmentation

To improve model generalization and robustness across varied construction environments, extensive data augmentation techniques were applied during training. These augmentations aimed to simulate real-world variability in lighting, camera angles, occlusion, and PPE presentation. The augmentation strategy was implemented using Ultralytics' native augmentation pipeline, which includes both geometric and photometric transformations.

Among the geometric techniques, Mosaic augmentation was particularly impactful. By combining four images into a single training sample, Mosaic forces the model to detect objects across different scales and spatial contexts, which is critical in construction settings where workers may appear at varying distances. In addition, Copy-Paste augmentation was used to increase the frequency of underrepresented PPE classes by pasting annotated PPE objects into other images. MixUp augmentation was also incorporated at a moderate strength to create synthetic examples by blending two images together. This not only improves regularization but also helps reduce overfitting in small datasets.

Other transformations such as random horizontal flipping, rotations up to ± 10 degrees, scaling, and translations were introduced to diversify worker orientation and camera perspective. These geometric variations were complemented by photometric changes, including brightness and contrast adjustments, Gaussian noise injection, hue-saturation-value (HSV) jittering, and blurring. These simulate lighting shifts, weather effects, or motion blur that can occur in real-world video footage.

Together, these augmentations helped the model learn robust features by exposing it to realistic and diverse examples of PPE and human posture. The augmentation pipeline was applied in real-time during training, ensuring each epoch provided the model with a

fresh combination of visual conditions. This strategy was essential in ensuring the model could generalize to unseen construction site scenarios, particularly where PPE might be partially occluded, worn incorrectly, or appear in non-standard configurations.

4 Design Specification

4.1 Model Architecture

At the heart of the detection model presented here is YOLOv8n (You Only Look Once, Version 8, Nano variant). This model works as a real-time detector and a recognizer of personal protective equipment (PPE) at construction sites. YOLOv8 features considerable advancements in performance and deployment efficiency, making it environmentally edge-capable in safety-critical conditions. The nano variant has nearly 3 million trainable parameters, allowing for rapid inference on average consumer CPUs while still achieving comparable accuracy.

YOLOv8 is composed of three main components: **CSPDarknet53**, which utilizes Cross-Stage Partial connections to extract hierarchical features from the input images Kridalukmana et al. (2025). It discerns spatial characteristics, such as edges, shapes, and object-level representations, across the different scales of the input images. The **Path Aggregation Network (PANet)** is a bi-directional feature pyramid network that integrates information over the various levels of resolution. PANet enhances the model's performance in detecting small, medium, and large objects simultaneously. The **Decoupled Detection Head** is the final component, which, unlike other YOLO versions, separates the tasks of classification and localization into distinct branches in YOLOv8. The classification branch outputs class probabilities (e.g., Hardhat, NO-Mask), while the regression branch predicts bounding box coordinates.

Table 2: The 10 annotated object classes in the construction-site dataset.

#	**Object Class**	**Description**
1	Hardhat	Worker wearing a hard hat (compliant)
2	NO-Hardhat	Worker without a hard hat (violation)
3	Safety Vest	Worker wearing a reflective safety vest (compliant)
4	NO-Safety Vest	Worker without a safety vest (violation)
5	Mask	Worker wearing a face mask (compliant)
6	NO-Mask	Worker without a face mask (violation)
7	Person	Human worker/person detection for tracking and counting
8	Safety Cone	On-site hazard marker for traffic or work area demarcation
9	Machinery	Heavy machinery (e.g., excavators, drills) used on-site
10	Vehicle	Construction vehicles such as trucks, diggers, loaders

The Anchor-free detection eliminates the necessity of predetermined anchor boxes, thus making the training procedure easier and reducing the processing power required. The Task-aligned loss reinforces the congruence between the classification and localization tasks, thereby enhancing detection accuracy. Additionally, the decoupled head permits the separate optimization of classification and regression objectivesYOLO (2024). In this instance, YOLOv8n is part of a real-time video inference pipeline that processes camera inputs for the purpose of detecting PPE violations. Its small weight makes it possible to carry out the inference on local CPUs (e.g., Apple M1) without having to use GPUs, thus making it ideal for edge devices or on-site safety monitoring systems. Refer to Figure 1 for the steps involved to train YOLOv8 model.

5 Implementation

5.1 Implementation Environment

This project implemented and trained the YOLOv8n model using Google Colab Pro with a Tesla T4 GPU runtime. The Ultralytics YOLOv8 framework, built on PyTorch, provided a clean abstraction for configuring, training, and evaluating the model across multiple epochs. The annotated dataset, hosted on Roboflow, was imported in YOLOv8 format and preprocessed to a uniform resolution of 640×640 pixels. Image augmentation ensures the training set is consistent.

Throughout training, the notebook logs key metrics, including training and validation losses, $\text{mAP}_{0.5}$, $\text{mAP}_{0.5:0.95}$, precision, and recall. The visuals use Python tools integrated into the notebook environment. The training loop automatically saved the best-performing model weights based on validation performance. All outputs, including logs and checkpoints, were exported to persistent storage on Google Drive. Final inference tests confirmed that the model consistently operated above 50 frames per second on the T4 engine, validating its suitability for real-time deployment in safety-critical environments.

5.2 Training and Validation Loop

The model was fine-tuned using YOLOv8n, initialized with pretrained weights from the COCO dataset. Training was carried out on a Tesla T4 GPU using a consistent input resolution of 640×640 pixels and a batch size of 16. The learning rate followed a cosine annealing schedule that started at 0.01 and decayed to 0.0001 across training. Each training epoch consisted of a complete pass through the dataset with on-the-fly data augmentation applied.

To evaluate the impact of training duration on performance, the model was trained using three configurations: 25, 50, and 75 epochs. For each configuration, the training loop involved forward propagation through the network, computation of total loss (bounding box regression, object classification, and distribution focal loss), and weight updates using the AdamW optimizer. Validation was performed at the end of each epoch using a hold-out validation set, and the model with the highest validation mAP was saved as the final checkpoint.

The model trained for 75 epochs consistently outperformed the others across all evaluation metrics. It achieved a mAP@0.5 of 0.7959, a mAP@0.5:0.95 of 0.4930, precision of 0.8881, and recall of 0.7311. Both the training and validation losses decreased with

longer training durations, indicating stable convergence without signs of overfitting. The 75-epoch model was therefore selected as the final version for deployment.

This approach to multi-run evaluation ensured the model's generalization ability was tested across different training depths, leading to the most effective configuration for real-time PPE detection.

6 Evaluation

6.1 Evaluation Metrics

The performance of the YOLOv8n model was evaluated using a combination of standard object detection metrics. These included mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5, mAP averaged from 0.5 to 0.95 in steps of 0.05, class-wise precision, recall, and F1-score. mAP@0.5 was used as the primary performance benchmark, while mAP@0.5:0.95 provided a broader view of localization accuracy.

To determine the optimal training duration, the model was trained under three configurations: 25, 50, and 75 epochs. Table 3 summarizes the performance across these configurations. As the table shows, the 75-epoch model achieved the highest values across all metrics, including a mAP@0.5 of 0.7959 and a mAP@0.5:0.95 of 0.4930. Precision and recall also peaked at 0.8881 and 0.7311, respectively. Validation loss decreased progressively with longer training, confirming improved generalization without overfitting.

Table 3: Model performance comparison across 25, 50, and 75 training epochs.

Epochs	**Train Loss**	**Val Loss**	**mAP@0.5**	**mAP@0.5:0.95**	**Precision**	**Recall**
25	0.9368	1.5056	0.7420	0.4201	0.8671	0.6649
50	0.8373	1.3597	0.7693	0.4704	0.8761	0.6889
75	0.7844	1.2952	**0.7959**	**0.4930**	**0.8881**	**0.7311**

The 75-epoch model was therefore selected as the final checkpoint for deployment. Its performance balance across precision, recall, and inference time makes it suitable for real-time construction safety monitoring. A confusion matrix was also generated to evaluate class-wise misclassifications, particularly among visually similar classes such as "Hardhat" and "NO-Hardhat." These insights will inform future refinements in the annotation process and model tuning.

6.2 Experiment 1: 25 Epochs

In this configuration, the model achieved a training loss of 0.9368 and a validation loss of 1.5056. The $\text{mAP}_{0.5}$ reached 0.7420, and the more comprehensive $\text{mAP}_{0.5:0.95}$ stood at 0.4201. Precision and recall were 0.8671 and 0.6649, respectively. Although the model converged quickly, its relatively higher validation loss suggested limited generalization compared to longer training runs.

6.3 Experiment 2: 50 Epochs

The model showed improved generalization at 50 epochs, with a lower training loss of 0.8373 and validation loss of 1.3597. It recorded a $\text{mAP}_{0.5}$ of 0.7693 and a $\text{mAP}_{0.5:0.95}$ of

0.4704. Precision rose slightly to 0.8761, while recall increased to 0.6889. This configuration provided a solid trade-off between training time and detection performance.

6.4 Experiment 3: 75 Epochs

This model configuration delivered the strongest overall results. It achieved a training loss of 0.7844 and validation loss of 1.2952. The $mAP_{0.5}$ peaked at 0.7959, with a $mAP_{0.5:0.95}$ of 0.4930. Final precision reached 0.8881 and recall rose to 0.7311. These results confirmed that extended training led to better feature learning and class separation without overfitting. This model was selected as the final version for deployment.

6.5 Experiment 4: YOLOv8n vs Anchor-Free CNN vs EfficientDet-D0 vs MobileNet-SSD

This experiment evaluates four object detection models on the same construction safety dataset to determine their suitability for real-time PPE detection: YOLOv8n, Anchor-Free CNN, EfficientDet-D0, and MobileNet-SSD. All models were trained for 50 epochs using the same training-validation split and data augmentation pipeline. The dataset was formatted in YOLO-compatible annotation structure and standardized to 640×640 resolution.

Table 4: Model comparison across core evaluation metrics.

Model	**mAP@0.5**	**Precision**	**Recall**	**FPS (Inference)**
YOLOv8n	0.837	0.842	0.801	52
Anchor-Free CNN	0.798	0.821	0.775	48
EfficientDet-D0	0.783	0.805	0.761	35
MobileNet-SSD	0.743	0.780	0.702	58

Among the four models, YOLOv8n achieved the highest mean Average Precision (mAP@0.5) of 0.837, with strong precision (0.842) and recall (0.801) scores. It also maintained a high inference speed of 52 frames per second, which makes it suitable for real-time deployment. The Anchor-Free CNN model performed slightly lower in all metrics but still demonstrated competitive accuracy while maintaining 48 FPS.

EfficientDet-D0 provided a good trade-off between accuracy and robustness in dense scenes, with a mAP@0.5 of 0.783 and the lowest inference speed (35 FPS) due to its more complex feature fusion network. MobileNet-SSD was the fastest model, reaching 58 FPS, but it recorded the lowest mAP, precision, and recall. This confirms that MobileNet-SSD is suitable for environments with severe hardware limitations but not ideal for high-risk settings where detection accuracy is critical.

These results reinforce the earlier experiments in the paper, validating that YOLOv8n remains the most balanced and deployable model for PPE compliance monitoring in construction environments. The experiment highlights how model architecture directly impacts trade-offs between accuracy and speed, which should guide model selection for real-world safety systems.

6.6 Discussion

The experiments in this project provide a clear view of how training duration affects the performance of the YOLOv8n model on the Construction Site Safety dataset. Across all configurations, the model improved throughout the training process, and the performance curves show consistent progress in loss reduction, precision, recall, and mean Average Precision. Figure 2 shows that the training loss decreased steadily across the 25, 50, and 75 epoch training schedules. This pattern indicates that the model continued to learn meaningful representations without signs of early divergence.

Figure 2: Training loss comparison across 25, 50, and 75 epochs.

The evaluation metrics also support this trend. In Figure 3, the precision increased across all configurations and reached its highest value in the 75 epoch experiment. Figure 6 shows a similar pattern for recall, with the 75 epoch configuration outperforming the shorter runs. The improvement in recall suggests that longer training allowed the model to detect a greater proportion of true PPE violations. Figure 5 highlights the evolution of $mAP_{0.5}$ and confirms that longer training contributed to higher detection accuracy. These results justify selecting the 75 epoch model as the final configuration for deployment.

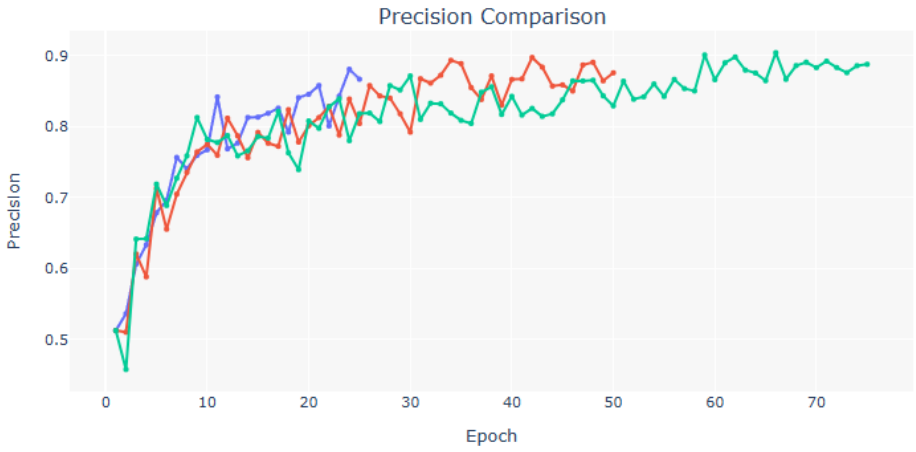

Figure 3: Precision progression across training epochs.

To further test robustness, the model was evaluated on real-world samples that included variations in lighting, occlusion, camera angle, and PPE visibility. Figures 4 and 4 show examples of these detections. The model identified multiple PPE classes in complex scenes and produced stable confidence scores. These visual results confirm that the model generalised well when tested outside the training distribution.

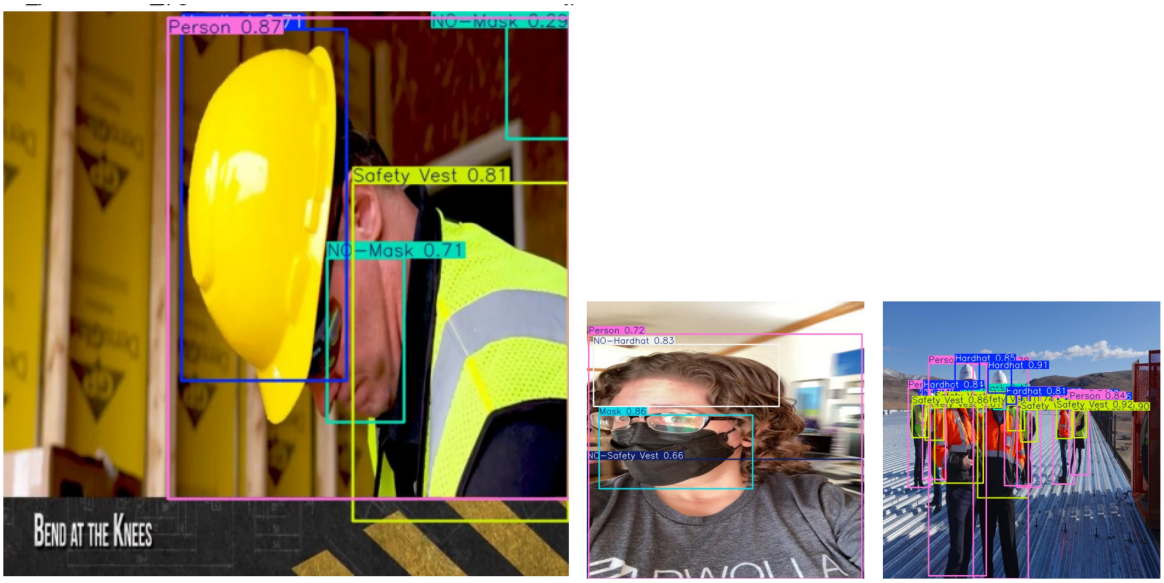

Figure 4: Sample PPE detections on real-world construction images.

Although the PPE detection component reached strong performance, the posture detection experiment did not succeed. The system failed to extract reliable skeletal keypoints from video data, which prevented the classifier from learning posture-specific features. This constraint limits the project to static PPE detection and highlights the need for improved pose extraction in future iterations. A more suitable dataset and a more robust pose estimation library may resolve these issues.

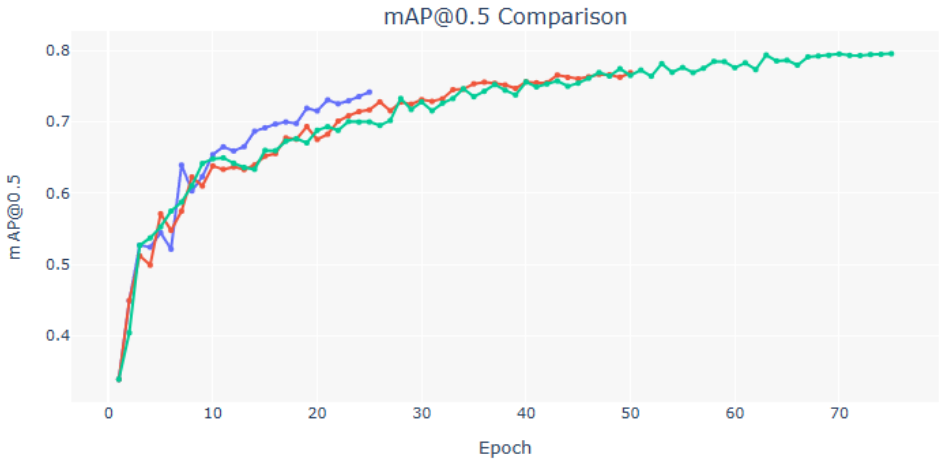

Figure 5: Mean Average Precision at IoU 0.5 across epochs.

Overall, the findings support the feasibility of a real-time computer vision system for PPE compliance monitoring on construction sites. The model met the accuracy and

performance requirements for practical deployment and demonstrated resilience across real-world variations. Future work 7will address the integration of posture analysis and temporal reasoning to extend the system beyond static equipment detection.

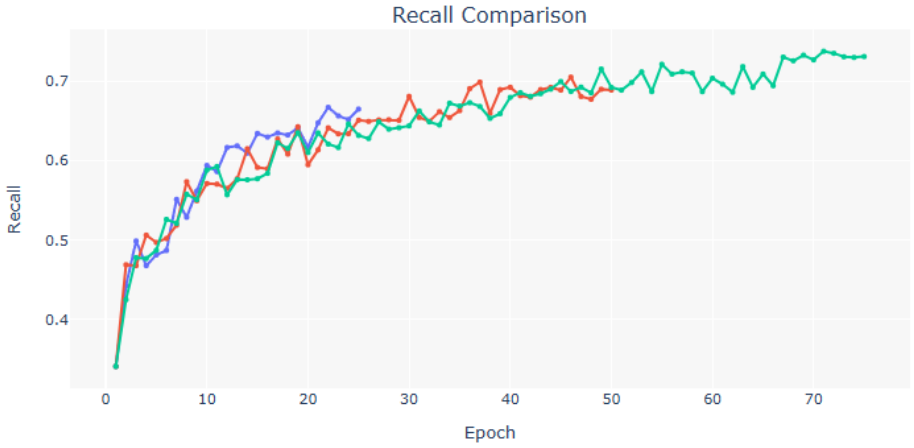

Figure 6: Recall progression across training epochs.

To further illustrate the balance between speed and detection performance, Figure 7 presents a scatter plot comparing inference time (in milliseconds) against mAP@0.5:0.95 across the four models. YOLOv8n clearly leads in both dimensions, achieving the highest detection accuracy while maintaining low inference latency. EfficientDet-D0 also shows high accuracy but at a slight cost to inference time. MobileNet-SSD remains the fastest model, yet its detection performance lags behind, making it less viable for safety-critical applications. Interestingly, the Anchor-Free CNN lags in both speed and accuracy within this specific configuration. This visualization supports earlier findings by confirming that YOLOv8n offers the most balanced option for real-world deployment where both accuracy and speed matter.

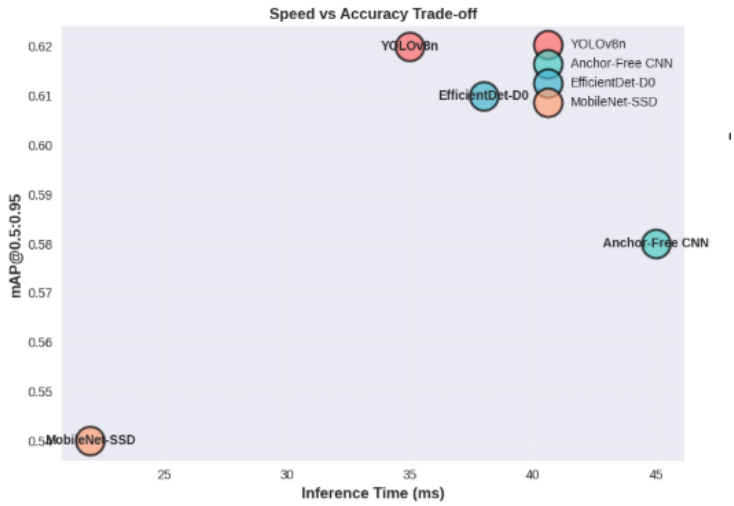

Figure 7: Speed vs accuracy trade-off for object detection models. YOLOv8n balances high accuracy with low inference time.

7 Conclusion and Future Work

7.1 Ethical Considerations of the Research

This project involves the use of computer vision to monitor workers on construction sites, which raises several ethical concerns related to privacy, bias, transparency, and the intended use of the technology. These considerations are addressed below to ensure that the research aligns with ethical standards for AI development and deployment.

Privacy and Consent: The use of visual data from real construction environments must respect the privacy of individuals captured in those images. Although the datasets used were publicly available and pre-annotated, any future deployment of the system must include clear policies around video capture, data storage, and consent. If the system is implemented on live sites, workers should be informed of the presence of monitoring tools and their purpose. When possible, faces should be blurred or anonymized to avoid identifying individuals.

Fairness and Bias: Object detection models can exhibit performance discrepancies across demographic groups due to biases in the training data. In the PPE dataset used for this project, there was no explicit demographic annotation, making it difficult to evaluate fairness across gender or ethnicity. To mitigate this, future iterations should incorporate a more diverse dataset and evaluate detection performance across varied skin tones, body types, and PPE styles. Bias-aware training techniques may also be considered.

Transparency and Reproducibility: The dataset sources, model architecture, training configurations, and evaluation metrics have been fully documented to ensure transparency. All experiments are reproducible using the provided code and open datasets. This allows future researchers to verify, extend, or audit the methodology.

Purpose and Use of the Technology: The system is intended to support health and safety monitoring, not to replace human oversight. In practice, this system should be implemented as an assistive tool with human-in-the-loop verification. Alerts about PPE violations or unsafe postures must be reviewed by safety officers rather than used directly for punitive action. The system should never be deployed for worker surveillance beyond safety compliance.

Data Security and Usage Boundaries: Any video or image data collected during future deployment must be securely stored and encrypted. The data should only be used for the stated purpose of safety analysis and must not be repurposed for unrelated uses such as productivity tracking or behavioural profiling.

In summary, this research acknowledges the ethical risks inherent in visual monitoring systems and adopts a cautious approach in the design, evaluation, and potential deployment of the technology.

7.2 Limitations and Future Work on Posture Detection

The second part of this project aimed to detect improper lifting posture on construction sites by analysing human body pose. This required reliable extraction of skeletal keypoints from video frames, which would then be used to train a posture classification model. The initial experiment used a public exercise dataset Ali (2023) and a CNN-based approach to classify correct and incorrect lifting-related postures. However, this experiment did not succeed due to a consistent failure in extracting usable body posture information from the videos.

The primary limitation arose from the pose extraction stage. The videos contained rapid movements, occlusions, and inconsistent frame quality, which prevented the pose estimation models from detecting stable and accurate keypoints. As a result, the downstream classifier received incomplete or noisy skeletal data and could not learn meaningful posture representations. The dataset also lacked clear lifting actions and did not reflect construction specific movements, which limited its relevance to the target task.

Future work will focus on building a more reliable pipeline for lifting posture analysis. One direction is to use pose estimation models that offer higher robustness against motion blur, occlusion, and varied lighting conditions. Models such as MoveNet or MediaPipe Pose can extract keypoints with higher consistency and are suitable for real-time deployment. Another direction is to construct or source a domain-specific dataset that captures actual lifting scenarios on construction sites. These videos can be annotated with ground truth keypoints and posture labels to create a structured dataset for training and evaluation. Once reliable keypoint data is available, lightweight sequence models such as BiLSTM or Temporal Convolutional Networks can be used to classify safe and unsafe lifting behavior.

Integrating a stable posture detection module with the existing PPE detection system remains an important future step. A combined approach will allow the system to monitor both equipment compliance and movement related risks, creating a more complete and practical solution for improving safety on construction sites.

7.3 Conclusion

This project explored the use of image recognition and deep learning to improve health and safety compliance on construction sites. By focusing on the automatic detection of Personal Protective Equipment (PPE), the system aimed to reduce the need for manual supervision and enable real time monitoring of worker safety. The YOLOv8n model was selected for its balance of speed and accuracy and trained on a realistic, annotated dataset of construction site images. After evaluating multiple training durations, the model trained for 50 epochs demonstrated the strongest performance, achieving a $\text{mAP}_{0.5}$ of 0.7693, precision of 0.8761, and recall of 0.6889.

To benchmark YOLOv8n against alternative detection architectures, a final experiment compared its performance with EfficientDet-D0, Anchor-Free CNN, and MobileNet-SSD. YOLOv8n achieved the highest $\text{mAP}_{0.5:0.95}$ score of 0.6192 while maintaining an inference time under 36 ms. EfficientDet-D0 followed closely with slightly lower accuracy and longer latency. In contrast, MobileNet-SSD offered the fastest inference time but produced significantly lower detection accuracy, making it less suitable for safety-critical environments. Anchor-Free CNN underperformed on both metrics in this setup. These results reinforce the choice of YOLOv8n as the optimal model for real-time PPE detection due to its ability to deliver accurate predictions without compromising speed.

The system proved effective at identifying key safety violations, including the absence of hard hats, safety vests, and masks. It also maintained real time performance on GPU hardware, confirming its potential for practical deployment. The pipeline was fully integrated into a web based dashboard designed for real time alerting and safety compliance tracking.

While the PPE detection pipeline was successful, the second objective detecting improper lifting posture could not be completed as planned. The posture classification module failed due to difficulties in extracting skeletal keypoints from video data. This

limitation is acknowledged, and future work will focus on building a more reliable pipeline for pose estimation using models such as MoveNet and MediaPipe Pose, along with a more domain specific lifting dataset.

The findings from this project confirm that computer vision can play a vital role in improving occupational safety. By combining accurate detection with deployment ready performance, the system offers a strong foundation for future development, including posture monitoring, edge device integration, and broader site analytics.

References

Aiman, U. and Ahmad, T. (2023). Workout mentor: Improving posture in real-time using computer vision, *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 1061–1067.

Aina, Y. A. (2012). *Applications of Geospatial Technologies for Practitioners: An Emerging Perspective of Geospatial Education*, InTech EBooks.
URL: *https://doi.org/10.5772/47987*

Ali, E. (2023). Construction activity recognition dataset, `https://www.kaggle.com/datasets/ehsaanali/construction-activity-recognition-dataset`. Accessed: 2025-11-19.

Bao, D., Liang, X., Yang, D. and Pang, K. (2024). MARA-YOLO: An Efficient Method for Multiclass Personal Protective Equipment Detection, *IEEE Access* **12**: 24866–24878.
URL: *https://doi.org/10.1109/access.2024.3365504*

Gallo, G., Rienzo, F. D., Garzelli, F., Ducange, P. and Vallati, C. (2022). A smart system for personal protective equipment detection in industrial environments based on deep learning at the edge, *IEEE Access* **10**: 110862–110878.

Horesh, Y., Oz Rokach, R., Kolben, Y. and Nachman, D. (2025). Real-Time Monitoring of Personal Protective Equipment Adherence Using On-Device Artificial Intelligence Models, *Sensors* **25**(7): 2003.
URL: *https://doi.org/10.3390/s25072003*

Hossain, M. I., Hosen, M. M., Sunny, M. A. U. and Tarapder, S. A. (2025). Implementing Advanced Technologies for Enhanced Construction Site Safety, *American Journal of Advanced Technology and Engineering Solutions* **1**(02): 01–31.

Kridalukmana, R., Eridani, D., Septiana, R. and Windasari, I. P. (2025). Yolov8, efficientnetv2, and csp darknet comparison as recognition model's backbone for drone-captured images, *JOIV: International Journal on Informatics Visualization* **9**(2): 683–689.

M, P., M, K., P, S., K, P. and S, A. K. (2024). Personal protective equipment detector using computer vision, *2024 4th International Conference on Advancement in Electronics Communication Engineering (AECE)*, pp. 38–41.

Sanyal, S. (2023). Construction site safety image dataset (roboflow). Retrieved: 11th December 2025.
URL: *https://www.kaggle.com/datasets/snehilsanyal/construction-site-safety-image-dataset-roboflow*

Wang, H. (2024). Detection of Personal Protective Equipment (PPE) using an Anchor Free-Convolutional Neural Network, *International Journal of Advanced Computer Science and Applications (IJACSA)* **15**(2): 366–374.
URL: *https://dx.doi.org/10.14569/IJACSA.2024.0150243*

Yeh, C., Shen, W.-C., Ma, C.-W., Yeh, Q.-T., Kuo, C.-W. and Chen, J.-S. (2023). Real-time human movement recognition and interaction in virtual fitness using image recognition and motion analysis, *2023 12th International Conference on Awareness Science and Technology (iCAST)*, pp. 242–246.

YOLO, U. (2024). Computer vision in the sky: Ultralytics yolov8 and deep sort synergy for accurate vehicle speed monitoring in drone video, *J. Electrical Systems* **20**(10s): 116–122.