Automated Construction Site Safety Violation Detection using Object Detection and Convolutional Neural Networks

Submitted in Partial Fulfillment of the Requirements in
CPE313 Data Science Track Elective 3:
Advanced Machine Learning and Deep Learning

Submitted by
Audrick Zander G. Cuadra
Marc Joseph S. Dolores

May 23 2025

Automated Construction Site Safety Violation Detection using Object Detection and Convolutional Neural Networks

Audrick Zander G. Cuadra
Computer Engineering
Technological Institute of the Philippines
Quezon City
qazgcuadra@tip.edu.ph

Marc Joseph S. Dolores
Computer Engineering
Technological Institute of the Philippines
Quezon City
qmjsdolores@tip.edu.ph

ABSTRACT

Construction sites are primarily known as risky areas. However, one of the major preventive factors that help in the prevention of danger to workers is the compliance of Personal Protective Equipment (PPE)[1]. Manual observation here is sometimes cumbersome and error-prone. This research is all about designing and comparing three CNN-based models for automatic recognition of PPE compliance by taking photographs while on-site. We worked on Custom CNN-LSTM, CNN-Bidirectional LSTM, and NASNetMobile-based transfer learning models, where all models were trained on a labeled dataset consisting of safety-compliant and non-compliant images. Additionally, object detection models such as YOLOv8, RT-DETR, and RetinaNET were trained to detect different PPE worn by each worker for compliance assessment. The findings of the research reveal that Custom CNN-LSTM achieved the highest validation accuracy of 92%, followed by CNN-Bidirectional LSTM 85%, and NASNetMobile-based transfer learning lagging behind at 83%. In addition, object detection models showcased results indicating that RT-DETR was the best-performing model, with a validation MAP50 score of 91%. This was then followed by RT-DETR with a score of 94%, and lastly RetinaNET (to be followed). Further performance analysis on practical applicability and deployment potential for each model showed that Custom CNN-LSTM and RT-DETR had the highest suitability for real-time detection of PPE from static images. CNN LSTM and object detection best models were both tested 20 times however object detection showed better results than the CNN LSTM. In object detection the accuracy reached 100% while CNN LSTM only got 95% which made object detection better. Also, object detection was more efficient than the CNN LSTM model but CNN LSTM is much faster. With the model improvements can be done that can be used to monitor construction sites and lessen PPE involved accidents.

Keywords: *PPE compliance, construction safety, deep learning, object detection*

I. Introduction

The construction industry plays a pivotal role in any country's economic and development sectors, as it provides not only infrastructure but also job opportunities [2]. Construction sites are primarily well-known as dangerous by nature, especially when safety measures are not strictly adhered to. Equipping proper Personal Protective Equipment (PPE) minimizes the risk of unprecedented accidents [3]. As a matter of fact, 56% of construction site mishaps are caused by the absence of PPE items [4]. These PPE items provide a first layer of protection for construction workers against common site hazards such as falling debris, sharp objects, and many others. The most commonly seen PPE items for construction workers include helmets, which protect an individual's head from concussions; safety vests, which enable workers to be easily seen; gloves, which protect an individual's hands against abrasions and cuts; and boots, which provide workers an edge against foot injuries and prevent slipping. Ensuring that each individual worker wears these items is crucial for safety compliance and minimizes workplace injuries.

This study aims to automate the process of PPE compliance detection with the implementation of object detection and Convolutional Neural Networks (CNN) to enhance efficiency and reliability. Object detection was used to identify each PPE item worn, while CNN is used to classify whether all workers are wearing complete PPE. Subsequently, it provides a performance evaluation for the best performing object detection model and CNN-LSTM model by subjecting them through the same 20 test videos.

Taking this into account, by ensuring that all workers are equipped with complete PPE before starting their labor promotes following safety compliance and reduces injuries brought workplace hazards. This aligns with the United Nations Sustainable Development Goal (UN SDG) 8, Target 8.8, which aims to protect labor rights and promote a safe and secure working space for all workers.

II. Related Work

The mandatory utilization of Personal Protective Equipment (PPE) is a significant concern in safety compliance for its direct influence in the reduction of workplace injuries. Various researchers have investigated the implementation of deep learning models to automate the monitoring of PPE compliance.

A. Object Detection

Vishwakarma et al. (2025) proposed a smart PPE real-time monitoring system implemented with YOLOv7 to guarantee safety compliance in real-time. Their work achieved an accuracy of 97.29%, 91.75% in precision, and 96.39% in recall [1]. Hao Wang et al. (2024) propose LMD-RTDETR, a lightweight model that is designed for PPE identification from small targets, thus making it more robust when faced with different environmental conditions for real-time detection [5]. Another study showcased the performance of RetinaNET reaching 74% in mAP50, 93% in precision, and 82% in recall; this indicates that the model needs a bit more fine-tuning before it can be fit for real-time detection [6].

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have been widely utilized for image-based classification tasks in PPE compliance systems due to their spatial feature extraction capabilities. Shao et al. (2021) proposed a CNN-based architecture to classify PPE usage from site images, achieving a classification accuracy of 91.3% in construction environments [11]. Another study by Silva-Rojas et al. (2021) developed a real-time helmet detection system using a lightweight CNN, which attained over 90% detection accuracy in video streams recorded on-site [12]. Jiang et al. (2020) enhanced traditional CNN models to better detect safety helmets and

vests under low visibility, reporting a precision of 92% and a recall of 89% [14]. Albahli and Alwan (2021) applied CNNs for PPE detection in industrial environments and achieved consistent accuracy across varied lighting and distance conditions, recording a mean accuracy of 88.7% [13]. In a related work, Nguyen et al. (2021) employed transfer learning with pre-trained CNNs and reached 90.1% classification accuracy for PPE compliance in real-world construction site images [15]. Lastly, a foundational approach using deep CNNs trained from scratch by Krizhevsky et al. (2017) set a benchmark in image classification, influencing many modern PPE detection architectures through its effective hierarchical feature learning strategy [10].

III. Methods

A. Dataset and Preprocessing

The dataset that was used in this study to train the different object detection models was derived from Roboflow which showcases the different PPE items, them being worn by different construction workers, and individuals who are not wearing proper PPE. The dataset provides five distinct annotated classes of PPE items, namely Helmet, Vest, Gloves, and Boots.

A.1 Object Detection Dataset

TABLE I. Dataset

Split	Images	Percentage
Training	989	71%
Validation	262	19%
Testing	148	11%

A.2 CNN dataset

TABLE II. Dataset

Split	Images	Percentage
Training	630	70%
Validation	270	30%
Testing	150	

B. Object Detection Models

Object detection models add depth to the system by pinpointing the PPE worn by personnel, which is crucial for Personal Protective Equipment (PPE) compliance monitoring system, as it can emphasize missing PPE items for each worker. The study implements and compares the performance of three object detection models: YOLOv8n, RTDETR-L, and RetinaNET.

B.1 **YOLOv8n**: a pre-trained YOLOv8 model was imported using ultralytics and was trained on a custom dataset from Roboflow at 15 epochs. Afterwards, the model was tuned with a

predetermined set of parameters such as learning rate, momentum, weight decay, etc. and the values of the best values of each predetermined set of parameters were displayed. Then, a pre-trained YOLOv8n was trained with the set of given parameters, giving a 91% validation mAP50. This lightweight model utilizes the lowest count of parameters, making it suitable for devices with low computing capabilities.

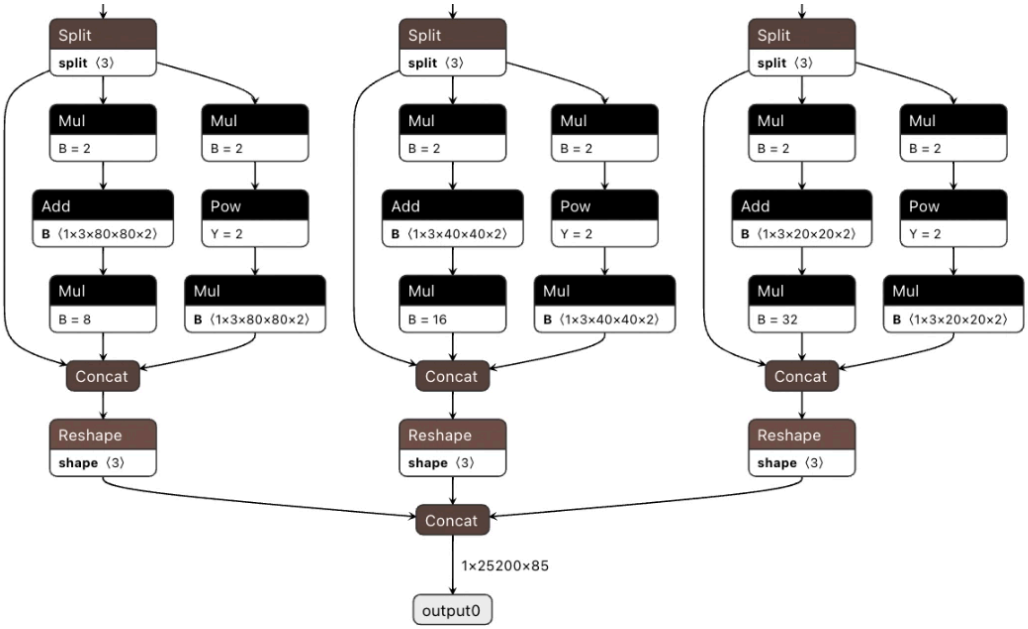


Figure 1. YOLOv8 Architecture [7]

B.2 **RT-DETR**: a pretrained RTDETR-L model was imported using ultralytics and was trained on the same dataset at 15 epochs. Afterwards, the model was tuned using the same predetermined parameters and was trained using the best value of the parameters, giving a 0.942 validation mAP50. This model utilizes a transformer-based architecture that has the capability of real-time detection at high accuracy [8].

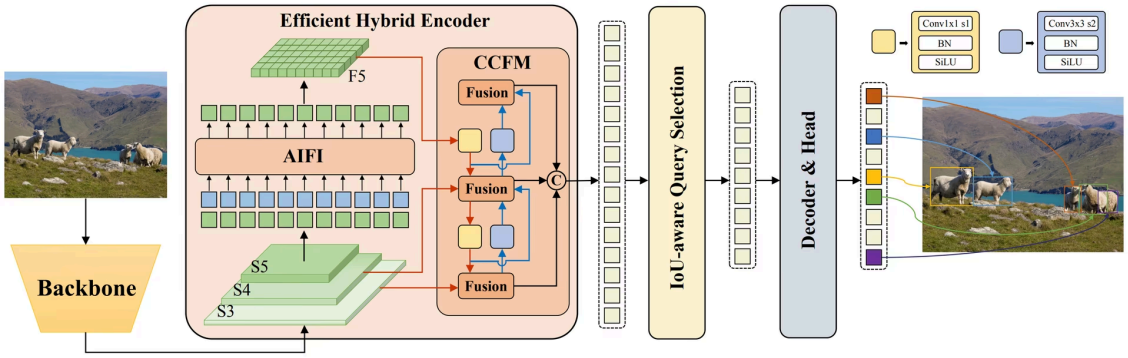


Figure 2. RTDETR Architecture [8]

B.3 **RetinaNET**: a pretrained RetinaNET model was imported from GitHub by cloning the repository. Afterwards, train.py, which came from the GitHub repo, was utilized to train the model on the custom dataset with ResNET50 as its backbone at 10 epochs. The model was evaluated using the evaluation.py that is located in the GitHub repository, giving a 0.891 mAP50 score. This model utilizes Focal Loss to better handle any existing class imbalance in the dataset [9], this results in providing an advantage towards PPE items that appears less frequently.

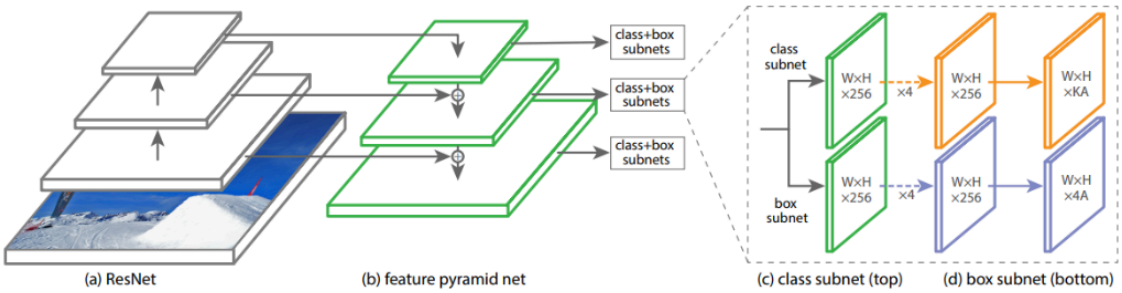


Figure 3. RTDETR Architecture [9]

C. CNN models

In the introductory stage of the project, a variety of advanced deep learning architectures for classifying compliance to personal protective equipment (PPE) were considered; however, the following four models were eventually eliminated due to their consistently poor performances and limitations in computation resources:

CoAtNet-0 (Transfer Learning): A hybrid convolution-transformer model that stagnated at approximately 50% accuracy. Despite transfer learning and fine-tuning, it failed to improve, largely due to GPU memory constraints and its high computational demand.

ResNet + EfficientNet Hybrid: This custom fusion model attempted to combine the residual learning capabilities of ResNet-50 with the efficiency of EfficientNet-B0. Even after applying advanced techniques such as gradual layer unfreezing and extensive data augmentation (rotation, flipping, zooming, and shifting), the model remained below 50% accuracy and showed no signs of convergence.

ConvNeXt-Small: A modern architecture that introduced improvements inspired by transformer models into a CNN framework. While it slightly outperformed the previous models, it peaked below 60% validation accuracy and required excessive training time, making it impractical for further development.

RegNetY (Transfer Learning): Known for its scalable and efficient structure, RegNetY showed early promise with accuracy reaching just under 60%. However, due to unstable validation performance and significant hardware resource demands, it was excluded from final evaluation.

Such limitations aimed towards lightweight and tailor-made CNN-based models, which conferred a better fit to our computing environment and were thus preferred for real-time analysis and deployment on embedded systems. Hence, three models were selected to be executed and evaluated: Custom CNN-LSTM, CNN-Bidirectional LSTM, and NASNetMobile (Transfer Learning).

To improve the operationality of these models, a sequence generator was incorporated into the preprocessing pipeline. This component transforms static input images into a structured sequence of spatial patches or feature vectors. The transference of the image into a temporal-like sequence provides models with LSTM and Bi-LSTM layers the ability to interpret

spatial dependencies across regions of the image, similarly to how time-series data is learned sequentially.

This technique is particularly good for detecting incomplete or subtle forms of PPE compliance amidst occlusions or convoluted arrangements in the images so as to know what is between items such as a helmet and a vest when either item is getting partially hidden. With the input from the sequence generator, the model can better extract and learn from these spatial patterns, which would not have been possible in the traditional case of using standard CNNs.

In addition to sequence formatting, the sequence generator also serves to augment the training set with data augmentation techniques to improve model generalization, among which include the following: Rotation (40 degrees), Horizontal flipping, Shifting (width and height), and Zooming. These augmentations smudge real camera variation and worker motion, thus ensuring robustness against orientation, position, and scale changes.

C.1 **The Custom CNN-LSTM** model combines convolutional layers for feature extraction with LSTM layers for temporal or sequential learning. The CNN layers process spatial features such as edges, shapes, and textures of PPE, while the LSTM layers capture dependencies between image regions that may indicate whether PPE is being worn correctly. This hybrid design allows the model to recognize subtle and spatially dependent patterns, which is especially useful in noisy or partially occluded images.

C.2 **CNN-Bidirectional LSTM,** This architecture extends the traditional CNN-LSTM model by using Bidirectional LSTM layers, which process data in both forward and backward directions. This allows the model to learn context from both preceding and succeeding regions of an image, resulting in a richer representation of PPE compliance. It is particularly effective in scenes with complex arrangements or occluded gear, as it analyzes visual dependencies from multiple directions.

C.3 **NASNetMobile** is a lightweight, high-performing convolutional neural network developed through neural architecture search (NAS). In this study, NASNetMobile was employed using transfer learning, where the model was pretrained on ImageNet and fine-tuned using our PPE compliance dataset. This approach provides strong generalization and requires fewer training epochs. Its compact architecture makes it ideal for deployment on resource-constrained devices such as surveillance cameras or embedded edge devices.

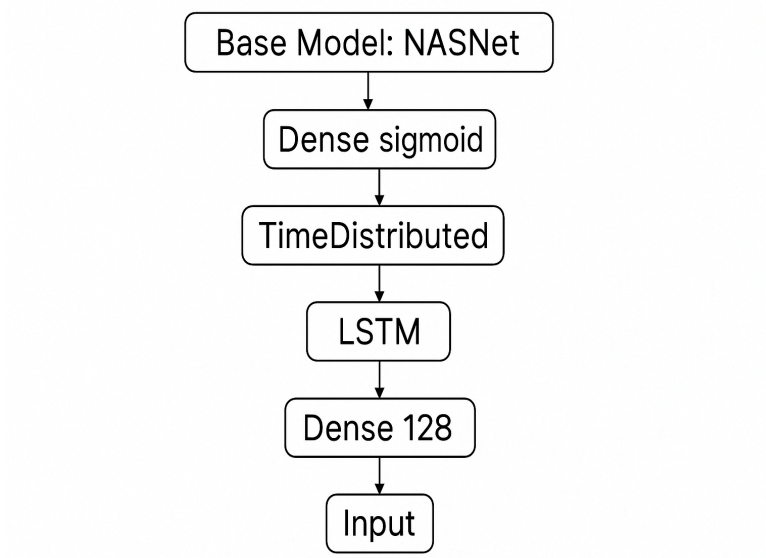


Figure 4. NASNetMobile Transfer Learning Model Architecture

D. Model Deployment

The model that achieved the highest performance metrics would be deployed through the use of Streamlit, an open-source application that enables data scientists to build simple interactive data applications.

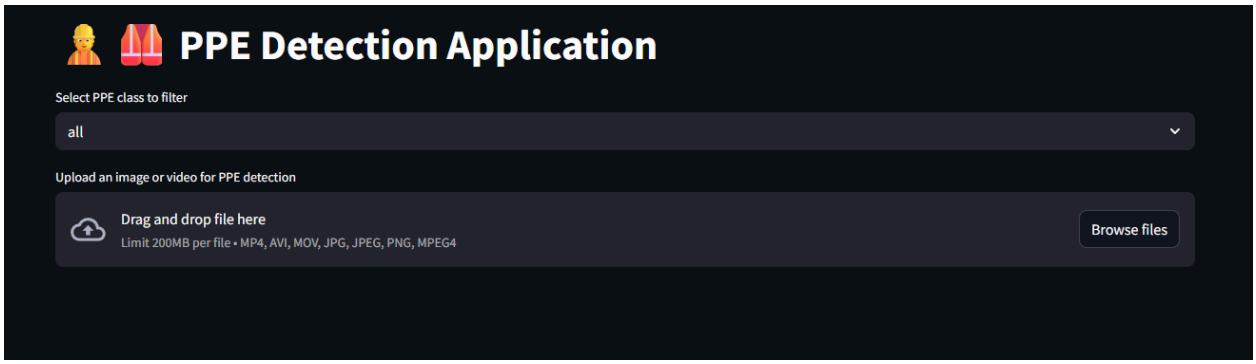


Figure 5. Interface of the Streamlit Application (Object Detection)

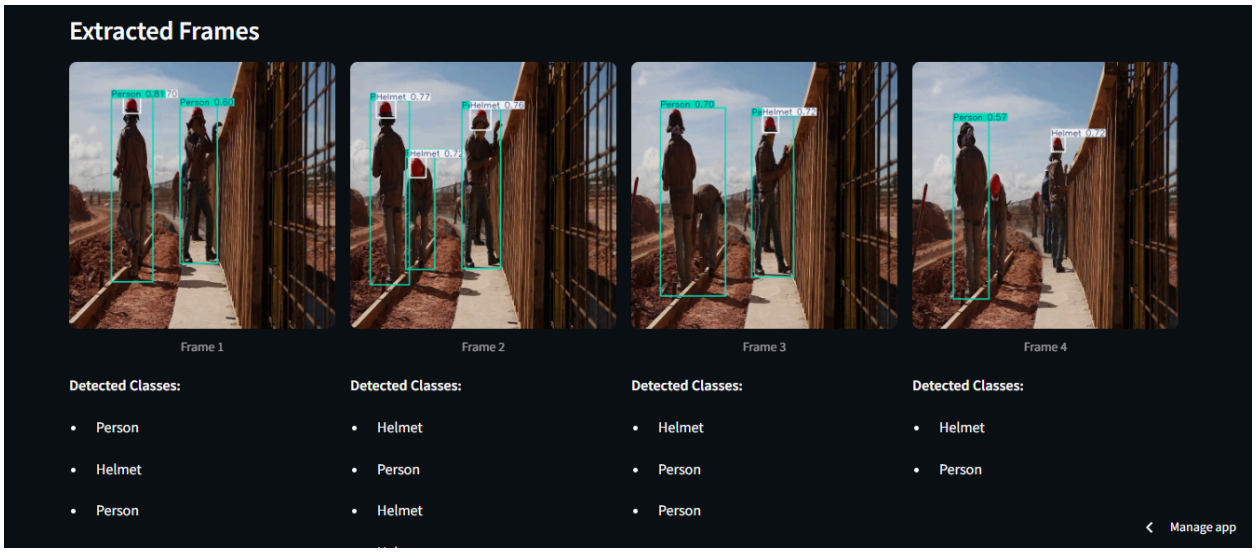


Figure 6. Output of the Streamlit Application (Object Detection)

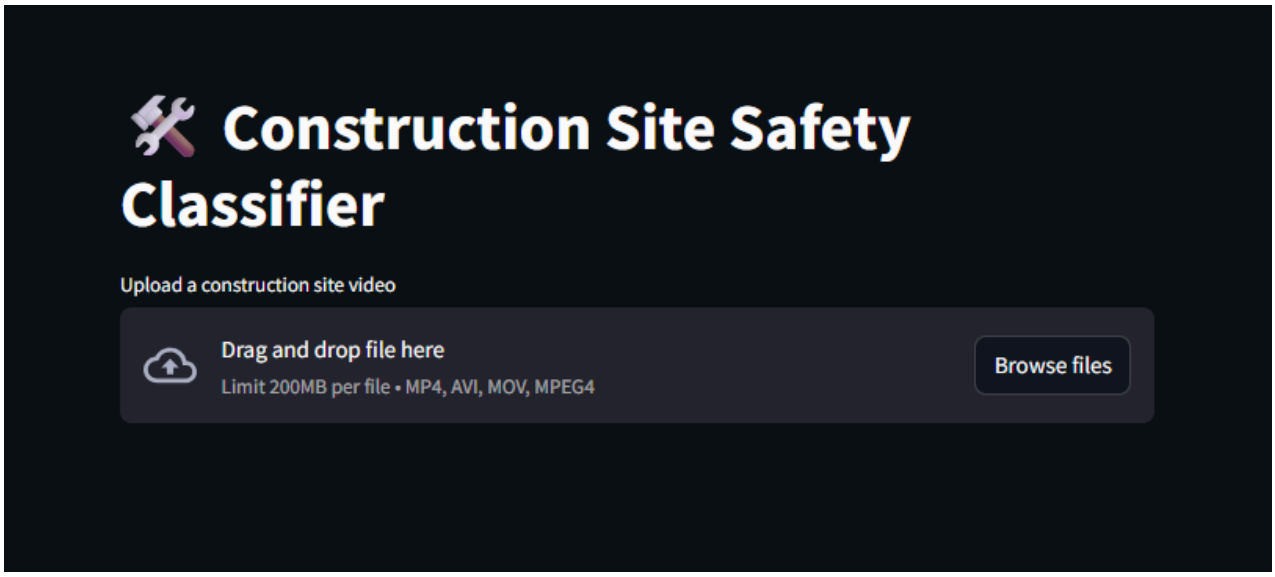


Figure 7. Interface of the Streamlit Application (CNN)

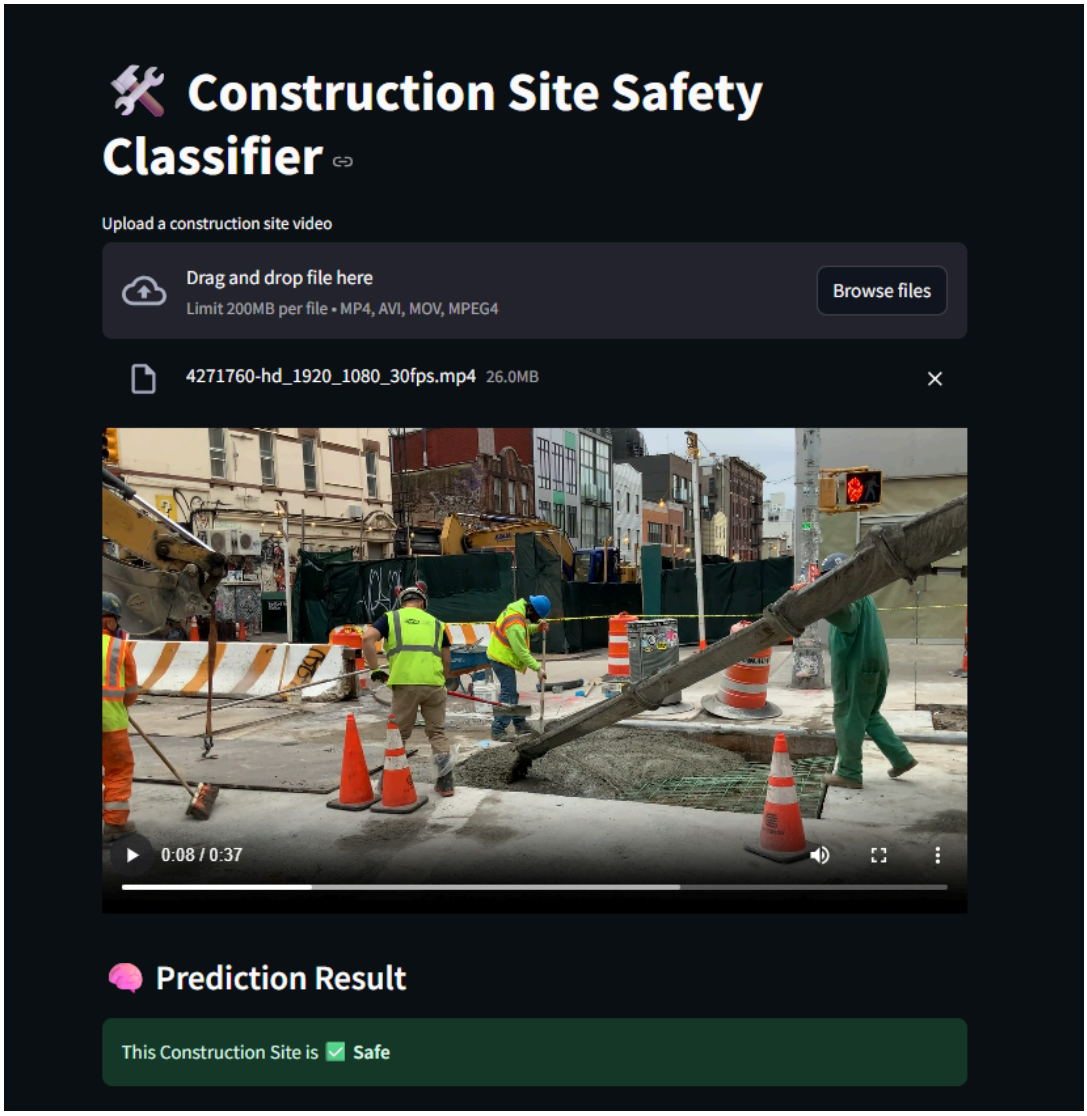


Figure 8. Interface of the Streamlit Application (CNN)

IV. Results and Discussion

TABLE I. Performance of the Three Object Detection Models

Validation Metric	Models		
	YOLOv8n	RTDETR-L	RetinaNET
mAP50	0.917	0.942	0.891
mAP50-95	0.712	0.734	0.575

The three object detection models were compared using validation mAP50 and mAP50-95 scores. RTDETR-L achieved the highest mAP50 of 94.2% and mAP50-95 of 73.4%. This suggests that the RTDETR-L is overall better at pinpointing each PPE item class.

TABLE II. Performance of the Three CNN Models

Validation Metric	Models		
	Custom CNN-LSTM	CNN-Bidirectional LSTM	NASNetMobile
Train Accuracy	94.37%	97.08%	96.33%
Validation Accuracy	92.05%	85.17%	83.53%

Three classification models were evaluated using training accuracy, validation accuracy, and loss. The Custom CNN-LSTM model performed best with a 92% validation accuracy, effectively capturing both spatial and sequential features. The CNN-Bidirectional LSTM followed at 85%, leveraging its ability to model forward and backward dependencies. The NASNetMobile-based transfer learning model achieved 83%, demonstrating strong feature extraction despite a limited dataset. Overall, the Custom CNN-LSTM proved most suitable for frame-level PPE compliance classification.

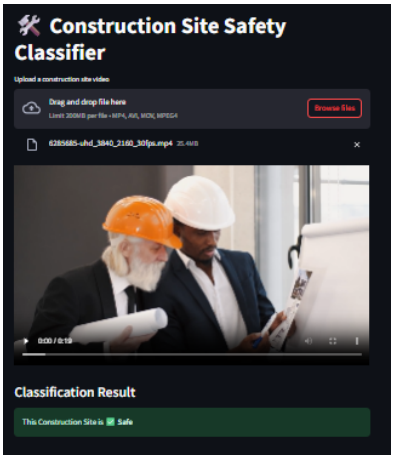
All three classification models were trained for 10 epochs with the same training/validation splits, preprocessing steps, and batch sizes. Evaluation metrics included training accuracy, validation accuracy, and validation loss. The Custom CNN-LSTM model exhibited the highest training and validation accuracy, reaching approximately 92%, with a relatively low validation loss. This suggests strong generalization from the training set to unseen data. The CNN-Bidirectional LSTM followed with a validation accuracy of 85%, indicating effective spatial-temporal learning but slightly less generalization capacity compared to the unidirectional CNN-LSTM. Lastly, the NASNetMobile-based transfer learning model achieved a validation accuracy of 83%, demonstrating solid performance and efficiency, especially given its lightweight architecture and reliance on pretrained weights.

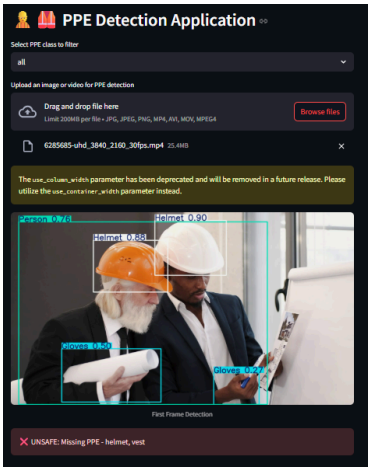
Throughout the training process, the Custom CNN-LSTM model showed consistent improvement in both training and validation accuracy, with loss steadily decreasing, reaching optimal performance around the eighth epoch. The CNN-Bidirectional LSTM also exhibited good learning behavior, although its accuracy plateaued slightly earlier than the unidirectional variant. NASNetMobile performed reliably with moderate improvements per epoch, though its accuracy gains were slightly lower—likely due to the fixed feature representations learned from pretraining on ImageNet, which might not fully adapt to the specific PPE dataset.

In terms of object detection, all models were trained in almost the exact same way with training/validation, method of loading the model, default batch size, and the same hyperparameters for tuning. The evaluation metrics of the model included mAP50, precision, and recall. First, RT-DETR was the highest-performing model, achieving (0.942) in mAP50, (0.918) in precision, and (0.883) in recall. This suggests that the model performs well in distinguishing each specified PPE on unforeseen data. Second, YOLOv8n, albeit lagging behind by a few points in terms of performance, still achieved an (0.917) in mAP, (0.898) in precision, and (0.835) in recall. This suggests that the model is still performing quite well in distinguishing each specified PPE on unforeseen data, even though lagging behind, which is expected since YOLOv8n is a lightweight model that is used for devices that have low processing power.

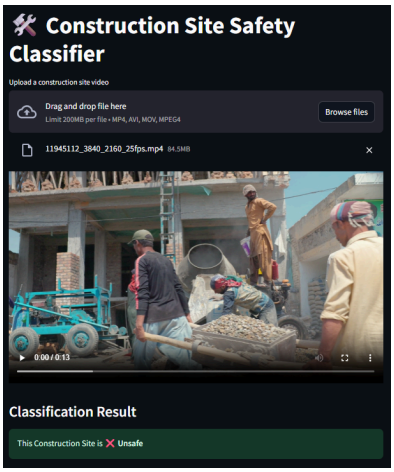
TABLE III. Test Cases on Comparison between two models

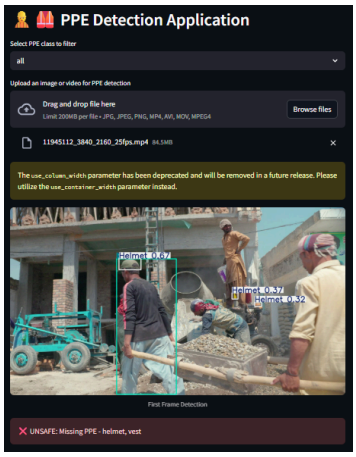
CNN-LSTM	Object Detection	Result
		Both Accurately Predicted as Unsafe
		Both Accurately Predicted as Unsafe



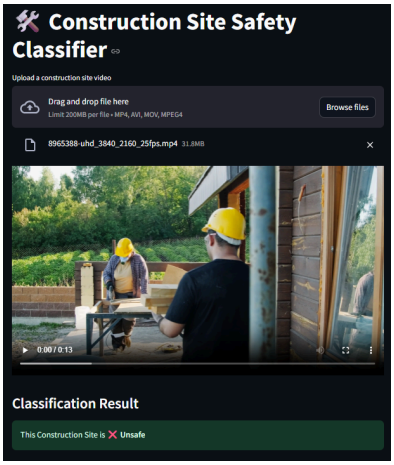


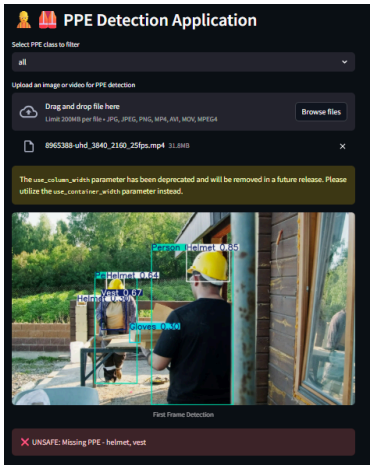
RTDETR Object Detection Accurately Predicted Unsafe while CNN-LSTM Misclassified Safe



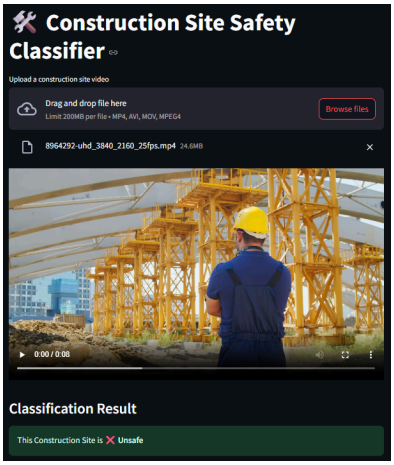


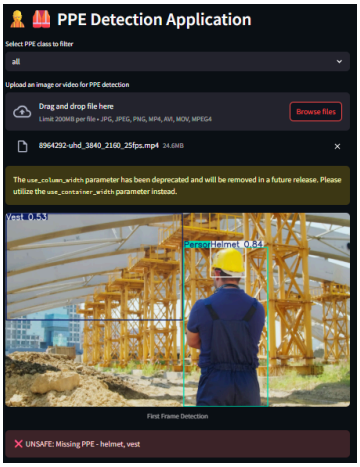
Both Accurately Predicted as Unsafe



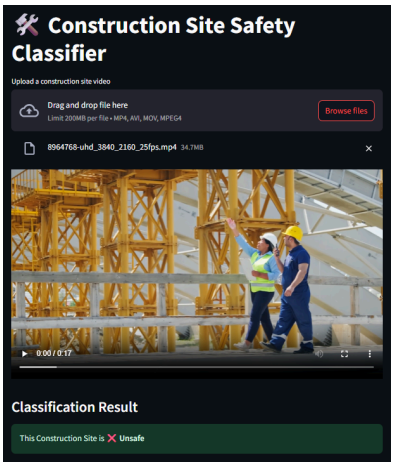


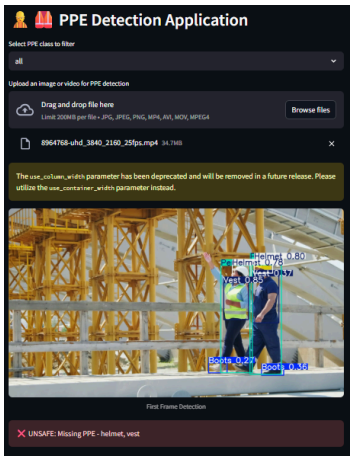
Both Accurately Predicted as Unsafe



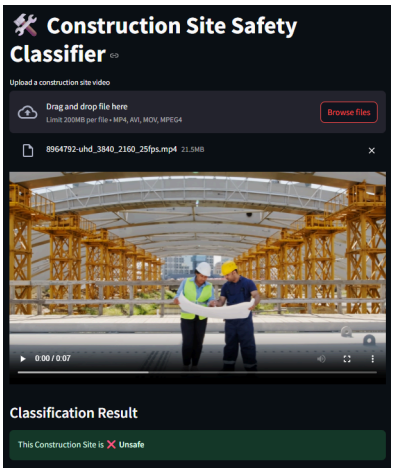


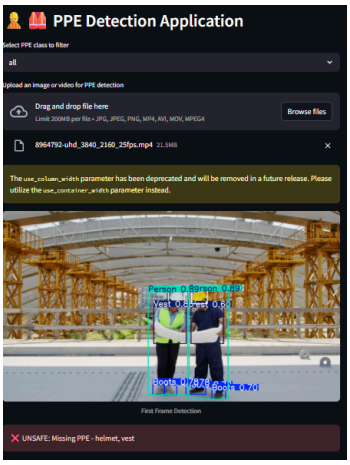
Both Accurately Predicted as Unsafe



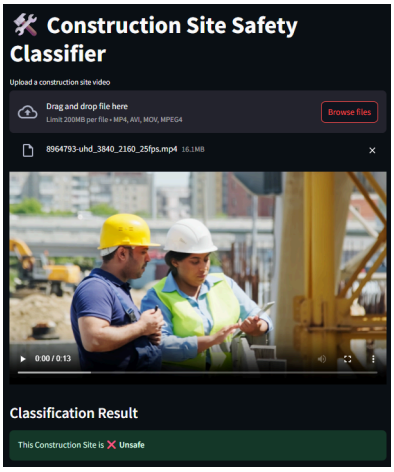


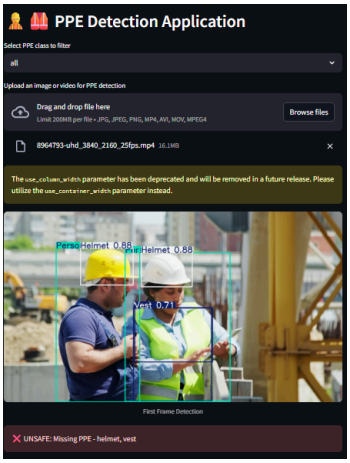
Both Accurately Predicted as Unsafe



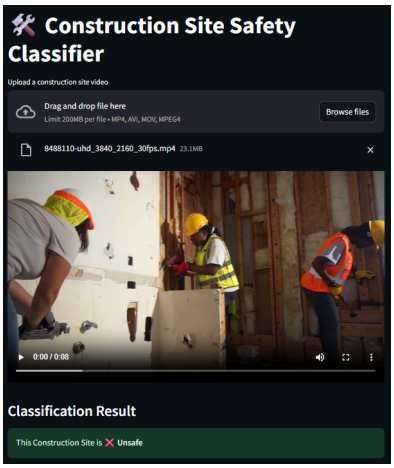


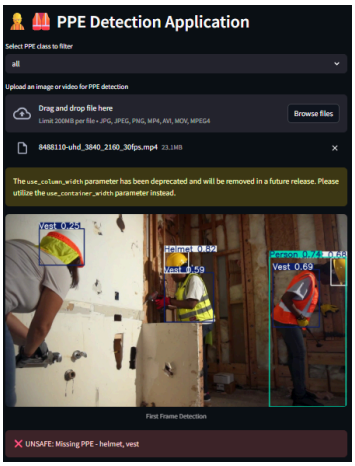
Both Accurately Predicted as Unsafe



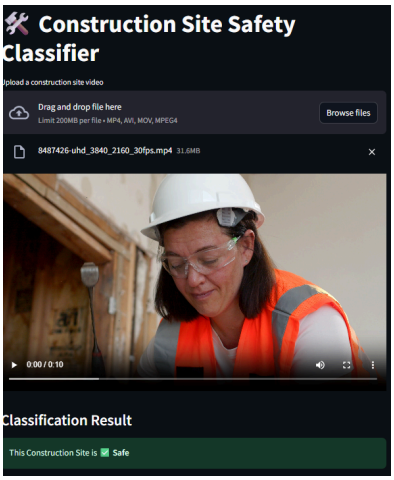


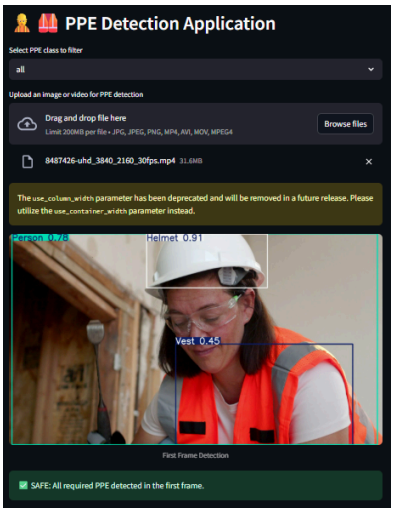
Both Accurately Predicted as Unsafe



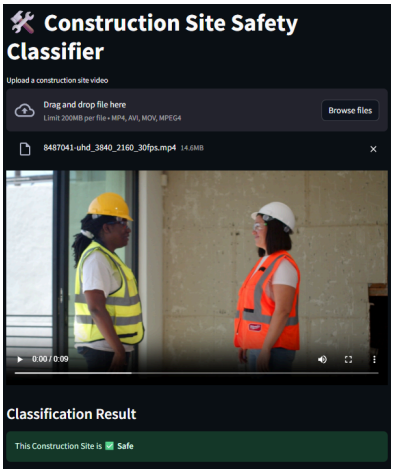


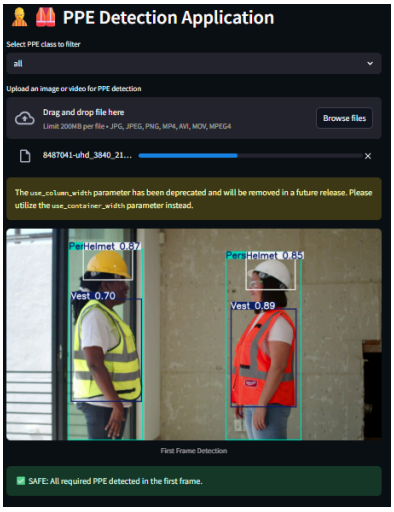
Both Accurately Predicted as Safe



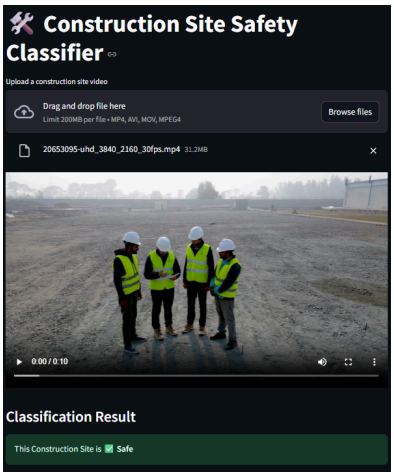


Both Accurately Predicted as Safe

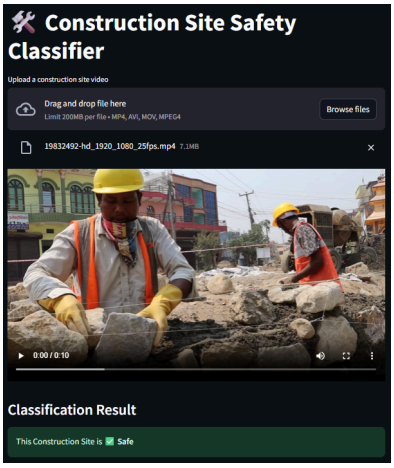


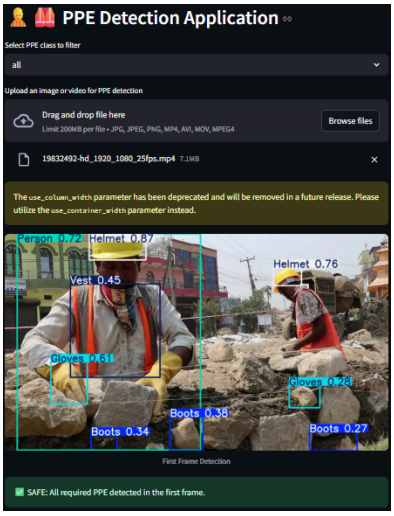


Both Accurately Predicted as Safe



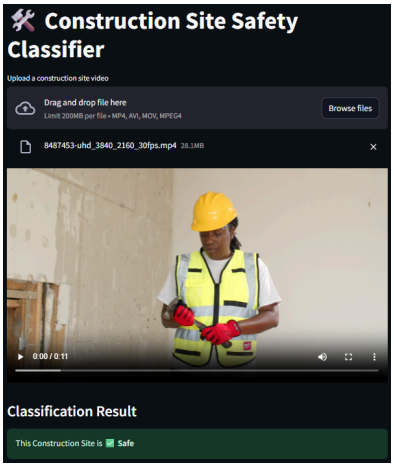
Both Accurately Predicted as Safe



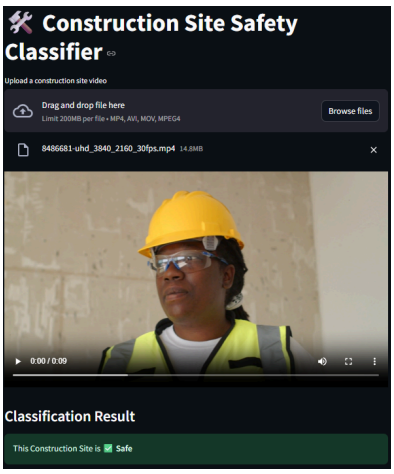


Both Accurately Predicted as Safe

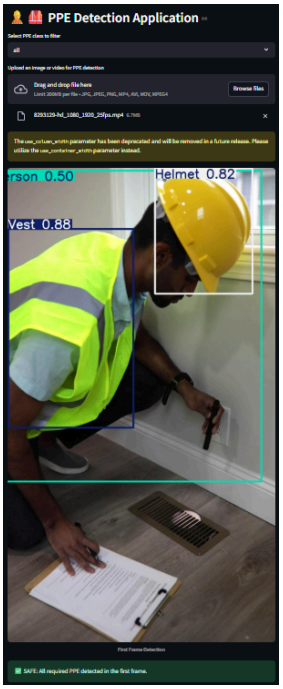
Both Accurately Predicted as Safe



Both Accurately Predicted as Safe

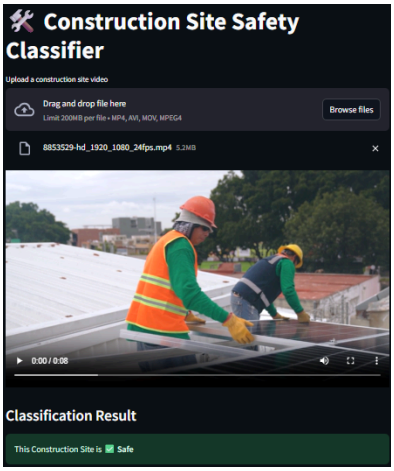
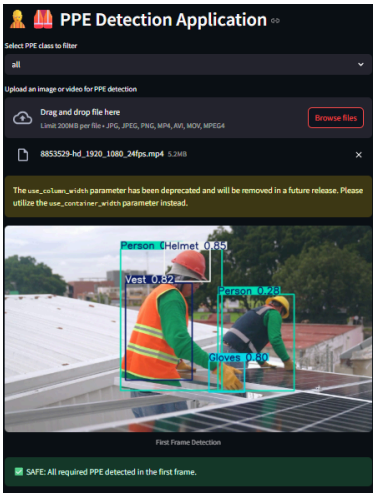


Both Accurately Predicted as Safe



Both Accurately Predicted as Safe

Both Accurately Predicted as Safe

		Both Accurately Predicted as Safe

After testing both CNN-LSTM classification model which results in 19/20 or 95% and RTDETR object detection model which results in 20/20 or 100% on the same 20 videos, It is indicated that RTDETR performed better than the classification in indicating a safe or unsafe construction environment through PPE detection of workers. Another metric is efficiency of the model which the RTDETR won as it is more efficient to use with medium size and yet only small resources expanded while in CNN-LSTM it do have small size but it expands medium to large resources, and the speed which the CNN-LSTM won as in the testing we did the CNN-LSTM do perform faster than the RTDETR. This suggests that RTDETR object detection is the best method for identifying PPE violations in the construction site workplace.

V. Conclusion and Recommendations

This paper showcased the different implementations of deep learning methods for an automated Personal Protective Equipment (PPE) compliance monitoring system. A CNN-LSTM architecture served as a robust video classification framework by leveraging spatial and temporal features extracted from sequences of frames. It evaluated overall site safety based on whether all personnel wore complete PPE attire, achieving the highest validation accuracy of 92%, followed by CNN-Bidirectional LSTM at 85%, and NASNetMobile-based transfer learning at 83%. This approach is well-suited for scenarios where a high-level assessment of site safety is prioritized over individual detection.

Additionally, object detection models such as YOLOv8, RT-DETR, and RetinaNet were trained to detect specific PPE worn by each worker for fine-grained compliance assessment. Among these, the RTDETR-L model demonstrated the best performance, achieving a mAP50 of 94.2% and a mAP50-95 of 73.4%, indicating its strong potential for real-time, instance-level detection. The researchers recommend expanding the annotated dataset to improve detection in cases of occlusion or poor visibility due to environmental conditions.

After 20 test cases have been performed. It can be concluded that Object Detection, specifically, RT-DETR object detection model yields the best performance in classifying a site as either safe or unsafe. It does not only classify safety evaluation but also provides the PPE items worn by a personnel in a construction place.

VI. Acknowledgements

We would like to give our sincere gratitude to Engr. Roman M. Richard for providing us with his expertise and guidance throughout the successful completion of this project.

VII. References

[1] P. K. Vishwakarma and M. R. Suyambu, "Smart Personal Protective Equipment (PPE) Systems: Integrating Deep Learning to Improve Workplace Safety," 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2025, pp. 1-6

[2] O. P. Giri, "Factors Causing Health and Safety Hazards at Construction Sites," Technical Journal, vol. 2, no. 1, pp. 68–74, Nov. 2020, doi: https://doi.org/10.3126/tj.v2i1.32841.

[3] M. I. B. Ahmed et al., "Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach", Sustain., 2023.

[4] G. Gallo, F. D. Rienzo, F. Garzelli, P. Ducange and C. Vallati, "A Smart System for Personal Protective Equipment Detection in Industrial Environments Based on Deep Learning at the Edge," in IEEE Access, vol. 10, pp. 110862-110878, 2022, doi: 10.1109/ACCESS.2022.3215148.

[5] H. Wang, J. Ma, W. Chen, Q. Han, and Z. Yao, "Personal Protective Equipment Detection for Industrial Environments: A Lightweight Model Based on RTDETR for Small Targets," Oct. 29, 2024.

[6] H. WANG, "Detection of Personal Protective Equipment (PPE) using an Anchor Free-Convolutional Neural Network," International journal of advanced computer science and applications/International journal of advanced computer science & applications, vol. 15, no. 2, Jan. 2024, doi: https://doi.org/10.14569/ijacsa.2024.0150239.

[7] J. Torres, "YOLOv8 Architecture: A Deep Dive into its Architecture - YOLOv8," Yolov8, Jan. 15, 2024. https://yolov8.org/yolov8-architecture/

[8] Ultralytics, "RT-DETR (Realtime Detection Transformer)," Ultralytics.com, 2023. https://docs.ultralytics.com/models/rtdetr/#overview

[9] "How RetinaNet works? | ArcGIS API for Python," *Arcgis.com*, 2016. https://developers.arcgis.com/python/latest/guide/how-retinanet-works/

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[11] M. Shao, Y. Liu, and D. Wang, "A Deep Convolutional Neural Network for PPE Detection and Compliance in Construction Environments," Automation in Construction, vol. 132, pp. 103922, 2021, doi: 10.1016/j.autcon.2021.103922.

[12] R. Silva-Rojas, F. Bolea, and J. J. Guerrero, "Real-Time Detection of Helmets on Construction Sites Using CNNs," IEEE Sensors Journal, vol. 21, no. 18, pp. 20659–20667, 2021, doi: 10.1109/JSEN.2021.3083835.

[13] A. Albahli and A. A. Y. Alwan, "CNN-Based Detection of PPE for Worker Safety in Industrial Environments," Computers, Materials & Continua, vol. 66, no. 2, pp. 1601–1615, 2021, doi: 10.32604/cmc.2021.013735.

[14] Y. Jiang, Z. Wang, and Y. Huang, "Helmet and Vest Detection for Construction Site Safety Monitoring Using Improved CNN," IEEE Access, vol. 8, pp. 145274–145283, 2020, doi: 10.1109/ACCESS.2020.3014843.

[15] T. P. Nguyen et al., "Automatic Detection of Construction Workers and Safety Equipment Using Deep CNN and Transfer Learning," Journal of Building Engineering, vol. 43, 102564, 2021, doi: 10.1016/j.jobe.2021.102564.

Evaluation Metrics

Student Outcome 7							
Criteria	Ratings						Pts
T.I.P. SO 7.1 Acquire and apply new knowledge from outside sources threshold: 4.2 pts	6 pts [Excellent] Educational interests and pursuits exist and flourish outside classroom requirements,knowledge and/or experiences are pursued independently and applies knowledge learned into practice	5 pts [Good] Educational interests and pursuits exist and flourish outside classroom requirements,knowledge and/or experiences are pursued independently	4 pts [Satisfactory] Look beyond classroom requirements, showing interest in pursuing knowledge independently	3 pts [Unsatisfactory] Begins to look beyond classroom requirements, showing interest in pursuing knowledge independently	2 pts [Poor] Relies on classroom instruction only	1 pts [Very Poor] No initiative or interest in acquiring new knowledge	6 pts
T.I.P. SO 7.2 Learn independently threshold: 4.2 pts	6 pts [Excellent] Completes an assigned task independently and practices continuous improvement	5 pts [Good] Completes an assigned task without supervision or guidance	4 pts [Satisfactory] Requires minimal guidance to complete an assigned task	3 pts [Unsatisfactory] Requires detailed or step-by-step instructions to complete a task	2 pts [Poor] Shows little interest to complete a task independently	1 pts [Very Poor] No interest to complete a task independently	6 pts
T.I.P. SO 7.3 Critical thinking in the broadest context of technological change threshold: 4.2 pts	6 pts [Excellent] Synthesizes and integrates information from a variety of sources; formulates a clear and precise perspective; draws appropriate conclusions	5 pts [Good] Evaluate information from a variety of sources; formulates a clear and precise perspective.	4 pts [Satisfactory] Analyze information from a variety of sources; formulates a clear and precise perspective.	3 pts [Unsatisfactory] Apply the gathered information to formulate the problem	2 pts [Poor] Gather and summarized the information from a variety of sources but failed to formulate the problem	1 pts [Very Poor] Gather information from a variety of sources	6 pts
T.I.P. SO 7.4 Creativity and adaptability to new and emerging technologies threshold: 4.2 pts	6 pts [Excellent] Ideas are combined in original and creative ways in line with the new and emerging technology trends to solve a problem or address an issue.	5 pts [Good] Ideas are creative and adapt the new knowledge to solve a problem or address an issue	4 pts [Satisfactory] Ideas are creative in solving a problem, or address an issue	3 pts [Unsatisfactory] Shows some creative ways to solve the problem	2 pts [Poor] Shows initiative and attempt to develop creative ideas to solve the problem	1 pts [Very Poor] Ideas are copied or restated from the sources consulted	6 pts
							Total Points: 24

Evaluated by:

_____

Engr. Roman M. Richard

Course Instructor