Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19276-8

1236: EXPLAINABLE ARTIFICIAL INTELLIGENCE SOLUTIONS FOR IN-THE-WILD HUMAN BEHAVIOR ANALYSIS

Unsafe-Net: YOLO v4 and ConvLSTM based computer vision system for real-time detection of unsafe behaviours in workplace

Oğuzhan Önal[1] · **Emre Dandıl**[2]

Received: 20 October 2023 / Revised: 9 January 2024 / Accepted: 18 April 2024

Abstract
Unsafe behaviour is a leading cause of death or injury in the workplace, including many accidents. Despite regular safety inspections in workplaces, many accidents occur as a result of breaches of occupational health and safety protocols. In these environments, despite efforts to prevent accidents and losses in hazardous environments, human error cannot be completely eliminated. In particular, in computer-based solutions, automated behaviour detection has low accuracy, is very costly, not real-time and requires a lot of time. In this study, we propose Unsafe-Net, a hybrid computer vision approach using deep learning models for real-time classification of unsafe behaviours in workplace. For the Unsafe-Net, a dataset is first specifically created by capturing 39 days of video footage from a factory. Using this dataset, YOLO v4 and ConvLSTM methods are combined for object detection and video understanding to achieve fast and accurate results. In the experimental studies, the classification accuracy of unsafe behaviours using the proposed Unsafe-Net method is 95.81% and the average time for action recognition from videos is 0.14 s. In addition, the Unsafe-Net has increased the real-time detection speed by reducing the average video duration to 1.87 s. In addition, the system is installed in a real-time working environment in the factory and employees are immediately alerted by the system, both audibly and visually, when unsafe behaviour occurs. As a result of the installation of the system in the factory environment, it has been determined that the recurrence rate of unsafe behaviour has been reduced by approximately 75%.

Keywords Occupational health and safety · Unsafe behaviour detection · Deep learning · Computer vision · YOLO v4 · ConvLSTM

✉ Emre Dandıl
emre.dandil@bilecik.edu.tr

[1] Department of Electronic and Automation, Vocational School, Bilecik Seyh Edebali University, Bilecik, Türkiye

[2] Department of Computer Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, Bilecik, Türkiye

Published online: 04 May 2024

1 Introduction

Unsafe behaviours are mainly operational errors such as violations of safety work permits and safety procedures [1]. In the world, as countries prioritise production and competition for economic development, various occupational health and safety risks arise from unsafe behaviours in the workplace. These risks mostly occur as occupational accidents. It is estimated that there are 2.3 million cases of occupational accidents and diseases and 0.3 million deaths each year worldwide [2]. The economic costs of work-related injuries and illnesses vary from 1.8 to 6.0% of GDP in country estimates, with an average of 4% according to the International Labour Organization (ILO) [3, 4].

In workplaces and work environments, ensuring that people work in a safe environment is a very challenging task due to dynamic and complex working conditions. In work environments, unsafe behaviours can occur when safety rules, instructions and standards are not followed [5]. Such actions can put employees at risk and lead to loss of performance in the workplace [6]. Despite the continuous updating of regulations and the increase in preventive measures, the occurrence of accidents, injuries and even fatalities in the workplace is a major problem [7]. Therefore, in order to improve safety performance in work environments, it is very important that unsafe behaviours can be reduced or prevented.

In recent years, advanced mechanisms have been introduced to prevent occupational hazards. In particular, the rules to be followed by employees in the industry are revised according to changing conditions. Companies organise various training sessions and advanced preventive measures are taken to ensure that employees fully comply with safety and health procedures. Traditionally, observation-based methods are preferred to identify and monitor unsafe behaviours of employees in the workplace. However, these methods have some limitations in terms of time-consuming, individualised and labour-intensive activities. Therefore, it is very important to identify unsafe behaviours in the workplace using automated systems.

It is very important to use personal protective equipment and to comply with occupational health and safety rules in the working environment of companies and production facilities. In addition, it is also necessary to pay close attention to behaviours that have been identified as hazardous in the facilities [8]. In terms of occupational health and safety in the workplace, activity recognition from video surveillance has recently been widely used to identify and classify dangerous behaviour. Human activity recognition from video surveillance systems is a current area of research that offers some of the computer vision applications such as content-based video analysis, network-based surveillance, user interface and monitoring [9, 10]. Early work in action recognition and video content understanding focused on the detection of simple human gestures such as hand waving and walking [11]. More recently, however, research has focused on more realistic and challenging problems involving complex activities with object interactions and multiple activities [12].

One of the most important aspects in making sense of surveillance video is the ability to recognise objects in complex work environments and to understand unsafe behaviour in this context. Video understanding is very useful for many different applications such as artificial intelligence, surveillance, video search and security. To perform video understanding, it is necessary to create a model that can take video images as input and recognise any actions performed by people in that video. These models are typically built using computer vision. Computer vision is widely used for object recognition, tracking, anomaly detection, activity detection and video interpretation and is based on deep learning.

It can be seen that the previous studies proposed for the detection of unsafe behaviours are mainly focused on construction sites. Most of the first studies proposed for the detection of unsafe behaviours and actions are statistically based and aim at training workers through video footage of accidents, rather than real-time approaches [13]. In addition, wireless sensor networks and various equipment are also used to detect unsafe behaviours in work environments [14]. Barro-Torres et al. [8] proposed a system using Zigbee and RFID to monitor the use of personal protective equipment in real time. In another study, Yu et al. [15] developed a real-time and image skeleton based method to detect the movement behaviour of construction workers. The method used Kinect and infrared sensors to capture the workers' movements. However, there are limitations to similar systems, such as custom manufacturing and pre-processing of the equipment.

There have also been studies using machine learning methods to detect unsafe behaviour in the workplace. Alwasel et al. [16] presented a method for determining the poses of safe and productive masons using machine learning in the study, stable measurement units and video cameras were used to collect kinematic data from masons and pose sets were determined. In another study, Wu and Zhao [17] focused only on detecting whether workers were wearing helmets or not and the colour of the helmets. In their study, a hierarchical support vector machine was created for classification and some accuracy was achieved.

Despite traditional precautions, accidents, injuries and even deaths cannot be prevented due to unsafe behaviour in the workplace. Workplaces are closed environments that require specific applications with more objects. Predicting hazardous behaviours in industrial facilities, workplaces and production sites can provide remedial measures to prevent and reduce accidents that may occur. Therefore, computer vision technologies can be used to provide solutions in such areas. The use of computer vision techniques in field observations is considered to be an effective automated tool for extracting safety-related information from images and videos, complementing existing manual applications [18]. To improve safety and productivity at work sites, computer vision-based inspections play an important role in detecting the presence of objects such as workers, facilities, equipment and materials. The most popular methods for motion detection from video often use deep learning models. In deep learning approaches, the entire process is automated, except for the generation of a labelled dataset. For example, in one of these studies, Wei et al. [19] presented a computer vision-based approach using deep learning to automatically identify a person performing unsafe behaviour from video. The study also applied real-time safety management in construction environments.

Recently, very successful results have been achieved with the use of deep learning methods in studies on object detection, object classification and video understanding [20]. Many approaches have been proposed in previous studies to detect unsafe behaviours in work environments, but after the widespread use of deep learning, the detection accuracy and speed of the methods have increased. Fang et al. [21] proposed a method based on deep learning to automatically inspect the personal protective equipment of workers in high structures, such as towers and chimneys, for fall hazards. Ding et al. [6] presented a hybrid deep learning approach using computer vision and pattern recognition approaches to detect unsafe behaviour. The study developed a model using convolutional neural networks (CNN) and long short-term memory (LTSM). Wu et al. [22] used the single shot multibox detector (SSD) based deep learning algorithm to determine whether the use of different coloured hard hats by construction workers is appropriate. Chen and Demachi [23] proposed a solution to identify inappropriate use of protective equipment by construction workers using deep learning based object detection and geometric relationship analysis. Kong et al. [24] proposed an LSTM-based computer

vision approach to automatically predict unsafe behaviour on construction sites from videos. In another study, Liu et al. [25] used deep learning based methods to identify different types of unsafe behaviours from digital images. The study followed sequential steps such as visual and textual feature extraction, recursive sub-query and bounding box generation. Fang et al. [26] proposed a content-based image retrieval approach to identify unsafe movements in construction sites. For automatic detection of workers and equipment at construction sites, Fang et al. [27], proposed a faster R-CNN based deep learning model. A large object database is also created in the study and the results are compared with state-of-the-art methods. Son et al. [28] presented a model based on deep residual networks and faster R-CNN deep learning architectures for the detection of construction workers in variable poses and backgrounds. Khan et al. [29] developed a Mask R-CNN and correlation based approach for mobile scaffold safety monitoring and detection of worker's unsafe behaviour. Yang et al. [30] introduced a transformer based deep learning model for unsafe action detection from video data in workplace. In the study, unsafe actions of construction workers were detected using feature encoders and feature fusion models.

Unlike other environments, workplaces such as factories are in constant motion, with complex processes taking place. Traditionally, a number of safety plans, rules and measures have been put in place to ensure workplace safety. In addition, in some cases, sites are continuously monitored while work is in progress to identify any unsafe behaviour that may occur. Moreover, even today, many researchers have proposed many approaches to detect, monitor and prevent unsafe behaviours in workplaces, supported by regular databases and relevant safety rules that depend on specific regulations. In particular, in recent years, computer vision and deep learning methods have been used to detect unsafe behaviour by monitoring workers using action recognition from workplace video. However, there are still research gaps in this area. For example, it is clear that computer vision-based real-time recognition is limited in existing approaches to detecting unsafe behaviour. In addition, many proposed approaches are known to increase costs and are not very successful in reducing detection time. Furthermore, for action recognition, video understanding and object detection for unsafe behaviour detection in the field, the original datasets in many studies are limited and the training models are generally weak. Furthermore, in most of the previously proposed studies, the number of classes in the video datasets prepared for unsafe behaviour detection is quite small. Therefore, there is still a need for computer vision-based methods that can detect unsafe behaviours with high performance in real time, using real data in real workplaces.

In the Unsafe-Net proposed in this study, unsafe behaviours are classified in real time using video data collected from the work environments of industrial production facilities. In this study, object detection and video understanding are provided by a hybrid computer vision approach using YOLO v4 and ConvLSTM deep learning models. In the experimental studies, videos from cameras installed in factories are instantly inspected by the proposed Unsafe-Net method and the video scene is made meaningful. In this way, unsafe behaviours in the workplace are detected and pre-processes for expert intervention are created. The study is also implemented a real-time pilot application for use in factories, alerting workers visually and audibly when unsafe behaviour occurs. The contributions of this study are outlined as below:

1. We propose Unsafe-Net, a hybrid computer vision approach using deep learning models for real-time classification of unsafe behaviours in workplace.

2. For the Unsafe-Net, a dataset is first specifically created by capturing 39 days of video footage from a factory.
3. Using this dataset, YOLO v4 and ConvLSTM methods are combined for object detection and video understanding to achieve fast and accurate results.
4. The classification accuracy of unsafe behaviours using the proposed Unsafe-Net method is 95.81% and the average time for action recognition from videos is 0.14 s.
5. The Unsafe-Net has increased the real-time detection speed by reducing the average video duration to 1.87 s.
6. The system is installed in a real-time working environment in the factory and employees are immediately alerted by the system, both audibly and visually, when unsafe behaviour occurred.
7. In a pilot study carried out in the factory, it has been determined that the recurrence rate of unsafe behaviour has been reduced by approximately 75%.

The remaining sections of the paper are structured as follows. In the Section 2, the dataset used in the study is described and the proposed methodology is detailed. In this Section, the architectures of the YOLO v4 and ConvLSTM deep learning frameworks are also presented. In addition, Section 2 describes the infrastructure of the warning and monitoring system. Experimental studies and results are analyzed in the Section 3 and discussion supported with qualitative and quantitative evaluations in Section 4. Lastly, conclusions and implications are presented in the Section 5.

2 Material and method

In this study, Unsafe-Net, a real-time computer vision system based on deep learning, is proposed to classify unsafe behaviours in factory work environments from collected video data. The study aims to prevent undesirable situations that may occur due to unsafe behaviours for complete control in a production facility. The block diagram for the working mechanism of the proposed system is shown in Fig. 1. The study started by creating a database of videos containing safe and unsafe behaviours from different work environments with moving machines and work traffic in the cooperating factory. In the study, YOLO v4 was used to simplify the video by detecting the objects containing unsafe behaviours, selecting the flowing images and excluding all other frames. Using YOLO v4, frames with unsafe behaviour are selected from the video and then a video segment is created for objects belonging to unsafe action classes within these frames. By ensuring that there is

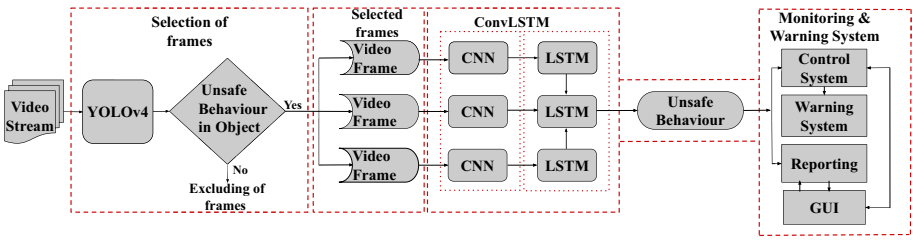

Fig. 1 Block diagram and mechanism of the proposed Unsafe-Net hybrid deep learning system based on YOLO v4 and ConvLSTM for real-time detection of unsafe behavior from videos

only one action in the same video, the accuracy of the classification with ConvLSTM is increased and the action detection time is reduced. In the Unsafe-Net architecture, a CNN is created to extract spatial features at a given time step for video understanding, followed by a Long Short Term Memory (LSTM) network to identify temporal relationships between frames. In addition, a pilot study was conducted at the work site, where live video frames from two different IP cameras were applied to the pre-trained YOLO v4 + ConvLSTM network, and when unsafe behaviours belonging to four different classes were detected, the electronic warning system designed in the study was activated and the violators were warned with audible and visual warnings each time. Finally, the effectiveness of the study was evaluated by analysing the surveillance videos before and after the installation of the developed system in the working environment of the factory.

2.1 Dataset

In this study, an original video dataset was prepared to be used with deep learning models for real-time automatic detection of unsafe behaviours in workplaces. Videos from the factory site were collected between 5 November and 13 December 2022 from "Kafaoğlu Metal Plastik Makine San. ve Tic. A.Ş." in Eskisehir, Turkiye. The video data was obtained from two different IP cameras for 39 days, totalling approximately 4000 h. The brand of the IP cameras is UNV and the model is IPC2122CR3-PF40-A. In addition, the IP cameras have a resolution of 1920 × 1080 and can record in Full HD. This video dataset is divided into two sub-groups, both for understanding the video content and for recognising the objects in the video content. A consent form for the protection and use of personal data was obtained from the employees in the videos. In addition, it was approved by the decision of the Ethics Committee of the Rectorate of Bilecik Seyh Edebali University dated 11.06.2021 and number 10 that there is no ethical contradiction in the conduct of the study.

During the creation of the dataset for the study, 4 different unsafe behaviours that are most common in the working environment were identified by taking the opinions of factory managers and workplace safety experts. These unsafe behaviours are: Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift. In addition, 4 safe behaviours, namely Safe Walkway, Authorized Intervention, Closed Panel Cover and Safe Carrying, were identified for classification and decision making in the developed real-time detection system. In the study, the class ID, class name and behaviour type information for the safe and unsafe behaviour classes are shown in Table 1. The unsafe behaviour Safe Walkway Violation indicates that workers go beyond the boundaries

Table 1 Class ID, class name and behavior type information for safe and unsafe behavior classes

Class ID	Class name (unsafe/safe behaviour)	Behaviour type
0	Safe Walkway Violation	Unsafe
1	Unauthorized Intervention	Unsafe
2	Opened Panel Cover	Unsafe
3	Carrying Overload with Forklift	Unsafe
4	Safe Walkway	Safe
5	Authorized Intervention	Safe
6	Closed Panel Cover	Safe
7	Safe Carrying	Safe

of the designated safe walkway at workplaces, while the safe and normal behaviour is Safe Walkway. On the other hand, the behaviour of people intervening in unauthorized situations at work sites and not wearing safety equipment is defined as Unauthorized Intervention, while the accepted class of this behaviour is Authorized Intervention. Similarly, leaving panel covers open/forgetting them is classified as Opened Panel Cover, an unsafe behaviour, while closed panel covers are classified as Closed Panel Cover, a safe behaviour. Finally, the Carrying Overload with Forklift class represents carrying 3 or more blocks with a forklift at workplaces and is an unsafe behaviour class with the label Carrying Overload with Forklift. Carrying 2 or less blocks with a forklift is a safe behaviour and the class label is Safe Carrying.

To create the dataset, videos of safe and unsafe behaviours by workers/employees were first recorded by 2 different cameras in the factory over 19 days. After collecting the data, video segments averaging 3–18 s were created by identifying the safe and unsafe behaviours in the presence of experts. Figure 2 shows the behaviours belonging to the safe and unsafe classes, showing the full angle of the cameras belonging to the classes identified for the video dataset. In Fig. 2a there is a safe behaviour of carrying 2 blocks and less with a forklift (Safe Carrying), while Fig. 2e shows an example of an unsafe employee behaviour of carrying 3 blocks and more with a forklift (Carrying Overload with Forklift). Figure 2b shows an example of an authorized intervention with a green vest (Authorized Intervention), while Fig. 2f shows an example of an unauthorized intervention with a red-black vest (Unauthorized Intervention). Similarly, Fig. 2c shows a safe behaviour for the Safe Walkway class along the green path in the work environment, while Fig. 2g shows unsafe behaviour as a worker for the Safe Walkway Violation class outside the green path. Finally, Fig. 2d shows a safe behaviour for the Closed Panel Cover class for a panel connected to a machine, while Fig. 2h shows an unsafe behaviour for the Opened Panel Cover class.

In the study, the number of training and test sets for the originally created dataset, consisting of video fragments collected between 5 November and 23 November 2022 for 8 classes including Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover, Carrying Overload with Forklift, Safe Walkway, Authorized Intervention, Closed Panel Cover and Safe Carrying, which include safe and unsafe behaviours from factory work environments, are given in Table 2. The dataset contains a total of 566 video segments for training, while a total of 125 video segments were prepared for test.

As shown in Table 3, to train of YOLO v4, frames from the videos in the training set were obtained and the objects in a total of 2262 images belonging to the classes were labelled according to the label structure of the YOLO v4 algorithm using LabelImg software [31] and a ground truth was created. In addition, to evaluate the success of YOLO v4 in detecting object, a total of 329 images belonging to unsafe behaviour classes in the video frames of the test set were labelled using the same method. The ground truth process was carried out in a box with an occupational health and safety expert working in the factory to determine the position of the object in the image. The generated label file contains values related to the x-y coordinates, height and width of the objects.

2.2 YOLO v4

In this study, the YOLO v4 algorithm is used to establish the relationship between object and unsafe motion for real-time unsafe behaviour detection. The frames with unsafe behaviours in the video dataset are selected by the YOLO v4 deep learning algorithm. Therefore, in this study, the frames that are not related to the identified unsafe behaviours are excluded with

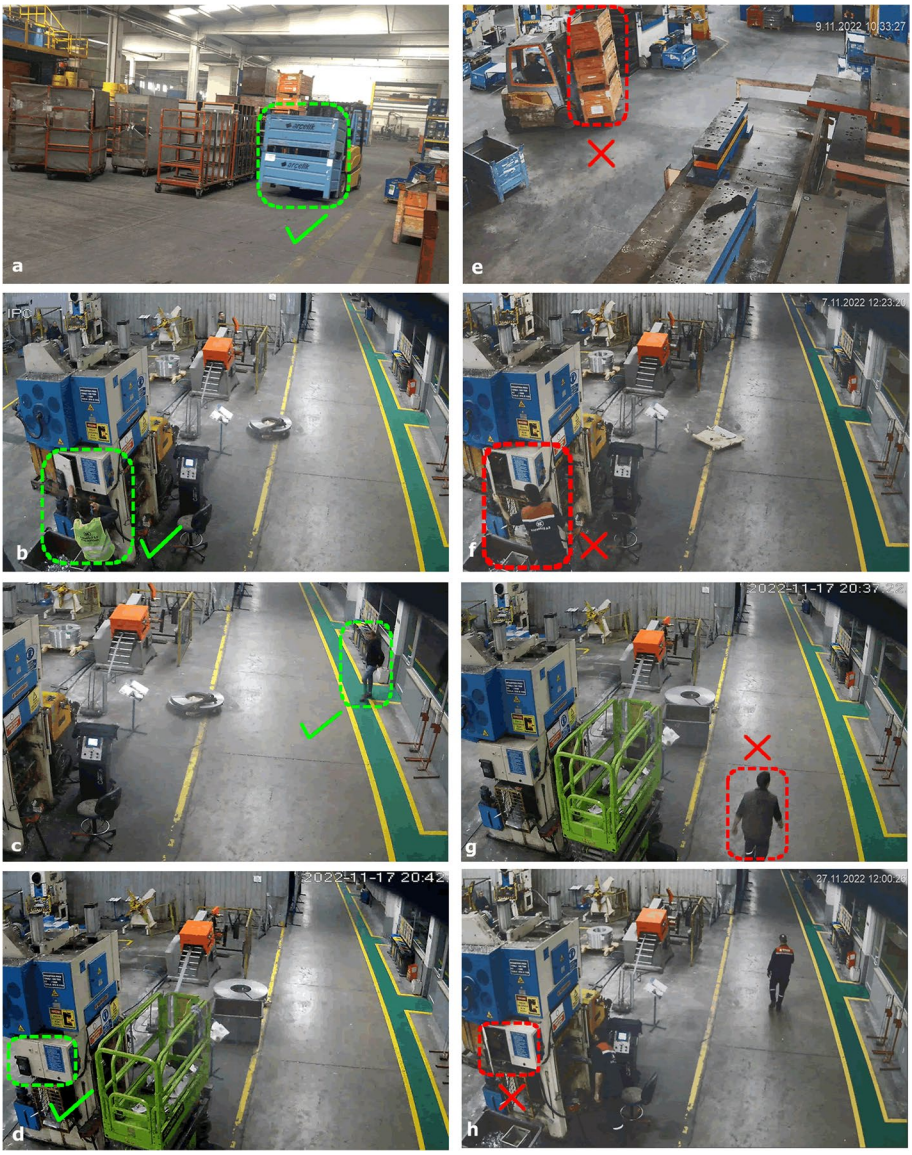

Fig. 2 Samples of behaviours belonging to safe and unsafe classes in the dataset. **a** Safe Carrying, **b** Authorized Intervention, **c** Safe Walkway, **d** Closed Panel Cover, **e** Carrying Overload with Forklift, **f** Unauthorized Intervention, **g** Safe Walkway Violation and **h** Opened Panel Cover

YOLO v4 to reduce the load on the video understanding architecture and increase its detection speed and accuracy. The YOLO v4 algorithm is a deep learning algorithm that has been widely used in real-time object recognition, especially to improve speed and detection [32, 33]. The YOLO v4 algorithm, whose block diagram is shown in Fig. 3, consists of three sub-layers. In this architecture, feature extraction is performed in the backbone layer, while the neck layer extracts information from adjacent feature maps with bottom-top and top-down

Table 2 Number of training and test sets consisting of video fragments collected between November 05–23, 2022

Class name (unsafe/safe behaviour)	Number of video for training set	Number of video for test set	Average video duration (s)
Safe Walkway Violation	178	32	9.7
Unauthorized Intervention	97	11	10.2
Opened Panel Cover	129	13	5.8
Carrying Overload with Forklift	48	8	6.9
Safe Walkway	50	25	9.2
Authorized Intervention	23	15	8.6
Closed Panel Cover	19	13	4.6
Safe Carrying	22	8	7.2
Total	**566**	**125**	

Table 3 The number of images labelled in training and test sets for object detection using YOLO v4

Unsafe/safe behaviour	Number of images labelled in training set	Number of images labelled in test set
Safe Walkway Violation	526	100
Unauthorized Intervention	180	23
Opened Panel Cover	590	60
Carrying Overload with Forklift	200	40
Safe Walkway	351	43
Authorized Intervention	82	12
Closed Panel Cover	252	40
Safe Carrying	81	11
Total	**2262**	**329**

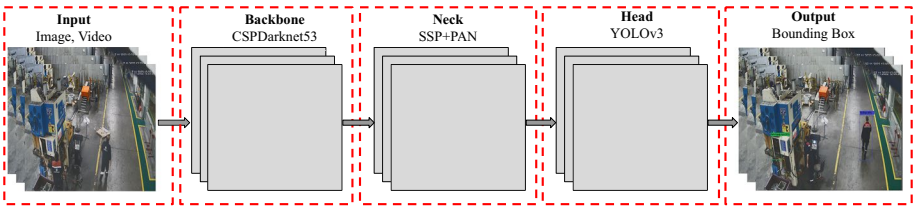

Fig. 3 Block diagram of the YOLO v4 algorithm used in the study for frame selection from video segments

flows to achieve higher performance in object prediction. In the YOLO v4 architecture, CSP-Darknet53 is used as the backbone, and spatial pyramid pooling (SPP) and path aggregation network (PAN) are used in the neck layer. In the last layer, the head, there are bounding boxes and the class of each box is estimated. The prediction procedure of the YOLO v3 algorithm is also used in the head layer.

2.3 ConvLSTM

In this study, the ConvLSTM model, which is a combination of CNN and LSTM networks, is used to classify unsafe behaviours. ConvLSTM is a hybrid deep learning architecture that combines the convolutional layers of CNN with LSTM [34]. ConvLSTM combines the shape information of images with temporal information such as behaviour and speed to analyse and predict video. Except for convolution operations and higher dimensional data representation, ConvLSTM is similar to the typical LSTM structure [35]. The block diagram of the ConvLSTM architecture, created by combining CNN and LSTM networks, is shown in Fig. 4. Here, $X_1, \ldots, X_t$ are the inputs of the network, while $C_1, \ldots, C_t$ represent the memory units of the ConvLSTM network. In addition, $H_1, \ldots, H_t$ represents the state of the hidden layers in the LSTM network. In the ConvLSTM network, i_t,f_t, O_t are used for the input gate, the forgetting gate and the output gate respectively. All variables in the network are in 3D tensor structure, where '*' stands for the convolution process and '°' for the Hadamard product. The calculation of i_t, f_t, O_t, C_t ve $H_t$is given in Eqs. (1), (2), (3), (4) and (5), where σ the sigmoid activation function and tan h is the hyperbolic tangent activation function. The new memory C_t and the output H_t are generated thanks to updating memory C_{t-1} using the current input X_t and the previous output H_{t-1} [36].

$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f\right) \tag{2}$$

$$O_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o\right) \tag{3}$$

$$C_t = f_t^{\circ} C_{t-1} + i_t^{\circ} \tan h\left(W_{xo} * X_t + W_{hc} * H_{t-1} + b_c\right) \tag{4}$$

$$H_t = O_t^{\circ} \tan h\left(C_t\right) \tag{5}$$

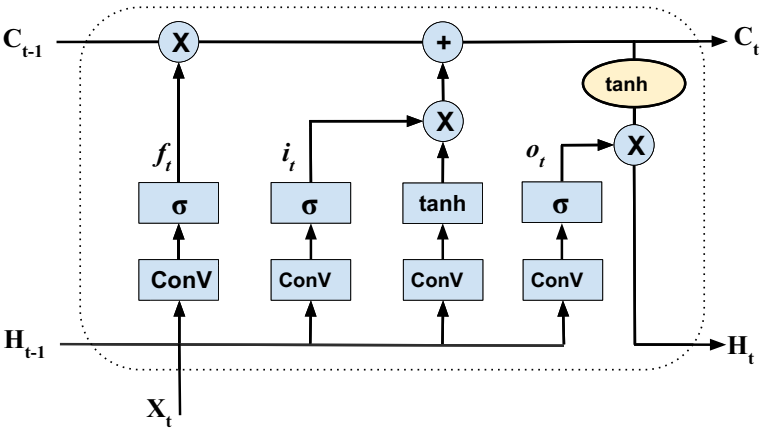

Fig. 4 Block diagram of the ConvLSTM architecture

Data from cameras or imaging devices is characterised as a time sequence. In such cases, it is more appropriate to use an LSTM-type model. In this type of architecture, the proposed model passes the previous hidden state to the next step of the sequence. Thus, the architecture is used for the network to hold information and make decisions about the data it has already seen [37]. ConvLSTM has a mechanism that allows learning from relatively small image datasets, with a maximum workload for operating image models in motion time. ConvLSTM units often operate more efficiently than traditional recurrent neural networks (RNNs) and make better use of input features. As a result, ConvLSTM is a particularly powerful approach for analysing and classifying real-time video images.

2.4 Proposed hybrid deep learning model

A simple LSTM structure is not compatible with self-modelling temporal data, as it only receives one-dimensional input. In this work, after detecting frames containing unsafe behaviour in videos using YOLO v4, the ConvLSTM model is used, which consists of a CNN to extract spatial features at a given time step of frame sequences from the video containing the labelled objects, followed by an LSTM network to classify the temporal relationships between frames. This not only improves the accuracy of detecting unsafe behaviour in video, but also makes the model lighter, enabling real-time detection and analysis of industrial operations. The infrastructure of the Unsafe-Net hybrid deep learning model approach, consisting of the proposed YOLO v4 and ConvLSTM deep learning architectures, is shown in Fig. 5. In the proposed Unsafe-Net model, YOLO v4 first identifies possible frames containing unsafe behaviour from video sequences and discards the rest of the video. As a result, a reduced video stream is obtained at the output of the YOLO v4 network compared to the input. These potentially unsafe behaviour frames are then passed to the ConvLSTM network, which is a combination of CNN and LSTM methods. The reduced video stream is classified in the ConvLSTM network to determine which class of unsafe behaviour occurs in the video.

The ConvLSTM cell is a layer that contains the convolutional operations in the network and is a convolutional LSTM embedded in the architecture. Thus, in this study, the spatial features of the data are detected by taking into account the temporal relationship. For video classification, the proposed approach effectively captures the spatial relationship in each frames and the temporal relationship in different frames. As a result of this convolutional structure, ConvLSTM is able to receive 3D inputs, namely width, height and num_of_channels, unlike a single LSTM network. Figure 6 shows the diagram of the internal structure of the architecture

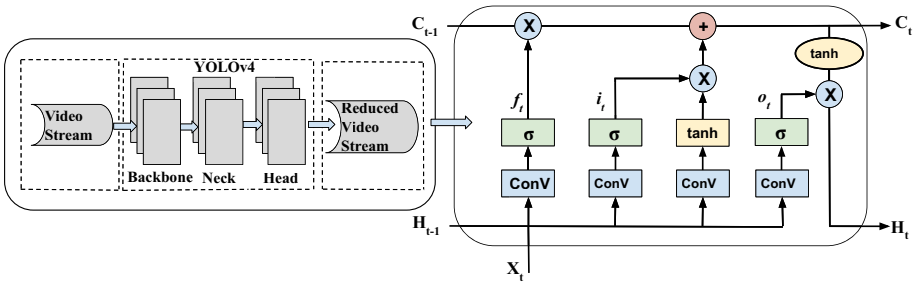

Fig. 5 Architecture of proposed Unsafe-Net deep learning model based on YOLO v4 and ConvLSTM for unsafe behaviour detection

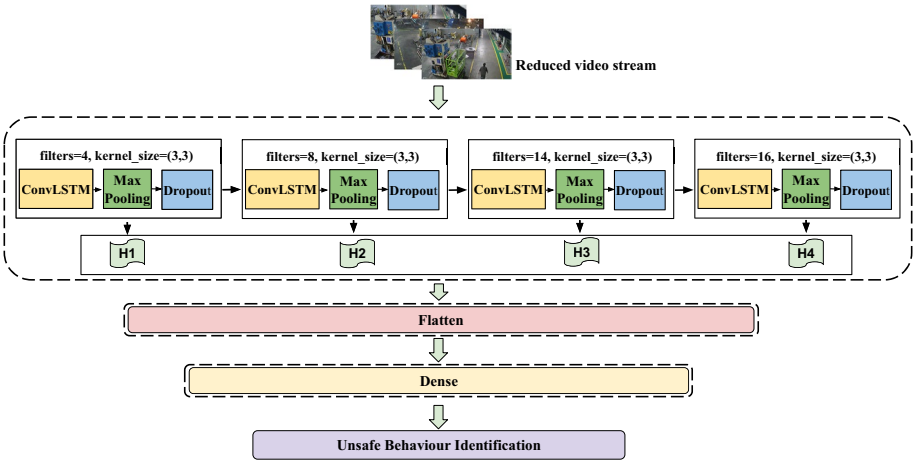

Fig. 6 Diagram of the internal structure of the architecture of the ConvLSTM structure in the Unsafe-Net deep learning model proposed for the detection of unsafe behaviour in the workplace

of the ConvLSTM structure in proposed Unsafe-Net learning model for the detection of unsafe behaviours in workplace. This architecture takes as input the videos from the YOLO v4 network containing the unsafe actions Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift. In the developed approach, 4 layers consisting of ConvLSTM, MaxPooling and Dropout components are used. The number of filters in these layers is 4, 8, 14 and 16 respectively. In addition, kernel_size=(3, 3), pool_size=(1, 2, 2) and dropout=0.2. In addition, *tanh* is used as the activation function. Finally, unsafe behaviour is identified by adding flatten and dense layers to the output of the network.

2.5 Warning and monitoring system

This study has also developed a real-time warning and monitoring system that is activated when unsafe behaviour is detected by the proposed method from the videos of the work environment in the factory. When this warning and monitoring system is activated, it sends an audible and visual warning directly to the area where the unsafe behaviour is taking place. Figure 7 shows the structure of the developed warning and monitoring system.

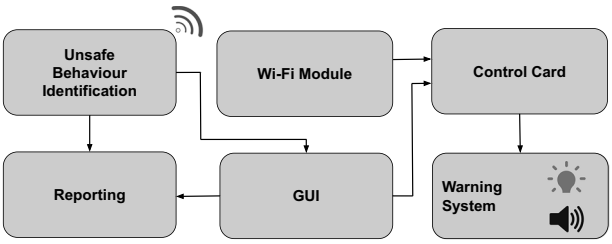

Fig. 7 The structure of the developed warning and monitoring system

In the developed warning and monitoring system, after the unsafe behaviour is detected by the software, the information that the unsafe behaviour has occurred is wirelessly transmitted to an Arduino Uno-based control card via an ESP8266 Wi-Fi module. The control card triggers audible and visual alerts based on the information received. In addition, automatic reporting and data logging is performed when unsafe behaviour occurs. In addition, both the reporting and the control card infrastructure can be accessed via a GUI program.

3 Experimental results

In this study, a system was developed that enables real-time detection of unsafe behaviours in work environments and immediate visual and audible warning to employees using Unsafe-Net, a deep learning based approach. The study used YOLO v4 and ConvLSTM algorithms to classify the safe and unsafe behaviours of workers in a factory located in an industrial area. With YOLO v4, the frames containing the behaviours in the video scene videos from the work site were presented as sequences to the ConvLSTM network, allowing the system to work quickly and with high accuracy. The experimental analyses performed in the study were carried out on a workstation with dual NVIDIA GeForce GTX 1080 GPU, Intel I9 9900 CPU, 16 GB RAM, 3 TB HDD and 500 GB SSD.

Some key performance metrics are used to evaluate the performance of the proposed Unsafe-Net architecture. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in these key metrics are the values from the confusion matrix table. For the analyses carried out in the experimental studies, the metrics Average Precision (AP), Mean Average Precision (mAP) and Intersection over Union (IoU) were measured to evaluate the success achieved in detecting safe and unsafe behaviours in work environments with the proposed YOLO v4 network. Precision (P) is the ratio of the values predicted as TP to the total positive values, while Recall (R) is the ratio of the values predicted as TP to the values actually known to be positive. In addition, AP is the total area under the P and R curves. Furthermore, IoU is the area of the intersection of prediction and ground truth divided by the area of the union of these two values. In addition, mAP is measured by averaging the AP values of the classes. In addition, the Accuracy (Acc) metric is used in the Unsafe-Net architecture to evaluate the classification results of unsafe behaviours. mAP, IoU and Acc metrics are given in Eqs. (6), (7) and (8), respectively.

$$Mean\ Average\ Precision(mAP) = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{6}$$

$$\text{Intersectionover Union(IoU)} = \frac{|\text{prediction} \cap \text{ground\_truth}|}{|\text{prediction} \cup \text{ground\_truth}|} \tag{7}$$

$$\text{Accuracy}(\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \tag{8}$$

In the study, the first task to be performed after preparing the video database is to perform object recognition of safe and unsafe behaviours from video parts using YOLO v4. Therefore, by selecting the frames that belong to the unsafe behaviour classes Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift, the frames that belong to the safe behaviour classes Safe Walkway,

Authorized Intervention, Closed Panel Cover and Safe Carrying and other parts that do not contain other unsafe behaviours are removed from the videos and the videos are prepared for classification with ConvLSTM. The change in the average Loss and mAP values over the iteration period as a result of training with YOLO v4 with 2262 images in the training set is shown in Fig. 8. For the training of the network to be successful, the Loss value, shown in blue, is expected to approach zero and the mAP value, shown in red, is expected to increase towards 100%. Here we can see that the average mAP value during the iteration period is around 97% and the Loss value is very close to zero. It can therefore be said that the YOLO v4 training has been successfully completed. The dimensions of the images used in YOLO v4 training were 416×416, the number of classes was 4, the batch size was 65, the subdivisions were 32, and the number of GPUs was 2.

The validation and testing of YOLO v4 was completed by using the image in 329 test sets. Using the @0.5 threshold for IoU, the AP results obtained for the Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift classes, as well as the TP and FP values, are shown in Table 4. It can be

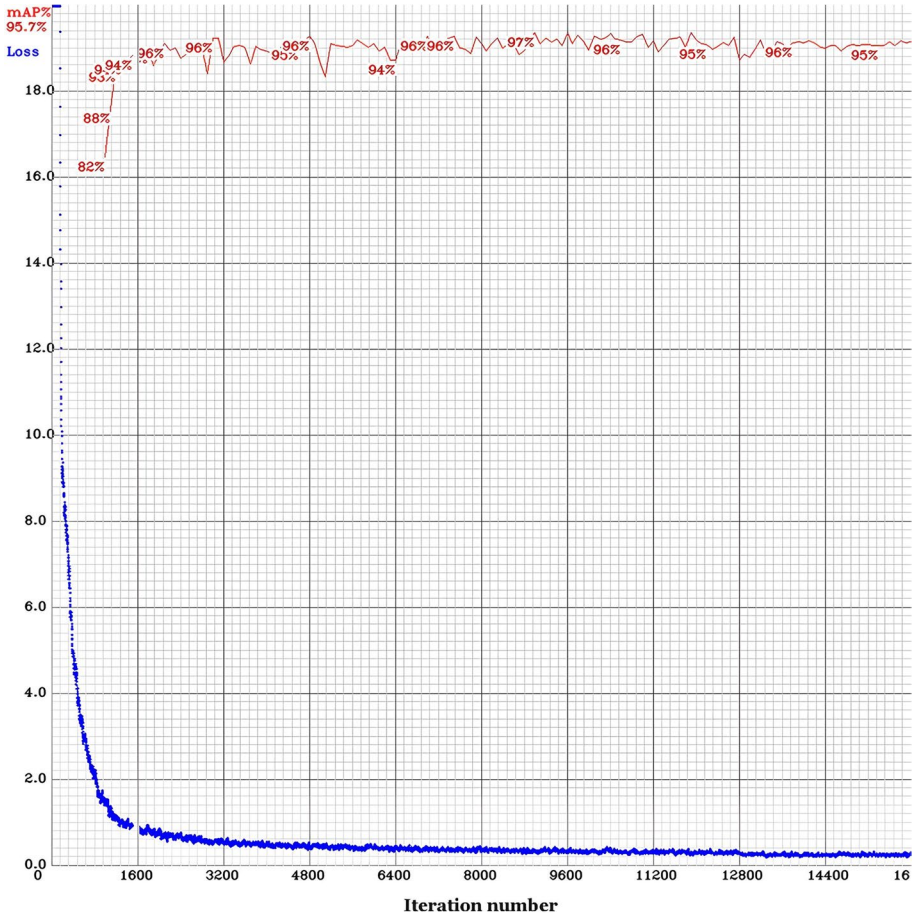

Fig. 8 The change in the average Loss and mAP values over the iteration period during the training of YOLO v4

Table 4 AP score and TP and FP values obtained for each class in the test process with the YOLO v4 network

Unsafe/safe behaviour	AP (%)	TP	FP
Safe Walkway Violation	97.89	97	12
Unauthorized Intervention	95.83	23	0
Opened Panel Cover	96.92	63	7
Carrying Overload with Forklift	97.50	40	2
Safe Walkway	90.82	36	5
Authorized Intervention	100	12	0
Closed Panel Cover	100	41	0
Safe Carrying	86.30	11	8

concluded that YOLO v4 successfully detects objects with unsafe behaviours. In addition, the total duration of the test process with YOLO v4 was 14 s on 329 images. Here, the AP scores achieved with YOLO v4 for the 7 safe and unsafe classes, with the exception of the Safe Walkway class, are successful with over 90% performance. In addition, the proposed YOLO v4 method achieved a score of 95.66% for mAP in the test set. In addition, it can be seen that the lowest AP score in the test set is in the Safe Carrying class with 86.30% and the TP value for this class is 11 while the FP value is 8. The reason for this is probably due to the similarity of object detection in the Carrying Overload with Forklift class in the case of 3 blocks and more and in the Safe Carrying class in the case of 2 blocks and less. On the other hand, the FP values for the Safe Walkway Violation and Opened Panel Cover classes were also higher than the other classes, although they were expected to be lower. In addition, the Closed Panel Cover and Authorized Intervention classes achieved AP scores of 100% and FP values of 0, resulting in high performance for these classes.

The results for the object successfully detected with YOLO v4 for the safe and unsafe behaviour classes on the videos in the test set are shown in Fig. 9. Figure 9a shows a video with the unsafe behaviour Safe Walkway Violation and frames belonging to the safe behaviour classes Authorized Intervention, Closed Panel Cover and Safe Walkway. Figure 9b shows another video with frames of the safe behaviour class Authorized Intervention and the unsafe behaviour class Opened Panel Cover. In this work, YOLO v4 is used both for detecting objects belonging to safe and unsafe behaviours and for generating reduced video segments for ConvLSTM in the Unsafe-Net architecture, where safe and irrelevant frames are excluded. Figure 9c shows a video in which only the frames belonging to the unsafe class behaviours Opened Panel Cover and Safe Walkway Violation remain and the irrelevant ones are excluded. Video fragments containing frames of a single unsafe class behaviour for a single action are then created and sent to the ConvLSTM network. For example, Fig. 10 shows some frames from a video consisting of 40 frames. This video initially consists of 3.5 s and 105 frames. After this video is passed through the YOLO v4 network, it can be seen that after the irrelevant frames are excluded, it is reduced to only 40 frames and the duration of the video is reduced to 1.33 s. This video is transmitted to the ConvLSTM network in three parts for the next stages, including the Safe Walkway Violation, Opened Panel Cover and Unauthorized Intervention unsafe behaviour classes and only one action in each video. These procedures are applied in the same way to each video in the test set.

In the Unsafe-Net architecture, videos containing dangerous behaviour are processed using the YOLO v4 model with the pre-trained weights and simplification is performed

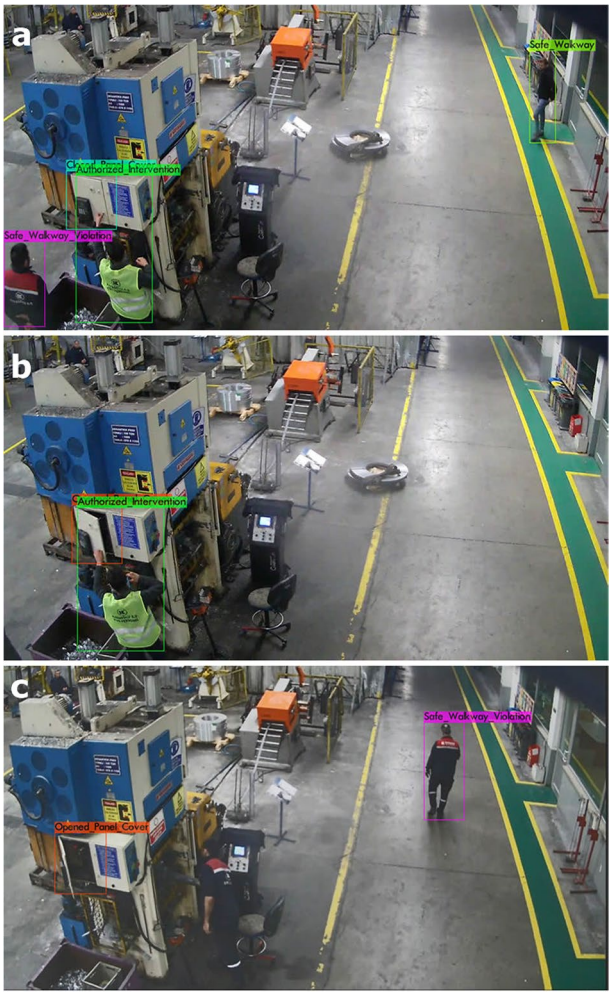

Fig. 9 Examples of objects successfully detected by YOLO v4 belonging to safe and unsafe behaviour classes in the test video

by discarding unrelated frames on the videos. This increases the classification accuracy and the speed of extracting meaning from real-time video images. After YOLO v4 is used for object recognition from videos containing unsafe behaviours and unnecessary parts are excluded, these videos with reduced size are classified by class type using ConvLSTM. For the real-time classification of unsafe behaviours in work environments with ConvLSTM, 452 videos were selected for the training set and 64 videos were selected for the test set. By using YOLO v4, frames that do not contain unsafe behaviour were excluded from the videos in the training and test sets, reducing both the size of the videos and the recognition time, as the process involves real-time detection. Figure 11 shows the variation of the Loss, Validation Loss (val_loss), Accuracy and Validation Accuracy (val_accuracy) values as a function of the number of iterations during the training period of the ConvLSTM network. It can be concluded that the loss value is very close to 0 and the accuracy metric is very close to 1, and as a result the training of the ConvLSTM network is very successful.

Fig. 10 A video fragment initially consisting of 105 frames, after passing through the YOLO v4 network, becomes a fragment consisting of 3 actions (unsafe behaviour) and 40 frames, and some sample frames in the reduced video

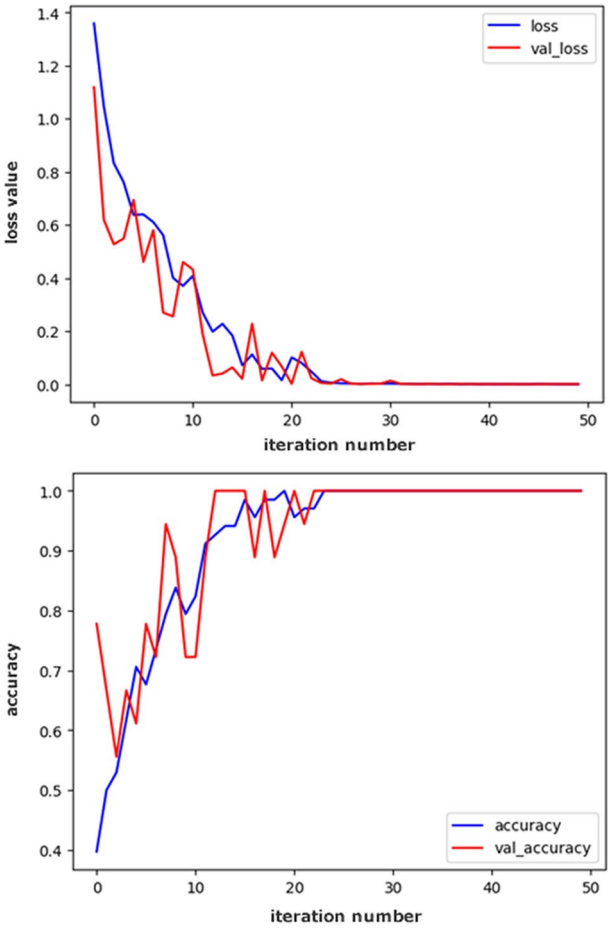

Fig. 11 The variation of the loss, validation loss (val_loss), accuracy and validation accuracy (val_accuracy) values during the training time of ConvLSTM network

Classification with Unsafe-Net identified 4 different unsafe behaviour/action classes, namely Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift, which are common in workplaces. These classes were assigned class ID values of 0, 1, 2 and 3 respectively. After training the ConvLSTM network, the performance of the Unsafe-Net hybrid architecture proposed in the study was evaluated by using 32 videos for the Safe Walkway Violation class, 11 videos for the Unauthorized Intervention class, 13 videos for the Opened Panel Cover class and 8 videos for the Carrying Overload with Forklift class in the test set and a total of 32 videos passing through the YOLO v4 network. Table 5 shows

Table 5 Performance of ConvLSTM network without video frame reduction with YOLO v4 for each unsafe behaviour class for the video test set

		Prediction			
	Class ID	0	1	2	3
Actual	0	31	0	0	1
	1	0	9	2	0
	2	0	3	10	0
	3	1	0	0	7

the performance of YOLO v4 and ConvLSTM network without video frame reduction for each unsafe behaviour class for the video test set. Here, 1 video for ID=1 Safe Walkway Violation class, 2 videos for ID=2 Unauthorized Intervention class, 3 videos for ID=3 Opened Panel Cover class and 1 video for ID=4 Carrying Overload with Forklift are classified as incorrect. On the other hand, Table 6 shows the performance of the Unsafe-Net approach for each unsafe behaviour class for the video test set. As can be seen, all videos are correctly classified for the Safe Walkway Violation and Carrying Overload with Forklift behaviour classes using the Unsafe-Net hybrid deep learning architecture proposed in this study. On the other hand, 1 video each was misclassified for Unauthorized Intervention and Opened Panel Cover classes using the Unsafe-Net architecture. Out of a total of 64 video data belonging to 4 classes in the test set in the dataset, 62 videos were successfully classified using the Unsafe-Net architecture.

Table 7 shows the performance of the ConvLSTM network without video frame reduction with YOLO v4 and the Unsafe-Net (YOLO v4+ConvLSTM) hybrid deep learning model proposed in this study for the Acc metric in each unsafe behaviour class for the video test set in the dataset. Accordingly, the average performance of the YOLO v4 and ConvLSTM network without video frame reduction was obtained as 85.78% Acc. In addition, 100% Acc was achieved for the classes of Safe Walkway Violation and Carrying Overload with Forklift. On the other hand, 90.91% Acc was achieved for the Unauthorized Intervention class, while 92.31% Acc was achieved for the Opened Panel Cover unsafe behaviour class. For all unsafe behaviour classes, the proposed Unsafe-Net deep learning architecture achieved an average Acc of 95.81%. It can be concluded that unsafe behaviours are successfully classified from videos captured from work environments using the proposed Unsafe-Net network.

In addition, some examples of correct and incorrect classifications for the proposed Unsafe-Net method are shown in Fig. 12. Figure 12a, b, c and d show that the proposed Unsafe-Net architecture successfully classifies Safe Walkway Violation, Unauthorized

Table 6 Performance of the Unsafe-Net (YOLO v4+ConvLSTM) network proposed in this study for each unsafe behaviour class for the video test set

		Prediction			
	Class ID	0	1	2	3
Actual	0	32	0	0	0
	1	0	10	1	0
	2	0	1	12	0
	3	0	0	0	8

Table 7 Comparison of the performance of the ConvLSTM network and the proposed Unsafe-Net model for the Acc metric in each unsafe behaviour class for the video test set

Class ID	Unsafe behaviour	Acc (%) via ConvLSTM	Acc (%) via Unsafe-Net (YOLO v4+ConvLSTM)
0	Safe Walkway Violation	96.88	100
1	Unauthorized Intervention	81.82	90.91
2	Opened Panel Cover	76.92	92.31
3	Carrying Overload with Forklift	87.50	100
Overall (%)		**85.78**	**95.81**

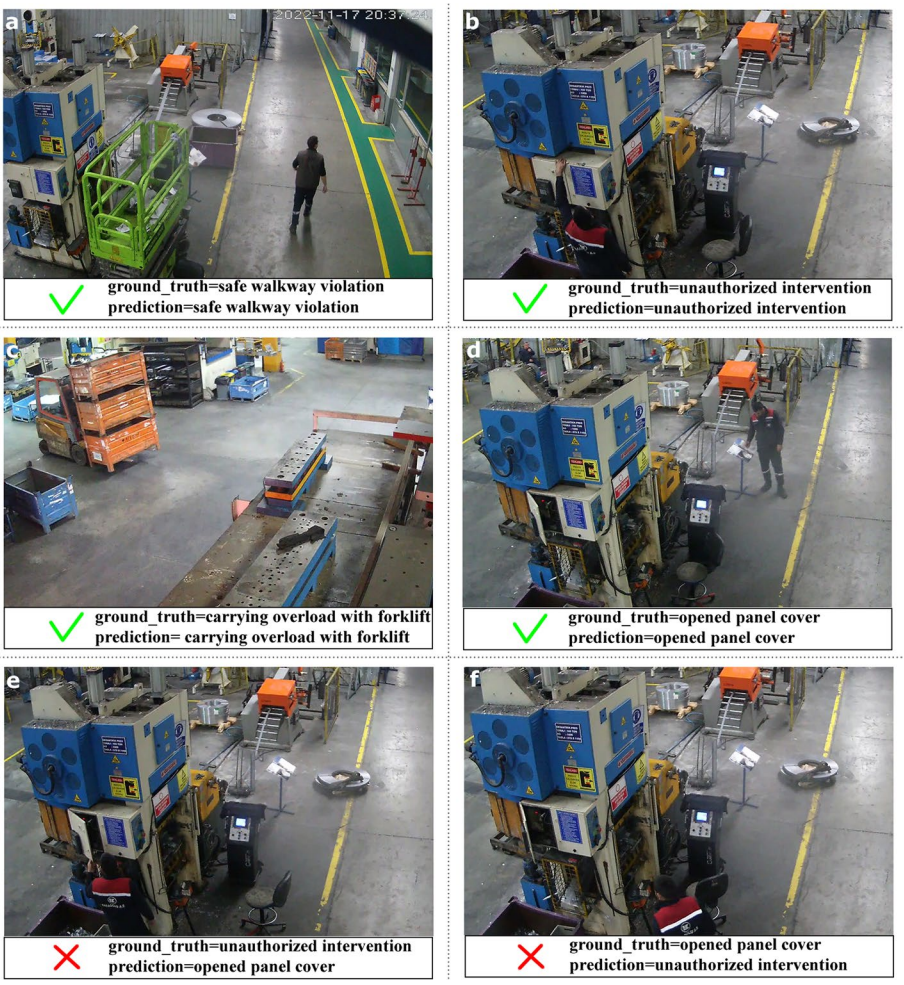

Fig. 12 True and false classification examples for the proposed Unsafe-Net method

Intervention, Opened Panel Cover and Carrying Overload with Forklift for 4 different unsafe action classes, thus matching the ground truth. On the other hand, Fig. 12e and f show that the unsafe behaviours Unauthorized Intervention and Opened Panel Cover are confused and misclassified.

In the Unsafe-Net deep learning architecture proposed in this study, a program with the GUI shown in Fig. 13 was developed using the Python programming language and PyQt Designer to monitor and control the real-time system. This program can be used to start and stop the operation of the real-time system, view unsafe activities in the working environment, access text document type reports generated by the system, some of the contents of which are shown in Fig. 14, and run the warning system.

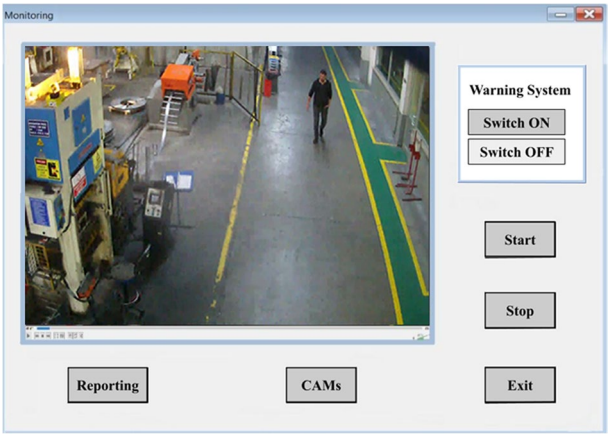

Fig. 13 Interface of the developed program for monitoring and controlling the system in Unsafe-Net architecture

```
21-Safe Walkway Violation              ZONE-1  09:35   25/11/2022.............
22-Unauthorized Intervention           ZONE-1  10:45   25/11/2022.............
23-Opened Panel Cover                  ZONE-1  10:54   25/11/2022.............
24-Carrying Overload with Forklift     ZONE-2  12:09   25/11/2022.............
25-Unauthorized Intervention           ZONE-1  13:05   25/11/2022.............
26-Opened Panel Cover                  ZONE-1  13:19   25/11/2022.............
27-Opened Panel Cover                  ZONE-1  13:22   25/11/2022.............
28-Carrying Overload with Forklift     ZONE-2  15:09   25/11/2022.............
29-Safe Walkway Violation              ZONE-1  15:35   25/11/2022.............
30-Unauthorized Intervention           ZONE-1  16:45   25/11/2022.............
31-Safe Walkway Violation              ZONE-1  16:54   25/11/2022.............
32-Safe Walkway Violation              ZONE-1  16:55   25/11/2022.............
33-Safe Walkway Violation              ZONE-1  17:19   25/11/2022.............
34-Opened Panel Cover                  ZONE-1  10:22   26/11/2022.............
35-Unauthorized Intervention           ZONE-1  10:45   26/11/2022.............
36-Safe Walkway Violation              ZONE-1  10:54   26/11/2022.............
37-Carrying Overload with Forklift     ZONE-2  11:09   26/11/2022.............
38-Unauthorized Intervention           ZONE-1  13:05   26/11/2022.............
39-Safe Walkway Violation              ZONE-1  13:19   26/11/2022.............
40-Opened Panel Cover                  ZONE-1  13:22   26/11/2022.............
41-Carrying Overload with Forklift     ZONE-2  14:09   26/11/2022.............
42-Safe Walkway Violation              ZONE-1  14:35   26/11/2022.............
43-Safe Walkway Violation              ZONE-1  14:45   26/11/2022.............
44-Unauthorized Intervention           ZONE-1  15:54   26/11/2022.............
45-Safe Walkway Violation              ZONE-1  16:09   26/11/2022.............
46-Unauthorized Intervention           ZONE-1  17:05   26/11/2022.............
47-Opened Panel Cover                  ZONE-1  10:19   27/11/2022.............
48-Opened Panel Cover                  ZONE-1  10:22   27/11/2022.............
49-Carrying Overload with Forklift     ZONE-2  11:09   27/11/2022.............
50-Safe Walkway Violation              ZONE-1  14:35   27/11/2022.............
51-Unauthorized Intervention           ZONE-1  16:08   27/11/2022.............
52-Carrying Overload with Forklift     ZONE-2  16:09   27/11/2022.............
53-Safe Walkway Violation              ZONE-1  16:35   27/11/2022.............
```

Fig. 14 Reporting of the content generated by the Unsafe-Net system in text document type when unsafe behaviour occurs in the factory

4 Discussion

In the Unsafe-Net system, when any of the unsafe behaviours are performed by the workers in the work environment, the system and the electronic control card communicate via the Wi-Fi network and audible and visual warnings are sent to the environment where the workers are located. As can be seen in Fig. 15, the unsafe behaviour

Fig. 15 Giving an audible and visual warning when unsafe behavior occurs

belonging to the Safe Walkway Violation class was successfully detected by the developed system and audible and visual warnings were given to the environment.

The Unsafe-Ne architecture proposed in this study for the real-time detection of unsafe actions in work environments has a two-stage execution. In the first stage, YOLO v4 is used to select the frames with unsafe actions in the videos and the others are excluded. Since the proposed system is real-time, the action detection time is very important. As can be seen in Table 8, the average video duration of the 64 videos in the video dataset with unsafe actions ranging from 3 to 18 s is 7.72 s. When the videos in the test set are classified with ConvLSTM without any reduction with YOLO v4, the average action recognition time for 4 different classes is 0.58 s. On the other hand, with the Unsafe-Net architecture proposed in this study, when the frames containing unsafe actions in the videos are selected and the others are excluded, the average video duration for the videos is reduced to 1.87 s. In addition, the action detection time for 64 videos in the test set is 0.14 s. Thus, the real-time execution of the system significantly accelerates the action recognition time.

This study also monitored the change in unsafe behaviours in the factory environment before and after the installation of the system with the Unsafe-Net architecture proposed in this study. As can be seen in Table 9, the number of behaviours detected daily by the occupational safety expert in the factory environment belonging to the unsafe classes Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift before the system was installed between 5 and 23 November 2022 is given. It can be seen that a total of 518 unsafe behaviours from 4 classes occurred in the first period of data collection.

In the proposed system, after the results obtained between 5 and 23 November 2022, the employees were informed about the system and the video footage was taken again in the following days and the number of unsafe actions related to unsafe behaviours was determined again. Table 10 shows the number of unsafe behaviours that occurred between 24 November 2022 and 13 December 2022 after the real-time system with

Table 8 The comparison of average video duration and average action identification time for ConvLSTM and proposed Unsafe-Net

Method	Average video duration (s)	Average action identification time (s)
ConvLSTM	7.72	0.58
Unsafe-Net	1.87	0.14

Table 9 Number of unsafe behaviours that occurred between 5 and 23 November 2022 prior to the installation of the real-time Unsafe-Net architecture in the factory

Date (D/M/Y)	Unauthorized intervention	Opened panel Cover	Carrying overload with forklift	Safe walkway violation
05/11/2022	6	7	3	10
06/11/2022	5	5	2	9
07/11/2022	7	9	2	9
08/11/2022	9	4	4	12
09/11/2022	4	5	5	9
10/11/2022	11	12	0	9
11/11/2022	3	6	1	14
12/11/2022	4	8	6	17
13/11/2022	0	11	3	8
14/11/2022	5	12	2	9
15/11/2022	6	6	2	12
16/11/2022	5	5	2	9
17/11/2022	8	3	3	11
18/11/2022	6	9	4	9
19/11/2022	7	8	1	12
20/11/2022	5	2	2	10
21/11/2022	4	11	4	9
22/11/2022	4	9	7	15
23/11/2022	9	10	5	17
Total	**108**	**142**	**58**	**210**

Unsafe-Net architecture was installed in the factory. Here, the performance of the system was re-evaluated by examining the 20-day unsafe behaviour report data, and the contribution of the system to occupational safety was examined. As can be seen from the Table 10, the unsafe behaviours of Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift were significantly reduced after the installation of the system proposed in the study. The number of unsafe actions, which was 518 in 4 different unsafe behaviour classes for the same period before the system was installed, was reduced to 134 after the system was installed, a significant reduction of approximately 75%. Here, the accuracy of the proposed real-time system is verified by first looking at the decision of the Unsafe-Net architecture for the 4 unsafe behaviour classes, and then the classification results were confirmed by the occupational safety expert in charge at the factory. In addition, when the classification result of the proposed system was compared with the surveillance camera monitoring in the same period, it was reported by the occupational safety expert that only one unsafe behaviour in the Opened Panel Cover class was included in the Unauthorized Intervention class.

Although this study proposes a novel real-time computer vision-based method for detection and classification of unsafe behaviours from a new video dataset collected from workplaces, it is considered to have some limitations. Firstly, the video dataset is generated from one workplace in a factory for only 39 days. In addition, the surveillance video

Table 10 The number of unsafe behaviours that occurred between 24 November and 13 December 2022 after system installation

Date (D/M/Y)	Unauthorized intervention	Opened the panel cover	Carrying overload with forklift	Safe walkway violation
24/11/2022	5	4	3	6
25/11/2022	4	3	2	5
26/11/2022	4	2	2	5
27/11/2022	3	2	2	5
28/11/2022	3	2	2	4
29/11/2022	2	2	0	3
30/11/2022	3	2	1	4
01/12/2022	1	1	2	3
02/12/2022	0	1	3	4
03/12/2022	1	0	2	3
04/12/2022	0	1	2	3
05/12/2022	0	1	3	1
06/12/2022	0	1	1	2
07/12/2022	1	2	1	1
08/12/2022	0	0	1	2
09/12/2022	0	0	0	2
10/12/2022	1	0	0	1
11/12/2022	0	0	0	1
12/12/2022	0	0	0	0
13/12/2022	0	0	0	0
Total	**28**	**24**	**27**	**55**

was collected from only two different cameras or angles from the workplaces. Thus, it is considered that a limited dataset was generated. Secondly, only 4 different classes of unsafe behaviour were identified for the study. However, as there are many unsafe behaviours in real workplaces, the number of classes is considered to be small. Thirdly, when extracting unsafe behaviour frames from the input videos with YOLO v4, it is observed that the number of FPs for some classes is high. To reduce this, it is clear that training with more and more classes of datasets is required. Finally, the pilot study carried out by installing the proposed system in a factory environment can be repeated for a longer period of time.

5 Conclusion

In this study, the unsafe behaviours of employees were identified via deep learning models using real-time surveillance video in industrial factories, and the system was able to instantly warn employees to prevent dangerous behaviours. In the study, the most common unsafe behaviours in the factory were determined by taking the opinions of company authorities and occupational safety experts, and data was collected in the form of surveillance video from security cameras for approximately 39 days. This

data was processed and labelled in accordance with object detection and video understanding studies, and an original dataset was created as part of the study. In the Unsafe-Net architecture proposed in the study, frames that do not contain objects related to unsafe behaviour are detected with the YOLO v4 deep learning model and separated from the videos before processing in the ConvLSTM model. With the proposed Unsafe-Net hybrid deep learning architecture, unsafe behaviours belonging to classes of the Safe Walkway Violation, Unauthorized Intervention, Opened Panel Cover and Carrying Overload with Forklift were detected with an average Acc rate of 95.81%. In addition, with the Unsafe-Net architecture, the average video duration for the videos in the test set was decreased to 1.87 s and the action detection time was reduced to 0.14 s, achieving significant results.

If any of the unsafe behaviours are performed by workers in the workplace, the warning and monitoring system in the infrastructure of the real-time Unsafe-Net-based system sends audible and visual alerts to the environment where the workers are located. In this way, workers are guided by the system to avoid unsafe behaviour. For this purpose, a real-time system based on the proposed Unsafe-Net architecture was installed in a factory for pilot application and the tendencies of employees in unsafe behaviour were monitored for a while. The experimental results showed that the number of unsafe actions in 4 different classes of unsafe behaviour decreased from 518 before the system was installed to 134 after the system was installed, thus reducing the number of unsafe behaviours in the factory by about 75%, which is a significant improvement. As a result, by combining the YOLO v4 deep learning model with CNN and LSTM algorithms, the Unsafe-Net approach to unsafe behaviour detection improves the accuracy of unsafe behaviour video understanding and reduces the action recognition time. In conclusion, the Unsafe-Net model proposed in this study is a robust framework for real-time detection of unsafe behaviours because (1) it has 95.81% accuracy in classifying unsafe behaviours from surveillance videos, (2) the average time for action detection from videos is 0.14 s, (3) the real-time detection speed is increased by reducing the average video duration to 1.87 s, and (4) it reduces the recurrence rate of unsafe behaviours by approximately 75% in factory environment monitoring.

In the YOLO v4 component of the proposed Unsafe-Net architecture, it takes approximately 1.2 s to reduce the duration of a 3-second video to 1 s and make it ready for ConvLSTM. Considering that the average action detection time in the Unsafe-Net architecture is 0.14 s, it takes approximately 1.34 s to detect unsafe actions from videos in real time. This time is considered sufficient for a real-time detection and identification system.

Acknowledgements We thank "Kafaoğlu Metal Plastik Makine San. ve Tic. A.Ş." for permission to collect and use the image/video data used in this study. In addition, for this study, the people in the images were informed about the study and their permission was obtained for the consent form.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This study was financially supported by the Scientific Research Projects Coordinatorship of Bilecik Seyh Edebali University with project number 2019-02.BŞEÜ.01–03.

Data availability The video dataset specially prepared for this study will be shared upon request. In addition, after the completion of some procedures, the dataset will be made publicly available.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhu M, Li G, Huang Q (2024) Recognizing unsafe behaviors of workers by frequency domain features of facial motion information. Multimed Tools Appl 83:8189–8205
2. Cavlak N, Turkoglu A, Kiliccioglu DB, Tokdemir M (2022) Fatal occupational injuries in eastern Turkey between 2000 and 2016. Medicine 11(2):766–769
3. Takala J, Hämäläinen P, Saarela KL, Yun LY, Manickam K, Jin TW, Heng P, Tjong C, Kheng LG, Lim S (2014) Global estimates of the burden of injury and illness at work in 2012. J Occup Environ Hyg 11(5):326–337
4. Chen H, Luo X, Zheng Z, Ke J (2019) A proactive workers' safety risk evaluation framework based on position and posture data fusion. Autom Constr 98:275–288
5. Önal O, Dandıl E (2021) Object detection for safe Working environments using YOLOv4 deep learning model. Avrupa Bilim ve Teknoloji Dergisi 26:343–351
6. Ding L, Fang W, Luo H, Love PE, Zhong B, Ouyang X (2018) A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory. Autom Constr 86:118–124
7. Ceylan H, Ceylan H (2012) Analysis of occupational accidents according to the sectors in Turkey. Gazi Univ J Sci 25(4):909–918
8. Barro-Torres S, Fernández-Caramés TM, Pérez-Iglesias HJ, Escudero CJ (2012) Real-time personal protective equipment monitoring system. Comput Commun 36(1):42–50
9. Onofri L, Soda P, Pechenizkiy M, Iannello G (2016) A survey on using domain and contextual knowledge for human activity recognition in video streams. Expert Syst Appl 63:97–111
10. Sánchez-Caballero A, Fuentes-Jiménez D, Losada-Gutiérrez C (2023) Real-time human action recognition using raw depth video-based recurrent neural networks. Multimed Tools Appl 82(11):16213–16235
11. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
12. Wolf C, Lombardi E, Mille J, Celiktutan O, Jiu M, Dogan E, Eren G, Baccouche M, Dellandréa E, Bichot C-E (2014) Evaluation of video activity localizations integrating quality and quantity measurements. Comput Vis Image Underst 127:14–30
13. Guo S, Luo H, Yong L (2015) A big data-based workers behavior observation in China metro construction. Procedia Eng 123:190–197
14. Luo X, O'Brien WJ, Leite F, Goulet JA (2014) Exploring approaches to improve the performance of autonomous monitoring with imperfect data in location-aware wireless sensor networks. Adv Eng Inform 28(4):287–296
15. Yu Y, Guo H, Ding Q, Li H, Skitmore M (2017) An experimental study of real-time identification of construction workers' unsafe behaviors. Autom Constr 82:193–206
16. Alwasel A, Sabet A, Nahangi M, Haas CT, Abdel-Rahman E (2017) Identifying poses of safe and productive masons using machine learning. Autom Constr 84:345–355
17. Wu H, Zhao J (2018) An intelligent vision-based approach for helmet identification for work safety. Comput Ind 100:267–277
18. Seo J, Han S, Lee S, Kim H (2015) Computer vision techniques for construction safety and health monitoring. Adv Eng Inform 29(2):239–251
19. Wei R, Love PE, Fang W, Luo H, Xu S (2019) Recognizing people's identity in construction sites with computer vision: a spatial and temporal attention pooling network. Adv Eng Inform 42:100981
20. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst 156:117–127

21. Fang Q, Li H, Luo X, Ding L, Luo H, Li C (2018) Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. Autom Constr 93:148–164
22. Wu J, Cai N, Chen W, Wang H, Wang G (2019) Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. Autom Constr 106:102894
23. Chen S, Demachi K (2021) Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. Autom Constr 125:103619
24. Kong T, Fang W, Love PE, Luo H, Xu S, Li H (2021) Computer vision and long short-term memory: learning to predict unsafe behaviour in construction. Adv Eng Inform 50:101400
25. Liu J, Fang W, Love PE, Hartmann T, Luo H, Wang L (2022) Detection and location of unsafe behaviour in digital images: a visual grounding approach. Adv Eng Inform 53:101688
26. Fang W, Love PE, Luo H, Xu S (2022) A deep learning fusion approach to retrieve images of people's unsafe behavior from construction sites. Dev Built Environ 12:100085
27. Fang W, Ding L, Zhong B, Love PE, Luo H (2018) Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach. Adv Eng Inform 37:139–149
28. Son H, Choi H, Seong H, Kim C (2019) Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks. Autom Constr 99:27–38
29. Khan N, Saleem MR, Lee D, Park M-W, Park C (2021) Utilizing safety rule correlation for mobile scaffolds monitoring leveraging deep convolution neural networks. Comput Ind 129:103448
30. Yang M, Wu C, Guo Y, Jiang R, Zhou F, Zhang J, Yang Z (2023) Transformer-based deep learning model and video dataset for unsafe action identification in construction projects. Autom Constr 146:104703
31. Tzutalin (2015) LabelImg. https://github.com/tzutalin/labelImg. Accessed 08.09.2023
32. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934
33. Long X, Deng K, Wang G, Zhang Y, Dang Q, Gao Y, Shen H, Ren J, Han S, Ding E (2020) PP-YOLO: an effective and efficient implementation of object detector. arXiv preprint arXiv:200712099
34. Zheng H, Lin F, Feng X, Chen Y (2020) A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction. IEEE Trans Intell Transp Syst 22(11):6910–6920
35. Pang S, Gao L (2022) Multihead attention mechanism guided ConvLSTM for pixel-level segmentation of ocean remote sensing images. Multimed Tools Appl 81(17):24627–24643
36. Zhang P, Chen L, Li Z, Xing J, Xing X, Yuan Z (2019) Automatic extraction of water and shadow from SAR images based on a multi-resolution dense encoder and decoder network. Sensors 19(16):3576
37. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv Neural Inf Process Syst(NIPS 2015) 28:1–9