Journal of Intelligent & Robotic Systems (2026) 112:5
https://doi.org/10.1007/s10846-025-02341-1

REGULAR PAPER

Intelligent Framework for Human-Robot Collaboration: Dynamic Ergonomics and Adaptive Decision-Making

Francesco Iodice[1] · Elena De Momi[1] · Arash Ajoudani[2]

Received: 14 May 2025 / Accepted: 25 November 2025

Abstract
The integration of collaborative robots into industrial environments has improved productivity but also highlights challenges in operator safety and ergonomics. This paper presents an integrated framework that combines advanced visual perception, continuous ergonomic monitoring, and adaptive Behaviour Tree (BT) decision-making. We adopt a supervisory human-robot collaboration paradigm in which the robot provides temporary, ergonomics-driven assistance only when real-time OWAS assessment indicates hazardous conditions (classes 3-4) and returns execution to the operator as risk subsides, preserving human primacy (97.4% human-led operations in our study). Our modular, scalable approach synthesizes deep learning models, advanced tracking, and dynamic ergonomic assessment. Experimental validation in controlled laboratory settings with industrial-grade sensing and simulation-in-the-loop actuation demonstrates strong performance across multiple dimensions: the perception module achieves 72.4% mAP@50:95; grasp-intention recognition reaches 92.5%; ergonomic risks are classified with 0.081 s mean pose-monitoring latency (95% CI [0.072, 0.093]); and BT policies trigger robotic interventions with 0.07 s decision-layer latency (tick-to-trigger)—approximately 56% faster than a representative prior HRC controller under comparable tasks—while the integrated end-to-end response averages 0.452 s (95% CI [0.283, 0.622]) while maintaining auditable, deterministic safety logic. This comprehensive solution provides a robust platform for enhancing human-robot collaboration in industrial environments by prioritizing ergonomic safety, operational efficiency, and real-time adaptability.

Keywords Human-robot collaboration · Real-time ergonomics · Visual perception · Adaptive decision-making · Intention recognition · Integrated safety framework

1 Introduction

Industrial automation has revolutionized manufacturing processes and enabled the integration of collaborative robotic systems, allowing robots and human operators to work together and enhancing the productivity and adaptability of production lines. This paradigm, termed Human-Robot Collaboration (HRC), has shown considerable productivity enhancements while also presenting intricate issues concerning safety, ergonomics, and flexibility in dynamic operational situations [1].

Recent perspectives in industrial HRI emphasize that safety, performance, and worker well-being must be jointly optimized within Industry 5.0 paradigms, highlighting the central role of human factors and cognitive ergonomics in collaborative cells [2]. In parallel, new experimental evidence shows that collaboration level measurably affects postural stability and biomechanical load during standing tasks, motivating adaptive systems that allocate roles based on real-time human state [3].

The prevention of Work-Related Musculoskeletal Disorders (WMSD) [4], which can arise due to incorrect postures or repetitive strain, is one of the most relevant issues. These disorders not only compromise the health of operators, but also negatively affect company productivity, increasing operating costs. As demonstrated by extensive studies on muscle fatigue in work environments [5], these issues require

✉ Francesco Iodice
francesco.iodice@polimi.it

Elena De Momi
elena.demomi@polimi.it

Arash Ajoudani
arash.ajoudani@iit.it

[1] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy

[2] Human-Robot Interfaces and Physical Interaction (HRI^2) Lab, Istituto Italiano di Tecnologia (IIT), Genoa, Italy

Published online: 09 December 2025

comprehensive monitoring solutions that can adapt to dynamic working conditions and provide timely interventions. Traditional methodologies such as the Ovako Working Posture Analysis System (OWAS) and Rapid Entire Body Assessment (REBA) offer useful tools for identifying ergonomic risks, but are inadequate for the continuous monitoring required in today's complex industrial environments.

We adopt a supervisory HRC paradigm in which the robot provides temporary, ergonomics-driven assistance when real-time OWAS assessment indicates hazardous conditions (classes 3-4), and returns execution to the operator as risk subsides. Human task primacy is thus preserved, with human state used as an implicit control signal for Behaviour Tree gating and role handover/resumption.

To the best of our knowledge, no existing framework in the literature synergistically integrates advanced technologies such as visual detection, continuous ergonomic monitoring, and adaptive decision-making through Behaviour Trees (BT) for collaborative industrial settings. Current approaches address these challenges in isolation or only partially, limiting their practical applicability [6]. This work aims to bridge this gap by proposing an innovative framework that combines state-of-the-art technologies to enhance safety, ergonomics, and efficiency in industrial environments.

The proposed framework, illustrated in Fig. 1, stands out for its non-invasive nature and its ability to adapt to complex and dynamic operational scenarios. It integrates advanced visual detection technologies (YOLO11 with Unscented Kalman Filter tracking and physical properties calculation, and OpenPose with 3D perspective projection) for comprehensive posture and object recognition, an action recognition module using temporal analysis for movement pattern interpretation, a modular decision-making system based on BTs with conditional and composite nodes for dynamic human-robot role allocation and ergonomic-based intervention, and continuous OWAS-based ergonomic assessment methods to prevent WMSD and other physical risk situations. This integration approach addresses the safety challenges highlighted in industrial collaboration studies [1], while providing the adaptability required for varied manufacturing tasks and the non-invasiveness that traditional sensor-based systems [7] often lack. This unified approach enables continuous monitoring and optimization of human-robot interactions, enhancing both safety and overall productivity.

The remainder of this paper is organized as follows: Section 2.8 reviews the current state of research in human-robot collaboration, identifying key limitations and gaps that motivate our integrated approach. Section 3 presents a comprehensive description of the proposed framework architecture, detailing the interaction between visual perception, ergonomic assessment, and adaptive decision-making modules. Section 4 reports experimental validation results across multiple scenarios, demonstrating system performance and comparative analysis. Section 5 addresses ethical considerations, privacy implications, and workforce impact of the proposed monitoring system. Finally, Section 6 summarizes key contributions and outlines directions for future research and industrial deployment.

2 Related Work

In recent years, the field of human-robot collaboration has seen significant developments in several areas, such as visual perception, ergonomic evaluation and decision-making models to ensure safe and efficient robotic intervention. As highlighted by Villani et al. [6], these advancements have created new possibilities for intuitive and safe human-robot interactions in industrial settings, yet the integration of these technologies remains a significant challenge. Despite these advances, current solutions often suffer from limited integration of available technologies or take a static approach to ergonomics and safety. The need for more dynamic and adaptive approaches has been emphasized in multiple studies on work-related musculoskeletal disorders [4, 5], which identify real-time monitoring and adaptation as key factors in preventing occupational injuries.

2.1 Visual Perception and Object Detection

Visual perception constitutes the foundational layer of human-robot collaborative systems, determining the quality of all subsequent processing stages. The evolution of object detection methodologies reveals a critical transition from accuracy-prioritizing two-stage detectors to efficiency-oriented single-stage architectures designed for time-sensitive applications. Early two-stage detection methods, such as R-CNN [9] and Fast R-CNN [10], achieved high accuracy but were computationally inefficient. Faster R-CNN [10] introduced region proposal networks to reduce computational requirements; however, their latency remains relatively high, limiting real-time application potential in industrial contexts (see Table 1).

Single-stage detection frameworks, initiated by the YOLO architecture [11], significantly reduced latency by formulating detection as a regression problem. Subsequent iterations addressed specific limitations: YOLOv2 [12] improved localization accuracy, YOLOv3 [13] enhanced multi-scale detection, and YOLOv4 [14] optimized performance across platforms. More recently, YOLOv8 incorporated segmentation capabilities (YOLO-seg), enabling more detailed scene understanding critical for industrial tasks. YOLOv9 introduced advanced attention mechanisms improving feature representation, whereas YOLOv10 prioritized latency optimization, compromising segmentation capabilities. YOLO11 balanced these trade-offs, offering enhanced segmentation

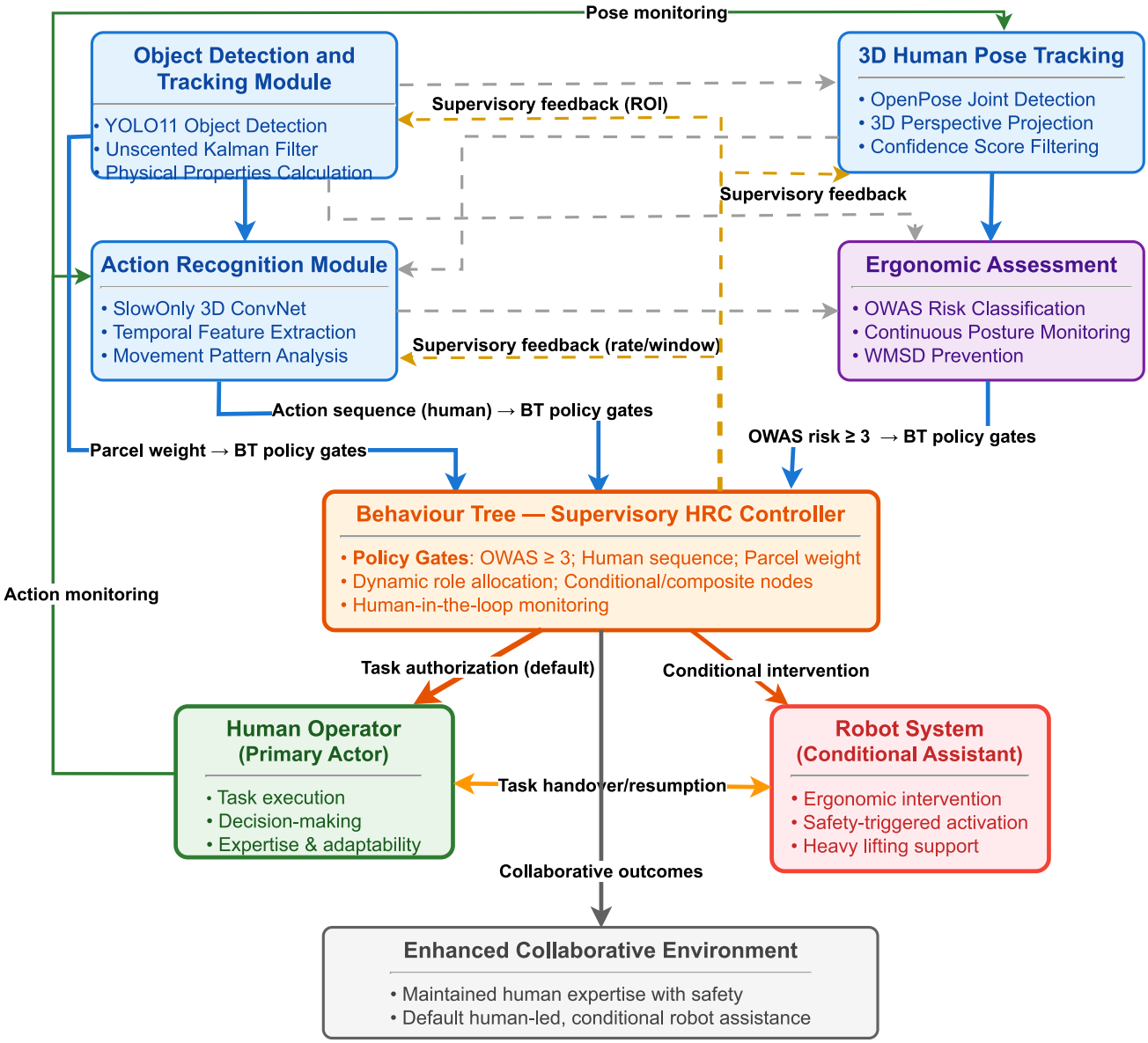

Fig. 1 Supervisory human-robot collaboration framework. Perception modules (blue) integrate YOLO11 object detection, OpenPose 3D pose tracking, and SlowOnly action recognition. Ergonomic assessment (purple) performs continuous OWAS risk classification. The Behaviour Tree controller (orange) implements policy gates based on OWAS thresholds (≥ 3), human action sequences, and parcel weight. Solid arrows indicate data flow; dashed lines represent supervisory feedback. The system maintains human primacy (green) with conditional robot assistance (red) triggered by ergonomic risk, enabling dynamic role allocation while preserving human decision-making authority

performance and low latency, making it highly suitable for real-time HRC applications. Recent developments include YOLO12; however, our comparison indicates only marginal accuracy gains offset by higher computational requirements [8], as detailed in Table 1.

Alternative architectures like EfficientDet [15] explored compound scaling to efficiently balance model complexity and detection accuracy, achieving good speed but lacking native segmentation capabilities. Mask R-CNN [16] extended Faster R-CNN by incorporating instance segmentation, providing high accuracy but at significant computational costs and latency unsuitable for real-time scenarios (Table 1).

Detection performance degradation under real-world industrial conditions such as occlusions and variable lighting was demonstrated by Dodge and Karam [17], and Michaelis et al. [18]. Our experimental validation confirms YOLO11's robustness in variable industrial environments, and its computational efficiency provides practical advantages over more recent alternatives such as YOLO12 for resource-constrained HRC deployments.

Table 1 Representative object-detection architectures on COCO

Model	Params (M)	FLOP (G)	mAPval 50:95 (%)	Latency (ms)	Segmentation
YOLOv8-X	68.2	257.8	53.9	16.86	Yes (YOLO-seg)
YOLOv9e	58.1	192.5	55.6	N/A	Yes (YOLO-seg)
YOLOv10-X	29.5	160.4	54.4	10.70	No
YOLO11-X	56.9	194.9	54.7	11.30	Yes (YOLO-seg)
YOLO12-X	59.1	199.0	55.2	11.79	Yes (YOLO-seg)
EfficientDet (D3)	21.5	46.0	45.4	3.82	No
Faster R-CNN	41.3	87.5	49.8	6.38	No
Mask R-CNN	44.5	170.0	52.0	12.8	Yes

The table summarizes parameters, FLOPs, mAP, latency, and segmentation support across YOLO (v8-v12), two-stage detectors, and hybrid models. YOLO11-X exhibits lower computational load than YOLO12-X (194.9 vs 199.0 GFLOPs) with comparable accuracy (54.7% vs 55.2% mAP), supporting its selection for resource-constrained HRC deployment [8]

2.2 Real-Time Tracking

Real-time tracking is crucial to ensure safety and fluidity in human-robot interactions. The Speed and Separation Monitoring (SSM) method proposed by Marvel et al. [19] monitors the speed and distance between human and robot to prevent collisions, but the lack of integration with advanced visual perception technologies limits its fluidity in complex environments . In contrast, Simple Online and Realtime Tracking (SORT) [20] offers a more responsive solution, using Kalman filtering [21] for rapid detection coupling, but suffers from reduced performance in the presence of occlusions or nonlinear movements, which are common challenges in dynamic industrial environments.

More recent developments, such as DeepSORT and ByteTrack [22], improve robustness under occlusion conditions. Our proposed approach integrates YOLO11 with an Unscented Kalman Filter (UKF) [23], extending traditional Kalman filter principles to effectively manage complex human movements typical of industrial environments.

2.3 Pose Detection and Action Recognition

Human pose and action recognition is essential for improving safety and efficiency in human-robot collaborations. While OpenPose [24] offers powerful pose estimation capabilities, its integration with ergonomic analysis systems remains challenging, as noted in human-robot collaboration surveys [6]. Although OpenPose [24] is widely used for real-time human joint detection, it does not offer dynamic ergonomic evaluation or decision-making based on classified actions.

Contemporary advances in monocular 3D human pose estimation demonstrate mature pipelines for vision-based posture analysis [25], validating the feasibility of non-invasive monitoring in industrial settings. Our framework leverages these capabilities by integrating continuous 3D pose estimation with temporal action recognition, enabling comprehensive movement analysis for safety and ergonomics applications.

In our previous work [26], we demonstrated that the SlowOnly network offers better performance in recognising slow, repetitive movements common in industrial settings than models such as SlowFast [27] and I3D [28].

Lasota et al. [29], in an extensive survey of safety methodologies in human-robot interactions, highlight the importance of human action recognition for collision avoidance. However, most approaches are based on two-dimensional models, which do not adequately address the temporal and spatial complexity of human actions. Our framework, which integrates OpenPose 3D and SlowOnly, offers a more robust and efficient three-dimensional analysis to improve safety and ergonomics.

Cherubini et al. [30] explore collision avoidance in production scenarios, but do not integrate a three-dimensional system for action recognition, nor neural networks optimised for temporal recognition. In contrast, our approach exploits SlowOnly to accurately recognise human actions, improving the management of physical interactions in real time. This attention to temporal patterns in human movement aligns with the safety priorities outlined in collaborative manufacturing studies [1, 6], which identify action prediction as a key component of proactive safety systems. Peternel et al. [31] propose a system for the management of muscle fatigue during human-robot collaborations, but do not exploit networks optimised for the detection of complex and repetitive movements. Our framework fills this gap, improving ergonomic adaptation and robotic response through more sophisticated action recognition capabilities that can detect subtle movement patterns associated with fatigue and ergonomic risk [5].

2.4 Dynamic Role Allocation with Behaviour Trees

Dynamic role allocation is essential for efficient collaboration, especially in industrial settings characterized by rapidly

changing conditions. Traditional methods, such as probabilistic models based on human demonstrations proposed by Rozo et al. [32], often exhibit limitations in adapting flexibly to dynamic environments. Similarly, Finite State Machines (FSMs) face scalability challenges due to rapidly increasing complexity [33], and Markov Decision Processes (MDPs) [34] often require extensive computational resources, limiting their practical application in real-time adaptive scenarios.

Empirical investigations into collaborative work roles demonstrate measurable impacts on cognitive ergonomics, recommending allocation strategies that balance task demands with operator cognitive-physical state [35]. Our BT architecture operationalizes these insights through policies driven by real-time human state assessment (pose/action/OWAS), enabling quantifiable role adaptation.

To overcome these limitations, our framework employs Behaviour Trees (BT) [36], which provide a modular, hierarchical, and adaptable decision-making structure. Behaviour Trees facilitate rapid adaptation of robotic tasks by continuously monitoring operational conditions. Compared to probabilistic models, FSMs, and MDPs, BTs significantly simplify the integration of multiple real-time input streams, enhancing responsiveness and adaptability.

Recent advancements in unified architectures for dynamic role allocation [37] underline the importance of such adaptable frameworks. Merlo et al. [38] have also proposed ergonomic-driven role allocation aimed at reducing musculoskeletal fatigue. Our work further develops this concept by integrating continuous ergonomic monitoring, significantly improving overall safety, flexibility, and operational responsiveness in dynamically evolving industrial environments [29].

2.5 Learning-Based Controllers for HRC and Comparative Rationale

Learning-based control is widely used in HRC through reinforcement learning (RL) and imitation learning from demonstration (IL/LfD). Comprehensive surveys report strong task performance but also practical hurdles for industrial deployment—sample inefficiency, reward design sensitivity, sim-to-real transfer, and safety verification concerns [39–41]. In collaborative settings, IL has enabled close-contact assistance from human demonstrations [42], while RL has explored safety-aware interaction for industrial cells [43, 44], including mutual adaptation in shared autonomy [45]. In contrast, Behaviour Trees (BTs) provide modularity, explicit safety gating, and predictable bounded-latency execution [36]. In our application, BT conditions encode ergonomic thresholds and perception-confidence checks that deterministically trigger safe fallbacks with sub-100 ms-class bounded latency (empirically characterized in Section 4). BTs and learning are complementary: learned skills can populate leaf nodes while the BT orchestrates decision flow and fallbacks, preserving interpretability [46]. Given the requirements for real-time ergonomic monitoring in safety-critical industrial deployment, our framework prioritizes BT-based decision-making for its transparency, modularity, and deterministic behaviour, while recognizing the potential for future hybrid architectures.

2.6 Real-Time Ergonomic Assessment

Traditional ergonomic assessment techniques, such as OWAS [47], RULA [48], REBA [49] and the NIOSH Lifting Equation [50], are based on manual observations and post-hoc analyses, and are unsuitable for continuous and dynamic monitoring of modern industrial environments. These methods have been systematically compared for their effectiveness in identifying potential work-related musculoskeletal disorders [51]. More recent studies, such as that of Ferraguti et al. [52] have proposed a solution to automate ergonomic assessment in HRC collaborations, but such approaches do not always succeed in continuously monitoring the physical condition of operators.

Lightweight Vision Transformer architectures achieve frame-level ergonomic classification at industrial-compatible latencies without explicit skeleton reconstruction [53]. Our framework extends beyond classification by integrating continuous OWAS analysis with immediate Behaviour Tree interventions, creating a closed-loop system for real-time ergonomic risk mitigation.

Our approach overcomes these limitations by integrating dynamic ergonomic analysis with advanced computer vision technologies such as OpenPose, enabling continuous monitoring of postures in real time [54]. This real-time analysis capability addresses a fundamental gap identified in traditional ergonomic evaluation methods [47, 50], which typically require manual observation and cannot adapt to rapidly changing work conditions. This approach not only prevents injuries related to incorrect postures as cataloged in traditional ergonomic assessment methods [48, 49], but also enables an immediate adaptive response, improving safety and reducing operator muscle fatigue through interventions that align with established ergonomic principles [51].

2.7 Constraints of Wearable Sensor Methodologies and Benefits of Computer Vision

Several studies have explored the use of wearable sensors to monitor ergonomic risk and assess operators' movements in work contexts. For example, Santopaolo et al. [55] used inertial sensors and machine learning to classify biomechanical risks related to lifting, while Donisi et al. [56] combined wearable sensors with the NIOSH Lifting Equation to provide a detailed assessment of ergonomic risks in lifting tasks.

Conforti et al. [7] have developed a system based on wearable sensors to monitor operators' movements, demonstrating the effectiveness of these systems for collecting detailed posture and movement data.

Despite their accuracy, wearable sensor-based approaches have significant limitations. Sensors may be invasive, interfering with operators' movements and requiring ongoing management for charging, calibration, and maintenance. Moreover, such systems can increase operational costs, especially in large-scale industrial environments where every worker must be equipped with physical devices.

Comprehensive evaluations of wearable technologies in industrial ergonomics corroborate persistent deployment challenges despite sensing-accuracy advantages [57–59]. Field implementations consistently report operator-acceptance issues, maintenance/calibration overhead, and scalability constraints. In our study, a non-invasive computer-vision pipeline enabled continuous, multi-workstation posture monitoring without per-operator hardware, aligning with industrial deployment constraints.

In contrast, our framework based on artificial vision offers a non-invasive solution for continuous ergonomic assessment. Using technologies such as YOLO11 [60] for object detection and OpenPose for human posture analysis, the system monitors operators' posture and movements in real-time without requiring the use of physical devices. This approach builds upon established object detection principles [11] while overcoming the limitations of traditional monitoring systems, providing a scalable solution that can be deployed across multiple workstations without additional hardware costs per operator. This allows a more natural assessment of ergonomic conditions, dynamically adapting to changes in operators' movements.

In addition, computer vision allows for scalable coverage in complex environments, monitoring multiple operators and robots simultaneously without the need for additional sensors. The system can identify incorrect postures or risky movements and intervene in real-time, reducing the risk of repetitive motion or incorrect posture-related injuries, as confirmed by previous studies on musculoskeletal disorders [5, 50].

In summary, although wearable sensors offer high accuracy, our computer vision-based approach has significant advantages in terms of flexibility, non-invasiveness and scalability, making it particularly suitable for dynamic industrial settings where real-time ergonomic assessment is required. This approach aligns with the evolution of industrial safety paradigms [1, 6] toward more integrated and adaptable solutions that can accommodate the full range of human-robot collaborative scenarios while minimizing disruption to existing workflows.

The landscape of recent ergonomics-centric HRC research demonstrates this diversification through contributions spanning real-time posture optimization via RULA [61], torque minimization during tool operations [62], dynamic role allocation with task-dependent indices [38], and preference-aware optimization frameworks [63]. While these works advance the state of the art in ergonomic enhancement, they typically report domain-specific improvements rather than integrated perception-decision-actuation latency metrics. Given this heterogeneity in evaluation approaches, we adopt Tortora et al. [64] as our primary quantitative baseline, which provides the most directly comparable end-to-end response time metric (0.16s) for lifting/transport scenarios analogous to our validation context.

2.8 Proposed Framework Improvements

Compared to previous works, particularly those that address individual aspects of human-robot collaboration [19, 24, 36] rather than providing an integrated solution, our framework has the following advantages:

- **Continuous Ergonomic Monitoring**: Real-time ergonomic assessments improve proactive injury prevention by dynamically identifying high-risk postures and movements.
- **Enhanced Action Recognition**: Integration of OpenPose 3D and SlowOnly enables accurate detection of subtle and repetitive actions indicative of fatigue or ergonomic risks.
- **Intelligent Task Adaptation**: Utilizing Behaviour Trees facilitates adaptable robot interventions based on real-time conditions, optimizing safety and efficiency.

3 Proposed Framework Architecture

The proposed framework, illustrated in Fig. 2, combines advanced computer vision techniques, human action recognition, and ergonomic evaluation, enabling synergy between human operator and robot in collaborative environments. This architecture enables continuous monitoring of operators' physical conditions, with an adaptive decision-making system managing dynamic task allocation between human and robot via a BT.

The modular design of our framework allows for component-level improvements and adaptation to various industrial contexts without requiring complete system redesign, addressing a key limitation of monolithic approaches identified in previous research [6]. Integration between different technologies, such as computer vision and ergonomic

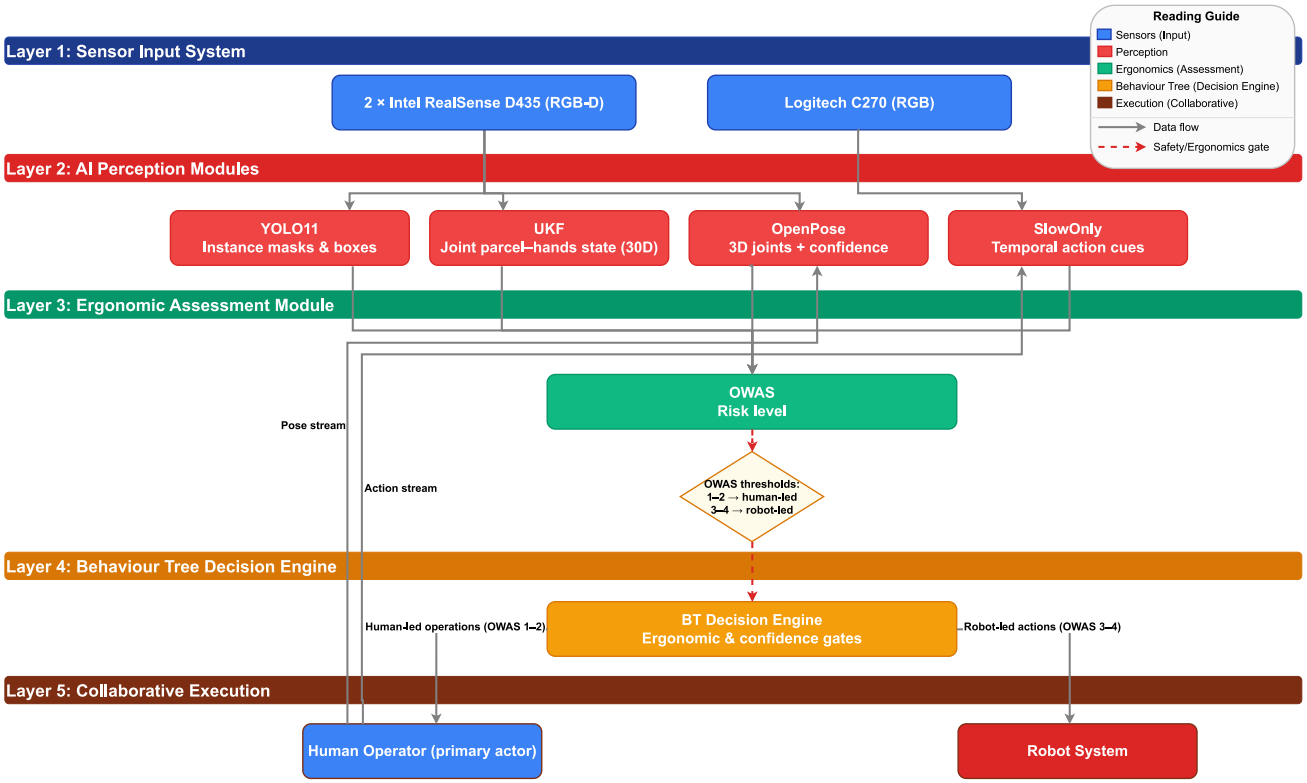

Fig. 2 System architecture overview. Five-layer structure: sensor input (Layer 1: 2× Intel RealSense D435 RGB-D cameras + 1× Logitech C270 RGB webcam), AI perception modules (Layer 2), OWAS ergonomic assessment (Layer 3), Behaviour Tree decision engine (Layer 4), and collaborative execution (Layer 5). OWAS-based allocation assigns classes 1-2 to human-led operations and classes 3-4 to robot-led interventions while maintaining human primacy

evaluation systems, is essential to improving operations safety and efficiency. The following subsections analyze each key module, explaining how they interact with each other and contribute to the overall operation of the framework.

3.1 Object Detection and Tracking Module

This module is responsible for detecting, tracking, and calculating the physical characteristics of objects in the environment, providing essential data for both the decision-making system and ergonomic evaluation.

3.1.1 Object Detection with YOLO11

For object detection and segmentation mask generation, YOLO11 is employed to provide spatial and semantic information essential for safe human-robot collaboration. For each detected instance, the output vector is

$$\mathbf{y} = (x, y, w, h, c, \mathbf{p}, \mathbf{m}),$$

where (x, y) denotes the bounding-box center, (w, h) its dimensions, c the detection confidence, $\mathbf{p}$ the class-probability vector, and $\mathbf{m}$ the binary segmentation mask.

Training optimizes a standard multi-task objective that balances localization, objectness, classification, and segmentation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{box}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}.$$

Here, $\mathcal{L}_{\text{box}}$ penalizes localization error, $\mathcal{L}_{\text{conf}}$ models objectness, $\mathcal{L}_{\text{cls}}$ addresses class prediction, and $\mathcal{L}_{\text{seg}}$ measures mask accuracy; λ_{conf}, λ_{cls}, and λ_{seg} weight the corresponding terms. Consistent with canonical YOLO11 formulations, default loss composition and weights are used without heuristic reweighting or task-specific tuning; training is end-to-end via backpropagation to favor reproducibility and maintain interoperability with downstream modules.

This formulation supports accurate instance separation in cluttered layouts and under partial occlusions, providing stable inputs to the subsequent tracking and ergonomic-evaluation pipelines.

3.1.2 Physical Property Calculation for Ergonomic Assessment

Using the data provided by YOLO11, the system performs the parallel calculation of the area, volume, and volumetric weight of objects, which are key inputs for ergonomic evaluation using the OWAS methodology.

Area The area of the object in the 2D projection is calculated as:

$$A_{\text{object}} = w \cdot h$$

where w and h are the width and height of the bounding box.

Volume The volume of the object is estimated by multiplying the area by the average depth D_{object} detected in the bounding box:

$$V_{\text{object}} = A_{\text{object}} \cdot (D_{\max} - D_{\text{object}})$$

where $D_{\max}$ represents the maximum distance and D_{object} is the average depth detected by Intel RealSense D435 camera.

Volumetric Weight Once the volume is estimated, the volumetric weight, P_{vol}, is calculated by dividing the volume by a standard conversion factor:

$$P_{\text{vol}} = \frac{V_{\text{object}}}{\text{conversion factor}}$$

where conversion factor is a standard value of 6000 used in logistics companies to calculate volumetric weight.

This input is critical to the calculation of the OWAS score, where the weight of objects handled by operators is one of the key factors in assessing ergonomic risk.

3.1.3 Unscented Kalman Filter for Movement Tracking

In parallel with calculating the physical properties of objects, the UKF is used to track the positions and movements of objects and operators' hands in real time. The UKF is particularly suitable for handling nonlinear dynamics typical of complex movements in industrial environments.

Tracking is handled through a state vector $\mathbf{x} \in \mathbb{R}^{30}$ that includes the positions, velocities, and orientations of objects and hands:

$$\mathbf{x} = \begin{bmatrix} \mathbf{p}_{\text{parcel}} \\ \mathbf{v}_{\text{parcel}} \\ \mathbf{q}_{\text{parcel}} \\ \mathbf{p}_{\text{hand\_left}} \\ \mathbf{v}_{\text{hand\_left}} \\ \mathbf{q}_{\text{hand\_left}} \\ \mathbf{p}_{\text{hand\_right}} \\ \mathbf{v}_{\text{hand\_right}} \\ \mathbf{q}_{\text{hand\_right}} \end{bmatrix}$$

Here, $\mathbf{p} \in \mathbb{R}^3$ denotes the position in 3D space, $\mathbf{v} \in \mathbb{R}^3$ the linear velocity in 3D space, and $\mathbf{q} \in \mathbb{R}^4$ the orientation expressed in quaternions to avoid the ambiguities that arise in the use of Euler angles. This unified 30-dimensional state preserves cross-correlations between parcel and hands during close manipulation, supporting identity consistency through brief occlusions typical of collaborative tasks. Relative to independent filters, the coupled formulation centralizes data association and confidence management across perception streams, simplifying integration with the decision layer.

The state update is conducted through a transition function $\mathbf{f} : \mathbb{R}^{30} \to \mathbb{R}^{30}$ that predicts the evolution of the variables over time. This function takes into account the linear and rotational dynamics of the system and is expressed as follows:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \mathbf{p}_{\text{parcel}} + \mathbf{v}_{\text{parcel}} \cdot dt \\ \mathbf{v}_{\text{parcel}} \\ \mathbf{q}_{\text{parcel}} + \frac{1}{2}\mathbf{q}_{\text{parcel}} \otimes \Omega(\boldsymbol{\omega}_{\text{parcel}}) \cdot dt \\ \mathbf{p}_{\text{hand\_left}} + \mathbf{v}_{\text{hand\_left}} \cdot dt \\ \mathbf{v}_{\text{hand\_left}} \\ \mathbf{q}_{\text{hand\_left}} + \frac{1}{2}\mathbf{q}_{\text{hand\_left}} \otimes \Omega(\boldsymbol{\omega}_{\text{hand\_left}}) \cdot dt \\ \mathbf{p}_{\text{hand\_right}} + \mathbf{v}_{\text{hand\_right}} \cdot dt \\ \mathbf{v}_{\text{hand\_right}} \\ \mathbf{q}_{\text{hand\_right}} + \frac{1}{2}\mathbf{q}_{\text{hand\_right}} \otimes \Omega(\boldsymbol{\omega}_{\text{hand\_right}}) \cdot dt \end{bmatrix}$$

In this formulation:

- $\mathbf{p} \in \mathbb{R}^3$ is the updated position as a function of linear velocity $\mathbf{v}$ and time interval $dt \in \mathbb{R}^+$,
- $\mathbf{v} \in \mathbb{R}^3$ represents velocity, held constant within the update interval dt. This choice is justified by the high update rate of the system (30 Hz), which makes speed changes between two consecutive updates negligible,
- $\mathbf{q} \in \mathbb{R}^4$ represents the updated orientation using the quaternion product $\otimes$ with the angular velocity $\boldsymbol{\omega} \in \mathbb{R}^3$ and time interval dt,
- $\Omega : \mathbb{R}^3 \to \mathbb{R}^4$ is a function that converts angular velocity into quaternion derivative format.

The predicted estimate of the state, after updating, is given by the transition function:

$$\hat{\mathbf{x}}_{k+1} = \mathbf{f}(\mathbf{x_k})$$

where $\hat{\mathbf{x}}_{k+1} \in \mathbb{R}^{30}$ represents the predicted state at time step $k+1$, and $\mathbf{x}_k \in \mathbb{R}^{30}$ is the current state at time step k.

State Correction After estimating the future state with the transition function, the UKF compares this estimate with actual observations obtained from depth sensors, such as the Intel RealSense D435 camera. The measurement function $\mathbf{h} : \mathbb{R}^{30} \rightarrow \mathbb{R}^{21}$ maps the estimated state onto observables, i.e., the position and orientation of objects and hands:

$$\mathbf{h}(\mathbf{x}) = \begin{bmatrix} \mathbf{p}_{\text{parcel}} \\ \mathbf{q}_{\text{parcel}} \\ \mathbf{p}_{\text{hand\_left}} \\ \mathbf{q}_{\text{hand\_left}} \\ \mathbf{p}_{\text{hand\_right}} \\ \mathbf{q}_{\text{hand\_right}} \end{bmatrix}$$

The measurement vector $\mathbf{z}_k \in \mathbb{R}^{21}$ consists of the direct observations of positions and orientations, excluding velocities which are not directly measurable. The measurement predicted at time k, $\hat{\mathbf{z}}_k \in \mathbb{R}^{21}$ is obtained by applying the measurement function to the state predicted at time k:

$$\hat{\mathbf{z}}_k = \mathbf{h}(\hat{\mathbf{x}}_k)$$

Subsequently, the new actual observations $\mathbf{z}_k$ are compared with those predicted estimates. The error between the predicted estimate $\hat{\mathbf{z}}_k$ and the actual observations $\mathbf{z}_k$ is used to update and correct the estimated state, via the Kalman gain $K \in \mathbb{R}^{30\times 21}$. The final correction of the state $\mathbf{x}_{k+1}$ is made with the following equation:

$$\mathbf{x}_{k+1} = \hat{\mathbf{x}}_{k+1} + K(\mathbf{z}_k - \hat{\mathbf{z}}_k)$$

The Kalman gain K is computed internally by the UKF algorithm and optimally weights the relative importance of the model prediction versus the new measurements, based on their respective uncertainties.

Covariance Matrices The stability and accuracy of the system strongly depend on a proper definition of the covariance matrices. The process noise covariance matrix $\mathbf{Q} \in \mathbb{R}^{30\times 30}$ represents the uncertainty in the system model, while the observation noise matrix $\mathbf{R} \in \mathbb{R}^{21\times 21}$ handles the uncertainty associated with sensor measurements:

$$\mathbf{Q} = \text{diag}([100, \ldots, 100]) \quad \text{and} \quad \mathbf{R} = \text{diag}([0.01, \ldots, 0.01])$$

The initial state covariance matrix $\mathbf{P} \in \mathbb{R}^{30\times 30}$, defined as a scaled identity matrix, also plays a key role in the accuracy of the first estimates:

$$\mathbf{P} = 0.01 \cdot \mathbf{I}_{30}$$

where $\mathbf{I}_{30} \in \mathbb{R}^{30\times 30}$ is the identity matrix of dimension 30.

The UKF uses Merwe scaled sigma points with standard parameterization ($\alpha = 10^{-3}$, $\beta = 2.0$, $\kappa = 0$) to capture nonlinearities, with numerical regularization ensuring stability. The correct selection of these parameters is essential to ensure that the UKF system maintains stability and accuracy even in the presence of uncertainties in the process or measurements, minimizing estimation error throughout the entire operational cycle.

This joint tracking approach provides temporally consistent motion estimates that support reliable hand-parcel association for intention recognition and reduce spurious safety interventions caused by transient detection failures.

3.2 3D Human Pose Tracking Module

The 3D Human Pose Tracking Module is responsible for detecting and analyzing the operator's posture in real-time, providing crucial data for ergonomic evaluation.

3.2.1 Pose Detection with OpenPose

OpenPose detects the joints of the human body in 3D using a perspective projection [65]. Each i-th joint is represented by the coordinates (x, y, z) and a confidence score c:

$$\mathbf{J} = \{(x_i, y_i, z_i, c_i)\}, \; i = 1, 2, \ldots, N$$

The collected data are subsequently analyzed by the SlowOnly module.

3.3 Action Recognition Module

The Action Recognition Module analyzes the operator's movements over time to identify specific actions and patterns, contributing to both ergonomic assessment and decision-making.

3.3.1 Temporal Analysis with SlowOnly

The SlowOnly model, a specialized three-dimensional convolutional network (3D ConvNet), analyzes the slow, repetitive movements of operators. The input is a sequence of video frames F_{in}, sampled at regular temporal intervals:

$$F_{\text{in}} = \left[I_t, I_{t+\Delta t}, I_{t+2\Delta t}, \ldots, I_{t+T\Delta t}\right]$$

Where:

- $I_t \in \mathbb{R}^{H \times W \times C}$ represents the video frame at time t, with H being the height, W the width, and C the number of channels of the frame
- $\Delta t \in \mathbb{R}^{+}$ is the temporal sampling interval between consecutive frames
- $T \in \mathbb{N}$ is the total number of frames in the sequence

Three-dimensional convolution applied to this sequence extracts both spatial and temporal features:

$$F_{\text{out}} = \text{Conv3D}(F_{\text{in}}, K, S)$$

Where:

- $F_{\text{out}} \in \mathbb{R}^{H' \times W' \times D \times N}$ is the output feature tensor
- $K \in \mathbb{R}^{k_h \times k_w \times k_t \times C \times N}$ represents the 3D convolutional kernels
- $S \in \mathbb{N}^3$ denotes the stride parameters (s_h, s_w, s_t)

The temporal features extracted by SlowOnly are then forwarded to the Ergonomic Assessment Module for evaluation of the operator's movements using the OWAS methodology.

3.4 Ergonomic Assessment Module

The Ergonomic Assessment Module integrates data from the previous modules to evaluate the operator's posture and working conditions according to established ergonomic standards.

3.4.1 Ovako Working Posture Analysis System

The Ovako Working Posture Analysis System (OWAS) classifies operators' postures into risk categories, allowing the system to monitor and intervene when postures are detected that could lead to musculoskeletal injuries.

Figure 3 illustrates the OWAS evaluation matrix used in our framework. This standardized tool provides a systematic approach to assessing ergonomic risks through a color-coded

Action		Digit
Back	Straight	1
	Bent	2
	Twisted	3
	Bent and twisted	4
Arms	Both arms below shoulder level	1
	One arm at or above shoulder level	2
	Both arms at or above shoulder level	3
Legs	Sitting	1
	Standing on two straight legs	2
	Standing on one straight leg	3
	Standing or squatting on two bent legs	4
	Standing or squatting on one bent leg	5
	Kneeling	6
	Walking	7
Load	Less or equal to 10 Kg	1
	Greater than 10 Kg and less or equal to 20 Kg	2
	Greater than 20 Kg	3

Back	Arms	Legs																					
		1			2			3			4			5			6			7			
		Load			**Load**			**Load**			**Load**			**Load**			**Load**			**Load**			
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	1	1	1	
	2	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	1	1	1	
	3	1	1	1	1	1	1	1	1	1	2	2	3	2	2	3	1	1	1	1	1	2	
2	1	2	2	3	2	2	3	2	2	3	3	3	3	3	3	3	2	2	2	2	3	3	
	2	2	2	3	2	2	3	2	3	3	3	4	4	3	4	4	3	3	4	2	3	4	
	3	3	3	4	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	2	3	4	
3	1	1	1	1	1	1	1	1	1	2	3	3	3	4	4	4	1	1	1	1	1	1	
	2	2	2	3	1	1	1	1	1	2	4	4	4	4	4	4	3	3	3	1	1	1	
	3	2	2	3	1	1	1	2	3	3	4	4	4	4	4	4	4	4	4	1	1	1	
4	1	2	3	3	2	2	3	2	2	3	4	4	4	4	4	4	4	4	4	2	3	4	
	2	3	3	4	2	3	4	3	3	4	4	4	4	4	4	4	4	4	4	2	3	4	
	3	4	4	4	2	3	4	3	3	4	4	4	4	4	4	4	4	4	4	2	3	4	

Interpretation of the result
1 - No action required
2 - Corrective actions required in the near future
3 - Corrective actions should be done as soon as possible
4 - Corrective actions for improvement required immediately

Fig. 3 OWAS (Ovako Working Posture Analysis System) evaluation matrix. The left section displays the coding system for back (1-4), arms (1-3), legs (1-7), and load (1-3) postures. The right section presents the action category matrix with color-coded risk levels (green: no action required; yellow: corrective actions required in the near future; orange: corrective actions should be done as soon as possible; red: corrective actions for improvement required immediately) based on combinations of back, arms, legs, and load values

matrix that indicates different levels of necessary intervention based on various posture combinations.

Classification of Postures The classification of operator postures, as shown in Fig. 3, is based on a systematic coding system that evaluates four key body segments:

1. **Posture of the back (1-4)**:
   - (1) Straight
   - (2) Bent
   - (3) Twisted
   - (4) Bent and twisted
2. **Posture of the arms (1-3)**:
   - (1) Both arms below shoulder level
   - (2) One arm at or above shoulder level
   - (3) Both arms at or above shoulder level
3. **Position of legs (1-7)**:
   - (1) Sitting
   - (2) Standing on two straight legs
   - (3) Standing on one straight leg
   - (4) Standing or squatting on two bent legs
   - (5) Standing or squatting on one bent leg
   - (6) Kneeling
   - (7) Walking
4. **Weight lifted (1-3)**:
   - (1) Less than or equal to 10 kg
   - (2) Greater than 10 kg and less than or equal to 20 kg
   - (3) Greater than 20 kg

The OWAS evaluation process combines these four codes to determine a risk level using the standardized evaluation matrix shown in Fig. 3. This matrix cross-references specific combinations of postures and load to assign one of four action categories:

- **Action Category 1**: No action required - normal posture without harmful effect on the musculoskeletal system.
- **Action Category 2**: Corrective actions required in the near future - posture has some harmful effect on the musculoskeletal system.
- **Action Category 3**: Corrective actions should be done as soon as possible - posture has a distinctly harmful effect on the musculoskeletal system.
- **Action Category 4**: Corrective actions for improvement required immediately - posture has an extremely harmful effect on the musculoskeletal system.

OWAS Score Calculation While the standard OWAS methodology determines the action category directly through the evaluation matrix, our implementation calculates a continuous weighted score to enable more nuanced ergonomic assessment. The formula for calculating this weighted OWAS score is:

$$P_{\text{owas}} = \alpha \cdot P_{\text{back}} + \beta \cdot P_{\text{arm}} + \gamma \cdot P_{\text{leg}} + \delta \cdot P_{\text{weight}}$$

where:

- $P_{\text{back}}, P_{\text{arm}}, P_{\text{leg}}, P_{\text{weight}}$ are the numerical codes assigned to each postural category (ranging from 1-4, 1-3, 1-7, and 1-3 respectively)
- α, β, γ, δ are weighting coefficients that modulate the relative contribution of each postural factor

The weighting coefficients (α, β, γ, δ) are derived from ergonomic research on work-related musculoskeletal disorders. In the original OWAS methodology [47], these weights reflect the relative impact of different body segments on overall postural strain. The exact values are calibrated based on:

- The prevalence of disorders in different body regions in industrial settings
- The biomechanical load associated with each posture type
- The time sustainability of various postures before discomfort onset
- The recovery time needed after maintaining specific postures

In our implementation, the primary intervention decisions rely directly on the standard OWAS evaluation matrix (Fig. 3), which determines action categories through validated lookup tables rather than explicit numerical computation. While the weighted score formula is presented for completeness and could enable future optimization studies, our current system uses the matrix-based classification for all intervention decisions, ensuring compliance with the established OWAS methodology.

The calculated weighted score is then mapped to the four action categories using threshold values that align with the OWAS evaluation matrix shown in Fig. 3. This approach allows our system to continuously monitor ergonomic risk levels and trigger appropriate interventions when hazardous postures are detected.

Framework Integration of Ergonomic Assessment The OWAS methodology is integrated directly into the framework

to monitor operators' posture in real time, establishing a computational pathway between sensor data and ergonomic risk assessment. The integration process consists of the following steps:

1. **Skeletal data acquisition**: OpenPose detects the 3D coordinates and confidence scores of key body joints, while SlowOnly analyzes temporal movement patterns to identify sustained postures.
2. **Posture mapping**: The acquired skeletal data is mapped to the corresponding OWAS coding system (back, arms, legs), while load parameters are derived from the volumetric calculations performed by the Visual Perception Module.
3. **OWAS score computation**: The weighted score is calculated using the formula previously described, with the calibrated coefficients reflecting the biomechanical impact of each postural component.
4. **Dynamic intervention**: When the computed ergonomic risk exceeds the predefined thresholds corresponding to OWAS action categories, the system initiates appropriate intervention strategies to mitigate the identified hazards.

This approach transforms the traditional manual OWAS assessment into an automated, continuous monitoring system capable of detecting ergonomic hazards in real-time industrial scenarios.

3.5 Decision-Making with Behaviour Trees

Within the proposed framework, dynamic role allocation between human operator and robot is managed through a BT. This decision structure allows the system to adapt robotic behaviour based on operational conditions, operator actions, and real-time sensory inputs. Unlike traditional approaches, such as finite-state diagrams or Markovian decision models, the BT offers greater flexibility and modularity, which is particularly relevant in dynamic industrial settings where human-robot collaboration requires fast and accurate decisions.

3.5.1 Structure and Execution of the Behaviour Tree

The BT implemented in the framework is hierarchically structured and composed of different types of nodes:

- **Action Nodes**: They perform concrete operations, such as "grabbing an object" or "moving to a destination". These nodes allow the robot to perform the required actions based on the detected conditions.
- **Conditional Nodes**: They evaluate whether or not the action can proceed by monitoring parameters such as operator biomechanical fatigue or completion of a sequence of human movements. For example, a conditional node might check whether the lifted package has an acceptable weight for handling.
- **Composite Nodes**: These nodes combine multiple actions and conditions, creating complex logic flows. Composite nodes used include:
  - *Sequence*: Nodes are executed in sequential order until all actions are completed correctly. If a node fails, execution stops.
  - *Fallback*: Nodes are executed sequentially until one succeeds. This node type is particularly useful for handling alternative situations where the primary action is unavailable or fails.

As illustrated in Fig. 4, the BT structure begins with a ROOT Fallback node at the top level, which provides the initial decision-making capability. This root node connects to a Main Sequence node that orchestrates the operational flow. From the Main Sequence, the tree branches into five parallel nodes that constitute the core decision-making components:

- *Go Homing* (Action): Verifies that the robot is in the correct starting position before initiating operations.
- *Check Human Sequence Completed* (Condition): Evaluates whether the operator has completed their intended sequence of movements.
- *Check Biomechanics Fatigue* (Condition): Monitors operator biomechanical fatigue based on ergonomic data collected in real time.
- *Check Parcel Weight* (Condition): Evaluates whether the weight of the package falls within acceptable parameters for handling.
- *Action Sequence Fallback*: A composite fallback node that determines the appropriate action path based on the outcome of preceding conditions.

The execution process begins at the ROOT Fallback node and proceeds through the Main Sequence. Based on the evaluation of the conditional nodes, the tree implements two primary execution pathways:

- The first pathway through the *No Walk With Parcel Sequence* directs robot-led actions when the operator's ergonomic condition indicates risk. This sequence leads to actions including *Move To Parcel*, *Move Arm to Grasp*, *Correct Grip*, *Grasp Parcel*, *Lift Parcel*, *Move to Container*, and *Release Parcel*.
- The second pathway through the *Walk With Parcel Sequence* manages human-led operations with robot support when ergonomic conditions are favorable. This

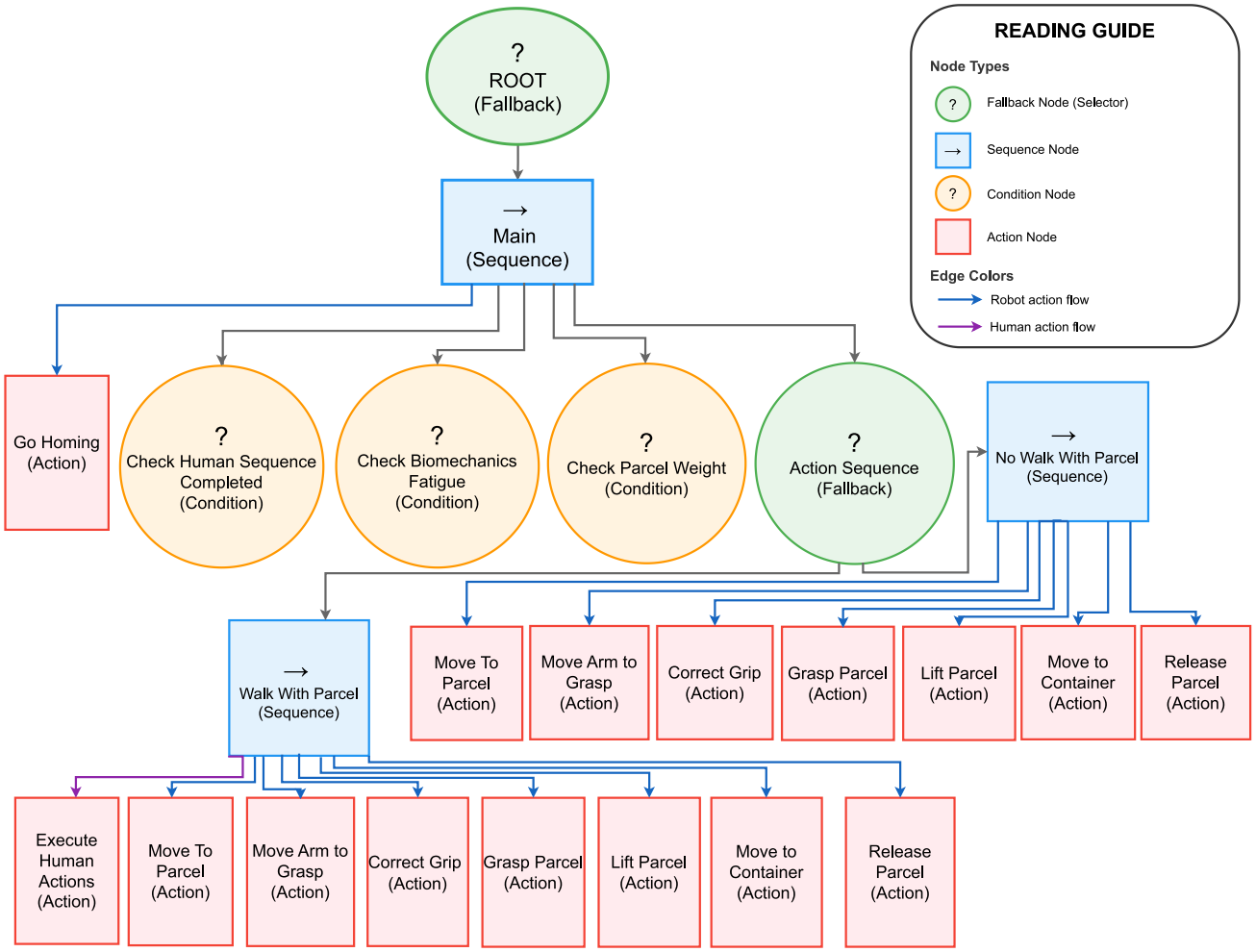

Fig. 4 Behaviour Tree for dynamic role allocation. ROOT fallback branches to parallel condition nodes evaluating human sequence completion, biomechanics fatigue, and parcel weight. Action Sequence directs to human-led (Walk With Parcel) or robot-led (No Walk With Parcel) operations. Edge colors distinguish robot (blue) from human (purple) action flows

includes monitoring through *Execute Human Actions* and providing assistance as needed.

When the OWAS assessment indicates elevated ergonomic risk (for example, due to improper posture or fatigue detected via the Check Biomechanics Fatigue node), the BT prioritizes robot intervention. Conversely, if the ergonomic assessment indicates favorable conditions for human operation, the tree allows for human-led execution with robot monitoring. This dynamic allocation ensures optimal workload distribution based on real-time ergonomic conditions.

Industrial robustness requires systematic handling of perception uncertainties and sensor degradation. The framework addresses these challenges through confidence-aware conditional nodes that monitor pose estimation quality and trigger safe fallbacks when degradation occurs, such as during prolonged occlusions or adverse lighting conditions. The integrated UKF provides temporal consistency by maintaining hand and parcel trajectories through brief detection gaps, while sustained confidence loss automatically activates safe operational modes until reliable observations recover. This layered approach ensures continuous operation while preventing actions based on unreliable sensory input, though quantitative characterization of confidence thresholds and recovery dynamics represents important future work for deployment-grade robustness.

3.5.2 Advantages of the Behaviour Tree Approach

Beyond generic modularity, the implemented BT provides deployment-oriented benefits within this framework.

- **Deterministic safety logic:** Conditional nodes evaluate OWAS-based ergonomic thresholds and perception-

confidence checks at each tick, triggering immediate fallbacks when risk or uncertainty rises. This explicit, threshold-driven gating supports the rapid decision-making capabilities evidenced in Section 4.

- **Perception-aware continuity:** BT conditions leverage temporally consistent estimates from the unified UKF state (see Section 3.1.3), which maintains hand-parcel association through brief detection gaps. This design reduces spurious safety interventions caused by transient perception failures.
- **Conflict-free multimodal integration:** Pose, action, object, and load assessment streams (Sections 3.1 to 3.4) are combined through structured conditional flows rather than additional arbitration heuristics. This architectural choice maintains interpretable decision logic while supporting system responsiveness.
- **Industrial deployment readiness:** Because transitions are driven by explicit conditions with clear decision paths, the system remains auditable and amenable to debugging during operation, facilitating integration in regulated industrial environments without relying on opaque policies.

4 Experiments and Results

To evaluate the overall effectiveness of the proposed framework in human-robot collaboration, experiments were designed and conducted in a controlled laboratory environment and in simulation. The laboratory experiments replicated realistic operational scenarios, focusing on ergonomic monitoring, action detection, visual perception, and overall integration of the framework components. In parallel, the Behaviour Tree (BT)-based decision-making module was tested in simulation to analyze its dynamic and adaptive management capability in a wide range of complex scenarios that are difficult to replicate in the laboratory.

4.1 Experimental Setup

All experiments were conducted using a standardized hardware and software infrastructure to ensure consistency across different test scenarios:

4.1.1 Hardware Components

1. **Visual Sensing**: Two Intel RealSense D435 depth cameras were deployed for object detection/tracking and human pose estimation respectively. An additional Logitech C270 webcam was used for capturing video streams for action recognition analysis.
2. **Robotic System**: The experiments utilized a collaborative robotic arm (Universal Robots UR16e) mounted on the Summit XL mobile platform, configured to respond to commands from the decision-making system.
3. **Computing Resources**: The framework operated on a distributed network comprising four workstations in a ROS-based master-slave configuration. Three workstations were dedicated to vision processing, equipped with different GPUs (NVIDIA GTX 970, NVIDIA GTX 1080Ti, and NVIDIA RTX A3000) to handle various perception tasks. The fourth workstation was the onboard computer of the mobile platform, which managed robot control operations.

4.1.2 Software Framework

The system ran on Ubuntu 20.04 with ROS Noetic providing the communication infrastructure between components. The YOLO11 detection module, OpenPose tracking, and SlowOnly action recognition algorithms were deployed across the vision processing nodes according to their computational requirements. Neural network models were initially trained offline, while experiment execution utilized the distributed GPU setup for real-time inference.

This heterogeneous computing architecture was designed to maintain system responsiveness while processing multiple data streams, enabling the ergonomic assessment and decision-making modules to operate effectively in dynamic conditions.

4.2 Experiment: Grasping Action Detection and Volumetric Weight Estimation with Variable Size Parcels

In this preliminary phase of the framework, the main objectives were the recognition of grasping intention and estimation of volumetric weight of parcels of varying sizes. The dataset used included 15,160 images annotated with the Segment Anything Model (SAM), ensuring accurate segmentation of hands and parcels into diverse configurations. Images were divided into training (80%), validation (10%) and testing (10%), ensuring a balanced distribution for robust evaluation.

Four segmentation models were analyzed: YOLOv8x-seg, YOLOv9e-seg, YOLO11x-seg, and YOLOv12x-seg, trained with the same configuration parameters and evaluated on two main classes, "*parcel*" and "*hand*".

For grasping intention recognition, 40 trials were conducted on dynamic sequences, combining the segmentation model with the Unscented Kalman Filter (UKF). For volumetric estimation, 4 parcels of varying sizes and shapes were tested, each in 10 trials for a total of 40 measurements. The estimated volumetric weights, obtained from the segmented masks and depth data provided by the Intel RealSense D435 sensor, were compared with the actual weights.

Table 2 Comparative analysis of object detection performance across YOLO model generations (Box metrics, validation on `best.pt`)

Model Architecture	mAP@50 (%)	mAP@50:95 (%)
YOLOv8x-seg	77.6	72.0
YOLOv9e-seg	77.7	**72.4**
YOLO11x-seg	**77.8**	**72.4**
YOLOv12x-seg	77.0	71.6

Note: Mean Average Precision metrics at IoU threshold of 0.5 (mAP@50) and across multiple IoU thresholds from 0.5 to 0.95 (mAP@50:95) demonstrate the comparative performance trajectory across model generations. Best results are highlighted in bold. All values are *Box* metrics; mask ("m...") metrics are not reported

The performance of the segmentation models was evaluated using mAP@50 (mean Average Precision with an IoU threshold of 0.5) and mAP@50:95 (mean Average Precision averaged over multiple IoU thresholds from 0.5 to 0.95 with a step size of 0.05). Precision, defined as the ratio of true positives to the sum of true positives and false positives, and recall, defined as the ratio of true positives to the sum of true positives and false negatives, were calculated for class-specific analysis alongside mAP values.

For volumetric weight estimation, percentage errors were calculated using the formula:

$$E_p = \frac{|P_{v,\mathrm{st}} - P_{v,\mathrm{re}}|}{P_{v,\mathrm{re}}} \times 100$$

where $P_{v,\mathrm{st}}$ and $P_{v,\mathrm{re}}$ represent estimated and actual volumetric weights, respectively.

As shown in Table 2, YOLO11x-seg achieved the highest mAP@50 of 77.8% and tied for best mAP@50:95 of 72.4% with YOLOv9e-seg. YOLOv8x-seg achieved 77.6% mAP@50 and 72.0% mAP@50:95. YOLOv12x-seg, despite being the most recent architecture, achieved 77.0% mAP@50 and 71.6% mAP@50:95, showing marginally lower performance compared to YOLO11x-seg.

Table 3 reveals class-specific performance differences. For parcel detection, YOLOv9e-seg achieved the highest mAP@50:95 of 86.9%, closely followed by YOLO11x-seg (86.8%) and YOLOv8x-seg (86.7%), while YOLOv12x-seg achieved 86.1%. For hand detection, YOLO11x-seg demonstrated superior performance with 58.0% mAP@50:95 and the highest recall of 61.1%, while precision ties with YOLOv8x-seg at 89.8%; YOLOv9e-seg and YOLOv12x-seg achieved 57.9% and 57.1% mAP@50:95, respectively.

The observed hand detection recall of 61.1% across all models represents a technical limitation that warrants discussion. This constraint stems from the inherent challenges of hand detection in industrial scenarios: high variability in hand poses during object manipulation, frequent occlusions when hands interact with parcels, and the relatively small size of hand features compared to parcels in the dataset. While YOLO11x-seg achieved the best hand recall of 61.1%, this limitation could potentially affect system responsiveness in scenarios requiring immediate hand detection. However, the framework mitigates this through multiple temporal integration mechanisms: the UKF maintains hand trajectories at 30 Hz (Section 3.1.3), the ergonomic assessment operates at 14.99 Hz with inter-arrival 0.066 s latency (Fig. 10), and SlowOnly analyzes 8-frame windows. These combined mechanisms enable 92.5% grasping intention recognition accuracy despite the moderate frame-level recall, demonstrating that temporal integration effectively compensates for instantaneous detection limitations.

For grasping intention recognition, the system correctly identified grasping in 37 cases out of 40, with an overall accuracy of 92.5%. The associated 95% Wilson confidence interval is [80.1%, 97.4%] (Fig. 5), corroborating the reliability of intention gating for downstream decisions.

The overall mean percent error for volumetric weight estimation over all trials was calculated as 17.58%, with greater variability observed in the extreme size parcels, attributable to the sensor's limitations in terms of spatial resolution and surface reflectivity. Because OWAS uses discrete load classes

Table 3 Class-specific performance (Box metrics, validation on `best.pt`)

Class	Metrics	YOLOv8x-seg	YOLOv9e-seg	YOLO11x-seg	YOLOv12x-seg
Parcel	Precision (%)	**88.0**	85.6	84.0	85.6
	Recall (%)	79.3	80.9	**81.1**	79.2
	mAP@50 (%)	**89.4**	**89.4**	89.3	88.8
	mAP@50:95 (%)	86.7	**86.9**	86.8	86.1
Hand	Precision (%)	**89.8**	89.6	**89.8**	89.1
	Recall (%)	59.9	60.9	**61.1**	59.3
	mAP@50 (%)	65.7	66.1	**66.3**	65.3
	mAP@50:95 (%)	57.3	57.9	**58.0**	57.1

Box metrics parsed from validation logs (first P/R/mAP block after `Instances`); mask metrics are not used. Best per row in bold (ties bolded). For hand detection, YOLO11x-seg attains the best recall (61.1%) and mAP@50:95 (58.0%), while precision ties with YOLOv8x-seg (89.8%)

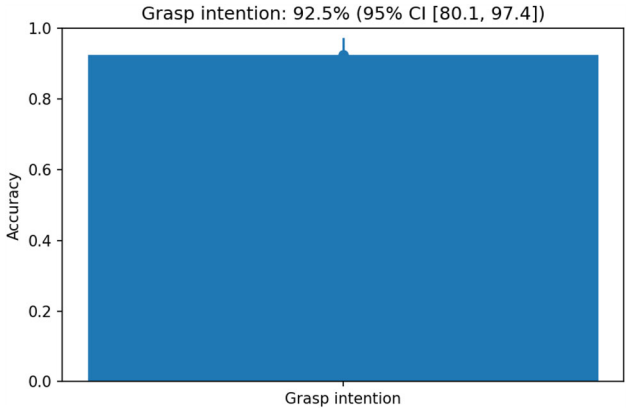

Fig. 5 Grasp-intention accuracy (37/40) with 95% Wilson CI [80.1%, 97.4%] — *grasping experiment*

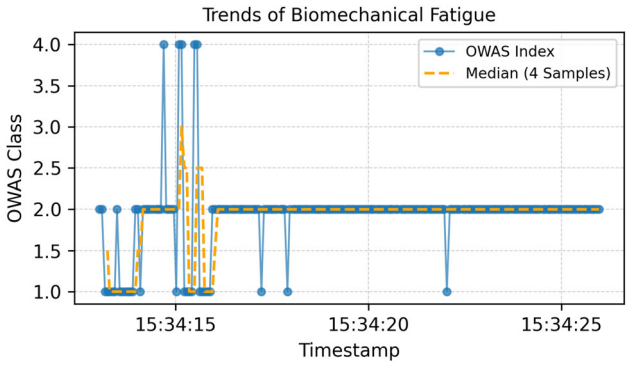

Fig. 6 Median OWAS class over time (13 s window) — *ergonomic experiment*: class 2 predominates with brief peaks to class 4

(≤ 10 kg, 10–20 kg, > 20 kg), within-class errors do not alter the BT policy; only boundary-crossing misestimates can change the load code used by the controller.

In conclusion, the comprehensive comparison across four YOLO architectures confirmed that YOLO11x-seg represents the optimal solution for this HRC application context. While YOLOv12x-seg incorporates the latest architectural improvements, our experimental validation demonstrates that YOLO11x-seg achieves superior accuracy for the specific demands of human-robot collaboration scenarios, particularly excelling in the critical task of hand detection. The marginally lower performance of YOLOv12x-seg (71.6% vs 72.4% mAP@50:95) supports the selection of YOLO11x-seg for resource-constrained industrial deployment scenarios where both accuracy and computational efficiency are essential.

4.3 Experiment: Ergonomic Monitoring, Risk Classification and Robotic Intervention via Behaviour Tree

The objective of this experiment is to evaluate the framework's ability to monitor operator postures in real time, classify ergonomic risk, and manage task transfer to the robot in the presence of critical conditions. This experiment validates the complete integrated framework for human-robot collaboration, demonstrating coordinated operation of all system components from perception through decision-making to robotic intervention. During testing, the operator performed movements typical of an industrial environment, such as bending and lifting, while the system monitored posture using OpenPose and classified ergonomic risk using the OWAS method.

When medium-high-risk (OWAS class 3) or high-risk (OWAS class 4) postures were detected, or volumetric weight in excess of allowable limits, the system generated a biomechanical fatigue message. This message was processed by a BT, which combined ergonomic data, action recognition results, and volumetric weight values to determine whether robotic intervention should be activated. This approach ensured modular and responsive decision making, allowing a smooth transition between the operator and the robot (Fig. 6).

The performance of this experiment was evaluated through several metrics: the distribution of postures across OWAS risk classes (1–4), system update rate (Hz) and latency (seconds) for pose monitoring, response time (seconds) to risky situations, and time trend analysis of posture risk levels. To strengthen statistical rigor, we report 95% Wilson confidence intervals for class proportions and 95% CIs for latency estimates, linked to Figs. 7, 8 and 9.

The data collected during the experiment showed that 12.8% of the postures belonged to OWAS class 1 (no risk), 84.6% to class 2 (moderate risk), no postures were classified in OWAS class 3 (medium-high risk), and 2.6% belonged to class 4 (high risk), as shown in Fig. 7. These results confirm the system's ability to distinguish between ergonomically safe and hazardous postures, demonstrating the effectiveness of the framework in monitoring and classifying risk. Compared to the approach based on OpenPose by Lin et al. [66],

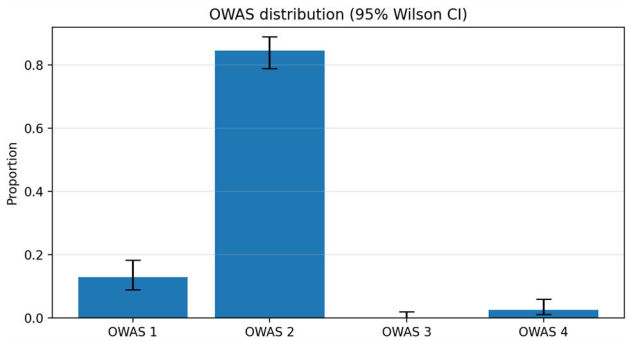

Fig. 7 OWAS class distribution with 95% Wilson CIs over a 13 s window (N=195): class 1 = 12.8% [8.84, 18.24], class 2 = 84.6% [78.89, 89.01], class 3 = 0.0% [0.00, 1.93], class 4 = 2.6% [1.10, 5.86]

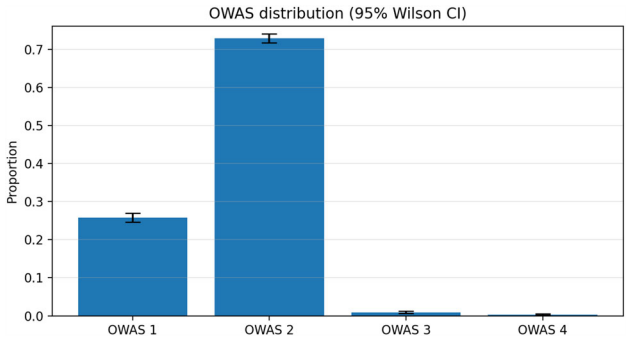

Fig. 8 OWAS class distribution over the full run (N=5597) with 95% Wilson CIs: class 1 = 25.8% [24.67, 26.96], class 2 = 72.95% [71.77, 74.10], class 3 = 0.91% [0.69, 1.20], class 4 = 0.34% [0.22, 0.53]

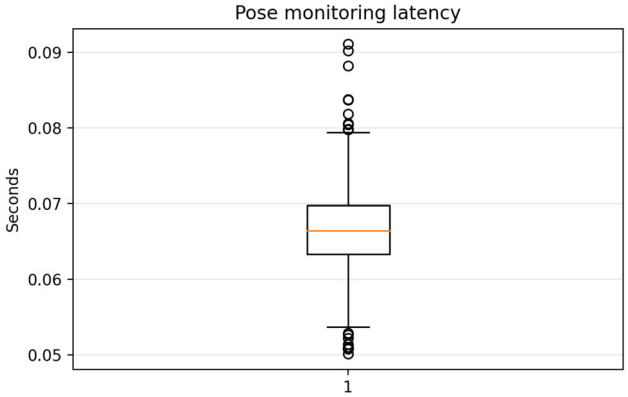

Fig. 10 Pose-monitoring latency (OWAS inter-arrival) in the same window: n=194, mean 0.067 s, median 0.066 s, 95% CI [0.066, 0.068]

our system demonstrates significantly higher precision in ergonomic evaluation. The classification of high-risk postures (OWAS class 4) stands at 2.6%, reducing the frequency of class-4 labels compared to the 10.4% reported in the comparative work. In the same window, human-led operation (OWAS classes 1–2) accounts for 97.4%, with ergonomics-driven assistance (classes 3–4) at 2.6%; over the full session these shares are 98.8% and 1.25%, respectively (Fig. 8).

Temporally, the system maintained an effective update period of 66–68 ms in the 13 s window (median 0.066 s; Fig. 10). Over the full run the mean OWAS inter-arrival is 0.081 s (95% CI [0.072, 0.093], n=5596), i.e., ~12.3 Hz (Fig. 11). The framework's responsiveness extends to critical condition detection, achieving an average response time of 0.07 s from risk identification to intervention triggering (decision layer). This represents a 56% improvement over the 0.16 s baseline reported by Tortora et al. [64] under comparable lifting/transport conditions, while maintaining equivalent intention-recognition accuracy (92.5% vs. 94.3%). For clarity, the 0.07 s figure refers to the *decision layer* (BT tick-to-trigger) under controlled conditions. Log-derived, *end-to-end* response latency—from the first OWAS sample meeting the intervention threshold (class ≥3) to the biomechanical-fatigue message that triggers the BT branch—averaged 0.452 s (95% CI [0.283, 0.622], n=14) over the full run (Fig. 12); within the 13 s window a single event measured 0.330 s (Fig. 9). These end-to-end values include perception and ROS synchronization and remain sub-second in deployment. Such rapid response capability is critical in dynamic industrial environments, where millisecond-scale delays can erode safety margins, supporting our choice of BTs for deterministic, bounded-latency ergonomic gating (Section 2.5).

An analysis of the postures over time windows of four samples (Fig. 6) revealed that the median of the postures remained in OWAS class 2 (moderate risk), with temporary peaks in OWAS class 4 during particularly heavy lifting or postures held for a long time. This underscores the system's ability to accurately detect critical moments, providing useful decision support for operations management.

These measurements demonstrate not only the framework's efficiency in detecting and analyzing critical conditions, but also its role in supporting seamless operational

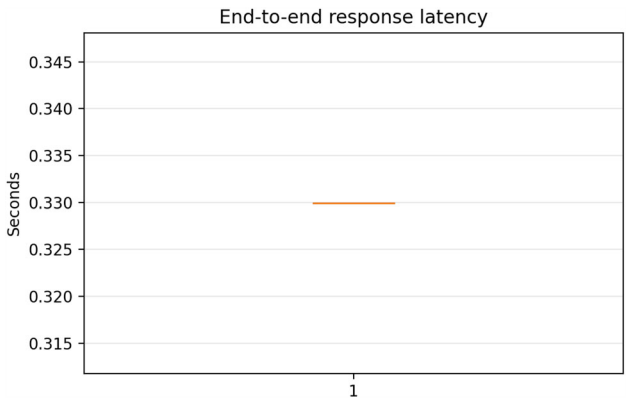

Fig. 9 End-to-end response time (OWAS ≥ threshold → fatigue signal) in the 13 s window (n=1)

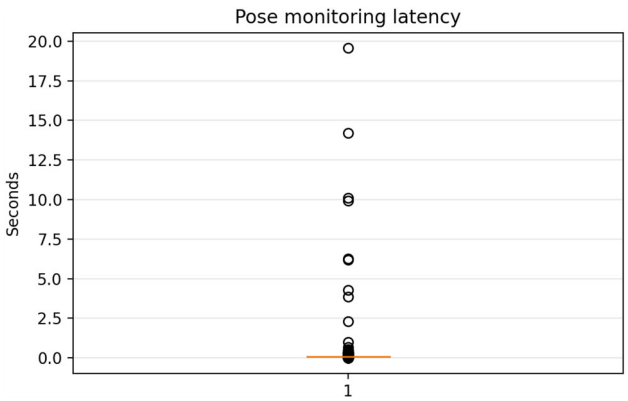

Fig. 11 Pose-monitoring latency (OWAS inter-arrival) over the full run: mean 0.081 s (95% CI [0.072, 0.093], n=5596)

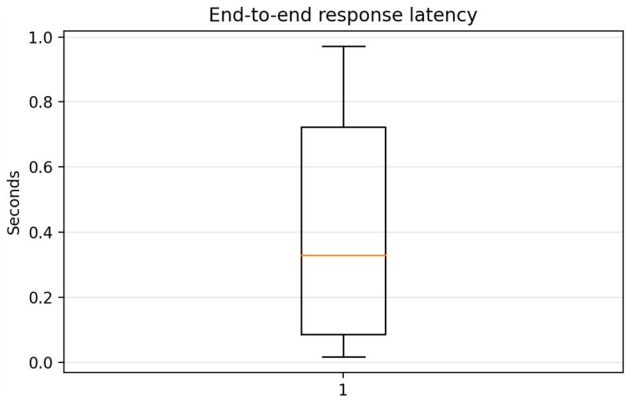

Fig. 12 End-to-end response time over the full run (n=14): median and IQR shown; see text for mean and 95% CI

transitions. The reduced overall latency, calculated through integrated pose monitoring and response time metrics, ensures smooth and responsive task transfer between operator and robot. This operational readiness, combined with the system's capacity to synthesize ergonomic and actionable data, represents a significant advancement in human-robot collaboration optimization. The demonstrated modular architecture, encompassing perception (YOLO11+UKF), action recognition (SlowOnly), ergonomic assessment (OWAS), and adaptive decision-making (BT), inherently supports adaptation to diverse collaborative scenarios beyond lifting/transport tasks. Extension to assembly, inspection, or welding applications would leverage this modularity through systematic component reconfiguration: retraining perception and action recognition modules for task-specific objects and behaviours, selecting appropriate ergonomic indices for the operational context, and adjusting BT decision thresholds accordingly. While such cross-domain validation represents valuable future work, the architectural foundations demonstrated here provide a scalable pathway for broader industrial deployment.

4.4 Experiment: Recognizing Actions with SlowOnly Model

This experiment evaluated the framework's ability to recognize and classify actions performed by an operator in a simulated industrial environment by exploiting the SlowOnly model. The model, based on a ResNet3D network with SlowOnly architecture and 50-level depth, was pre-trained on the HRI30 dataset, specifically developed to recognize actions in human-robot interaction scenarios. During the experiment, the model was configured to classify three classes of actions: lifting, carrying, and repositioning packages.

Video sequences were preprocessed to 8-frame clips, uniformly sampled with a time interval of four frames, and normalized using ImageNet mean values and standard deviations. Each clip was subjected to resizing, center cropping, and augmentation operations with random color changes and horizontal flips to improve the robustness of the model. The optimization process used the AdamW algorithm with an initial learning rate of 0.001, adjusted through a Cosine Annealing strategy with warmup for the first 1,000 steps. Training was conducted on an NVIDIA RTX 3090 GPU, using a batch size of 16 videos per GPU for 200 epochs, with the cross-entropy loss function.

The experiment's performance was assessed through classification loss during training, top-1 accuracy on the validation set, comparison with alternative architectures (I3D and TSM), and model robustness to occlusions and variability.

During the validation phase, the model achieved an accuracy of 95.83% demonstrating an excellent ability to generalize over the test data. The behaviour of loss during training shows rapid initial convergence, with stabilization at 0.129 toward the later epochs. This result reflects the effectiveness of the model setup and training pipeline in gradually reducing the classification error.

The use of short 8-frame clips proved particularly effective in capturing action movements without introducing significant computational complexity, while integration with the BT ensured that the classified data were exploited for real-time robotic decisions.

This model and framework configuration was further validated by the previous evaluation on HRI30, where SlowOnly was shown to outperform alternative architectures such as I3D and TSM in action recognition in industrial settings. Comparative analysis [26] revealed that SlowOnly achieved a top-1 accuracy of 86.55% compared to 78.43% for I3D and 73.91% for TSM when evaluated on the HRI30 dataset. The model's robustness to occlusions and its high accuracy make it an ideal choice for collaborative human-robot scenarios, where accurate and timely action recognition is crucial.

Experimental outcomes establish the effectiveness of the proposed approach, which combines action recognition, ergonomic analysis, and volumetric evaluation, integrating them into a modular and responsive decision-making framework. The framework showed a remarkable ability to adapt to operational variables, providing decisive support for safety and efficiency in collaborative industrial operations.

5 Ethical, Privacy, and Workforce Considerations

This work focuses on ergonomic risk mitigation rather than worker surveillance. The decision layer operates on posture-centric abstractions (skeletal joints from OpenPose, action labels, OWAS risk classes) and on object masks from YOLO-based perception, rather than on biometric identification. The

manuscript does not include identifiable images or videos, and all procedures were conducted under an approved protocol with informed consent (see *Declarations*). From a privacy standpoint, data used by the decision layer consist of skeletal coordinates, action recognition outputs, and OWAS-derived indicators, which limit exposure of personally identifiable information and constrain use to occupational safety purposes as described in the consent materials. Participants were informed about the study aims and procedures as part of the ethics-approved protocol. Potential psychosocial concerns associated with continuous monitoring are acknowledged and are mitigated by exclusive safety-oriented processing and by the use of de-identified posture descriptors. A broader assessment of long-term organizational impact falls outside the scope of this technical validation and is identified as a direction for subsequent field deployments. Governance considerations are informed by collaborative robot safety norms (ISO 10218 and ISO/TS 15066), occupational health and safety management (ISO 45001), and AI risk management frameworks (ISO/IEC 23894) together with architectural guidance for AI-enabled systems (ISO/IEC 23053), where applicable to the described posture-centric pipeline.

6 Conclusions and Future Work

In this work, an innovative framework for human-robot collaboration was proposed, combining advanced visual sensing technologies, real-time ergonomic monitoring, and Behaviour Tree (BT)-based adaptive decision-making. Experimental results show that the system significantly improves operator safety and efficiency of operations, providing a scalable and modular solution for complex collaborative environments.

To the best of our knowledge, this is the first framework to synergistically integrate advanced visual perception, real-time ergonomic assessment, and adaptive decision-making for industrial HRC. Compared with traditional methods based on static rules or post-hoc ergonomic analysis, our approach introduces greater flexibility and adaptability, improving both the safety and well-being of operators. The noninvasive nature of the system, which eliminates the need for wearable physical sensors, also helps to reduce operational costs and improve the naturalness of human-robot interaction.

Despite these encouraging results, our study has limitations. The experiments were conducted in controlled laboratory conditions and require extension to real industrial scenarios to assess robustness under production constraints. In addition, beyond the reported confidence intervals, performance metrics would benefit from formal significance testing and larger-sample validation. The modular architecture enables ablation-style analyses to quantify individual component contributions to overall system performance. Given the safety-critical, integrated nature of the framework, we avoid destructive ablations (e.g., disabling OWAS gating or BT fallbacks) that would misrepresent intended use; instead, the manuscript already includes a non-destructive, controlled component replacement for the detector under identical settings, and node-level instrumentation for module-wise latency profiling is planned as follow-up. Finally, the absence of direct comparisons with established technologies on shared benchmarks presents an opportunity to further position the framework within the state of the art.

To address the limitations that have emerged and broaden the impact of the framework, future work will focus on:

- **Validation in Real Environments**: Conducting field tests in operational industrial settings to demonstrate the robustness and scalability of the system on a large scale.
- **Statistical Validation**: Implementing comprehensive statistical analysis for robust performance characterization.
- **Ablation and Profiling**: pursuing non-destructive component replacement under identical settings (as done for the detector) and adding node-level instrumentation for module-wise latency profiling to quantify per-module contributions and bottlenecks without violating the intended safety envelope.
- **Comparisons with Existing Solutions**: Conduct quantitative comparisons with established academic technologies using standard benchmarks to further validate the framework's contribution.
- **End-User Engagement**: Collaborate with industrial operators and supervisors to gather feedback on the usability, efficiency, and acceptability of the framework, improving its configuration based on specific user needs.
- **Expansion for Complex Scenarios**: Extend the system to handle multi-operator and multi-robot interactions, improving collaborative efficiency in dynamic and high-variability environments.
- **Integration of Advanced Sensors**: Experiment with next-generation sensors, such as high-resolution RGB-D cameras and multi-sensor configurations, to further improve the accuracy of visual monitoring and volumetric estimation.
- **Cross-domain Applications**: Explore the framework's adaptability in non-industrial contexts such as healthcare, smart homes, rehabilitation centers, and logistics.

In conclusion, the proposed framework represents a major breakthrough in human-robot collaboration, combining safety, efficiency and adaptability in a single integrated solution. Its modular and scalable architecture makes it

particularly suitable for a wide range of industrial applications, including logistics, manufacturing, and healthcare. With validation in real-world environments and enrichment with additional functionality, the system has the potential to become a benchmark solution, improving the productivity and well-being of operators in modern manufacturing settings.

Author Contributions F.I. conceived and designed the study, developed the methodology, performed the experiments, collected and analyzed the data, and wrote the first draft of the manuscript. A.A. supervised the research and provided resources. E.D.M. provided substantial feedback on the structure and content of the manuscript, suggesting critical improvements, additions, and refinements. All authors read and approved the final manuscript.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. The initial data collection was conducted while the first author was supported by ERC-StG Ergo-Lean (Grant Agreement No. 850932).

Data Availability The datasets generated during and analyzed during the current study are not publicly available due to proprietary restrictions but are available from the corresponding author upon reasonable request and with permission of the affiliated institutions.

Materials Availability Not applicable.

Code Availability The custom code used for the implementation of the proposed framework is available from the corresponding author upon reasonable request, subject to intellectual property considerations.

Declarations

Ethics approval and consent to participate This study was conducted in accordance with the Declaration of Helsinki. The research protocol and data collection procedures were approved by the Ethics Committee of Azienda Sanitaria Locale (ASL) Genovese N.3 (Protocol IIT HRII ERGOLEAN 156/2020). Informed consent was obtained from all individual participants involved in the study.

Consent for publication Not applicable, as the manuscript does not contain any individual person's identifiable data, images, or videos.

Conflicts of Interest/Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Pedrocchi, N., Vicentini, F., Tosatti, A., Molinari-Tosatti, L.: Safe human-robot cooperation in an industrial environment. IEEE/ASME Trans. Mechatron. **19**(1), 151–160 (2013)
2. Billing, E., Fraboni, F., Gualtieri, L., Rosen, P.H., Thorvald, P.: Human factors and cognitive ergonomics in advanced industrial human-robot interaction. Frontiers Media SA (2025)
3. Bibbo, D., Corvini, G., Schmid, M., Ranaldi, S., Conforto, S.: The impact of human-robot collaboration levels on postural stability during working tasks performed while standing: Experimental study. JMIR Hum. Factors **12**(1), 64892 (2025)
4. Rahman, M.H., Ghasemi, A., Dai, F., Ryu, J.: Review of emerging technologies for reducing ergonomic hazards in construction workplaces. Buildings **13**(12), 2967 (2023)
5. Mahdavi, N., et al.: A review of work environment risk factors influencing muscle fatigue. Int. J. Ind. Ergon. **80**, 103028 (2020)
6. Villani, V., Pini, F., Leali, F., Secchi, C.: Survey on human-robot collaboration in industrial settings: safety, intuitive interfaces and applications. IEEE Robot. Autom. Lett. **3**(1), 1201–1208 (2018)
7. Conforti, I., et al.: Measuring biomechanical risk in lifting load tasks through wearable system and machine-learning approach. Sensors **20**(6), 1557 (2020)
8. Tian, Y., Ye, Q., Doermann, D.: YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. (2025)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 91–99 (2015)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
12. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
13. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv:1804.02767. (2018)
14. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934. (2020)
15. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. arXiv:2004.01655. (2020)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
17. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2016)
18. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: autonomous driving when winter is coming. arXiv:1907.07484. (2019)
19. Marvel, J., Norcross, R., Falco, J.: Implementing speed and separation monitoring in collaborative robot workcells. IEEE Trans. Autom. Sci. Eng. **12**(3), 969–976 (2015)
20. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. arXiv:1703.07402. (2017)
21. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35–45 (1960)

22. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: multi-object tracking by associating every detection box. In: European Conference on Computer Vision, pp. 1–21. Springer (2022)
23. Wan, E.A., Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation. In: Proceedings of the IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium, pp. 153–158 (2000)
24. Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1302–1310 (2017)
25. Bataineh, A.M., Mohamed, A.S.A.: Monocular 3D human pose estimation for REBA ergonomics: a critical review of recent advances. Comput. Mater. Continua. **84**(1) (2025)
26. Iodice, F., De Momi, E., Ajoudani, A.: HRI30: an action recognition dataset for industrial human-robot interaction. In: 26th International Conference on Pattern Recognition (ICPR), pp. 4941–4947 (2022)
27. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 2004–2013 (2019)
28. Peng, Y., Lee, J., Watanabe, S.: I3D: transformer architectures with input-dependent dynamic depth for speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
29. Lasota, P.A., Fong, T., Shah, J.A.: A survey of methods for safe human-robot interaction. Ann. Rev. Control Robot. Auton. Syst. **1**, 123–149 (2017)
30. Cherubini, A., Passama, R., Crosnier, A., Lasnier, A., Fraisse, P.: Collaborative manufacturing with physical human-robot interaction. IEEE Trans. Autom. Sci. Eng. **13**(1), 118–129 (2016)
31. Peternel, L., Tsagarakis, N.G., Caldwell, D.G.: Robot adaptation to human physical fatigue in human-robot co-manipulation. Auton. Robot. **42**(5), 1011–1021 (2018)
32. Rozo, L., Jimenez, P., Torras, C., Alenya, G.: Learning physical collaborative robot behaviors from human demonstrations. IEEE Trans. Rob. **32**(3), 513–527 (2016)
33. Lamon, E., De Franco, A., Peternel, L., Ajoudani, A.: A capability-aware role allocation approach to industrial assembly tasks. IEEE Robot. Autom. Lett. **4**(4), 3378–3385 (2019)
34. Iovino, M., Scukins, E., Styrud, J., Ögren, P., Smith, C.: A survey of behavior trees in robotics and AI. Robot. Auton. Syst. **154**, 104096 (2022)
35. Segura, P., Lobato-Calleros, O., Soria-Arguello, I., Hernández-Martínez, E.G.: Work roles in human-robot collaborative systems: effects on cognitive ergonomics for the manufacturing industry. Appl. Sci. **15**(2), 744 (2025)
36. Colledanchise, M., Ögren, P.: Behavior Trees in Robotics and AI: An Introduction. CRC Press (2018)
37. Lamon, E., Fusaro, F., De Momi, E., Ajoudani, A.: A unified architecture for dynamic role allocation and collaborative task planning in mixed human-robot teams. arXiv:2301.08038. (2023)
38. Merlo, E., Lamon, E., Fusaro, F., Lorenzini, M., Carfì, A., Mastrogiovanni, F., Ajoudani, A.: An ergonomic role allocation framework for dynamic human-robot collaborative tasks. J. Manuf. Syst. **67**, 111–121 (2023)
39. Kober, J., Bagnell, J.A., Peters, J.: Reinforcement learning in robotics: a survey. Int. J. Robot. Res. **32**(11), 1238–1274 (2013)
40. Ravichandar, H., Polydoros, A.S., Chernova, S., Billard, A.: Recent advances in robot learning from demonstration. Ann. Rev. Control Robot. Autonom. Syst. **3**(1), 297–330 (2020)
41. Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. **16**(1), 1437–1480 (2015)
42. Pignat, E., Calinon, S.: Learning adaptive dressing assistance from human demonstration. Robot. Auton. Syst. **93**, 61–75 (2017)
43. El-Shamouty, M., Wu, X., Yang, S., Albus, M., Huber, M.F.: Towards safe human-robot collaboration using deep reinforcement learning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 4899–4905. IEEE (2020)
44. Liu, Q., Liu, Z., Xiong, B., Xu, W., Liu, Y.: Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function. Adv. Eng. Inform. **49**, 101360 (2021)
45. Shafti, A., Tjomsland, J., Dudley, W., Faisal, A.A.: Real-world human-robot collaborative reinforcement learning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 11161–11166. IEEE (2020)
46. Colledanchise, M., Parasuraman, R., Ögren, P.: Learning of behavior trees for autonomous agents. IEEE Trans. Games **11**(2), 183–189 (2018)
47. Karhu, O., Kansi, P., Kuorinka, I.: Correcting working postures in industry: a practical method for analysis. Appl. Ergon. **8**(4), 199–201 (1977)
48. McAtamney, L., Corlett, E.N.: RULA: a survey method for the investigation of work-related upper limb disorders. Appl. Ergon. **24**(2), 91–99 (1993)
49. Hignett, S., McAtamney, L.: Rapid entire body assessment (REBA). Appl. Ergon. **31**(2), 201–205 (2000)
50. Waters, T.R., Putz-Anderson, V., Garg, A., Fine, L.J.: Revised NIOSH equation for the design and evaluation of manual lifting tasks. Ergonomics **36**(7), 749–776 (1993)
51. Kee, D.: Comparison of OWAS, RULA and REBA for assessing potential work-related musculoskeletal disorders. Int. J. Ind. Ergon. **83**, 103140 (2021)
52. Ferraguti, F., Villa, R., Landi, C.T., Zanchettin, A.M., Rocco, P., Secchi, C.: A unified architecture for physical and ergonomic human-robot collaboration. Robotica **38**(4), 669–683 (2020)
53. Cruciata, L., Contino, S., Ciccarelli, M., Pirrone, R., Mostarda, L., Papetti, A., Piangerelli, M.: Lightweight vision transformer for frame-level ergonomic posture classification in industrial workflows. Sensors **25**(15), 4750 (2025)
54. David, G.: Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. Appl. Ergon. **36**(4), 463–473 (2005)
55. Santopaolo, A., Lorenzini, M., Privitera, L., Varrecchia, T., Chini, G., Ranavolo, A., Ariano, P., Ajoudani, A.: Biomechanical risk assessment of human lifting tasks via supervised classification of multiple sensor data. In: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), pp. 746–751. IEEE (2022)
56. Donisi, L., et al.: Work-related risk assessment according to the revised NIOSH lifting equation: a preliminary study using a wearable inertial sensor and machine learning. Sensors **21**(8), 2593 (2021)
57. Naranjo, J.E., Mora, C.A., Bustamante Villagómez, D.F., Mancheno Falconi, M.G., Garcia, M.V.: Wearable sensors in industrial ergonomics: enhancing safety and productivity in industry 4.0. Sensors **25**(5), 1526 (2025)
58. Peters, M., Potthast, W., Wischniewski, S., Komnik, I.: Wearable sensors for classification of load-handling tasks with machine learning algorithms in occupational safety and health: a systematic literature review. Ergonomics, 1–21 (2025)
59. Babangida, A.A., Caraballo-Arias, Y., Decataldo, F., Violante, F.S.: Advancing occupational medicine through wearable technology: A review of sensor systems for biomechanical risk assessment and work-related musculoskeletal disorder prevention. ACS Sensors (2025)
60. Khanam, R., Hussain, M.: YOLOv11: an overview of key architectural enhancements. arXiv:2410.17725. (2024)
61. Shafti, A., Ataka, A., Lazpita, B.U., Shiva, A., Wurdemann, H.A., Althoefer, K.: Real-time robot-assisted ergonomics. In: 2019 Inter-

national Conference on Robotics and Automation (ICRA), pp. 1975–1981. IEEE (2019)
62. Kim, W., Peternel, L., Lorenzini, M., Babič, J., Ajoudani, A.: A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools. Robot. Comput.-Integr. Manuf. **68**, 102084 (2021)
63. Falerni, M.M., Pomponi, V., Karimi, H.R., Nicora, M.L., Dao, L.A., Malosio, M., Roveda, L.: A framework for human-robot collaboration enhanced by preference learning and ergonomics. Robot. Comput.-Integr. Manuf. **89**, 102781 (2024)
64. Tortora, S., Michieletto, S., Stival, F., Menegatti, E.: Fast human motion prediction for human-robot collaboration with wearable interface. In: 2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), pp. 457–462. IEEE (2019)
65. Iodice, F., Wu, Y., Kim, W., Zhao, F., De Momi, E., Ajoudani, A.: Learning cooperative dynamic manipulation skills from human demonstration videos. Mechatronics **85**, 102807 (2022)
66. Lin, P.-C., Chen, Y.-J., Chen, W.-S., Lee, Y.-J.: Automatic real-time occupational posture evaluation and select corresponding ergonomic assessments. Sci. Rep. **12**(1), 2139 (2022)

Francesco Iodice MSc in Artificial Intelligence and Robotics at Sapienza University of Rome in 2019. Since 2020, he is pursuing a Ph.D. in Bioengineering at Politecnico di Milano, Neuroengineering and Medical Robotics Laboratory (NearLab), working in collaboration with Istituto Italiano di Tecnologia (IIT), under the supervision of Dr. Arash Ajoudani at Human-Robot Interfaces and Physical Interaction (HRI2) laboratory. His research concerns Computer Vision , Robotics and Machine Learning, and aims to improve human ergonomics in highly dynamic human-robot-environment interactions.

Elena De Momi MSc in Biomedical Engineering in 2002, PhD in Bioengineering in 2006, currently Assistant Professor in Electronic Information and Bioengineering Department (DEIB) of Politecnico di Milano. She was co-founder of the Neuroengineering and Medical Robotics Laboratory, in 2008, being responsible of the Medical Robotics section. She has been an Associate Editor of the Journal of Medical Robotics Research and of the International Journal of Advanced Robotic Systems. In 2016 she has been an Associated Editor of the IEEE International Conference on Robotics and Automation. Her academic interests include image-processing, virtual environments, augmented reality and simulators, teleoperation, haptics, medical robotics, neuromechanics. She participated to several EU funded projects in the field of Surgical Robotics (ROBOCAST, ACTIVE and EuRoSurge, where she was PI for partner POLIM). She is currently PI for POLIMI of the EDEN2020 project, aimed at developing a neurosurgery drug delivery system. She has been evaluator and reviewer for the European Commission in FP6 and FP7.

Arash Ajoudani is a tenured senior scientist at the Italian Institute of Technology (IIT), where he leads the Human-Robot Interfaces and physical Interaction (HRI^2) laboratory. He received his PhD degree in Robotics and Automation from University of Pisa and IIT in 2014. He is a recipient of the European Research Council (ERC) starting grant 2019 (Ergo-Lean), the coordinator of the Horizon-2020 project SOPHIA, and the co-coordinator of the Horizon-2020 project CONCERT. He is a recipient of the IEEE Robotics and Automation Society (RAS) Early Career Award 2021, and winner of the Amazon Research Awards 2019, of the Solution Award 2019 (MECSPE2019), of the KUKA Innovation Award 2018, of the WeRob best poster award 2018, and of the best student paper award at ROBIO 2013. His PhD thesis was a finalist for the Georges Giralt PhD award 2015 - best European PhD thesis in robotics. He was also a finalist for the Solution Award 2020 (MECSPE2020), the best conference paper award at Humanoids 2018, for the best interactive paper award at Humanoids 2016, for the best oral presentation award at Automatica (SIDRA) 2014, and for the best manipulation paper award at ICRA 2012.

He is the author of the book "Transferring Human Impedance Regulation Skills to Robots" in the Springer Tracts in Advanced Robotics (STAR), and several publications in journals, international conferences, and book chapters. He is currently serving as an IEEE RAS AdCom member (2022–2024), the executive manager of the IEEE-RAS Young Reviewers' Program (YRP), and as chair and representative of the IEEE-RAS Young Professionals Committee. He has been serving as a member of scientific advisory committee and as an associate editor for several international journals and conferences such as IEEE RAL, ICRA, IROS, ICORR, etc. He is a scholar of the European Lab for Learning and Intelligent Systems (ELLIS). His main research interests are in physical human-robot interaction, mobile manipulation, robust and adaptive control, assistive robotics, and tele-robotics.