The International Journal of Advanced Manufacturing Technology
https://doi.org/10.1007/s00170-025-17168-1

ORIGINAL ARTICLE

A novel approach for monocular RGB-based ergonomics monitoring in industrial workspaces employing synthetic datasets to train a deep learning model

Thomas Agostinelli[1] · **Andrea Generosi**[1,2] · **Maura Mengoni**[1]

Received: 5 November 2025 / Accepted: 4 December 2025

Abstract

Physical ergonomics is critical for worker well-being, safety, and productivity, in Industry 4.0 and in the emerging Industry 5.0 paradigm, which emphasize human-machine collaboration. Advances in motion capture technologies, especially marker-less systems using RGB cameras, offer new possibilities for real-time ergonomic assessments. These systems effectively address the intrusiveness issues inherent to marker-based approaches and partially overcome challenges such as lighting variations, occlusions and calibration commonly associated with RGB-D camera-based systems. Nonetheless, their accuracy can be suboptimal due to limitations arising from perspective distortions. To overcome this issue, the present paper introduces an "inductive" RGB camera-based approach to ergonomic assessment, leveraging Deep Learning models trained on a synthetic dataset of 14,400 unique postures captured from 24 viewpoints in Unity 3D. Unlike traditional methods reliant on measured angles, this approach predicts ergonomic evaluations directly from posture observations. We implemented the proposed approach on a prototypal system that allowed us to experiment and collect quantitative results. We analysed the system's performance through comparative testing, juxtaposing predictions from our system with measurements obtained from an inertial system worn by individuals while performing tasks in a controlled laboratory environment. Experimentations highlighted the satisfactory performance of the proposed inductive approach and related technology, highlighting its potential for delivering more accurate real-time ergonomic assessments.

Keywords RGB motion capture · Deep learning · Ergonomics · Synthetic datasets · Human centred manufacturing

1 Introduction

The dynamic nature of modern production, with frequent workstation changes, task variability, and the integration of collaborative robots, challenges the maintenance of proper posture and the prevention of work-related musculoskeletal disorders (MSDs) [1].

MSDs primarily affect the spine and upper limbs and are classified as cumulative disorders, arising from repeated exposure to discomfort-inducing conditions. These disorders are influenced by various risk factors, categorized into primary (e.g., awkward postures, excessive force exertion, and repetitive motions) and secondary (e.g., mechanical contact stress, lighting, and environmental or climatic conditions). Ergonomic interventions aim to reduce MSD incidence by addressing high physical demands and awkward working postures [2]: by addressing both primary and secondary risk factors, industries can achieve improved worker well-being, enhanced operational efficiency, and significant reductions in healthcare expenditures.

Addressing these challenges requires tailored solutions that account for individual worker variability, behaviors, and specific occupational risks: the integration of Human Factors into business processes - aimed at optimizing the interaction between humans and technological systems

✉ Andrea Generosi
andrea.generosi@unipegaso.it; a.generosi@univpm.it

1 Department of Industrial Engineering and Mathematical Sciences, Università Politecnica delle Marche, via Brecce Bianche 12, Ancona, AN, Italy

2 Department of Science and Information Technology, Università Pegaso, piazza Trieste e Trento 48, Napoli, NA, Italy

Published online: 29 January 2026

- thus becomes crucial. In particular, physical ergonomics, one of the core pillars of Human Factors, enables the monitoring of human spatial interaction with complex technological systems, such as those found in industrial settings, and supports corrective interventions on posture and workstation design to prevent MSDs.

Therefore, ergonomics must be a key factor in shared workspaces between humans and robots [3] and humans and machines, both during the design phase of these spaces and during daily work through continuous monitoring.

Achieving the two goals of reducing work-related MSDs and of improving human-robot collaboration requires the development of robust and cost-effective tools capable of directly monitoring working postures, continuously evaluating ergonomic risks during activities, notifying workers of incorrect postures, and providing managers with actionable insights to improve individual workspaces and working conditions.

1.1 Techniques for physical ergonomic assessment

Physical ergonomics can be assessed using two primary methodologies: subjective methods and observational methods.

Subjective methods rely on rating scales applied to responses collected through questionnaires and interviews directed at workers. These methods gather and analyse qualitative data to evaluate factors such as comfort, usability, and perceived risk of musculoskeletal disorders (MSDs). Examples of tools under this category include the Nordic Musculoskeletal Questionnaire (NMQ) [4], the Cornell Musculoskeletal Discomfort Questionnaire (CMDQ) [5], and the Quick Exposure Check (QEC) [6], all designed to collect data on perceived musculoskeletal discomfort during work.

Observational methods, on the other hand, involve the direct or indirect observation (e.g., through video recordings) of workers' postures and movements during their activities, without relying on self-assessment. These methods are particularly effective in identifying risky behaviours and non-ergonomic postures.

Some observational tools address specific risk factors, such as the NIOSH Lifting Eq [7]., the Manual Handling Assessment Chart [8], and the Assessment of Repetitive Tasks (ART) [9]. Other tools focus on characterizing workers' postures by analysing the angles formed by their body segments—particularly the neck, upper limbs, back, and lower limbs. Prominent examples include RULA (Rapid Upper Limb Assessment) [10], REBA (Rapid Entire Body Assessment) [11], OCRA (Occupational Repetitive Action) [12] and the Ovako Working Analysis System (OWAS) [13].

Among the observational methods, RULA and REBA are particularly widely adopted. These methods assign scores to each body segment analysed, based on the angles that the segment assumes within the observed posture. By quantifying the ergonomic risk associated with specific postures, they enable targeted interventions to improve workplace ergonomics.

1.2 Measurement technologies in observational methods

The main challenge in assessing RULA, REBA and OCRA indexes lies in acquiring the specific joint angles that define workers' postures during their activities. In the past, most ergonomic assessments relied on manual analysis of photos and videos captured by RGB digital cameras to extract joint angles. As a result, ergonomic evaluation was costly, time-consuming, and often influenced by the evaluator's subjectivity [14].

To address the issue of joint angle acquisition and automate the analysis, research focused on the use of Motion Capture (MoCap) technologies [15–17], which can be categorized into sensor-based and optical systems [18].

Sensor-based systems (e.g., Xsens [19], Vicon Blue Trident [20]) are typically classified into electromagnetic, acoustic, and inertial systems [21]. These systems are not affected by occlusion problems but generally do not exhibit very high accuracy. Additionally, they can be highly invasive, requiring operators to wear a series of devices that may be cumbersome and interfere with regular work activities.

Optical systems, on the other hand, can be marker-based or markerless. Marker-based optical systems (e.g., Vicon Nexus [22], OptiTrack [23]) are usually very precise, enabling accurate and quantitative acquisition of joint angles. However, they are expensive in terms of both economic costs and installation and calibration times, and have limited applicability for monitoring factory environments due to various constraints, such as lighting conditions and electromagnetic interference [24], which restrict their use primarily to laboratory settings [25] and hence design purposes. Furthermore, these systems require operators to wear special markers placed according to specific guidelines [25] and undergo calibration procedures. These activities necessitate the involvement of specialized professionals and are time-intensive, making them impractical for daily use in real working conditions.

1.3 AI-based motion capture

Low-cost RGB body-tracking technologies rely on capturing video recordings with a simple RGB camera and processing them using machine learning algorithms [26] to predict joint positions and calculate joint angles [27, 28]. They don't need the use of markers to recognize operators' poses, therefore they are known as “markerless optical systems”.

In recent years, research on measurement systems for ergonomics has largely explored the applicability of these technologies. One of the most promising technology is OpenPose, developed by researchers at Carnegie Mellon University [26], and it represents the first real-time multi-person system for jointly detecting the human body, hands, face, and feet (137 key points per person: 70 key points for the face, 25 for the body/feet, and 2 × 21 key points for the hands) in a single image. It is an open-source software based on a convolutional neural network (CNN) within the OpenCV library [29], initially written in C + + and Caffe, and freely available for non-commercial use.

OpenPose therefore allows key points for the body, feet, face, and hands to be tracked. The library can process 2D images using RGB cameras or combine RGBD cameras with depth sensors to synchronize outputs, resulting in a unified set of key points with 3D coordinates. Moreover, it is not significantly impacted by occlusion problems, as it can track the entire body skeleton even when some points or segments are temporarily occluded [30].

Several studies have validated the accuracy of skeleton tracking using RGB cameras with OpenPose, comparing its results to those of the Vicon system. The limb positioning error index relative to wearable sensor systems was negligible, confirming the potential of OpenPose [31, 32]. In many studies, OpenPose has been evaluated as a useful tool for studying worker posture ergonomics using methods such as RULA, REBA, and OCRA [25, 31, 33–35].

However, deep learning algorithms like OpenPose, used for tracking individuals through RGB images, typically require hardware with high computational performance, making good CPU and GPU capabilities essential [36]. Among the various open source projects that make use of the models provided by OpenPose is Tf-pose-estimation [27], in which Tensorflow is exploited to ensure a computationally lighter implementation.

Google has also introduced its framework, MediaPipe [28], which includes tools for detecting body features such as pose, hand, and iris detection. Its standout feature is the hand-detection module, whose accuracy has been validated in various studies [37–39].

Motion Capture systems based on Deep Learning techniques, such as OpenPose, Tf-pose-estimation, or MediaPipe, offer innovative and accessible solutions for analysing human motion.

However, all mentioned technologies have inherent limitations, primarily stemming from their inability to capture accurate depth information and spatial positioning [40]. Unlike multi-camera setups or systems employing active depth sensors, single-camera systems struggle with perspective distortion, especially when the subject moves along the depth axis or is positioned far from the camera’s optical centre. This limitation arises because monocular systems lack the stereoscopic information necessary to infer three-dimensional spatial relationships with high precision.

Adopting a multi-camera setup with Deep Learning MoCap systems helps mitigating the perspective distortion issue but does not resolve it. Moreover, the correct positioning of a multi-camera setup is often hindered by the layout of the manufacturing site [40].

State-of-the-art deep learning models, such as OpenPose and MediaPipe, have made significant advancements in human pose estimation using single RGB cameras. These frameworks excel at identifying key body landmarks in two dimensions and have been adapted for a variety of tasks, including industrial ergonomics assessments, human-robot collaboration, and worker fatigue monitoring. Nevertheless, the reliance on two-dimensional representations introduces ambiguities in estimating joint angles, distances, and motion trajectories in three-dimensional space. Errors are particularly pronounced in scenarios where occlusions occur, where body parts are aligned along the depth axis, or where movements deviate significantly from the dataset distribution used during model training. Furthermore, factors such as varying lighting conditions, camera resolution, and subject clothing can further degrade performance.

Efforts to mitigate these limitations have focused on hybrid approaches, integrating monocular RGB data with depth estimation techniques derived from convolutional neural networks (CNNs), recurrent neural networks (RNNs), or attention mechanisms. Methods such as monocular depth estimation, inverse kinematics, and probabilistic pose reconstruction have been explored to reconstruct 3D poses from 2D data [41]. Additionally, synthetic datasets have played a crucial role in addressing the scarcity of annotated multi-view data by simulating diverse poses, lighting conditions, occlusions, and camera angles. These datasets enable the training of more robust pose estimation models capable of generalizing across different environments [42].

Despite these advancements, monocular Motion Capture systems remain susceptible to errors in spatial accuracy, particularly in dynamic environments where subjects frequently move relative to the camera’s optical axis. Moreover, the dependency on large-scale annotated datasets and computationally expensive training procedures remains a significant challenge for widespread adoption in real-world industrial scenarios. Researchers are increasingly turning to hybrid pipelines that combine monocular MoCap with additional sensor data, such as IMU (Inertial Measurement Unit) readings, to enhance depth estimation accuracy and motion tracking fidelity. Studies such as [43] with the VIBE model and [44] on integrating IMU data into deep learning pose estimation frameworks have demonstrated significant improvements in spatial accuracy and motion reconstruction stability. These approaches use sensor fusion techniques to compensate for the depth ambiguities inherent in monocular systems, resulting in more robust estimations in dynamic environments. However, these systems are not tested in a real manufacturing environment yet.

1.4 Synthetic datasets for deep learning

Due to the demand for ever larger datasets for Supervised training Deep Learning models, synthetic datasets have thus become an increasingly relevant resource in Deep Learning (DL). These artificially generated data aim to replicate the characteristics of real-world data and are used to train and evaluate DL models. Their utility is particularly evident in addressing issues such as the need to protect privacy while maintaining the statistical structure of the original data [45] and in improving the quality of model training by providing a large, diverse, and bias-free training dataset [46–48]. This approach enhances models' generalization capabilities across different scenarios and reduces the risk of overfitting, a phenomenon where the model becomes too tailored to the patterns in the training data but fails to generalize to new situations.

A critical aspect of using synthetic datasets is evident in image processing systems, where artificial data generation is much more efficient and less costly than the manual collection of real data [49]. Various industries can benefit from synthetic datasets, unified by the need for large quantities of labelled data.

A growing use case is in human recognition, including Human Activity Recognition (HAR) [50, 51], Face Recognition (FR) [52], and ergonomics in the automotive sector, where synthetic datasets are generated in 3D environments to simulate real-world scenarios, such as monitoring drivers and passengers [53, 54].

The generation of synthetic datasets relies on advanced technologies, primarily following two approaches. The first approach uses Machine Learning models, such as Variational Autoencoders (VAE), which learn the distribution of original data and generate new data by sampling its features [55], or Generative Adversarial Networks (GAN), which produce high-quality synthetic images [56]. The second, more traditional approach involves simulation tools or computer graphics software, such as Blender [49] and Unity [57], widely used for creating synthetic datasets derived from urban scenarios featuring pedestrians, vehicles, and signage for Autonomous Driving applications [58, 59], or from indoor scenarios for Human Activity Recognition [60].

The aim of this paper is to present a new approach to low-cost physical ergonomics measurement that eliminates the need for calibration procedures and/or for stereo systems with multiple views or even RGBD cameras. This approach leverages standard RGB cameras and Deep Learning to address the issue of perspective distortion common to uncalibrated RGB cameras.

The proposed approach, termed "inductive" and contrasted with the traditional "analytical" approach, utilizes the inductive reasoning capability of Deep Learning models to learn the correlation between an observed posture (regardless of the viewpoint) and the corresponding ergonomic evaluation as determined by the RULA method and make predictions.

To train the model on a sufficiently large database of postures and evaluations capable of covering the widest possible range of viewpoints, a synthetic dataset was generated using the methodology described in the following paragraph.

The rest of the paper is structured as follows: Sect. 2 showcases the methodology that was followed to create the synthetic dataset and to train the model based on the inductive approach, while Sect. 3 illustrates the experimentations conducted to assess the model performance and the results of such validations. Section 4 discusses the results and finally Sect. 5 draws some conclusion while also discussing some limitations.

2 The inductive approach

Traditional approach to using MoCap systems based on Deep Learning for measuring physical ergonomics will be referred as "analytical" due to the use of mathematical equations for calculating ergonomic indexes from the joint angles extracted by RGB cameras. In this approach (illustrated in Fig. 1), the pose estimation technology analyses video footage, detects human figures within the frames, and extracts the coordinates of key points on the human body. Practically it assists, or in some cases even replace, the direct observation or manual video analysis typically performed by an ergonomist.

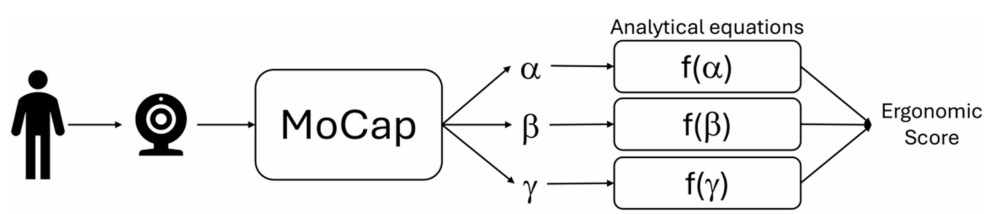


Fig. 1 Representative scheme of the analytical approach

Once the coordinates of the joint angles are extracted, an algorithm computes the ergonomic indexes according to the chosen evaluation method (e.g., RULA, REBA, OCRA). This approach is termed “analytical” because it involves calculating the ergonomic index by using the computed angles as input variables in the mathematical equations defined by the adopted evaluation methodology.

The analytical approach, however, as demonstrated in [41], suffers from perspective distortion, particularly difficult to resolve without adopting calibration. It has been previously shown that the accuracy of these tools is highest when subjects are oriented perpendicularly to the cameras capturing them. Conversely, accuracy degrades significantly as subjects rotate relative to the cameras, as the system no longer evaluates the true ergonomic angle but rather the apparent ergonomic angle - i.e., the projection of that angle onto the plane parallel to the camera sensor. This introduces a reading error that can become substantial, compromising the quality of the ergonomic assessment.

Adopting calibration procedures could partially or entirely address this issue, but at the cost of adding additional (often complex) steps, reducing the simplicity and cost-effectiveness of the tool, which would require calibration by qualified personnel.

Analysing the analytical approach, it is evident that ergonomic evaluation is achieved through a model that could be described as “bottom-up,” where individual components (ergonomic angles) are observed and combined to derive the evaluation.

However, a human observer, by watching another person, can reasonably assess whether the second person’s posture is “comfortable” or “uncomfortable,” safe or potentially harmful. Certainly, the human observer does not measure ergonomic angles nor analyse them one by one. Instead, they observe the entire posture and, using reasoning, judge its adequacy. This method follows a “top-down” model, where information is derived from the whole through reasoning. This inductive approach is illustrated in Fig. 2.

The “limitation” of the human observer is the inability to assign an objective score to a posture being observed. However, by taking inspiration from this information-processing model and leveraging Artificial Intelligence, it is possible to replicate this process in a more accurate and quantifiable manner.

The goal is to train a Machine Learning model capable of associating a given posture observed in an image with its corresponding ergonomic evaluation, without performing any analytical calculations. The final tool will still utilize Motion Capture models such as OpenPose or tf-pose-estimation to extract key points from images depicting a human subject. However, the difference from the analytical approach is that the ergonomic angles derived from these key points will not be used for mathematical calculations of RULA indices. Instead, they will all be used simultaneously as features to predict the ergonomic index by observing the entire posture.

To achieve this, it is necessary to create a dataset of postures, each labelled with its corresponding ergonomic evaluation, to be used for training the model using a supervised approach. The approach, from dataset creation to model training and validation, follows several phases that will be detailed below, as illustrated in Fig. 3:

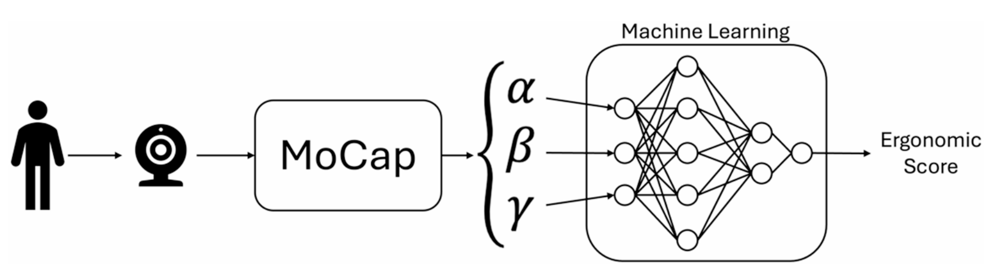
- Step 1: Virtual Human Mannequin creation.
- Step 2: Mannequin posing and coordinate extraction.
- Step 3: Data augmentation and dataset post-processing.
- Step 4: Creation of dataset labels.
- Step 5: Training and selection of the Machine Learning model.

2.1 Synthetic dataset creation methodology

The first phase of the methodology involves creating a virtual mannequin that will later be used for extracting posture coordinates. It is crucial to use a mannequin with accurate and realistic anthropometric proportions that can be posed without introducing artifacts in the mesh or skeleton, thereby avoiding distortions in subsequent coordinate extraction phases due to unnatural proportions or postures.

Once the digital mannequins are created, they should be posed to replicate the postures required for the analysis. This step involves configuring the mannequins to reflect all postures assessed by ergonomic evaluation methods such as RULA or REBA. These methods evaluate ergonomic risk by assessing specific anatomical angles (e.g., shoulder flexion/extension, neck flexion) and by categorizing them into predefined ranges, each associated with a specific risk score.

Fig. 2 Representative diagram of the inductive approach



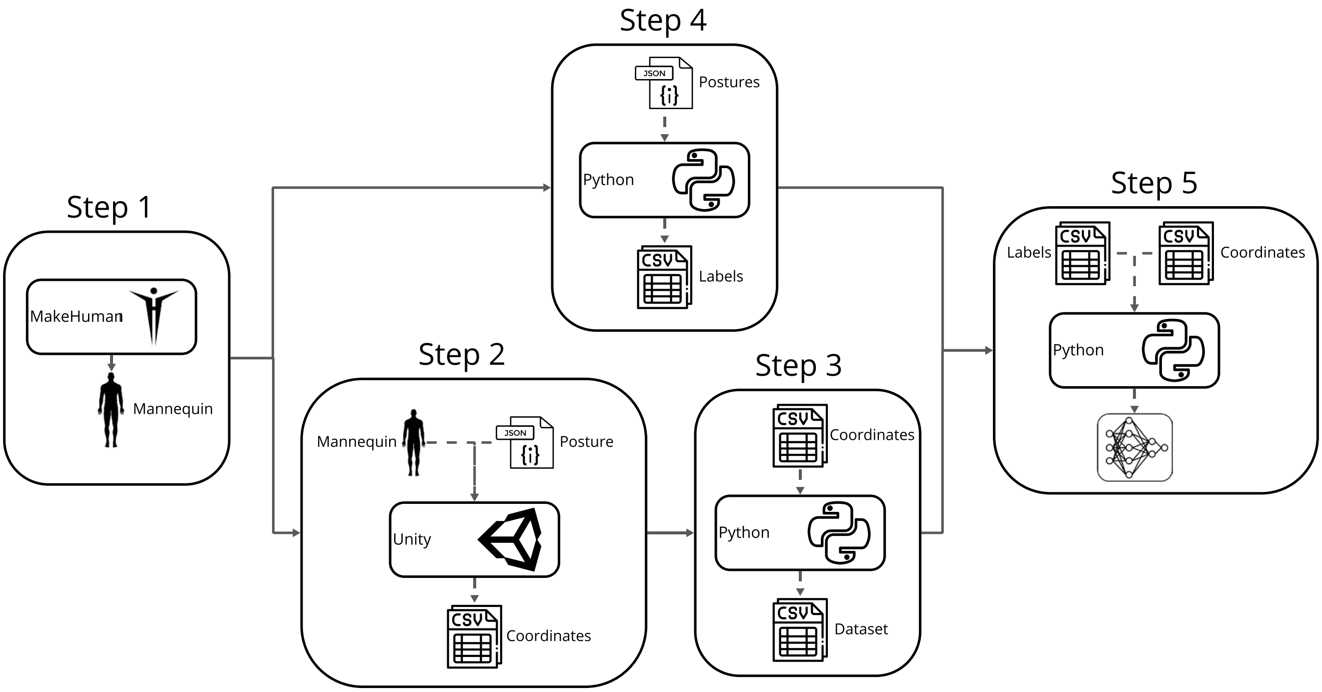


Fig. 3 Implementation scheme of the proposed methodology

In this way, risk categories are defined for each limb. By combining these categories, all assessable postures defined by the chosen method are generated and subsequently used to pose the virtual mannequin.

For instance, the RULA method specifies 5 risk categories for shoulder flexion/extension, 2 for shoulder abduction, 3 for elbow flexion, 3 for elbow flexion/extension, and 4 for flexion/extension of both the neck and the back.

Given that the human figure typically includes 2 shoulders, 2 forearms, and 2 wrists, the theoretical number of posture combinations totals 36 classes per limb configuration. However, due to the limitations of RGB motion capture systems in reliably tracking wrist positions, wrist angles are excluded from the analysis.

With this simplification, the total number of posture combinations (i.e., the number of 3-dimensional angles collections needed to pose the virtual mannequins) for each mannequin is as follows:

$$5 * 5 * 2 * 2 * 3 * 3 * 4 * 4 = 14400$$

Once the mannequins are posed, the coordinates of their limbs must be extracted. To enhance the robustness of the Machine Learning model trained on these postures, it is essential to ensure that the postures are extracted from multiple viewpoints. This accounts for the fact that the camera's positioning relative to the subject being evaluated cannot be predetermined. However, coordinates extracted from the virtual mannequins' postures represent an ideal condition. During deployment, the model may encounter two scenarios deviating from the ideal condition:

- *Occlusion*: The upstream MoCap model feeding into the ergonomic evaluation model might struggle to detect a certain key point due to occlusions or self-occlusions. As a result, the evaluation model could have to assess a posture with one or more missing key points.
- *Acquisition Uncertainty*: Since MoCap models provide an estimated key point position rather than a precise measurement, the actual position of a key point lies within a small neighbourhood of the estimated position.

Hence, the initial dataset containing the extracted coordinates should be augmented, in order to account for deviation from the ideal conditions. This step involves duplicating the dataset introducing subtle variations to the duplicated data, e.g., removing random key points to simulate occlusions or moving them away from their exact location to simulate acquisition uncertainty. Finally, the augmented dataset should be post-processed (i.e., normalized) in order to eliminate the "image size" factor.

Finally, each record in the posture datasets should be labelled with its corresponding true ergonomic evaluation according to the adopted ergonomic assessment method (e.g., RULA, REBA). The true ergonomic evaluation can be obtained by evaluating the 3-dimensional angles collections previously employed to pose the mannequins through an ergonomic methodology of choice. The full explanation of the implementation of this methodology is illustrated in Appendix A.

2.2 Machine learning model selection and training

The labelled dataset was used to train several Machine Learning models. The models were implemented in Python using the scikit-learn v1.5.1 [61] library. Specifically, the explored models were:

- Stochastic Gradient Descent Classifier (SGDClassifier).
- Linear Support Vector Classifier (LinearSVC).
- Kernel Approximation.
- KNeighbors Classifier.
- Multi-layer Perceptron (MLP).

For all five models, hyperparameter tuning (i.e., configuring the models' settings) was performed using the Grid Search technique to automate the process of finding the best configuration for each model.

The models were trained in a multilabel fashion, as they were trained on 6 labels simultaneously. The dataset was pre-processed by applying normalization using a StandardScaler and 80% of the dataset was used for training while the remaining 20% was reserved for validation. The comparison between the models, evaluated in terms of accuracy, is presented in Table 1.

The comparison shows that the MLP model is the most performant among the five. The best combination of parameters was found to be as follows:

- hidden_layer_sizes = 100.
- max_iter = 3000.
- activation = 'relu'.
- solver = 'adam'.

3 Experimental setup

The trained model underwent validation in two distinct phases. The first phase involves a preliminary validation, using data from a synthetic dataset generated with a virtual mannequin that differs from the one used to create the training dataset. This validation process compared the model's predicted angle values against the actual angles of the mannequin. The methodology for obtaining these ground-truth angles was consistent with the approach used to create the training dataset.

The second and final phase of validation utilized data collected from real human subjects instead of virtual mannequins. This phase involved comparing the ergonomic evaluations predicted by the tool, which implements the inductive approach, with those derived from direct measurements taken on real participants. To assess both static and dynamic performance, the validation was divided into two distinct tests: a static test and a dynamic test, with detailed descriptions provided later. Despite the inherent differences between static and dynamic postures, the same validation procedure was applied to both tests.

3.1 Preliminary validation on synthetic data

The preliminary validation of the model was performed by testing the trained model with the data derived from a synthetic dataset different from the training dataset. The dataset was generated and processed using the same procedure described earlier but starting with a different virtual mannequin than those initially used.

The preliminary validation yielded the following accuracy values:

- Right Upper Arm: 73.40%.
- Left Upper Arm: 72.72%.
- Right Lower Arm: 79.76%.
- Left Lower Arm: 79.77%.
- Neck: 57.37%.
- Trunk: 78.01%.
- Average: 73.50%.

3.2 Experimental procedure for the final validation

For both tests, 18 XSens WMT (Wireless Motion Tracker) sensors were applied to the subjects after obtaining their informed consent regarding the procedures. The XSens WMT inertial MoCap system allows for recording user movements and exporting anthropometric measurements, body segments, and joint angles. This enables accurate and objective assessment of body movements for each recorded frame. All output data are specifically suited for calculating the main anatomical joint angles, which are then used

Table 1 Accuracy obtained from the examined models

	Upper Arm dx [%]	Upper Arm sx [%]	Lower Arm dx [%]	Lower Arm sx [%]	Neck [%]	Trunk [%]	Media [%]
SGDClassifier	54.27	54.46	66.32	65.75	70.53	69.89	63.54
LinearSVC	54.52	54.00	65.84	66.26	69.86	71.92	63.69
Kern. Approx.	39.67	40.08	66.10	65.94	30.15	29.29	45.20
KNeighbours	75.11	74.39	66.17	65.05	73.30	85.76	73.30
MLP	**94.14**	**94.05**	**89.47**	**89.95**	**90.55**	**96.33**	**92.14**

for ergonomic evaluation. The following body parts were specifically chosen, considering the left and right sides of the body separately:

- Upper Arm: assessing flexion/extension and abduction angles.
- Lower Arm: assessing elbow flexion angle.
- Neck: assessing flexion/extension angle.
- Trunk: assessing flexion/extension angle.

The angles were calculated using algorithms specifically developed by the sensor manufacturer for processing the data recorded by the MoCap system.

Simultaneously, the subjects were recorded using standard RGB cameras (Microsoft LifeCam Studio, set to record at 720p and 30 FPS) to obtain videos for evaluation by the system under study. The positioning of the cameras relative to the subjects will be discussed in later subsections.

Given the difference in acquisition frame rates between the two systems (30 FPS for the cameras and 60 FPS for the XSens system), a synchronization procedure was required. The XSens system was started first. Then, one of the cameras was used to record the screen of the PC running the XSens software, specifically the section displaying the current frame number. Once this was captured, the camera was repositioned to the predetermined acquisition point, and the static or dynamic postures were initiated. The recording was not stopped until all poses were completed. This setup allowed the XSens frame corresponding to the first frame of each video to be read.

Knowing that the two systems have a frame rate ratio of 2:1, synchronization of the two acquisitions was achieved. If the initial XSens frame in the first frame of a video is fx_0, any subsequent XSens frame fx corresponding to a given video frame fv can be calculated using the following formula:

$$fx = 2fv + fx_0 \tag{1}$$

In this way, once the video segments to be analysed were extracted, it was possible to identify the corresponding XSens segments, establishing a correlation between the two systems.

The angular measurements provided by XSens were then processed by a Python script to compute ergonomic evaluations frame-by-frame. Finally, given the 2:1 frame rate ratio between the XSens system and the system under evaluation, the frame-by-frame evaluations were aggregated into second-by-second evaluations by averaging the values within the same second.

For the video system under evaluation, the video segments were first processed by the AI-based RGB MoCap system, tf-pose-estimation. This allowed the extraction of key point coordinates from the video frames. The set of coordinates then underwent the same procedures of normalization, feature extraction, and scaling with a StandardScaler as the training dataset, after which it was provided as input to the evaluation system to generate ergonomic assessments frame-by-frame. These assessments were similarly reduced to second-by-second evaluations using arithmetic averaging. This ensured the same number of ergonomic evaluations for both systems, enabling direct comparison.

The arithmetic averaging assumed that the movements in the dynamic test were relatively slow rather than rapid, allowing for minimal (if any) variation in ergonomic evaluations between frames within the same second. This assumption is particularly valid for the XSens system at 60 FPS (16 ms between frames), where the acquisition is much faster than the average human reaction time (around 250 ms), but it also holds true for the video system at 30 FPS (33 ms). For the static test, where subjects remained stationary, this issue did not arise.

The comparison between the trained model and the XSens system for validation purposes was performed using accuracy, calculated as the ratio between the correct predictions (TP, True Positives) and the total observations (N):

$$accuracy = \frac{TP}{N} \bullet 100 \tag{2}$$

Finally, it is important to note that the model predicts the partial results of the RULA method (i.e., Right and Left Upper Arm, Right and Left Lower Arm, Neck, and Trunk). To obtain the ergonomic evaluation in terms of Grandscore and RULA Risk Level, the modifiers discussed before must be added. Therefore, the validation would not be complete without this final step.

To achieve this, a comparison was conducted between the Grandscore and Risk Level values obtained through the XSens system and the trained model. All modifiers and the partial score related to the wrists (currently not covered by the model) were set to their respective minimum allowable values for both systems. A summary of the values predicted by the system and those manually input is presented in Table 2.

3.2.1 Static postures

For the validation with static postures, 15 subjects were involved, comprising 8 men and 7 women aged between 27 and 37 years (mean age 31.3 years) with varying heights and weights. The subjects posed in 19 different postures; for each posture, only one angle was imposed and measured using an orthopaedic goniometer before data acquisition began, while the other angles were left free. This ensured that each posture always had at least one angle consistent across participants.

The postures (shown in Fig. 4), named after the imposed angle, were recorded from either a frontal or lateral view depending on the case, as detailed in Table 3. The camera was positioned 2 m away from the subject, mounted on a tripod, and set at a height of 1.5 m to capture the entire human figure. The change between frontal and lateral views was achieved by rotating the subject rather than moving the camera, which remained stationary. The layout is depicted in Fig. 5.

The second-by-second measurements obtained through the previously described procedure were further averaged. Specifically, all values corresponding to the same posture were averaged together, as no differences were observed within the same second due to the static nature of the postures under examination. As a result, $N = 19$ ergonomic evaluations were obtained for each subject.

3.2.2 Dynamic postures

The validation with dynamic postures involved the same 15 subjects from the static posture test, recorded while performing a simulated work task in a laboratory setting. The task required the subject, standing in front of a table, to move to their right to pick up an empty box placed on a support 1 m above the ground. Once the box was picked up, it was placed on the table. The subject then opened the flaps of the box and simulated operations inside it, first inserting their right arm, then their left arm, bending over to view the interior of the box. After completing these tasks, the box was closed and returned to its original position.

Next, the subject picked up a second box, this time placed to their left, performing similar operations to those carried out with the first box. Once the sequence of simulated tasks was completed, the closed box was placed on top of the first box. Finally, the subject returned to a standing, upright posture in front of the table.

The simulated task (lasting 48 s and repeated twice) was recorded simultaneously from three different viewpoints (shown in Fig. 6), with the videos synchronized by turning on a light bulb used as a “clapperboard.”.

The first camera, with a frontal view, was placed on the other side of the desk, 2 m away from the subject and 1.5 m above the ground. The presence of the desk between the subject and the camera obstructed the view of the lower body from the waist down, introducing significant occlusion.

The second camera, with a lateral view, was positioned at the same distance and height as the first but to the subject’s right. The absence of visual obstacles allowed a full view of the subject during all phases of the task.

The third camera, with a diagonal view, was placed at an intermediate point between the frontal and lateral views. This perspective introduced less occlusion than the first camera (typically only the body portion below the knees was occluded) but allowed simultaneous observation of both lateral angles (e.g., trunk flexion) and frontal angles (e.g., shoulder abduction), albeit from a viewpoint that was not orthogonal to any anatomical plane. The other two cameras, positioned perpendicular to the coronal and sagittal planes respectively, only captured angles expressed within these planes.

The layout of the dynamic test is shown in Fig. 7. The measurements obtained from the XSens and the trained model were reduced to second-by-second measurements following the procedure described earlier, resulting in $N = 48$ samples, one for each second of video duration.

Table 2 Summary of the predicted values and those manually entered

Modifiers/Parameters	Value
Upper arm position	Predicted by the model
Raised shoulder	0
Abducted shoulder	Predicted by the model
Supported arm	0
Lower arm position	Predicted by the model
Arms across the midline	0
Wrist position	1
Wrist lateral bending	0
Wrist twist	0
Upper body muscle use	0
Upper body load	0
Neck position	Predicted by the model
Neck twist	0
Neck lateral bending	0
Back position	Predicted by the model
Back twist	0
Back lateral bending	0
Supported legs	1
Lower body muscle use	0
Lower body load	0

3.3 Results

This section provides the results for both the angular and the ergonomic assessment validations.

3.3.1 Angular assessment validation

Results for both the static and the dynamic tests are presented in Table 4, and the confusion matrices related to such results are presented in Appendix B. The static posture data from the three subjects were aggregated for analysis since they were recorded from the same viewpoints. The results of these analyses reveal that the model demonstrates good

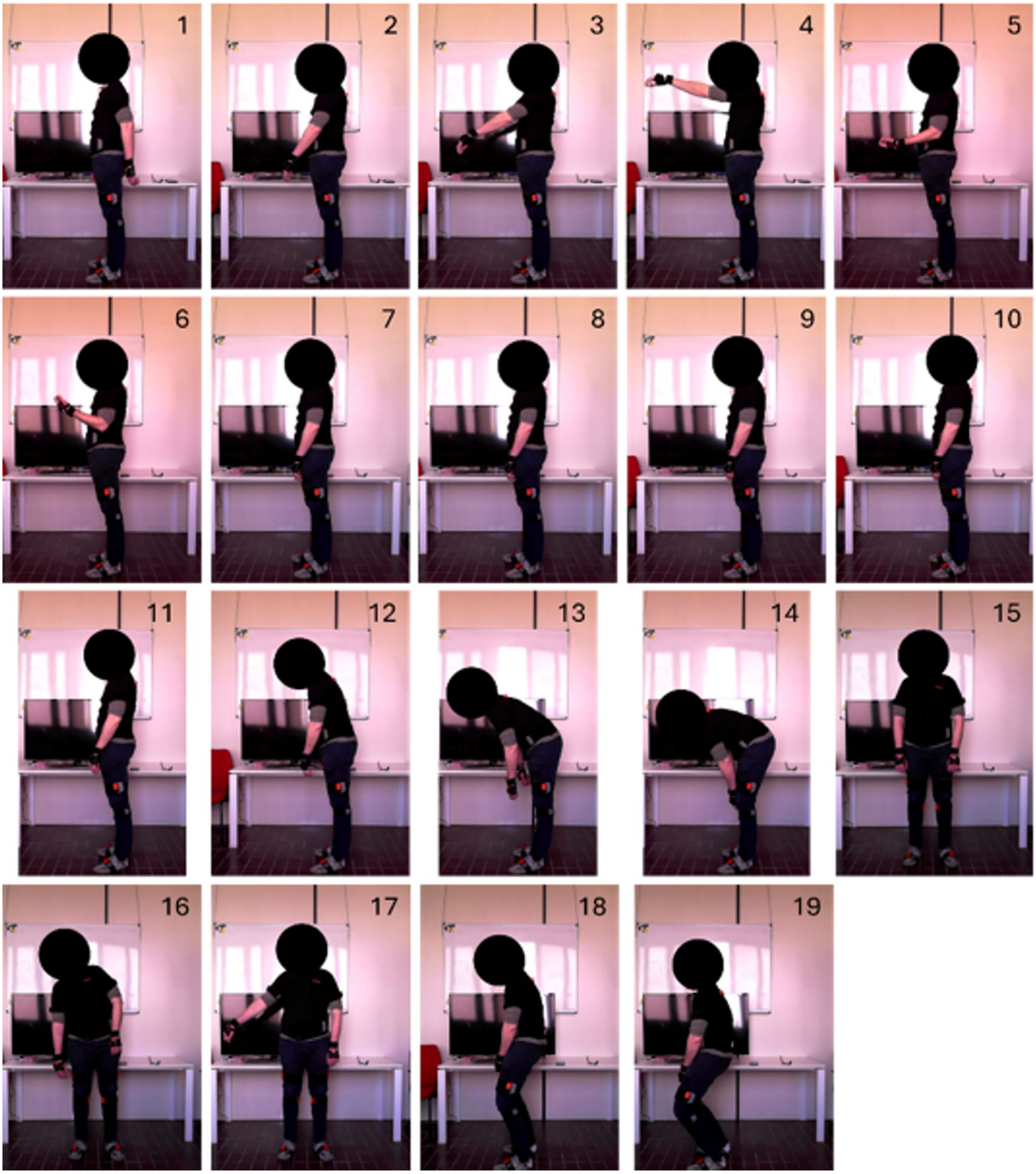


Fig. 4 Postures considered in static validation

accuracy in estimating elbow angles, with values of 87.43% for the right elbow and 95.22% for the left elbow.

The accuracy for the left shoulder, which was consistently in view, is 92.62%. In contrast, for the right shoulder, which was completely occluded in 16 of the 19 postures analysed, the accuracy is lower, at 74.86%. The neck evaluation shows marginally sufficient accuracy, at 50.55%, confirming the issues already identified during the preliminary validation

Table 3 Details of the examined postures

Postures	Frontal	Lateral
shoulderExt20		x
shoulderFlex20		x
shoulderFlex45		x
shoulderFlex90		x
sshoulderAbd45	x	
elbowFlex60		x
elbowFlex100		x
neckExt10		x
neckFlex0		x
neckFlex10		x
neckFlex20		x
neckBend20	x	
backFlex0		x
backFlex20		x
backFlex60		x
backFlex90		x
backBend20		x
kneeExt30		x
kneeExt60	x	

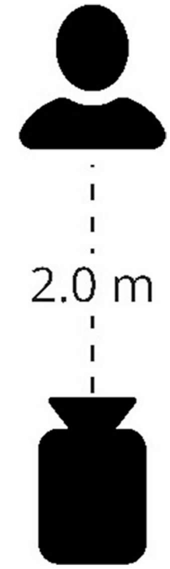
phase. Finally, the accuracy for the back is high, at 90.71%. Overall, the model's performance for static postures is good, with an average accuracy across all segments of 81.90%.

For the analysis of data from the dynamic tests, each viewpoint (diagonal, frontal, and lateral) was analysed separately, though data from the two different videos were aggregated. The best average performance was observed for the frontal view, with an average accuracy of 75.68%. The system achieved good overall accuracy, with values of 88.27% and 78.57% for the right and left shoulders, respectively, and 90.31% and 98.98% for the right and left elbows, respectively. However, the accuracy for the neck, at 50.51%, is only marginally sufficient, while the back shows even lower accuracy, at 47.45%.

Following the frontal view, the lateral view ranks second in performance with an average accuracy of 72.24%. The accuracy values for the right and left shoulders (73.01% and 80.98%, respectively) are good, as are those for the right and left elbows (87.40% and 82.78%, respectively). In this case, the accuracy for the back is also high, at 86.38%. However, the neck accuracy is notably poor, with a value of just 22.88%.

The last viewpoint analysed is the diagonal view, whose results reveal the lowest average accuracy (71.15%). In this case, the segment with the best performance is the back, with an accuracy of 90.86%, while the lowest accuracy is once again observed for the neck, at 25.27%. The calculated accuracy for the elbows remains high (89.25% for the right and 83.87% for the left), as does the accuracy for the left shoulder, at 81.18%. However, the right shoulder shows a modest accuracy of 56.45%.

Fig. 5 Layout of the static test



Overall, the results demonstrate satisfactory performance of the model, aligning in most cases with the observations made during the preliminary validation phase.

3.3.2 Ergonomic assessment validation

Regarding the final verification with RULA Grandscore and Risk Level, the static postures resulted in accurate RULA Grandscore predictions at 74.73%, with a corresponding accuracy for RULA Risk Level at 77.87% (Fig. 8).

For dynamic postures, the diagonal view produced the best results, with a Grandscore accuracy of 76.88% and a Risk Level accuracy of 87.63% (Fig. 9). This was followed by the lateral view, with a Grandscore accuracy of 74.73% and a Risk Level accuracy of 77.87% (Fig. 10). Lastly, the frontal view showed the lowest performance, with a Grandscore accuracy of 31.12% and a Risk Level accuracy of 54.59% (Fig. 11).

4 Discussion

The preliminary validation of the model demonstrated good overall model performance. However, it also highlighted a clear limitation: the model's inability to make accurate predictions for the Neck segment. This limitation may be attributed to two factors: (i) the movements of the training mannequin's neck are limited in amplitude, given the narrow range of motion considered by the RULA method. This reduces the variability of training data related to the neck compared to other segments and (ii) the eyes are always positioned relatively close together, especially when viewed from a non-frontal perspective. As a result, the midpoint between the eyes (used to calculate the neck angle) remains

Fig. 6 Framing considered in the test with dynamic postures

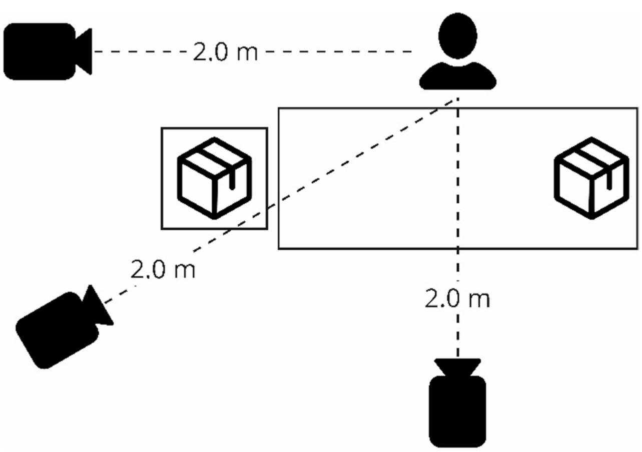


Fig. 7 Test layout with dynamic postures

confined to a small neighbourhood of a fixed position, further limiting the variability of neck-related data. Supporting this hypothesis, the standard deviation of neck angles in the training dataset is 28.22°, the lowest among all segments (Back = 32.41°, Right Shoulder = 44.76°, Left Shoulder = 44.77°, Right Elbow = 50.41°, Left Elbow = 50.59°).

The validation with real data from static postures confirmed the good performance observed during the preliminary validation (even surpassing it) and once again highlighted the model's inability to predict accurate values for the neck.

For the validation with real data from dynamic postures, different considerations arise depending on the viewpoint analysed. The frontal viewpoint appears to lead to better overall performance, with the highest average accuracy among the three viewpoints. However, examining the individual segment scores reveals that this higher average is driven by slightly better scores for shoulders and elbows compared to the diagonal and lateral viewpoints, as well as a higher neck score. What stands out most is the score for the back, which is the lowest of all (both compared to the other two viewpoints and to the static tests). There are two potential reasons for this low score: (i) the frontal viewpoint does not allow the system to capture meaningful back angle values, which cannot be accurately estimated from this perspective and (ii) the frontal viewpoint was the most occluded.

The human figure was indeed not visible below the waist, making it difficult for the upstream MoCap model to estimate the position of the knees, which are necessary for calculating the back angle.

Conversely, the higher neck score for the frontal viewpoint among the three perspectives seems to confirm the earlier hypothesis regarding the proximity of the eyes. The frontal viewpoint allows for a clearer separation between the right and left eyes, resulting in improved precision in detecting their positions.

The other two viewpoints, diagonal and lateral, show higher scores for shoulders, elbows, and back compared to the frontal view. What drastically lowers the average accuracy for these two viewpoints is the extremely low score for the neck. The fact that the lowest neck angle accuracies were observed in the two viewpoints where distinguishing between the two eyes is most challenging, once again supports the previously formulated hypothesis. This is further

Table 4 Results of the angular analysis for static and dynamic tests

Test	Right Upper Arm	Left Upper Arm	Right Lower Arm	Left Lower Arm	Neck	Trunk
Static	74.86%	92.62%	87.43%	95.22%	50.55%	90.71%
Frontal	88.27%	78.57%	90.31%	98.98%	50.51%	47.45%
Lateral	73.01%	80.98%	87.40%	82.78%	22.88%	86.38%
Diagonal	56.45%	81.18%	89.25%	83.87%	25.27%	90.86%

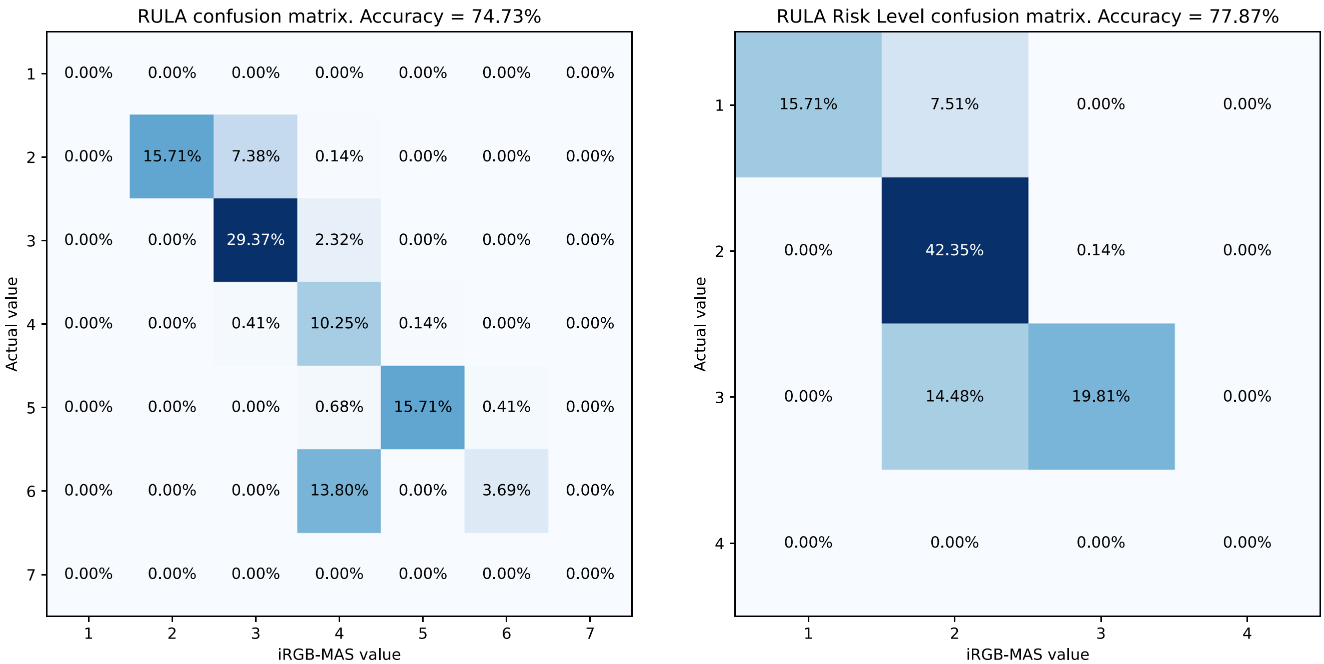


Fig. 8 Final RULA values for static postures

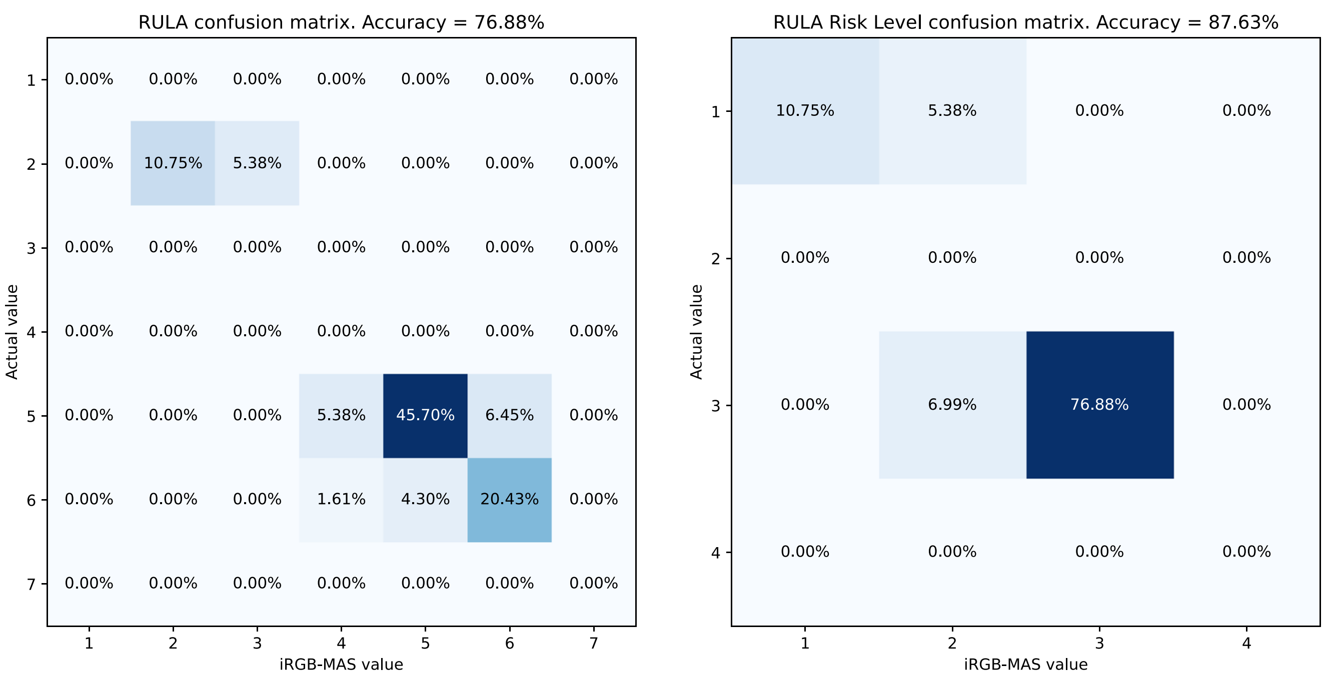


Fig. 9 Final RULA values for dynamic postures, diagonally framed

evidenced by the lateral viewpoint, where the eyes overlap the most, yielding the lowest neck accuracy, albeit slightly.

Regarding the accuracy of the RULA Grandscore and Risk Level, an interesting phenomenon emerges: the viewpoint producing the generally best partial results (the frontal one) results in the worst performance. While the lateral and diagonal viewpoints have neck accuracies roughly half that of the frontal viewpoint, they exhibit back accuracies around 90%, which is approximately double that of the frontal view.

Analysing the formulation of the RULA method reveals that neck and back angles have equal weight in the final evaluation. Consequently, it becomes apparent that it is preferable to have at least one of these angles with high accuracy, even at the expense of the other, rather than having two similar values with accuracies around 50%.

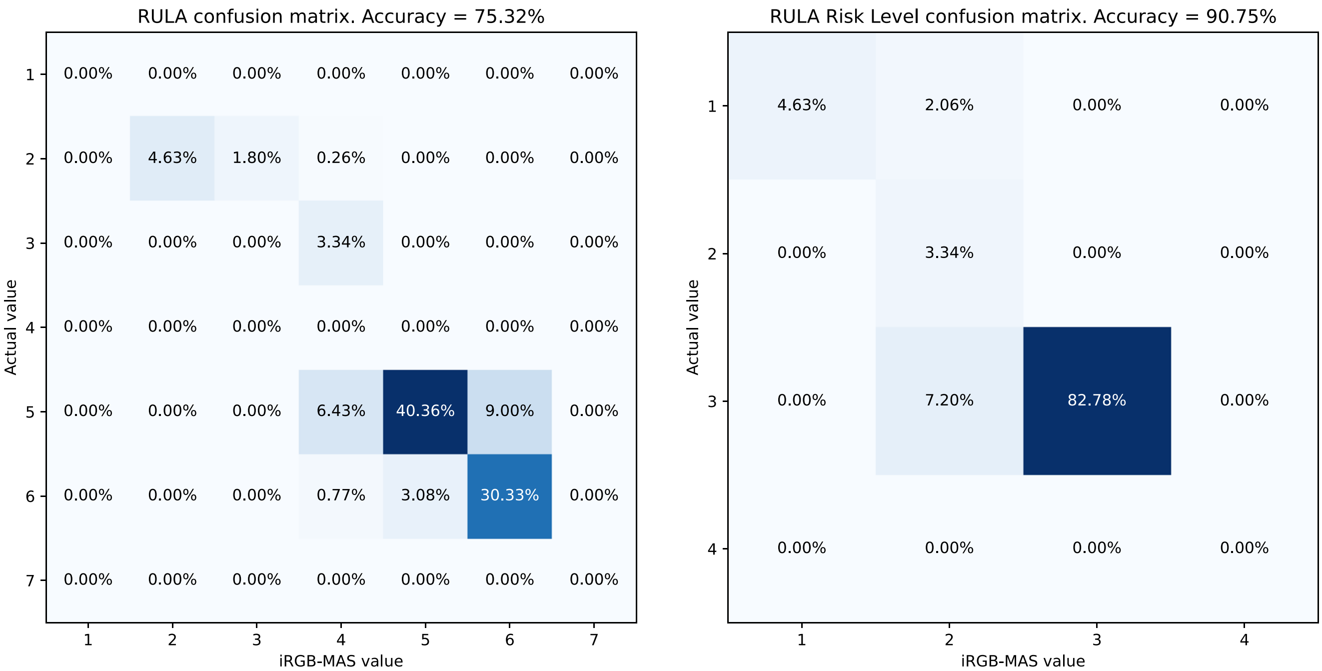


Fig. 10 Final RULA values for dynamic, laterally framed postures

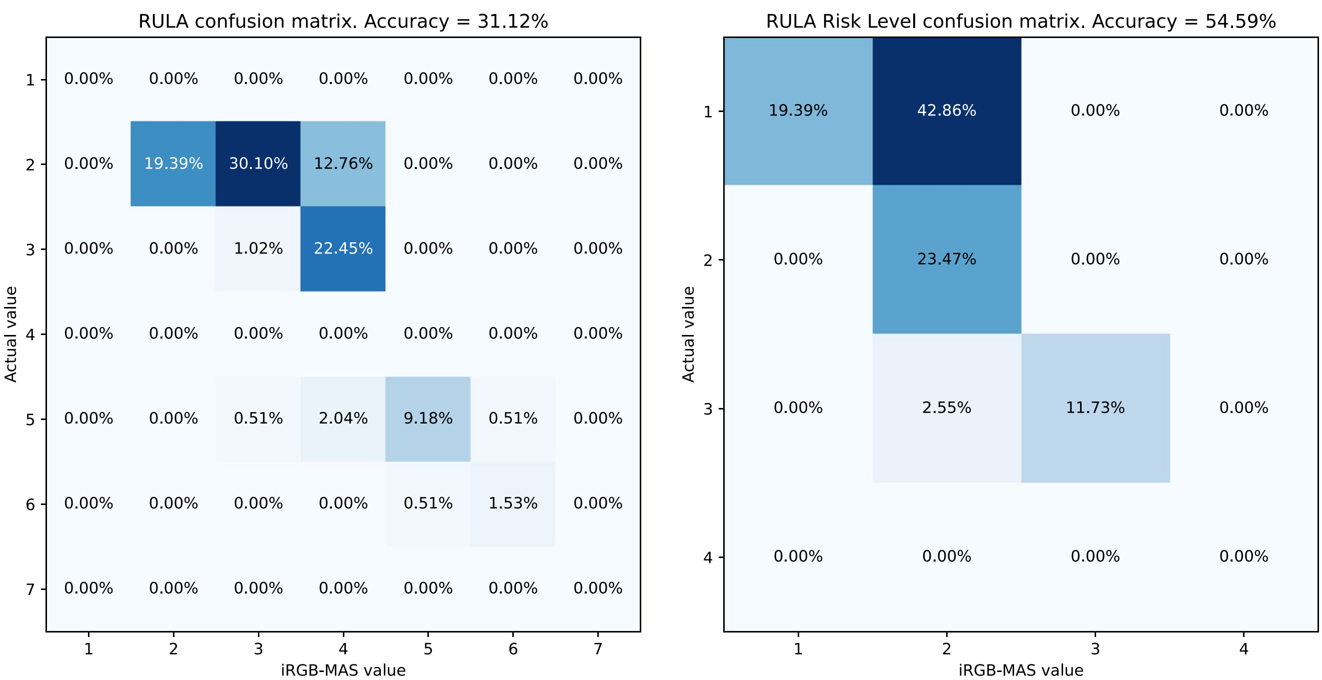


Fig. 11 Final RULA values for dynamic postures, frontally framed

As a result, the slightly higher accuracy for neck and back in the diagonal viewpoint compared to the lateral viewpoint leads to higher accuracies for the Grandscore and Risk Level, with the Risk Level approaching 90%.

However, it should be noted that obtaining a frontal viewpoint is often impractical in an industrial manufacturing context. As demonstrated in [49], the easiest viewpoints to achieve are the lateral (to some extent) and especially the diagonal, which is also the most comprehensive in terms of captured angles. Despite only capturing apparent angles (due to perspective distortion), it is therefore not surprising that the diagonal viewpoint produces the best results.

5 Conclusion

The dynamic nature of modern production, with frequent workstation changes, task variability, and the integration of collaborative robots, challenges the maintenance of proper posture and the prevention of work-related musculoskeletal disorders (MSDs). As the workforce ages, considering the variability of working capacities will become even more critical. Human Factors must be integrated into the design of machines and systems to reduce ergonomic risks and improve efficiency. Amongst the Human Factors, ergonomics specifically addresses the measurement of physical ergonomics, providing tools for ergonomic validation during workstation design and for monitoring work postures to enhance the physical ergonomics of existing workstations.

The objective of this paper was to develop a novel “inductive” approach to applying Deep Learning technologies for monitoring physical ergonomics in industrial manufacturing contexts as an alternative to the more traditional “analytical” approach widely adopted in the literature. The goal of the inductive approach is to address the main issue of the analytical approach, i.e., perspective distortion.

A system based on the inductive approach observes posture in a holistic manner: the angles extracted from the posture are not used as inputs for a set of equations characterizing the selected evaluation method. Instead, they serve as an objective description of the posture as a whole.

Such system, therefore, does not perform mathematical calculations on the angles but “observes” them and based on this observation predicts the ergonomic evaluation. The advantage of this approach is that, if the system is trained on a representative dataset of postures captured from a sufficient variety of angles, the issue of perspective distortion becomes negligible. Additionally, with appropriate data augmentation procedures applied to the dataset, the problem of occlusion can also be mitigated.

Training such a model requires a large and balanced training dataset capable of enabling the learning of the correlation between an observed posture and its corresponding ergonomic evaluation. However, obtaining such a dataset is complex and scarcely feasible when considering the need to sample postures of real individuals, for example, using wearable sensors, due to the vast amount of data that would need to be collected.

The innovative elements of this paper are then the development of the so-called “inductive approach” when measuring physical ergonomics in a real manufacturing environment and the adoption of a synthetic dataset to train a Machine Learning model based on such approach. The validation of the obtained model highlighted good results, achieving nearly 90% prediction accuracy and using a single viewpoint.

Potential future developments include:

I. transitioning to frameworks better suited for production environments, such as TensorFlow. This shift would enable the use of advanced optimization algorithms for training the neural network, potentially improving its robustness and accuracy.
II. testing the system in a real industrial manufacturing environment to assess its actual performance.
III. developing a system that accounts not only for varying viewpoints but also for the distance between the camera and the subject.

Furthermore, the model has only been tested in a laboratory setting. Its actual performance will become evident once tested in real industrial manufacturing environments. Although qualitative testing has indicated some robustness to occlusion for certain viewpoints, this aspect requires further investigation.

Lastly, while the training dataset included a broad representation of viewpoints, giving the model some robustness to variable perspectives (though certain viewpoints remain “preferential” due to their ability to provide a more complete observation of the posture), the issue of variable camera-to-subject distance remains unresolved. In the case of viewpoints, imagining the subject enclosed in an observability sphere, the sphere can be divided into sectors, each corresponding to a viewpoint. If these viewpoints overlap slightly, the resulting dataset can reasonably approximate the range of possible perspectives from which a posture might be observed.

The situation is different for camera-to-subject distance, as it is not feasible to define in advance the lower and upper limits of the distance at which the subject will be observed. In other words, the range of the sphere’s radius cannot be predetermined, making it challenging to build a dataset covering all scenarios. However, since we have trained our model on body angles and not limbs length, it is possible to assume that the distance won’t substantially impact the model’s performance, given that angles are invariant with the distance.

Appendix A - Synthetic dataset creation

The creation of virtual mannequins was carried out using the MakeHuman software [62], specifically through its MassProduce plugin. This tool generated 100 mannequins of various genders and skin tones, limited to the age range of the working population. A selection of the mannequins is shown in Fig. 12.

This procedure ensures that the variability of anthropometric proportions of the human body is respected in a real application context. Once created, each mannequin was equipped with the "CMU" rig, available among those offered in MakeHuman. This rig was chosen because it replicates the virtual skeleton produced by MoCap tools such as OpenPose or tf-pose-estimation We then developed a Python script to handle the generation of the 3-dimensional angles collections (i.e., the postures). The script combines the relevant risk categories and then generates random angles for each class in each posture. Finally, it writes the postures and their corresponding angles to a JSON file. Each mannequin had its own JSON file containing different angles for each posture.

Fig. 12 Sample of the mannequins created with Mass produces

The mannequin posing, camera management, and coordinate acquisition were implemented in Unity [56], a graphics engine commonly utilized for creating video games and interactive applications. In Unity, the virtual mannequins previously rigged in MakeHuman were imported in FBX format, and the following procedures were performed for each of the mannequins. A set of C# scripts handles all the required functions. The overall system architecture is shown in Fig. 13 Specifically, upon importing the mannequin ("Character"), two scripts are activated: #DrawPlane and #ComputeAngles.

The first script provides the calculation functions (and drawing functions, solely for debugging purposes) for the three anatomical planes perpendicular to each other that pass through the human body (visible in Figs. 14 and 15):

- Sagittal Plane: divides the human body into right and left sections. Its normal is represented by the red line in Fig. 15;
- Coronal Plane: divides the human body into ventral and dorsal sections. Its normal is represented by the blue line in Fig. 15;
- Transverse Plane: divides the human body into upper and lower sections. Its normal is represented by the green line in Fig. 15;

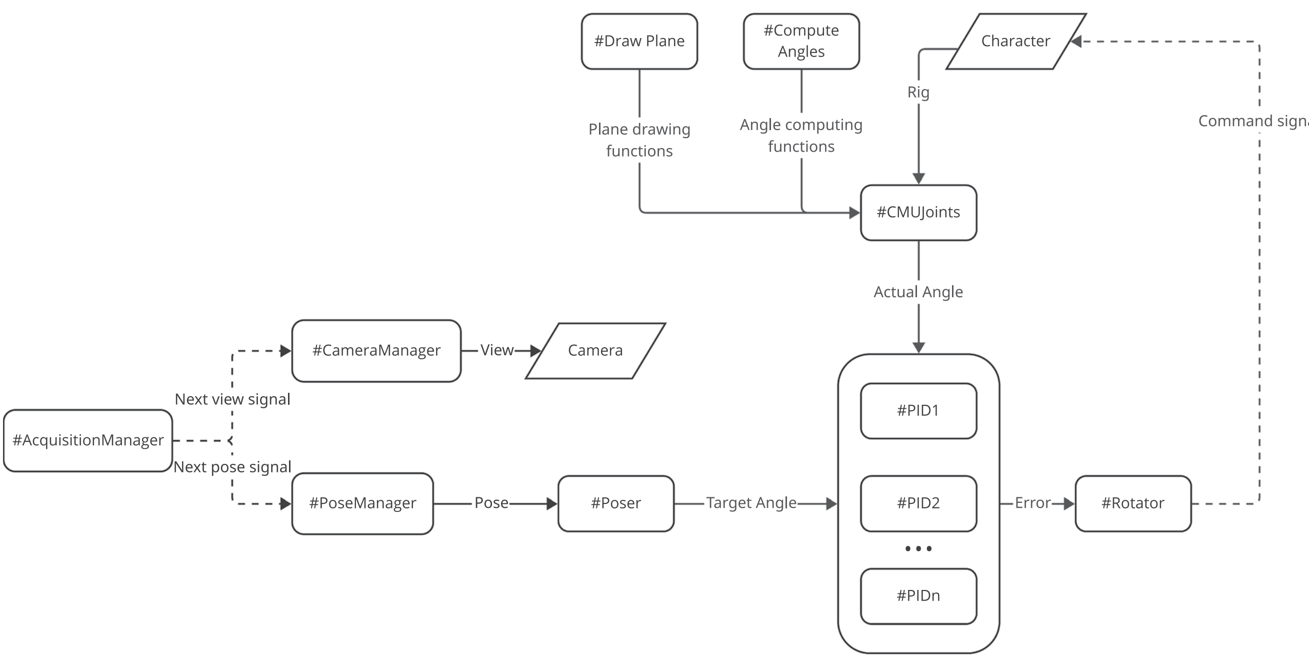


Fig. 13 Flowchart of the application developed in Unity

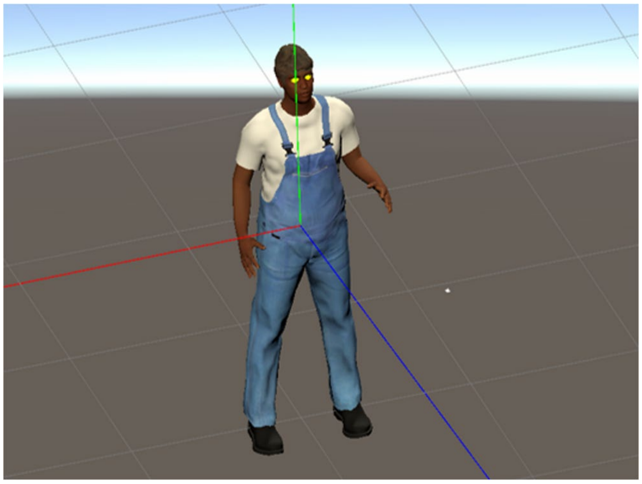


Fig. 14 Normal vectors of the anatomical planes as drawn in Unity

The second script calculates ergonomic angles by projecting them onto the anatomical planes. Subsequently, the #CMUJoints script retrieves the positions of the CMU model's key points from the mannequin and calculates the actual ergonomic angles using the functions provided by the two previous scripts. The output of this script is thus a set of actual ergonomic angles characterizing the mannequin's current posture.

Simultaneously, the #PoseManager script reads target postures from a JSON file containing the target angles and communicates them to the #Poser script. The #Poser script transmits the corresponding target angle to PID modules (one for each controllable joint of the mannequin). Each #PID script calculates the error between the target angle (provided by #Poser) and the actual ergonomic angle

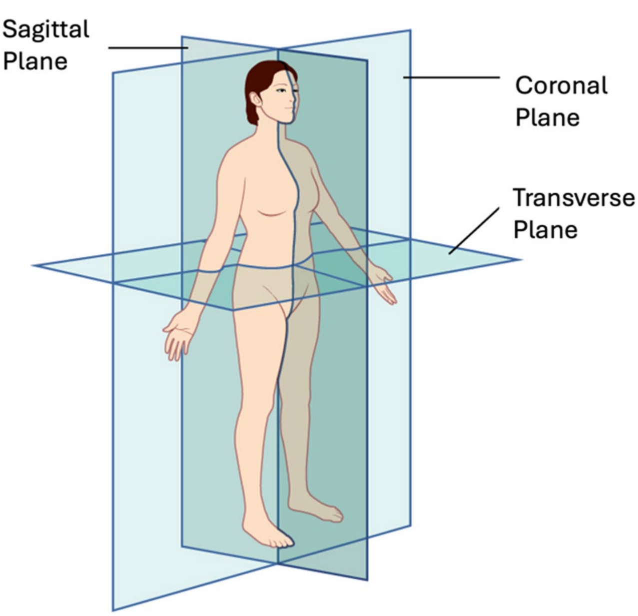


Fig. 15 Representation of anatomical planes

(calculated by #CMUJoints) using a proportional-integral-derivative (PID) feedback mechanism.

The adoption of a PID control system was motivated by the fact that, while the target angles are expressed in the coordinates of the anatomical planes, Unity can control the mannequin's joints around the X, Y, and Z directions of a local reference system attached to each joint. Rather than implementing an analytical rotation system based on transformation equations, a closed-loop control was preferred.

In this process, the errors calculated by the various PID modules are communicated to the #Rotator script, which is responsible for rotating the limbs until the target posture is achieved.

The management of viewpoints is handled by the #CameraManager script. This script, inspired by [63], performs roto-translations of the camera around the virtual mannequin, which is thus placed at the centre of an "observability sphere" (Fig. 16). The camera captures the virtual mannequin from three heights:

- At a height corresponding to the mannequin's neck;
- At a height H1 corresponding to the mannequin's knees;
- At a height H2 above the neck equal to the distance between the mannequin's knees and neck.

For each height, the camera completes a full rotation around the mannequin in 45° increments, while maintaining a constant distance from the virtual mannequin. For each increment, the camera captures the 2-dimensional coordinates of the virtual mannequin's limbs. Consequently, with 8 images captured per height, the total number of captures per posture amounts to 24. Figure 17 shows an example of the virtual mannequin from the 24 captures for the first posture, while Fig. 18 displays the same example with the skeleton of the same mannequin.

Finally, the #AcquisitionManager script handles the following tasks:

- Instructing #PoseManager to retrieve the next posture from the JSON file, thereby triggering the mannequin posing process.
- Instructing #CameraManager to move to the next viewpoint.
- Acquiring the 2D coordinates of the virtual mannequin's key points within the resulting image from the current viewpoint.

As previously mentioned, each JSON file (one for each mannequin) contained 14400 different postures each. Considering the 24 images captured for each posture, the resulting dataset comprised 345,600 records, one for each combination of posture and viewpoint, acquired over approximately 40 h for each mannequin. At the end of the

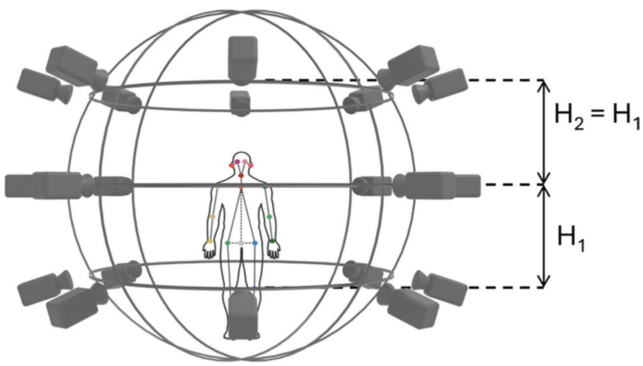


Fig. 16 Sphere of observability

acquisition process, the datasets were saved in CSV format for further processing.

The operations for data augmentation, normalization, feature extraction, and creation of training labels were performed using Python scripts. To prepare the evaluation model for the "occlusion" scenario, the dataset was duplicated, and for each duplicated record, one key point (randomly chosen) was removed, simulating the absence of a key point due to occlusion. To prepare the model for the "acquisition uncertainty" scenario, the original dataset was duplicated again, with each coordinate shifted horizontally (for x-coordinates) or vertically (for y-coordinates) by a random amount within the interval $[0; \epsilon]$, with $\epsilon = 2px$.

Fig. 17 Example with the virtual mannequin of the 24 shots of the first posture

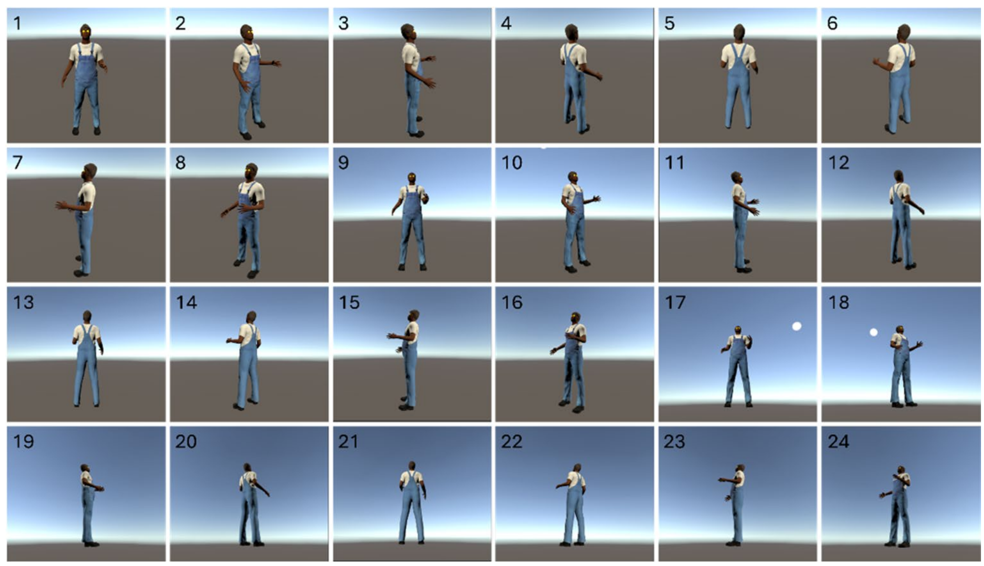
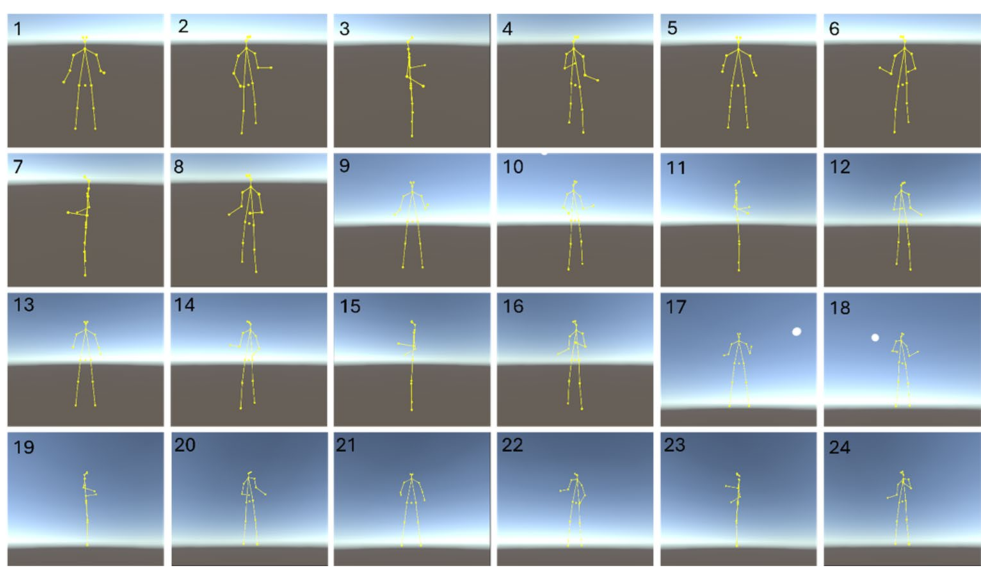


Fig. 18 Example with the skeleton relative to the virtual mannequin of the 24 shots relative to the first posture



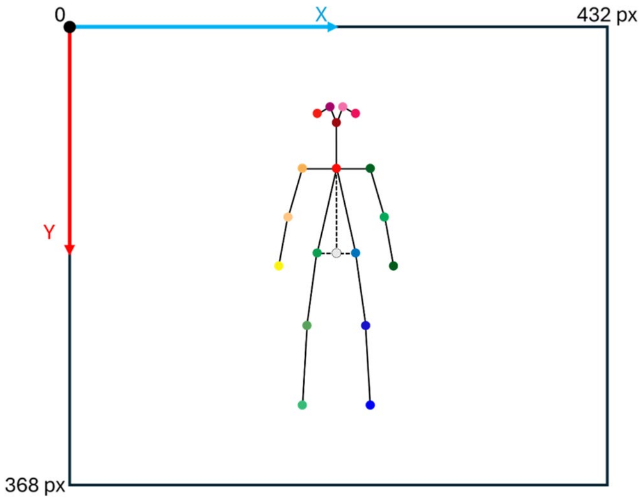


Fig. 19 The image reference system

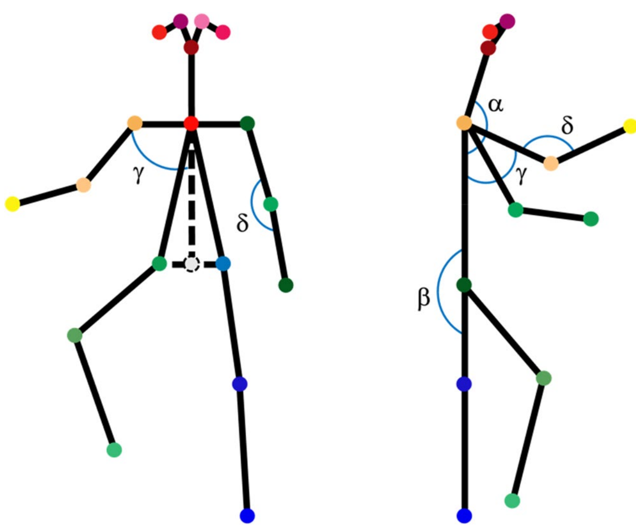


Fig. 20 Representation of the angles constituting the training features

At the end of the data augmentation phase, each original dataset was thus tripled. This approach was inspired by the one described in [64].

Subsequently, to eliminate the "image size" factor, the datasets were normalized. Specifically, the image size captured in Unity is 432 × 368px, chosen to match the cropped image size produced by tf-pose-estimation, with the reference system located at the top-left corner of the image, the x-axis pointing to the right, and the y-axis pointing downward (Fig. 19). The coordinates are thus within the ranges:

$$\begin{cases} 0 \leq \ x \leq \ 432 \\ 0 \leq \ y \leq \ 368 \end{cases}$$

The normalization of coordinates to a $[0; 1]$ range on both axes allows the image size to be disregarded, making the ergonomic evaluation model compatible with MoCap models other than OpenPose or tf-pose-estimation. Normalization was performed using the following equations:

$$\begin{cases} x^{'} \ = x/432 \\ y^{'} \ = y/368 \end{cases} \tag{3}$$

Finally, based on the normalized coordinates, the following apparent ergonomic angles were calculated for each posture during feature extraction (as shown in Fig. 20):

- Neck angle α : measured between the segment connecting the midpoint between the hips and the base of the neck, and the segment connecting the base of the neck to the midpoint between the eyes.
- Back angle β : measured between the segment connecting the midpoint between the knees and the midpoint between the hips, and the segment connecting the midpoint between the hips to the base of the neck.
- Right and left arm angles γ : measured between the segment connecting the midpoint between the hips and the base of the neck, and the segment connecting the shoulder to the elbow.
- Right and left elbow angles δ : measured between the segment connecting the shoulder to the elbow, and the segment connecting the elbow to the wrist.

These angles constitute the 6 features of the training datasets for the Machine Learning models.

To create labels for the dataset we adopted the RULA method. This ergonomic assessment method, once scores are assigned to the various joints, requires the application of modifiers to such scores. After applying the modifiers and consulting lookup tables, the final ergonomic evaluation is obtained. Acquiring these modifiers is either entirely impossible to automate with MoCap alone (e.g., the weight of a lifted object) or pertains to movements (e.g., neck torsion) that are not easily detectable with current technologies. Therefore, the choice was made to monitor only the ergonomic scores that can be directly measured by the tool (i.e., the scores associated with ergonomic angles), leaving it to the ergonomist to apply the modifiers later to obtain the final evaluation. It follows that the tool, configured this way, cannot immediately provide a complete ergonomic evaluation but rather only a measure of the incorrectness of the posture itself. Nevertheless, this measure is valuable in a work operations monitoring context: if a purely angular posture discrepancy is detected, the ergonomist is promptly notified and can intervene by focusing solely on the flagged posture. This allows the ergonomist to determine whether the issue is due to an incorrect posture adopted by the worker or a poorly designed workstation.

The dataset records are thus labelled not with a single label (i.e., the RULA Grandscore) but with a set of labels, namely:

- Left and right Upper Arm;
- Left and right Lower Arm;
- Neck;
- Trunk;

for a total of 6 labels per record.

Naturally, these ergonomic evaluations should not be calculated based on the apparent ergonomic angles but rather on the real ones. These angles were the same ones previously used for posing the virtual mannequin (i.e., the 3-dimensional angles collections), ensuring that each posture is assigned a correct ergonomic evaluation regardless of the camera's viewpoint.

By combining the datasets obtained after feature extraction with the labels, the training dataset was created, comprising $345600\ postures{*}3\ repetitions = 1036800\ records$, each characterized by 6 features and 6 labels. Considering that we created 100 different mannequins, and that each mannequin has its own features dataset and labels dataset, by combining all of them we obtained a master training dataset of 10368000 records. The master dataset is the one on which we trained the neural network.

Appendix B

Fig. 21 Confusion matrix for the static angular accuracy test

Fig. 22 Confusion matrix for the dynamic angular accuracy test (shot with frontal viewpoint)

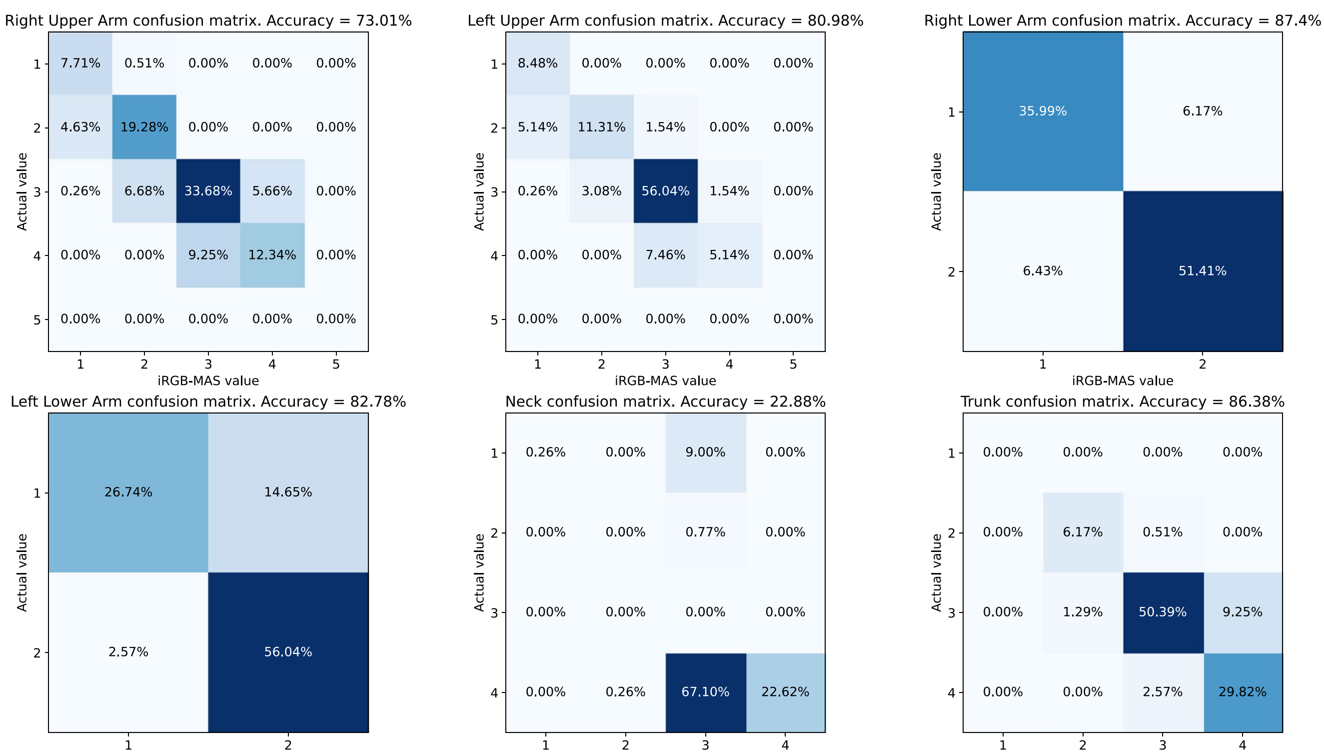


Fig. 23 Confusion matrix for the dynamic angular accuracy test (shot with lateral viewpoint)

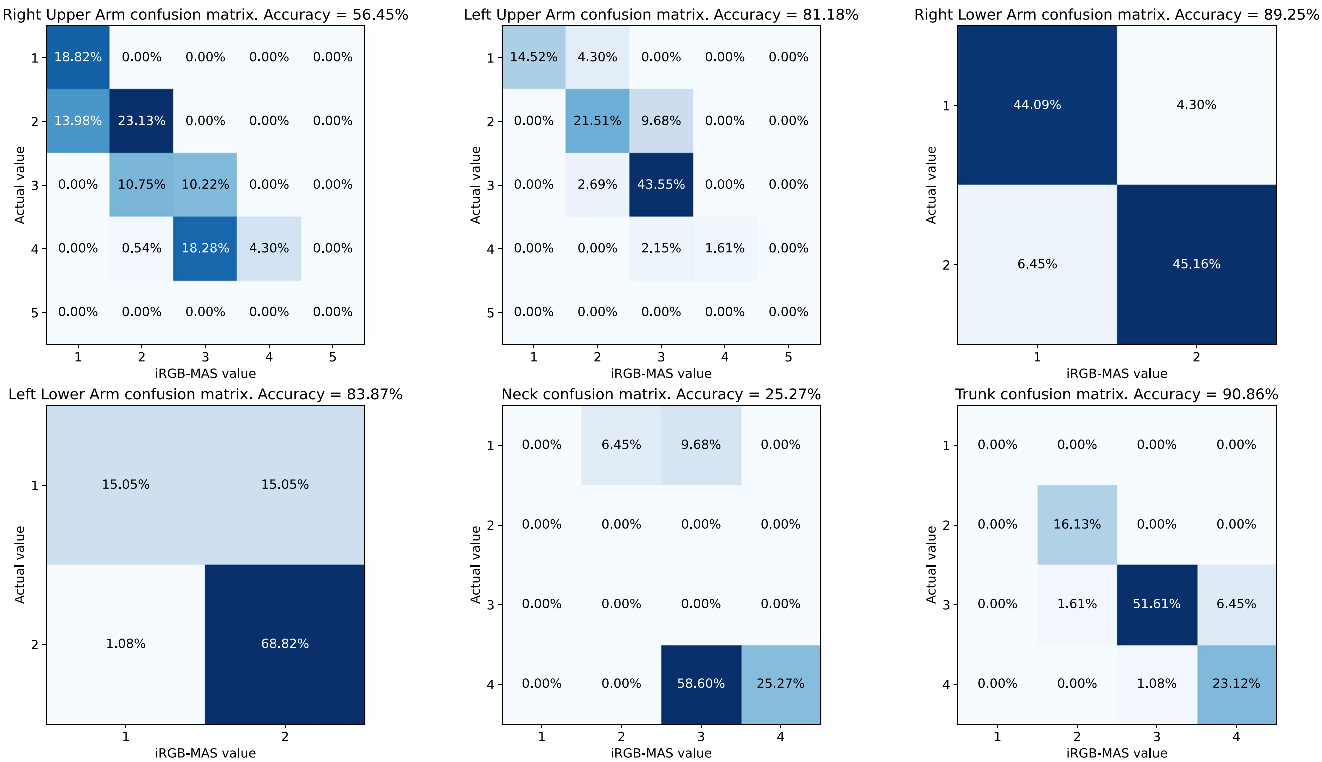


Fig. 24 Confusion matrix for the dynamic angular accuracy test (shot with diagonal viewpoint)

Authors' contributions Thomas Agostinelli: Conceptualization, Methodology, Software, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Visualization Andrea Generosi: Methodology, Software, Validation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing Maura Mengoni: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Università Politecnica delle Marche (Ancona, Italy), protocol code 0021586 and date of approval February 7, 2023.

Compliance with ethical standards Not applicable.

Informed consent Informed consent was obtained from all subjects involved in the study.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Ciccarelli M, Papetti A, Germani M (2023) Exploring how new industrial paradigms affect the workforce: a literature review of operator 4.0. J Manuf Syst 70:464–483
2. Dennerlein JT (2017) Ergonomics and Musculoskeletal Issues. In: Quah SR (ed) International Encyclopedia of Public Health (Second Edition). Academic Press, Oxford, pp 577–584. https://doi.org/10.1016/B978-0-12-803678-5.00139-9
3. Simões AC, Pinto A, Santos J, Pinheiro S, Romero D (2022) Designing human-robot collaboration (HRC) workspaces in industrial settings: a systematic literature review. J Manuf Syst 62:28–43
4. Kuorinka I et al (1987) Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergon 18(3):233–237
5. Hedge A, Morimoto S, McCrobie D (1999) 'Cornell musculoskeletal discomfort questionnaire', *Ergonomics*, Accessed: Oct.

07, 2024. [Online]. Available: https://psycnet.apa.org/doiLanding?doi=10.1037/t60061-000
6. David G, Woods V, Li G, Buckle P (2008) The development of the quick exposure check (QEC) for assessing exposure to risk factors for work-related musculoskeletal disorders. Appl Ergon 39(1):57–69
7. Waters TR, Putz-Anderson V, Garg A, Fine LJ (1993) Revised NIOSH equation for the design and evaluation of manual lifting tasks. Ergonomics 36(7):749–776. https://doi.org/10.1080/00140139308967940
8. Monnington SC, Quarrie CJ, Pinder AD, Morris LA (2003) 'Development of manual handling assessment charts (MAC) for health and safety inspectors'. Contemp Ergon, pp. 3–8
9. 'HSE - ART tool' Accessed: Oct. 07, 2024. [Online]. Available: https://www.hse.gov.uk/msd/uld/art/index.htm
10. McAtamney L, Corlett EN (1993) RULA: a survey method for the investigation of work-related upper limb disorders. Appl Ergon 24(2):91–99
11. Hignett S, McAtamney L (2000) Rapid entire body assessment (REBA). Appl Ergon 31(2):201–205
12. Occhipinti E (1998) OCRA: a concise index for the assessment of exposure to repetitive movements of the upper limbs. Ergonomics 41(9):1290–1311. https://doi.org/10.1080/001401398186315
13. Karhu O, Kansi P, Kuorinka I (1977) Correcting working postures in industry: a practical method for analysis. Appl Ergon 8(4):199–201
14. Lowe BD, Dempsey PG, Jones EM (2019) Ergonomics assessment methods used by ergonomics professionals. Appl Ergon 81:102882
15. Battini D, Persona A, Sgarbossa F (2014) Innovative real-time system to integrate ergonomic evaluations into warehouse design and management. Comput Ind Eng 77:1–10
16. Huang C, Kim W, Zhang Y, Xiong S (2020) Development and validation of a wearable inertial sensors-based automated system for assessing work-related musculoskeletal disorders in the workspace. Int J Environ Res Public Health 17(17):6050
17. Vignais N, Miezal M, Bleser G, Mura K, Gorecky D, Marin F (2013) Innovative system for real-time ergonomic feedback in industrial manufacturing. Appl Ergon 44(4):566–574
18. Mengoni M, Raponi D, Ceccacci S (2016) A method to identify VR-based set-up to foster elderly in design evaluation. Int J Intell Eng Inf 4(1):46. https://doi.org/10.1504/IJIEI.2016.074501
19. 'Motion Capture' Accessed: Jan. 26, 2024. [Online]. Available: https://www.movella.com/products/motion-capture
20. 'Blue Trident IMU | Inertial Sensor by Vicon | Biomechanic Tracking', Vicon. Accessed: Jan. 26, 2024. [Online]. Available: https://www.vicon.com/hardware/blue-trident/
21. Welch G, Foxlin E (2002) Motion tracking: no silver bullet, but a respectable arsenal. IEEE Comput Graph Appl 22(6):24–38
22. 'Nexus | Software For Motion Capture In Life Sciences', Vicon. Accessed: Jan. 26, 2024. [Online]. Available: https://www.vicon.com/software/nexus/
23. 'Motion Capture Systems' OptiTrack. Accessed: Jan. 26, 2024. [Online]. Available: http://optitrack.com/index.html
24. Manghisi VM, Uva AE, Fiorentino M, Gattullo M, Boccaccio A, Evangelista A (2020) Automatic ergonomic postural risk monitoring on the factory shopfloor–the ergosentinel tool. Procedia Manuf 42:97–103
25. Schall MC, Sesek RF, Cavuoto LA (2018) Barriers to the adoption of wearable sensors in the workplace: a survey of occupational safety and health professionals. Hum Factors 60(3):351–362. https://doi.org/10.1177/0018720817753907
26. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) 'Realtime multi-person 2d pose estimation using part affinity fields', in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291–7299. Accessed: Jan. 26, 2024. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Cao_Realtime_Multi-Person_2D_CVPR_2017_paper.html
27. 'GitHub - tryagainconcepts/tf-pose-estimation Deep Pose Estimation implemented using Tensorflow with Custom Architectures for fast inference.' Accessed: Jan. 30, 2024. [Online]. Available: https://github.com/tryagainconcepts/tf-pose-estimation
28. 'MediaPipe' Google for Developers. Accessed: Jan. 30, 2024. [Online]. Available: https://developers.google.com/mediapipe
29. Bradski G (2000) 'The OpenCV library'. Dr Dobb's J Soft Tools
30. Simon T, Joo H, Matthews I, Sheikh Y (2017) 'Hand keypoint detection in single images using multiview bootstrapping', in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1145–1153. Accessed: Jan. 30, 2024. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Simon_Hand_Keypoint_Detection_CVPR_2017_paper.html
31. He Ling W, Yun-Ju LEE (2019) 'Occupational evaluation with Rapid Entire Body Assessment (REBA) via imaging processing in field', in Human Factors Society Conference, Accessed: Jan. 30, 2024. [Online]. Available: https://orbit.dtu.dk/files/197181254/NES_2019_Proceedings_v07_FINAL.pdf#page=157
32. Ota M et al (2020) Verification of reliability and validity of motion analysis systems during bilateral squat using human pose tracking algorithm. Gait Posture 80:62–67
33. De Freitas PVA et al (2019) 'An ergonomic evaluation method using a mobile depth sensor and pose estimation', in Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, Rio de Janeiro Brazil: ACM, Oct. pp. 445–452. https://doi.org/10.1145/3323503.3349550
34. Deshpande AM, Kumar R, Kumar M 'IoT based AI Application for Posture Recognition to Reduce Workplace Injuries'
35. MassirisFernández M, Fernández JÁ, Bajo JM, Delrieux CA (2020) Ergonomic risk assessment based on computer vision and machine learning. Comput Ind Eng 149:106816
36. da Silva Neto JG, Teixeira JMXN, Teichrieb V (2020) 'Analyzing embedded pose estimation solutions for human behaviour understanding', in Anais Estendidos do XXII Simpósio de Realidade Virtual e Aumentada, SBC, pp. 30–34. Accessed: Jan. 30, 2024. [Online]. Available: https://sol.sbc.org.br/index.php/svr_estendido/article/view/12951
37. Chunduru V, Roy M, Chittawadigi RG 'Hand tracking in 3d space using mediapipe and pnp method for intuitive control of virtual globe', in (2021) IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, 2021, pp. 1–6. Accessed: Jan. 30, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9641587/
38. Halder A, Tayade A (2021) Real-time vernacular sign language recognition using mediapipe and machine learning. J. Homepage Www Ijrpr Com ISSN 2582:7421
39. Zhang F et al 'MediaPipe Hands: On-device Real-time Hand Tracking', Jun. 17, 2020, *arXiv*: arXiv:2006.10214. Accessed: Jan. 30, 2024. [Online]. Available: http://arxiv.org/abs/2006.10214
40. Agostinelli T, Generosi A, Ceccacci S, Mengoni M (2024) Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments. Sci Rep 14(1):27785
41. Khan F, Salahuddin S, Javidnia H (2020) Deep learning-based monocular depth estimation methods—a state-of-the-art review. Sensors 20(8):2272. https://doi.org/10.3390/s20082272
42. Ehret T (2023) Monocular depth estimation: a review of the 2022 state of the art. Image Process On Line 13:38–56. https://doi.org/10.5201/ipol.2023.459
43. Von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018. Lecture

Notes in Computer Science, vol 11214. Springer International Publishing, vol. 11214., Cham, pp 614–631. doi: https://doi.org/10.1007/978-3-030-01249-6_37.

44. Zhang Z, Hu L, Deng X, Xia S (2021) ‘Sequential 3D Human Pose Estimation Using Adaptive Point Cloud Sampling Strategy’, in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, Aug. pp. 1330–1337. https://doi.org/10.24963/ijcai.2021/184
45. Dahman D ‘ DisguiseData A Proposed Method for Creating Synthetic Data Based on The Geometric Structure of The Original’, 2024, Accessed: Oct. 18, 2024. [Online]. Available: https://www.researchsquare.com/article/rs-4978385/latest
46. Kim S, Jeon Y (2024) ‘Development of a Python Library to Generate Synthetic Datasets for Artificial Intelligence Education.’, Int. J. Adv. Sci. Eng. Inf. Technol., vol. 14, no. 3, Accessed: Oct. 18, 2024. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=20885334&AN=178289184&h=mHxTb38M3oxZoAz5zlpYOOeJA5DJ5hbz6zDMsip%2FsWWhncuVa1ogAIdBAyw0f92i7p8MEXSTICMoznPM0Y4%2FoQ%3D%3D&crl=c
47. Fragkouli S-C et al (2024) ‘Synthetic data: how could it be used in infectious disease research?’, Future Microbiol., vol. 19, no. 17, pp. 1439–1444, Nov. https://doi.org/10.1080/17460913.2024.2400853
48. Kenul E et al (2024) ‘ReplicantTM framework for synthetic data generation’, in Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications II, SPIE, pp. 445–459. Accessed: Oct. 18, 2024. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13035/130351E/ReplicantTM-framework-for-synthetic-data-generation/https://doi.org/10.1117/12.3013826.short
49. Bothma BC, Luwes N, ‘Exploring the use of Synthetic Training Data for the Classification of Electronic Components in Artificial Intelligence Systems’, in (2024) 16th International Conference on Human System Interaction (HSI), IEEE, 2024, pp. 1–6. Accessed: Oct. 18, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10613537/?casa_token=5H-S4h70NtQAAAAA:cB5b5LE0fSumNd_1Nn3MtUgNQtjRio1SLwWkcB2CoWMVOd5COmfFz7zATtkWNWz0qw5qxYtr
50. Schneider D (2024) IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 13038–13045. Accessed: Oct. 18, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10611486/?casa_token=EeJQqdhzoq4AAAAA:Fl2q5HPPDTAJENXCo_hE5n1yevTHK0Z5F6oD_RTY3KWnxiQyq2Zx04gpQ3vqrImOOvLiypXZ
51. Liaquat M, Nugent C, Cleland I (2023) ‘Using Synthetic Data to Improve the Accuracy of Human Activity Recognition’, in Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023), vol. 842, J. Bravo and G. Urzáiz, Eds., in Lecture Notes in Networks and Systems, vol. 842., Cham: Springer Nature Switzerland, pp. 167–172. https://doi.org/10.1007/978-3-031-48642-5_16
52. Geissbühler D, Shahreza HO, Marcel S (2024) ‘Synthetic Face Datasets Generation via Latent Space Exploration from Brownian Identity Diffusion’, Apr. 30, arXiv: arXiv:2405.00228. Accessed: Oct. 18, 2024. [Online]. Available: http://arxiv.org/abs/2405.00228
53. Canas PN, Ortega JD, Nieto M, Otaegui O ‘Virtual passengers for real car solutions: synthetic datasets’, May 13, 2022, arXiv: arXiv:2205.06556. Accessed: Oct. 18, 2024. [Online]. Available: http://arxiv.org/abs/2205.06556
54. Delussu R, Putzu L, Fumera G (2024) ‘Synthetic Data for Video Surveillance Applications of Computer Vision: A Review’, Int. J. Comput. Vis., vol. 132, no. 10, pp. 4473–4509, Oct. https://doi.org/10.1007/s11263-024-02102-x
55. Umer F, Adnan N (2024) Generative artificial intelligence: synthetic datasets in dentistry. BDJ Open 10(1):13
56. ‘Unity Real- Time Development Platform | 3D, 2D, VR & AR Engine’, Unity. Accessed: Jan. 31, 2024. [Online]. Available: https://unity.com
57. B. M (2024) Improving AI model performance by augmenting synthetic data. Int J Multidiscip Res 6(2):18972. https://doi.org/10.36948/ijfmr.2024.v06i02.18972
58. Khalzaa K, Karungaru S, Terada K ‘Creation and Testing of Synthetic Datasets for Training Road Scenes Algorithms’, in (2023) 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2023, pp. 1587–1592. Accessed: Oct. 18, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10309307/?casa_token=bOikfOWnTRsAAAAA:toMPT3P-JKQS-1ZnrNFhJREcyz75T9yNRYeq2NbFGqvYTaZJdDhtCNgHm3serTVWaNUlS7Do
59. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) ‘The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes’, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3234–3243. Accessed: Oct. 18, 2024. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Ros_The_SYNTHIA_Dataset_CVPR_2016_paper.html
60. Savva M, Chang AX, Hanrahan P, Fisher M, Nießner M (2014) Scenegrok: inferring action maps in 3D environments. ACM Trans Graph 33(6):1–10. https://doi.org/10.1145/2661229.2661230
61. ‘scikit-learn machine learning in Python — scikit-learn 1.5.1 documentation’. Accessed: Sep. 04, 2024. [Online]. Available: https://scikit-learn.org/stable/
62. ‘www.makehumancommunity.org’ Accessed: Aug. 29, 2024. [Online]. Available: http://www.makehumancommunity.org/
63. Yan X, Li H, Wang C, Seo J, Zhang H, Wang H (2017) Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. Adv Eng Inf 34:152–163
64. Li L, Martin T, Xu X (2020) A novel vision-based real-time method for evaluating postural risk factors associated with musculoskeletal disorders. Appl Ergon 87:103138