SPECTRE: a Deep Learning network for Posture Recognition in Manufacturing

Marianna Ciccarelli[2], Flavio Corradini[1], Michele Germani[2], Giacomo Menchi[2], Leonardo Mostarda[1], Alessandra Papetti[2] and Marco Piangerelli[1*]

[1]Computer Science Division, University of Camerino, via Madonna delle Carceri, Camerino, 62032, Italy.
[2]Department of Industrial Engineering and Mathematical Sciences, Poytechnic University of Marche, via Brecce Bianche, Ancona, 60131, Italy.

*Corresponding author(s). E-mail(s): marco.piangerelli@unicam.it;
Contributing authors: m.ciccarelli@staff.univpm.it; flavio.corradini@unicam.it; m.germani@staff.univpm.it; g.menchi@staff.univpm.it; leonardo.mostarda@unicam.it; a.papetti@staff.univpm.it;

Abstract

Work-related musculoskeletal disorders are a very impactful problem, both socially and economically, in the manufacturing sector. To control their effect, standardised methods and technologies for ergonomic assessment have been developed. The main technologies used are inertial sensors and vision-based systems. The former are accurate and reliable, but invasive and not affordable for many companies. The latter use machine learning algorithms to detect human pose and assess ergonomic risks. In this paper, using data collecting by reproducing the working environment in LUBE, the major Italian kitchen manufacturer, we propose SPECTRE (Sensor-independent Parallel dEep ConvoluTional leaRning nEtwork): a fully sensor-independent learning model based on convolutional networks to classify postures in the workplace. This system assesses ergonomic risks in major body segments through Deep Learning with a minimal impact. SPECTRE's performance is evaluated using established metrics for imbalanced data (precision, recall, F1-score and

area under the precision-recall curve). Overall, SPECTRE shows good performance and, thanks to an agnostic explainable machine learning method, is able to extrapolate which patterns are significant in the input.

Keywords: Computer Vision, Deep Learning, Ergonomic Risks, Human-Centered Manufacturing, Posture Recognition, Work-related musculoskeletal disorders.

1 Introduction

One of the major challenges for health in manufacturing environments is finding ways to prevent musculoskeletal disorders. Work-related musculoskeletal disorders (WMSDs) are the most prevalent occupational health problem affecting roughly three out of every five workers in the EU-28 of all sectors and occupations [1]. Its incidence is rapidly increasing due to workforce ageing. WMSDs have a multifactorial nature (i.e., physical, organisational, and psychosocial risk factors) and affect several anatomical regions such as the back, neck, shoulder, and wrist. In addition to pain, functional limitations, impairment, absence from work, etc. they have a significant socio-economic impact on companies, society at large, and workers' personal lives [2]. In particular, the manufacturing sector shows high levels of sick leave and an high rate of absenteeism due to WMDS. The back and upper limbs (e.g. wrists and elbow) are the most affected body areas. Moreover, according to data by economic sectors, the manufacturing sector suffers the highest economic losses due to MSDs. For instance in Germany, there are about EUR 6.45 million loss of production and EUR 10.63 million loss of gross value added [1]. The need for awareness, regulatory pressure, and workers' complaints have led to the development and spread of numerous standardised methods and tools (Ovako Working posture Analysing System, Rapid Entire Body Assessment, Rapid Upper Limb Assessment, etc.) for assessing the risks of WMSDs. They were designed for use by ergonomists, health and safety inspectors, occupational doctors, etc. and they usually require the assignment of scores based on the direct observation of workers while performing their work or video recordings. Some methods also require a discussion with stakeholders to better interpret results, understand the causes, hypothesise interventions, and define how to put them into practice. However, they often need a discussion with workers to arrive at the most objective scores possible [3]. The subjectivity or the evaluator bias are the main limitations of these approaches, in addition to the monitoring of limited periods of time (temporal instants or snapshots). From these considerations arises the need for objective evaluation tools (i.e., direct measurement), which allow for a long duration of data collection and are more accurate. They could be used to improve human ergonomics in dynamic scenarios, providing real-time feedback to workers or adapting working conditions (e.g., human-robot collaboration). They could be sensor-based or vision-based

systems. The former refers to the emerging use of wearable inertial sensing technology in occupational ergonomics. It includes several sensors such as accelerometers, inclinometers, gyroscopes, magnetometers, and inertial measurement units (IMUs). Their use in lab settings prevails, whereas applied industrial settings still lag. Lim and D'Souza, in their review [4], point out the following interesting issues to deal with: full-body measurement (17 body-worn inertial sensors) can be obtrusive and affect wearability; inertial sensors tend to lack the context of the performed tasks needing the incorporation of additional methods (e.g., direct observations, self-reported measures); and few studies offer real-time feedback functionality. Moreover, accurate and complete motion capture systems, including the relative software, could be too expensive to be affordable, for example, by small and medium-sized enterprises. The latter include software (tools) that allow real-time *detection of joints and body parts* from digital images and videos [5] and skeleton-free approaches that predict *body joint angles* from a single depth image [6]. These systems usually employ Machine Learning (ML) or Deep Learning (DL) algorithms to predict the human pose. Although these systems have proved to be less invasive and energy-independent (no need for batteries), the accuracy of the calculation of the joint angles is not adequate despite the initial promising results. To further improve accuracy, researchers should enhance (vision-based) models and look to implement personalised ML/DL models and support prevention activities [7]. Moreover, existing research works recognise vision system setup, data fusion algorithms, and self or object occlusion as the main problems to be faced when considering a real scenario [8, 9]. Occlusion cases are due to the workstation layout, the movement of operators and production systems (e.g., robots), or their interaction. This issue can be partially overcome by multi-view capture systems; however, they require a complicated cameras calibration and synchronisation process, as well as high-performance computing. Despite the progress and use of ML techniques for primary prevention of WMSDs will likely continue to increase at a rapid pace and the development of real-time worker risk monitoring systems seems to be the most popular area of research [7], features coming from vision-based systems are rarely fed to a ML algorithm for assessing the risk related to WMSDs. In this context, the present work aims at giving a further contribution to the state of the art by proposing SPECTRE (Sensor-independent Parallel dEep ConvoluTional leaRning nEtwork): a completely sensor-independent learning model based on a parallel architecture for identifying and classifying postures in working environments. SPECTRE uses a vision library only to segment frames (in pre-processing) and, once trained, it runs without special cameras thus being used by any company. The major contributions are summarised as follows:

- the application of DL to data collected simulating a real manufacturing scenario in a controlled environment, also addressing the problem of occlusion

- the use of an agnostic explainable Machine Learning (xML) approach, during the testing phase, to understand how the networks recognise the frame's labels, i.e. which are the significant/meaningful pixels
- the assessment of DL-aided ergonomic risks related to the main body segments, in addition to the global risk index
- the development of a low-cost smart enterprise system for WMSDs prevention, enhancing its accessibility and applicability.

The paper is organised as follows. Section 2 provides an overview about the state of the art of sensors-based and AI-based solutions for the WMSDs; Section 3 presents the case study and the adopted solution; results are given in Section 4 and, finally, Section 5 critically reviews the work highlighting both strengths and weakness and suggesting future works.

2 Related work

The literature highlights that the adoption of objective evaluation methods and tools for ergonomic risk assessment is increasingly needed. For this reason, different solutions, in terms of hardware and software, have been investigated.

2.1 Sensor-based solutions

IMUs are one of the most common devices used in manufacturing contexts for collecting data from workers. IMUs are wearable devices, composed of multiple sensors (i.e., tri-axial accelerometers, gyroscopes, and magnetometers), that can capture and record movements and postures recreating the position and the orientation of the body segments they are attached to. In the last few years, these systems have improved in terms of accuracy and precision, so they have been widely used for ergonomic assessment [10]. Several studies experimented the use of inertial motion capture systems for ergonomic evaluation in real work environments. [11] evaluated the musculoskeletal risks in a banana harvesting activity through objective measures using inertial sensor motion capture (Xsens). [12] presented an IMU-based system to assess ergonomic risk in real-time according to Rapid Upper Limb Assessment (RULA) method, also providing visual and auditory feedback to workers. [13] proposed a full-body integrated system for the ergonomics evaluation in warehouse environments based on inertial sensors. [14] developed a wearable system for ergonomic risk assessment for the upper body part. The system is composed of IMUs and electromyography sensors to calculate both joint angles and muscles' strain. Thanks to the wide employment of inertial motion capture systems, these devices can be considered reference systems for ergonomic assessment. However, they are invasive and obtrusive for the operator (thus they cannot be worn for the entire work shift) [15], they have limited battery life, and their cost is not always affordable for companies [16]. Moreover, using the motion

capture system in real working environments and for dynamics tasks, recurring calibrations of IMUs may be necessary to assure reliable and accurate measurements [17].

2.2 Vision-based solution

In recent times, vision-based solutions are slowly starting to join the more classic sensor-based methods. These solutions can be classified according to different aspects such as space (2D and 3D), sensing-modalities, pipelines (single-person and multi-person), learning methods, etc. Different technologies are available to detect the human body from images or videos and estimate skeleton and joints. [18] used OpenCV coupled with Haarcascade and Adaboost to quickly evaluate human features. The result, in this case, is quite accurate but only two-dimensional images can be elaborated. As such, the posture analysis proves rather difficult. One of the most common alternatives is OpenPose, as described by [19]. OpenPose allows combining the output from several cameras in order to obtain three-dimensional skeleton tracking. This proves to be better than the previous method but occasionally it has some flaws when coupling data from different cameras and the accuracy for assessing human kinematics still remains unknown [20]. [21] achieved an improved identification by using VoxelPose, which directly operates in a 3D space and as such avoids the coupling problem described before. Nonetheless, the accuracy still proves to be not good enough for a reliable ergonomic evaluation.

In addition, some works, developed recently, focused on the detection of the user's skeleton through video processing and DL. These works, based on Convolutional Neural Networks (CNNs), aimed at designing a skeleton detection and tracking system and integrating it with a recommendation system for postures [22–24]. However, video processing using CNNs requires significant computational resources to provide a real-time response. Moreover, vision-based approaches have been criticized due to limited site coverage by cameras and the high likelihood of occlusions as pointed out by two recently published works [25, 26].

Vision-based systems are becoming an advance for marker-less ergonomic assessment, since they allow evaluating postures by images and videos taken from common RGB cameras, enhancing their accessibility and feasibility. Both motion capture systems based on RGB (e.g., standard webcams) and RGB-D cameras are low cost if compared with marker and sensors solutions. Ergonomic analyses do not require high accuracy for tracking the human body, since a small deviation in detecting joint position generally does not change the calculated index. For these reasons, vision-based systems could be sufficiently accurate and affordable to perform an ergonomic assessment [27]. These systems mainly rely on an approach that combines the following steps to obtain a full-fledged ergonomic evaluation:

- Skeleton identification, i.e., cameras use ML/DL to detect the human skeleton and its joints locations

- Posture analysis, i.e., the detected skeleton and its joints angles are used for the actual ergonomic evaluation.

Table 1 summarizes the most interesting articles on the topic specifying input modalities and methods. For example, [28] estimated RULA body posture scores from 2D kinematic joint locations obtained from a deep learning algorithm using Euclidean distance and the cosine of the angle between 2D vectors. Some papers perform pose estimation on the RGB images using OpenPose. [29] presented a framework for skeleton-based posture recognition by applying a 3D CNN.

Other works focus on activity recognition using monocular RGB cameras, which represent the most common approach. [30] developed a temporal CNN that uses spatio-temporal features to analyze and recognize human activities through a short video as input. Similaliry, [31] proposed a fast model that fuses spatial and temporal features to recognize human action. Their system extracts temporal information using RGB images achieving high performance. The rapid development of 3D data capture devices (RGB-D cameras) is leading to testing their application even for action recognition. [32] proposed a system of skeletal data-based CNN classifiers for action recognition. The system is composed of six 1-channel CNN classifiers and each is built with one unique posture-related feature vector extracted from the time series skeletal data, recorded by the Microsoft Kinect.

As shown in Table 1 most of the articles mainly concern activity recognition and pose estimation, without carrying out the ergonomic evaluation. The works that consider the ergonomic risk index (e.g., RULA) use body points' positions and joint angles. The proposed approach wants to stand out by adopting a skeleton-free approach for ergonomic evaluation, without the joints angles calculation. Based on the classification proposed by [33], SPECTRE can be defined as a one-stage 3D pose estimation approach, which regresses the 3D pose directly from the image through a parallel CNN architecture.

Table 1 Comparison of vision-based systems for different applications

Article	Input	Method	Application
[29]	RGB videos + OpenPose	3D CNN	Writing posture recognition
[31]	RGB images	CNN	Activity recognition
[30]	RGB video frames	CNN	Activity recognition
[34]	RGB videos, depth videos, 3D joints coordinates	Multi-stage adaptive regression	Activity recognition
[35]	RGB	CNN with OFF	Activity recognition
[36]	RGB camera + OpenPose	CNN	Tool-dependent activity recognition Pose estimation
[32]	Microsoft Kinect	CNN	Assembly activity recognition
[8]	Microsoft Kinect + 3 RGB camera + OpenPose	Joints angles	RULA/REBA scores
[37]	RGB images + OpenPose	CNN + DNN	Pose estimation RULA scores
[5]	RGB camera	CNN + Joints angles	Pose estimation RULA scores
[28]	Video images	CNN + Joints angles	Pose estimation RULA scores
[16]	Microsoft Kinect	Joint angles	OWAS score
[24]	Webcam	CNN	Upper body posture recognition
[26]	Video images	SVM	Posture classification

3 Case study scenario

Our approach has been developed considering the main manual activities that characterise the working environment in LUBE, the major Italian kitchen manufacturer. Specifically, in this case study, some manual operations that the worker generally carries out in a collaborative robotics cell were reproduced in the laboratory: manual handling of products, assembly, and quality inspection. The goal was to collect data and then use it to train the neural network. This approach allows considering different scenarios in LUBE workplaces and also generalising the method to use it in multiple working environments.

3.1 Dataset images (frames) acquisition and labelling

Firstly, the neural network needs to be trained using a wide dataset of different human postures. For this reason, two different recordings were captured at the same time: a motion capture system for movements acquisition that operates as a ground reference, and a camera for video recording to provide data to the neural network. To synchronize the two acquisitions, both systems were set to record at the same framerate (60 fps) and the first few frames of the video recording also showed the frame counter of the motion capture software. In this way, it was possible to match the two recordings after the acquisition. **Fig.** 1 shows the labelling process for the classification of collected postures. During the acquisition, the user was equipped with 18 Xsens MTw (Wireless

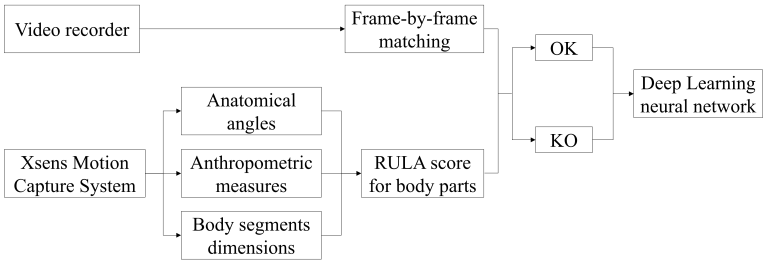

Fig. 1 Frame labelling process

Motion Tracker) for full-body monitoring. The Xsens MVN inertial motion capture system allows recording the movements of the user and exporting anthropometric measures, body segments, and joint angles. In this way, it is possible to evaluate the movements of the body for each recorded frame in an accurate and objective manner. Indeed, all the output data is functional for the main anatomical joint angles calculation, which are used for the ergonomic assessment. Specifically, the following body parts have been chosen, considering left and right body sides separately:

- Upper Arm: considering flexion and abduction
- Lower Arm: considering flexion and hand position related to the body's midline

- Wrist: considering flexion, deviation, and rotation
- Neck: considering flexion, lateral bending, and rotation
- Trunk: considering flexion, lateral bending, and rotation.

These angles and positions were calculated using specifically developed algorithms that allow elaborating data recorded by the motion capture system. For example, the hand location is determined by calculating the position of the wrist related to the shoulder. This is achieved by using the composition of the joint angles rotation matrices of shoulder and elbow and the measures of the related segments. For each body part, a specific threshold, based on the RULA method [38], has been defined to evaluate the related ergonomic risk. The body part is classified as "KO"/"OK" if the score is higher/lower than the following thresholds:

- Upper arm: 4
- Lower arm: 3
- Wrist: 4
- Neck: 4
- Trunk: 4

Simplifying, "OK" means "ergonomic position" and, on the contrary, "KO" means "non-ergonimic position". The classification resulting from the RULA assessment has been coupled with each video recorded frame in order to have an image data-set of which postures are correct and which are not. The classification process, which has been scripted, is divided in five parts:

1. Each video frame is extracted from the video itself.
2. The initial and final video frames in which no posture is performed are removed.
3. The corresponding frame row from the Xsens output file is selected.
4. The video frame is saved as picture with the frame number and the Xsens output (OK or KO) in its name.
5. The operation is repeated for each body parts, which are supposed to have different classifications.

As such, eight classification groups are created (shown in table 2) and each of them will be used to train a separate DL network.

3.2 SPECTRE architecture

SPECTRE, the architecture we propose in this paper, is fully shown in **Fig.** 2. As shown in **Fig.** 2-a we see two sequentially connected layers: the first layer, or segmentation layer, is given by the Python library Mediapipe and represent the network for pre-processing the frame, the second layer, or parallel convolutional layer, consists of 8 parallel CNNs - one for each body part we want to monitor - and is used for binary posture classification, see **Fig.** 2-b. Each network in the parallel CNN architecture, shown in **Fig.** 2, is made by 5 convolutional layers (CONV), five max-pooling layers (MAX POOL) and six

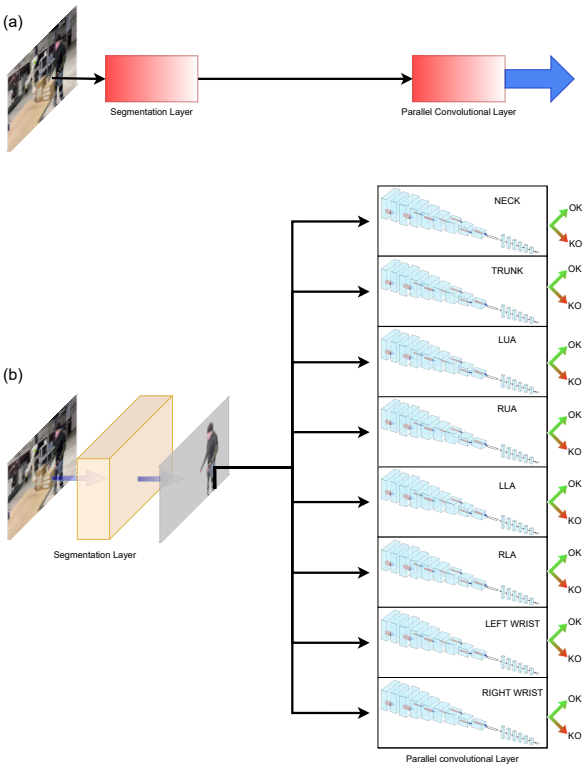

Fig. 2 SPECTRE architecture.

dense layers (DENSE). CONV layers are described by a triple N @ $W \times H$ and by a 2D vector (k_x, k_y). N is the number of kernels, W the layer width, H is its height and k_x and k_y are the kernels size; MAX POOL layers are characterised only by the above triple whereas DENSE layer only by its number of neuron X. Rectified linear unit (Relu) activation functions are used in each layer, but in the last one (DENSE), where a sigmoid activation function is used. Networks were implemented using Python 3.8 and TensorFlow 2.7. Each network is trained separately for 50 epochs using a batch of 32 images and a binary cross-entropy as loss function. Tests were performed in a platform using 6 GPUs NVIDIA A100 with 1TB RAM.

4 Results

In this section, we present the results of our work both in terms of evaluation metrics and xML for all CNNs in SPECTRE.

(a)

(b)

Fig. 3 (a). A frame extracted by the video. (b). The same frame after the segmentation procedure

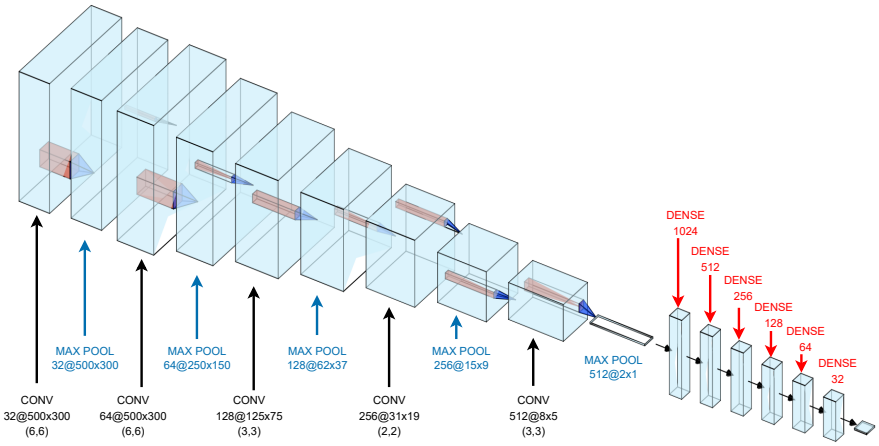

Fig. 4 CNN architecture. Three types of layer are present: convolutional (CONV), max-pooling (MAX POOL) or fully connected (DENSE). A CONV layer is described by a tuple N@widthxheight and by a

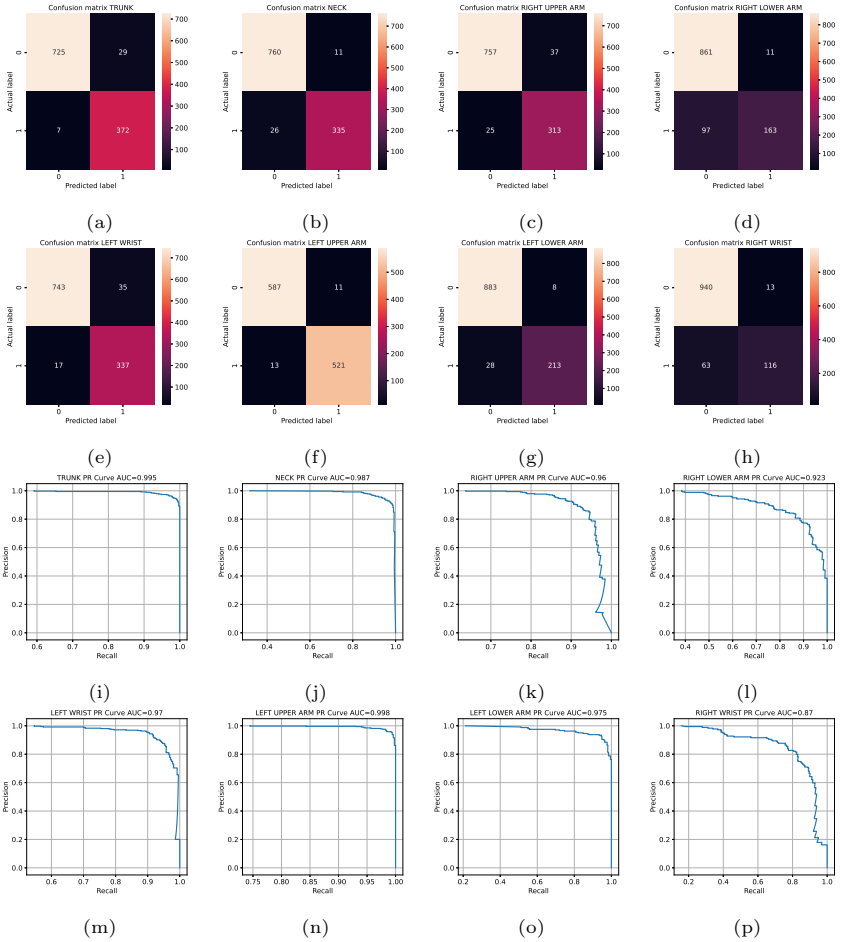

Fig. 5 (a)-(h). Confusion Matrices for each convolutional neural network in SPECTRE. (i)-(p) PR Curves for each convolutional neural network in SPECTRE.

4.1 Data and Evaluation Metrics

The labelling system in **Fig.** 1 produced a dataset consisting of 4601 frames (size 500×300), labelled "OK" or "KO" depending on the score and on the body part being considered. Hence, it is worth noticing that for each body part we have an unbalanced distribution between the two classes.

Each frame is then segmented in order to avoid possible interference of the background in the training phase; at the same time skeleton is extracted and superimposed to the segmented area. The result of described procedure is shown in **Fig.** 3. When the segmentation procedure is not successful in recognising the human figure in the picture or to superimposed the skeleton we discard that frame. The final data-set consists in 4527 frames. We used

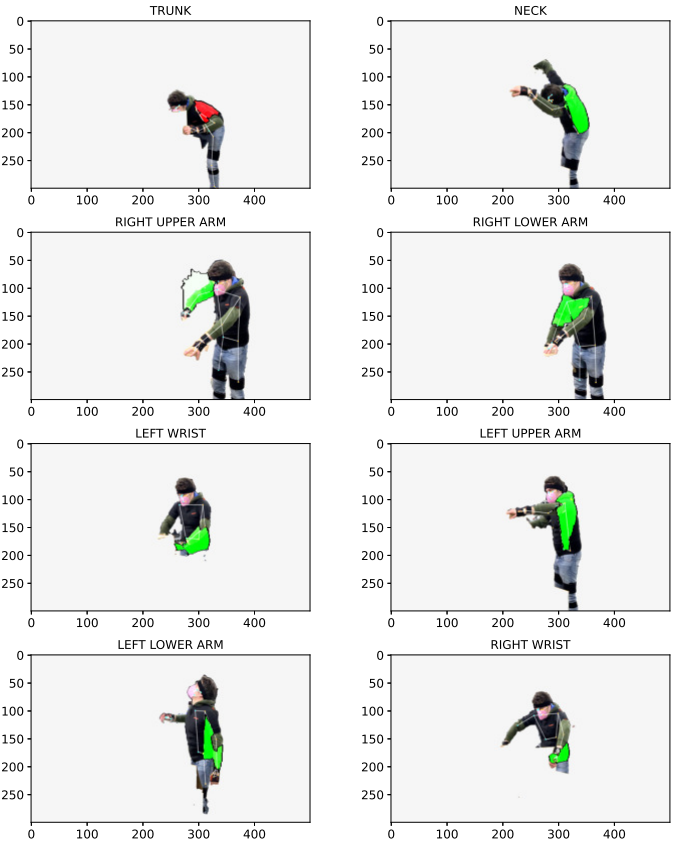

Fig. 6 LIME explanation of predictions. Each picture shows which part (which pixels) is important for the final prediction. Both the x-axis and y-axis represent the pixels that make up each picture. Green pixels represent those pixels that increase the probability for that picture to be classifies as "OK", on the contrary red pixels are involved in the decreasing of the probability for the "OK" postures. For instance, by looking at the LEFT WRIST frame it is possible to infer that those pixels increase the probability for the picture to be classified as "OK" that, actually means the picture is representing a correct position. On the contrary in TRUNK picture, the red pixels decrease the probability for that picture to be classified as a correct position.

a stratify 5-fold validation for checking the network model chosen. Then, the data-set is splitted in two parts: the 75% is used for training whereas the remaining 25% for testing. The split is done for each body part in a stratified fashion, i.e. the proportion between the two classes is preserved both in training and in testing.

In order to evaluate our model we referred to the classical confusion matrix containing the the numbers of True positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN) obtained during the test phase; however,

Table 2 Precision, Recall and F1-Score for each body area

Body part	Class	Precision	Recall	F1-score
LUA	OK=0	0.92 ± 0.09	0.97 ± 0.016	0.94 ± 0.046
	KO=1	0.96 ± 0.017	0.89 ± 0.138	0.92 ± 0.078
LLA	OK=0	0.98 ± 0.014	0.94 ± 0.045	0.96 ± 0.019
	KO=1	0.84 ± 0.113	0.92 ± 0.09	0.88 ± 0.055
LW	OK=0	0.87 ± 0.086	0.96 ± 0.035	0.91 ± 0.037
	KO=1	0.91 ± 0.07	0.66 ± 0.268	0.73 ± 0.193
RUA	OK=0	0.96 ± 0.055	0.9 ± 0.107	0.9 ± 0.052
	KO=1	0.83 ± 0.152	0.9 ± 0.156	0.84 ± 0.085
RLA	OK=0	0.88 ± 0.083	0.98 ± 0.022	0.9 ± 0.04
	KO=1	0.89 ± 0.086	0.53 ± 0.373	0.59 ± 0.359
RW	OK=0	0.87 ± 0.062	0.99 ± 0.022	0.92 ± 0.027
	KO=1	0.16 ± 0.253	0.18 ± 0.402	0.17 ± 0.376
NECK	OK=0	0.99 ± 0.008	0.95 ± 0.047	0.97 ± 0.024
	KO=1	0.90 ± 0.074	0.98 ± 0.016	0.94 ± 0.042
TRUNK	OK=0	0.98 ± 0.01	0.96 ± 0.022	0.97 ± 0.01
	KO=1	0.93 ± 0.032	0.96 ± 0.021	0.94 ± 0.019

LLA = Left Lower Arm, LUA = Left Upper Arm, LW = Left Wrist, RUA = Right Upper Arm; RLA = Right Lower Arm, RW = Right Wrist

such an unbalanced scenario, given its criticality, need more specific metrics as shown in [39–41]:

- Precision $= TP/(TP + FP)$
- Recall $= TP/(TP + FN)$
- F1-score $= 2 \cdot (\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall})$

Moreover we used Area Under the Precision-Recall Curve (AUPRC). The PR Curve shows the trade-off between precision and recall for different thresholds (of class prediction). A high area under the curve represents both high recall and high precision. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). Table 2 and **Fig.** 5 show the values of the above metrics: the former reports the results for the 5-folds validation whereas the latter displays the results of the best model. Table A general look shows that all body parts on the right side of the body (RUA, RLA, RW). Going into more detail, RW is the body part with lowest scores; on the contrary, LUA is the one with the highest scores.

4.2 Trusting the ML

xML is used in order to gain awareness of the obtained results and to check if SPECTRE learnt spurious associations or the information content is really related to the posture of the subject in the frame. We used the Local Interpretable Model-Agnostic Explanations (LIME) method for explaining SPECTRE [42].

4.3 LIME

Probably, "why such a prediction? and which variables are mostly involved the prediction?" are the FAQs about ML model results. LIME was developed for attempting to answer such questions. It is *model-agnostic*, i.e. it is able to explain any model by treating it as a black box, and *locally-faithful*. The main idea behind LIME is "reading" (explaining) the model perturbing the features values, and weighted them using a proximity function, in the neighborhood of the features to be explained. This creates a linear model that is able to understand the impact of the output [42, 43]. Those models are called surrogate models. LIME focuses on training *local* surrogate models to explain individual predictions thus providing local model interpretability. Other model interpretability techniques only answer the question above taking into account the entire data-set. For instance, feature importance explains on a data-set level which features are important but it is hard to diagnose specific model predictions. The idea behind LIME is quite intuitive. First of all, one needs to forget about the training data and imagine to have a black box model to be fed by input data points thus getting the predictions of the model. The final goal is to understand why the machine learning model made a certain prediction. LIME tests what happens to the predictions when one perturbs data into the machine learning model, e.g. modifying numerical values in tabular data or varying the pixels in images. In order to do that, LIME generates a new data-set consisting of perturbed samples and obtains the corresponding predictions of the black box model. On this new data-set LIME then trains an interpretable model (a linear model or a decision tree) approximating the black-box one and which is weighted by the proximity of the sampled instances to the instance of interest. The intuition is that it is less complicated to approximate a black-box model by a simple model locally, i.e. in the neighborhood of the prediction we want to explain, instead of approximating a model globally. In particular, the use of LIME to explain image predictions is based on creating image variations, not at pixel-level, but using "superpixels". Superpixels are groups of pixels grouped according to their color and obtained by segmenting the picture. The variations are created by randomly excluding some of superpixels, i.e. turning them off simply by replacing them using gray pixels.

4.4 Explaining SPECTRE

Fig. 6 shows the explanation of the predictions made by each CNN in SPECTRE. Green means that part of the image increases the probability for the label and red means a decrease.

The first achievement is given by the groups of pixels (superpixels) involved in the predictions: only the areas belonging to the segmented visible figure contribute to predict the status of the considered body part (excluding the environment). Green pixels indicate that the position of the considered body part increases the probability for the "OK", on the contrary red pixels mean a decrease for the same probability. The second noteworthy result is that the

system also infers the condition of one body part by taking advantage of the position of the other body parts. This condition is visible, for example in RUA, LUA, NECK, TRUNK, as depicted in **Fig.** 6. The most controversial results are those concerning RW. In fact, the system is not able to identify RW's position properly, even by deriving it from the positions of other areas of the body. This is probably due to the fact that, while the other joints are somewhat dependent, the posture of the wrists depends less on the relative position of the elements of the joint chain.

5 Discussion and Conclusions

SPECTRE, a parallel CNN for identifying and classifying postures in working environments, is presented. The proposed solution does not rely on ML/DL for identifying body joints and anatomical angles but it exploits the power of DL to recognise patterns in data able to directly check whether the workers' posture is correct or not. SPECTRE works independently for each body part of interest. This way, it is possible to identify which body part is mainly exposed to risk and suggest a healthier posture. Moreover, the chosen body parts are the same as those used for assessing the ergonomic overall risk according to the RULA method, that is why using SPECTRE it is also possible to obtain an overall risk score. The usage of LIME is extremely interesting since it allows to be aware what the system is looking at and trust the prediction. Indeed, other methodologies exist but they are not designed to deal with a large number of features [44]. Our solution is designed to be easy to use and affordable for small and medium companies. Indeed, the absence of wearable sensors and the possibility to use SPECTRE in any working environment increase usability and lead to a reduction in instrumentation costs (hardware and software). According to our knowledge, this is the first attempt to use DL for preventing WMSDs in manufacturing environments. Moreover, it is worth mentioning that the unbalanced data-set is an intrinsic condition in such a scenario: it is much more likely for a person to be in an ergonomic position than in a non-ergonomic one. Given that, we chose to not balance the two classes ("OK" and "KO" using methods such as the synthetic minority over-sampling technique (SMOTE) [45], but we managed the unbalancing using suitable metrics as described in 4.1. At the moment, our approach is deliberately not in real-time because we focused on long-lasting postures that are potentially more dangerous. Nonetheless, in the future, a real-time solution could be of interest, as reported in [24]. Finally, we are aware that some limitations emerge in our study; although the experiments were designed to be as realistic as possible, they were conducted in controlled environments, so testing SPECTRE in the field, i.e. in a real working environment, could be important to increase its performance; likewise, the possibility of optimizing the camera angle, by collecting more data, in order to get a better view of the hidden body parts (as the RW problem pointed out in Section 4.4) could be investigated, as well as the ability to overcome the limitation of single figure detection, in order to

allow our system to detect multiple subjects, could be a valuable improvement especially for crowded workplaces.

Acknowledgment

This work is partly funded by the URRÁ project "Usability of Robots and Reconfigurbility of processes: enabling technologies and use cases", on the topics of User-Centered Manufacturing and Industry 4.0, which is part of the project EU ERDF, POR MARCHE Region FESR 2014/2020–AXIS 1–Specific Objective 2–ACTION 2.1, "HD3Flab-Human Digital Flexible Factory of the Future Laboratory", coordinated by the Polytechnic University of Marche.

Declarations

Availability of data and code

The datasets and the code used and/or analysed during the current study are available from the corresponding author on reasonable request.

Author's contributions

Marianna Ciccarelli: Conceptulization of this study, Methodology, Data curation, Writing Paper. **Flavio Corradini**: Funding acquisition, Writing - Rerview & Editing. **Michele Germani**: Funding acquisition, Writing - Rerview & Editing. **Giacomo Menchi**: Data curation, Software. **Leonardo Mostarda**: Supervision Writing - Review & Editing. **Alessandra Papetti**: Conceptualization of this study, Methodology, Investigation, Writing paper. **Marco Piangerelli**: Conceptualization of this study, Methodology, Formal analysis, Software, Writing Paper.

Conflict of interest

All authors declare that they have no conflicts of interest.

References

[1] European Agency for Safety and Health at Work. Work-related musculoskeletal disorders: prevalence, costs and demographics in the EU. In: Luxembourg: Publications Office of the European Union; 2019. .

[2] Korhan O, Memon AA. Introductory chapter: work-related musculoskeletal disorders. In: Work-related musculoskeletal disorders. IntechOpen; 2019. .

[3] Malchaire J, Gauthy R, Piette A, Strambi F. A classification of methods for assessing and/or preventing the risks of musculoskeletal disorders. ETUI, European Trade Union Institute; 2011. .

[4] Lim S, D'Souza C. A narrative review on contemporary and emerging uses of inertial sensing in occupational ergonomics. International Journal of Industrial Ergonomics. 2020;76:102937. https://doi.org/https://doi.org/10.1016/j.ergon.2020.102937.

[5] Fernández MM, Álvaro Fernández J, Bajo JM, Delrieux CA. Ergonomic risk assessment based on computer vision and machine learning. Computers & Industrial Engineering. 2020;149:106816. https://doi.org/https://doi.org/10.1016/j.cie.2020.106816.

[6] Abobakr A, Nahavandi D, Hossny M, Iskander J, Attia M, Nahavandi S, et al. RGB-D ergonomic assessment system of adopted working postures. Applied Ergonomics. 2019;80:75–88. https://doi.org/https://doi.org/10.1016/j.apergo.2019.05.004.

[7] Chan VCH, Ross GB, Clouthier AL, Fischer SL, Graham RB. The role of machine learning in the primary prevention of work-related musculoskeletal disorders: A scoping review. Applied Ergonomics. 2022;98:103574. https://doi.org/https://doi.org/10.1016/j.apergo.2021.103574.

[8] Kim W, Sung J, Saakes D, Huang C, Xiong S. Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). International Journal of Industrial Ergonomics. 2021;84:103164. https://doi.org/https://doi.org/10.1016/j.ergon.2021.103164.

[9] Bibi S, Anjum N, Sher M. Automated multi-feature human interaction recognition in complex environment. Computers in Industry. 2018;99:282–293. https://doi.org/https://doi.org/10.1016/j.compind.2018.03.015.

[10] Battini D, Berti N, Finco S, Guidolin M, Reggiani M, Tagliapietra L. WEM-Platform: A real-time platform for full-body ergonomic assessment and feedback in manufacturing and logistics systems. Computers & Industrial Engineering. 2022;164:107881. https://doi.org/https://doi.org/10.1016/j.cie.2021.107881.

[11] Merino G, da Silva L, Mattos D, Guimarães B, Merino E. Ergonomic evaluation of the musculoskeletal risks in a banana harvesting activity through qualitative and quantitative measures, with emphasis on motion capture (Xsens) and EMG. International Journal of Industrial Ergonomics. 2019;69:80–89. https://doi.org/https://doi.org/10.1016/j.ergon.2018.10.004.

[12] Vignais N, Miezal M, Bleser G, Mura K, Gorecky D, Marin F. Innovative system for real-time ergonomic feedback in industrial manufacturing. Applied Ergonomics. 2013;44(4):566–574. https://doi.org/https://doi.org/10.1016/j.apergo.2012.11.008.

[13] Battini D, Persona A, Sgarbossa F. Innovative real-time system to integrate ergonomic evaluations into warehouse design and management. Computers & Industrial Engineering. 2014 11;77. https://doi.org/10.1016/j.cie.2014.08.018.

[14] Peppoloni L, Filippeschi A, Ruffaldi E, Avizzano CA. A novel wearable system for the online assessment of risk for biomechanical load in repetitive efforts. International Journal of Industrial Ergonomics. 2016;52:1–11. New Approaches and Interventions to Prevent Work Related Musculoskeletal Disorders. https://doi.org/https://doi.org/10.1016/j.ergon.2015.07.002.

[15] Yadav SK, Tiwari K, Pandey HM, Akbar SA. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Knowledge-Based Systems. 2021;223:106970. https://doi.org/https://doi.org/10.1016/j.knosys.2021.106970.

[16] Diego-Mas JA, Poveda-Bautista R, Garzon-Leal D. Using RGB-D sensors and evolutionary algorithms for the optimization of workstation layouts. Applied Ergonomics. 2017;65:530–540. https://doi.org/https://doi.org/10.1016/j.apergo.2017.01.012.

[17] Vlasic D, Adelsberger R, Vannucci G, Barnwell J, Gross M, Matusik W, et al. Practical Motion Capture in Everyday Surroundings. ACM Trans Graph. 2007;26(3). https://doi.org/https://doi.org/10.1145/1276377.1276421.

[18] Damle R, Gurjar A, Joshi A, Nagre K. Human Body Skeleton Detection and Tracking. In: Human Body Skeleton Detection and Tracking. vol. 3; 2015. p. 222–225.

[19] Slembrouck M, Luong HQ, Gerlo J, Schütte K, Cauwelaert DV, Clercq DD, et al. Multiview 3D Markerless Human Pose Estimation from OpenPose Skeletons. In: Advanced Concepts for Intelligent Vision Systems; 2020. p. 166–178.

[20] Clark RA, Mentiplay BF, Hough E, Pua YH. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. Gait & Posture. 2019;68:193–200. https://doi.org/https://doi.org/10.1016/j.gaitpost.2018.11.029.

[21] Tu H, Wang C, Zeng W. End-to-End Estimation of Multi-Person 3D Poses from Multiple Cameras. CoRR. 2020;abs/2004.06239. https://doi.org/https://doi.org/10.1007/978-3-030-58604-1_29. 2004.06239.

[22] Li C, Zhong Q, Xie D, Pu S. Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE; 2017. p. 597–600.

[23] Yoshikawa Y, Shishido H, Suita M, Kameda Y, Kitahara I. Shot detection using skeleton position in badminton videos. In: International Workshop on Advanced Imaging Technology (IWAIT) 2021. vol. 11766. International Society for Optics and Photonics; 2021. p. 117661K.

[24] Piñero-Fuentes E, Canas-Moreno S, Rios-Navarro A, Domínguez-Morales M, Sevillano JL, Linares-Barranco A. A Deep-Learning Based Posture Detection System for Preventing Telework-Related Musculoskeletal Disorders. Sensors. 2021;21(15). https://doi.org/10.3390/s21155236.

[25] Xiao B, Xiao H, Wang J, Chen Y. Vision-based method for tracking workers by integrating deep learning instance segmentation in off-site construction. Automation in Construction. 2022;136:104148. https://doi.org/https://doi.org/10.1016/j.autcon.2022.104148.

[26] Seo J, Lee S. Automated postural ergonomic risk assessment using vision-based posture classification. Automation in Construction. 2021;128:103725. https://doi.org/https://doi.org/10.1016/j.autcon.2021.103725.

[27] Regazzoni D, Vecchi GD, Rizzi C. RGB cams vs RGB-D sensors: Low cost motion capture technologies performances and limitations. Journal of Manufacturing Systems. 2014;33(4):719–728. https://doi.org/https://doi.org/10.1016/j.jmsy.2014.07.011.

[28] Nayak GK, Kim E. Development of a fully automated RULA assessment system based on computer vision. International Journal of Industrial Ergonomics. 2021;86:103218. https://doi.org/https://doi.org/10.1016/j.ergon.2021.103218.

[29] Liu J, Wang Y, Liu Y, Xiang S, Pan C. 3D PostureNet: A unified framework for skeleton-based posture recognition. Pattern Recognition Letters. 2020;140:143–149. https://doi.org/https://doi.org/10.1016/j.patrec.2020.09.029.

[30] Andrade-Ambriz YA, Ledesma S, Ibarra-Manzano MA, Oros-Flores MI, Almanza-Ojeda DL. Human activity recognition using temporal convolutional neural network architecture. Expert Systems with Applications. 2022;191:116287. https://doi.org/https://doi.org/10.1016/j.eswa.2021.116287.

[31] Zhu S, Fang Z, Wang Y, Yu J, Du J. Multimodal activity recognition with local block CNN and attention-based spatial weighted CNN. Journal of

Visual Communication and Image Representation. 2019;60:38–43. https://doi.org/https://doi.org/10.1016/j.jvcir.2018.12.026.

[32] Al-Amin M, Qin R, Moniruzzaman M, Yin Z, Tao W, Leu M. An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly. Journal of Intelligent Manufacturing. 2021 07;https://doi.org/10.1007/s10845-021-01815-x.

[33] Gamra MB, Akhloufi MA. A review of deep learning techniques for 2D and 3D human pose estimation. Image and Vision Computing. 2021;114:104282. https://doi.org/https://doi.org/10.1016/j.imavis.2021.104282.

[34] Liu B, Cai H, Ju Z, Liu H. Multi-stage adaptive regression for online activity recognition. Pattern Recognition. 2020;98:107053. https://doi.org/https://doi.org/10.1016/j.patcog.2019.107053.

[35] Xu H, Bazavan EG, Zanfir A, Freeman B, Sukthankar R, Sminchisescu C. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 6184–6193.

[36] Chen C, Wang T, Li D, Hong J. Repetitive assembly action recognition based on object detection and pose estimation. Journal of Manufacturing Systems. 2020;55:325–333. https://doi.org/https://doi.org/10.1016/j.jmsy.2020.04.018.

[37] Li L, Martin T, Xu X. A novel vision-based real-time method for evaluating postural risk factors associated with musculoskeletal disorders. Applied Ergonomics. 2020;87:103138. https://doi.org/https://doi.org/10.1016/j.apergo.2020.103138.

[38] McAtamney L, Nigel Corlett E. RULA: a survey method for the investigation of work-related upper limb disorders. Applied Ergonomics. 1993;24(2):91–99. https://doi.org/https://doi.org/10.1016/0003-6870(93)90080-S.

[39] Mancini A, Vito L, Marcelli E, Piangerelli M, De Leone R, Pucciarelli S, et al. Machine learning models predicting multidrug resistant urinary tract infections using "DsaaS". BMC bioinformatics. 2020;21(10):1–12. https://doi.org/https://doi.org/10.1186/s12859-020-03566-7.

[40] Bordoni L, Petracci I, Pelikant-Malecka I, Radulska A, Piangerelli M, Samulak JJ, et al. Mitochondrial DNA copy number and trimethylamine levels in the blood: New insights on cardiovascular disease biomarkers. The FASEB Journal. 2021;35(7):e21694. https://doi.org/https://doi.org/10.1096/fj.202100056R.

[41] Lopez M, Beurton-Aimar M, Diallo G, Maabout S. A simple yet effective approach for log based critical errors prediction. Computers in Industry. 2022;137:103605. https://doi.org/https://doi.org/10.1016/j.compind.2021.103605.

[42] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135–1144.

[43] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1–42. https://doi.org/https://doi.org/10.1145/3236009.

[44] Biecek P, Burzykowski T. Explanatory Model Analysis. Chapman and Hall/CRC, New York; 2021.

[45] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357. https://doi.org/https://doi.org/10.1613/jair.953.