forests

MDPI

Article

Human and Machine Reliability in Postural Assessment of Forest Operations by OWAS Method: Level of Agreement and Time Resources

Gabriel Osei Forkuo [1], Marina Viorela Marcu [1], Nopparat Kaakkurivaara [2], Tomi Kaakkurivaara [2] and Stelian Alexandru Borz [1,*]

[1] Department of Forest Engineering, Forest Management Planning and Terrestrial Measurements, Faculty of Silviculture and Forest Engineering, Transilvania University of Brasov, Şirul Beethoven 1, 500123 Brasov, Romania; gabriel.forkuo@unitbv.ro (G.O.F.); viorela.marcu@unitbv.ro (M.V.M.)
[2] Department of Forest Engineering, Faculty of Forestry, Kasetsart University, 50 Ngamwongwan Rd., Lad Yao, Chatuchak, Bangkok 10900, Thailand; ffornrm@ku.ac.th (N.K.); ffortmk@ku.ac.th (T.K.)
* Correspondence: stelian.borz@unitbv.ro

Abstract: In forest operations, traditional ergonomic studies have been carried out by assessing body posture manually, but such assessments may suffer in terms of efficiency and reliability. Advancements in machine learning provided the opportunity to overcome many of the limitations of the manual approach. This study evaluated the intra- and inter-reliability of postural assessments in manual and motor-manual forest operations using the Ovako Working Posture Analysing System (OWAS)—which is one of the most used methods in forest operations ergonomics—by considering the predictions of a deep learning model as reference data and the rating inputs of three raters done in two replicates, over 100 images. The results indicated moderate to almost perfect intra-rater agreement (Cohen's kappa = 0.48–1.00) and slight to substantial agreement (Cohen's kappa = 0.02–0.64) among human raters. Inter-rater agreement between pairwise human-model datasets ranged from poor to fair (Cohen's kappa = −0.03–0.34) and from fair to moderate when integrating all the human ratings with those of the model (Fleiss' kappa = 0.28–0.49). The deep learning (DL) model highly outperformed human raters in assessment speed, requiring just one second per image, which, on average, was 19 to 53 times faster compared to human ratings. These findings highlight the efficiency and potential of integrating DL algorithms into OWAS assessments, offering a rapid and resource-efficient alternative while maintaining comparable reliability. However, challenges remain regarding subjective interpretations of complex postures. Future research should focus on refining algorithm parameters, enhancing human rater training, and expanding annotated datasets to improve alignment between model outputs and human assessments, advancing postural assessments in forest operations.

Keywords: wood harvesting; ergonomics; reliability; comparison; variability; human rater; machine learning; consistency

Academic Editor: Gianni Picchi

Received: 8 April 2025
Revised: 24 April 2025
Accepted: 28 April 2025
Published: 29 April 2025

Citation: Forkuo, G.O.; Marcu, M.V.; Kaakkurivaara, N.; Kaakkurivaara, T.; Borz, S.A. Human and Machine Reliability in Postural Assessment of Forest Operations by OWAS Method: Level of Agreement and Time Resources. *Forests* **2025**, *16*, 759. https://doi.org/10.3390/f16050759

1. Introduction

Wood procurement is an important industrial sector with significant unexplored potential for achieving the goals of sustainable economies, societies, and environments. The renewability of the resource [1,2], its neutrality in terms of environmental pollution [3–5], the potential for a circular bioeconomy [6,7], the creation of employment opportunities [8–11],

particularly in rural areas, and its contributions to global and local gross domestic products [12,13] all support the development of a wood-based bioeconomy in many parts of the world.

In this context, forest operations play a challenging role because the decisions made and the operations implemented must balance economic, environmental, and social aspects [14,15]. Additionally, there are various methods for wood harvesting that can be applied under the same local conditions. The availability of cheap labor, the characteristics of local forest management, and the lack of state-of-the-art, fully mechanized harvesting systems often lead to a dominance of manual labor in such operations [16–18].

On the other hand, manual and motor-manual wood harvesting presents significant challenges from ergonomic and safety perspectives [19–22]. In these operations, there is a high prevalence of work-related musculoskeletal disorders [22–24], which have serious economic consequences [25–27]. Therefore, objective assessments are necessary to correlate the occurrence of musculoskeletal disorders with relevant factors. This approach aims to enable informed decision-making for postural assessment as a preventive tool. However, variability in anthropometrics [28–30], work habits [23,31], the characteristics of work objects and local conditions [32,33], along with the diverse methods available for application [34–36], complicate objective evaluations of postural conditions at the population level. Furthermore, existing studies have utilized a limited range of conditions and datasets to describe the postural conditions in these work environments [24,37–39].

The Ovako Working Posture Analysing System (OWAS) method is widely accepted and used in forest operations as a tool for evaluating postural conditions [23,24,31,40]. It was developed by a steel industry company to describe workloads during the overhauling of iron smelting ovens [41]. This ergonomic assessment tool identifies the most common back postures (4 postures), arm positions (3 postures), leg positions (7 postures), and the level of force being exerted (3 categories). This structure allows for up to 252 possible combinations of postures, which are classified into four action categories that indicate a need for ergonomic interventions. Each posture adopted by a worker is represented by a unique 4-digit code derived from the classification of postures for each body part and the load handled [42].

The OWAS method involves observing work tasks, coding the postures engaged during the tasks, assigning risk categories, and proposing corrective actions [25,43]. Observations are typically collected as 'snapshots', with sampling conducted at fixed time intervals [31,42]. However, studies have indicated that the agreement between OWAS results and direct technical measurements for time spent in bent postures is relatively low [44], potentially due to discrepancies in sampling strategies used between methods [45,46]. When compared to other methods, such as the NIOSH (National Institute for Occupational Safety and Health) lifting equation, OWAS results demonstrated significant differences due to the differing approaches of these methods [42,47]. Research has indicated that observations made using the Rapid Entire Body Assessment (REBA) method have shown moderate alignment with those of the OWAS method [36,48]. However, REBA tends to classify a greater number of postures as having a higher level of risk [36,48]. Similarly, comparisons between the Rapid Upper Limb Assessment (RULA) method and OWAS have revealed a moderate level of correspondence [36,48]. Consequently, it remains unclear which method more accurately reflects the underlying risks of musculoskeletal disorders associated with various tasks, highlighting a critical gap in our understanding of ergonomic evaluations using traditional methods [36,48,49].

The reliability of the OWAS method has been confirmed through extensive analysis conducted by a group of trained engineers [25], demonstrating good intra- [48,50] and inter-observer repeatability [41,50–52]. Similarly, a study by [53] highlighted the OWAS

method's high inter-rater reliability for assessing physical workloads, with Cohen's kappa coefficients ranging from 0.75 to 0.90 across various tasks. However, a notable gap exists in the lack of scientific studies examining the reliability of automated OWAS models in comparison to traditional methods, particularly in the context of postural assessment in forest operations. Moreover, the OWAS method is characterized by its simplicity and versatility, making it accessible for personnel across various domains, including health, engineering, and industry, without requiring highly specialized training [41]. It is well-documented and has been supported by different computer programs that facilitate its application, allowing researchers to save time and improve workflow efficiency [25]. These programs have already been implemented in several studies [54,55].

While OWAS offers several benefits, including ease of use and good repeatability, it is not without limitations [25]. Some authors have pointed out that it does not differentiate between the right and left upper limbs and fails to evaluate critical areas, such as the neck, elbows, and wrists [25]. Additionally, OWAS coding may be overly simplistic for shoulders, may require excessive time for implementation, and does not adequately address the repetition or duration of sequential postures [35,42]. However, considering its current features, the OWAS method is likely to see increased usage in future evaluations of ergonomic conditions. Its capacity to assess diverse postures and workloads, combined with ongoing advancements in automated applications, promises to enhance its relevance in various work settings as the need for ergonomic assessments continues to grow.

The posture of work, on the other hand, may change in a very short time [31,35,56]. A given task or operation can be described as a sequence of postures assumed by an individual during work, where each posture has a specific duration and repetition pattern. Dynamic work is more likely to provide a diverse postural profile, with individual postures changing rapidly in the time domain [31,56]. This is typical of manual and motor-manual wood harvesting operations [28,57,58], making it difficult to characterize a task using a limited dataset obtained through sampling. In fact, ref. [46] found that the use of the OWAS method to produce reliable results requires very fine sampling. Additionally, ref. [45] reported similar findings when comparing the reliability of random and systematic sampling to produce an initial dataset for analysis. These findings imply that extensive datasets are required to produce reliable results, a situation further complicated by the available expertise for annotation and, most importantly, by the resources in terms of time and money needed to conduct the analyses. One can also question the intra- and inter-rater reliability of estimates when using observational postural assessment methods such as OWAS [25,42,53].

To overcome these limitations, a system capable of collecting and analyzing extensive datasets with minimal resources while maintaining a high level of reliability is required. A potential solution lies in the use of intelligent computer vision-based deep learning (DL) algorithms, as once they are effectively trained, they can streamline the postural analysis process. For instance, ref. [59] compared the performance of four deep-learning neural networks using a comprehensive annotated set of images depicting manual and motor-manual operations and concluded that the ResNet-50 algorithm can provide highly accurate predictions through transfer learning (96.34% classification accuracy), making it competitive with the results that an expert labeler may provide [42,48,53]. ResNet-50 is a deep Convolutional Neural Network (CNN) featuring 50 layers. Its key innovation, skip connections, mitigates the vanishing gradient problem, enabling effective training of deeper networks [60]. This makes ResNet-50 highly effective for complex image classification tasks [60], such as postural assessment [59].

It is unclear, however, whether further improvements in classification accuracy are possible through finer tuning of the algorithm. Since there are many options and hyperpa-

rameters that can be adjusted, exploring the potential for improvement through trial and error would be challenging in terms of resources, including computational ones. However, keeping scalability in mind, the effort would be worthwhile if it leads to a model capable of improving classification accuracy by even 1%. Subsequently, exploring the performance of a finely tuned model on incoming data is important because the classification results may highlight potential oversights in the model and provide new data for re-training, validation, and improvement. This is particularly relevant given the general lack of purpose-based annotated datasets [61,62], which limits the available data and hinders the generalization ability of models.

Lastly, since a new method for solving a given classification problem is under testing, it is essential to evaluate how its outputs align with human expertise. In other words, the outputs of the machine learning model should be assessed for reliability in comparison to those of human experts to explore any important mechanisms behind reliability. Before this evaluation, it is also necessary to examine how the same expert assesses the same data and whether those assessments are consistent. Furthermore, the consistency of ratings given by different experts for the same data is also crucial.

The goal of this study was to assess the reliability of human raters in the postural assessment of manual and partly mechanized wood harvesting operations using the OWAS method and to compare their assessments with those made by a deep learning model developed for postural classification. This was achieved through three specific objectives, which were: (i) to assess the intra- and inter-rater reliability of human assessments in postural classification, (ii) to evaluate the deviation of human ratings from the ground truth data (model ratings), and (iii) to estimate the time efficiency of human ratings compared to machine ratings.

2. Materials and Methods

2.1. Deep Learning Model Used as a Reference

The ResNet-50 [60] model, which is a deep convolutional neural network, was used for fine-tuning due to its proven effectiveness and robustness in various image classification tasks [60,63,64]. Additionally, ResNet-50 is known for its skip connections, which mitigate the vanishing gradient problem and enable the training of deeper networks [63]. This model was selected over others, such as GoogLeNet [65], MobileNet-v2 [66], and ShuffleNet [67], based on experimental results from Forkuo and Borz [59], which showed that ResNet-50 achieved the highest classification accuracy while maintaining a favorable balance between accuracy and computational efficiency [59]. In particular, the DL model was trained, tested, and validated using a very large and diverse image dataset containing 23,000 images showing a variety of workers engaged in different forest operations; specifically, the images used were labeled by considering the context shown in them, by analyzing the video sequences from which they were extracted, with the goal of making better decisions regarding instances in which movement was in question, thereby providing the model with some degree of prior knowledge about such events observed in the images [59].

2.2. Dataset and Posture Rating by Human Experts

For this study, a separate dataset was compiled that accurately reflects the postures and movements of forest workers across various operations. Sampling a dataset that encompasses a similar domain is crucial for ensuring the effectiveness of the DL models used for postural classification, as the domain significantly influences the models' performance, particularly with respect to factors such as picture crowding, occlusion, and the variability of postures and action categories. A total of 100 images were randomly selected from an image data repository curated by the authors, covering various operations. Addition-

ally, the final dataset incorporated various important features that represent the different environments in which forest workers operate.

Three human raters (hereafter referred to as R1, R2, and R3) were selected based on their previous expertise with the method to evaluate the 100 images using the OWAS method (Table 1). The ratings for postures and action categories were performed manually, allowing for detailed evaluations by the raters. To facilitate this process, a structured file was created to capture the codes for each rated image regarding back, arms, and leg postures, as well as to assess the level of force exertion. Additionally, the action category was documented, and a specific column was reserved for recording the time each rater spent completing the rating of each image, measured to the nearest second. This template ensured a standardized approach for data rating and storage, with each rater filling in the necessary details after assessing each image based on a standardized guideline detailing the OWAS method, which was provided to each rater to ensure uniformity in assessments across all images.

Table 1. Description of the OWAS codes and categories used in the study.

Feature	Abbreviation in the Study	Number of Categories According to OWAS	Description
Back	B	4	Describes the posture of the back starting from a neutral straight posture and ending with the back being bent and twisted
Arms	A	3	Describes the posture of the arms starting from a neutral posture with both arms below shoulder level and ending with both arms being at or above the shoulder level
Legs	L	7	Describes the posture of the legs by seven categories starting from a neutral sitting posture and ending with legs being engaged in walking or moving
Force exertion	F	3	Describes the level of force exertion starting with handling loads or exerting forces less than 10 kg and ending with handling loads or exerting forces over 20 kg
Action category	AC	4	Indicates the level of postural risk by the urgency of the ergonomic interventions required, starting from no intervention required and ending with intervention required immediately

The rating process was conducted for each of the 100 images in two replications (hereafter referred to as r1 and r2), without providing the raters with any prior knowledge about the dataset. In the first replication, each rater was instructed to review and assess the entire image dataset. After completing the ratings, each rater stored the postural information, action category, and rating time in an Excel spreadsheet named with the rater's identifier and the replication number. Upon completion, the rater sent the file to the lead researcher and was required to delete all rating information from their computer. The second assessment (r2) was carried out after one month to prevent doing the rating based on the experience gained in the first round. In addition, the raters were not informed in advance that the same image dataset would be used for the second rating, and the order of showing the images was the same.

2.3. Reliability Assessment

Several datasets were used in the process of reliability assessment, as shown in Table 2. The intra-rater reliability assessment was based on the datasets produced by the same rater, comparing the results of the first (r1) and second (r2) replications. The pairwise inter-rater reliability assessment considered the data from all raters and replications, with the constraint of comparing data from the same replication. For example, the R1r1 dataset was compared against R2r1, then against R3r1, followed by a comparison between the R2r1 and R3r1 datasets, resulting in three assessments of intra-rater reliability. The same procedure was used for the datasets from the second replication (r2). The overall inter-rater reliability assessment was based on the replication-based data from all the raters. Initially, R1r1, R2r1, and R3r1 were used as datasets for assessment. Subsequently, the same assessment was carried out on the data sourced from the three raters in the second replication (r2).

Table 2. Description of the datasets used in the assessment.

Rater No.	Replication No.	Abbreviation of the Dataset	Description of the Dataset
R1	r1	R1r1	Ratings of the first rater in the first replication
R1	r2	R1r2	Ratings of the first rater in the second replication
R2	r1	R2r1	Ratings of the second rater in the first replication
R2	r2	R2r2	Ratings of the second rater in the second replication
R3	r1	R3r1	Ratings of the third rater in the first replication
R3	r2	R3r2	Ratings of the third rater in the second replication
RM	-	RM	Rating of the deep learning model

The DL model was utilized to produce a reference dataset (hereafter referred to as RM, Table 2) that was deemed suitable to represent the ground truth data, a decision which was based on the amount of data used to build it and the excellent classification results it provided [59]. To achieve this, the model was fed the image dataset and allowed to make its own predictions. The resulting data were then stored in a new Excel sheet. Subsequently, the RM dataset was employed to assess the reliability of replication-based human raters using a pairwise approach. For instance, the data sourced from each rater for each replication were compared to the predictions made by the model. Finally, overall reliability was evaluated by comparing the replication-based data from all human raters against the predictions made by the model.

In all the assessments conducted, five data subsets were used, representing the back (hereafter B), arms (hereafter A), and legs (hereafter L) postures, level of force exertion (hereafter F), and action category (hereafter AC). This allowed for the evaluation of the magnitude of agreement at two levels: specifically, at the level of postural code and action category.

2.4. Reliability Metrics Used for Assessment

In this study, inter-rater and intra-rater reliability were assessed using Cohen's kappa [68] and Fleiss' kappa [69], respectively. Kappa statistic is the most widely used mea-

sure for quantifying the level of agreement between two or more raters while accounting for the potential for chance agreement [70,71]. As such, Cohen's kappa is a chance-corrected statistic utilized to assess the agreement level rather than merely measuring association in ratings [72]. It is commonly used to measure agreement between two raters on categorical items while accounting for chance agreement, and it is calculated by assessing the level of agreement between the two raters and comparing it to the expected level of agreement by chance [68,72]. This statistic ranges from -1 to 1, with values closer to 1 indicating near-perfect agreement, values around 0 reflecting no agreement beyond random chance, and negative values suggesting worse-than-chance agreement [68,70]. For pair-wise inter-rater reliability, comparisons were made between the human raters themselves and between the human raters and the predictions of the DL model. Cohen's kappa facilitated a robust analysis [70] of how consistently both human and machine assessments aligned, establishing a reliable framework for evaluating working postures. The interpretation of kappa values in this study followed established criteria [68,71,73] for classifying levels of agreement: values ≤ 0 indicate no agreement, 0.01–0.20 denote slight agreement, 0.21–0.40 signify fair agreement, 0.41–0.60 reflect moderate agreement, 0.61–0.80 represent substantial agreement, and values from 0.81 to 1.00 indicate almost perfect agreement [70,74]. To calculate the percent agreement, the number of agreements was divided by the total number of scores [75], serving as a direct measure rather than an estimate [70].

Moreover, Fleiss' kappa, which is a modified version of Cohen's kappa and used for measuring agreement among multiple raters [69,70,76], was employed to thoroughly assess inter-rater reliability among the four raters (three human raters and the deep learning model) as they rated the same data, allowing for a comprehensive analysis of how consistently the ratings converged across the entire panel of raters. The Fleiss' kappa statistic measures the overall agreement while accounting for the level of agreement that could occur by random chance [76,77]. This approach mirrors the methods used in the postural analysis by Lins et al. [53], who applied both Cohen's and Fleiss' kappa statistics in assessing OWAS inter-rater reliability. Fleiss' kappa is particularly relevant in the context of postural assessments and has been similarly applied by Widyanti [75] and De Bruijn et al. [50] in their studies of inter-rater reliability for different observational techniques.

2.5. Time Assessment

Significant variations in image-based assessment time were anticipated at both the intra- and inter-rater assessment levels. Additionally, the DL model-based assessment was expected to require only a small amount of time to predict the images from the dataset. For each human rater and replication, the time taken to rate a given image was recorded to the nearest second. For the human raters, the rating time was the time in which the rating was done and included elements such as opening a given image, making the judgment on the body posture and level of force exertion, identifying the action category, and noting down the results of the assessment in the Excel sheet. For the DL model, the time required to rate each image was documented programmatically and measured to the nearest second. The time consumption assessment was conducted at the image level, as the time measurements pertained to an end-to-end image rating. Then, comparisons were made using appropriate statistical methods, as described in Section 2.6.

2.6. Statistical Analysis and Software Used

A first assessment of agreement was conducted visually by employing the multi-dimensional scaling (MDS) algorithm in Orange Visual Programming software version 3.38.1 (https://orangedatamining.com/ (accessed on 17 February 2025)) [78]. MDS is a powerful statistical tool used to map high-dimensional data in a bivariate plot, aiming to

understand the relationships among the data [79,80]. This approach is particularly useful for revealing important information regarding the similarity in data by transforming the distances among data pairs into a configuration of points mapped in Cartesian space. Orange Visual Programming Software version 3.38.1 facilitates MDS in a visual manner, allowing users to set feature and target variables. Two MDS analyses were performed, using as features the codes attributed to the back, arms, legs, level of force exertion, and action category. The first MDS focused on the similarity among the ratings of R1, R2, and R3, while the second MDS examined the similarity among the ratings of R1, R2, R3, and RM.

The kappa statistics were calculated using Python (v3.12), implemented in the PyCharm Community Edition 2025 [81] environment. These metrics were presented in tables, and their magnitudes were evaluated against commonly used scales to determine the degree of agreement. Finally, time consumption data were analyzed using a statistical comparison approach to detect significant differences in ratings from the same rater and between ratings from different raters. Accordingly, the time consumption for the first (r1) and second (r2) replications of each rater (R1, R2, R3) was compared, along with the inter-rater time consumption for the first and second replications. To determine the most appropriate statistical test, a normality check of the data was performed using the Shapiro-Wilk test ($\alpha = 0.05$, $p > 0.05$). All statistical comparisons were conducted in Microsoft Excel with the Real Statistics add-in [82]. The results of the time consumption data, along with the relevant metrics from the normality checks and statistical comparison tests, were reported in table form.

3. Results and Discussion

3.1. Overall Feature-Based Agreement

Figure 1 shows the results of multi-dimensional scaling based on target variables, which considered the human raters (R1, R2, and R3) and used the codes given by the raters in r1 and r2 as features. In terms of overall agreement, the expectation was that the data points from the ratings would overlap significantly for both intra- and inter-rater assessments. As illustrated, this overlap occurred to some extent, indicating several agreements at the image level; however, many data points remained dispersed when considering the replication number.

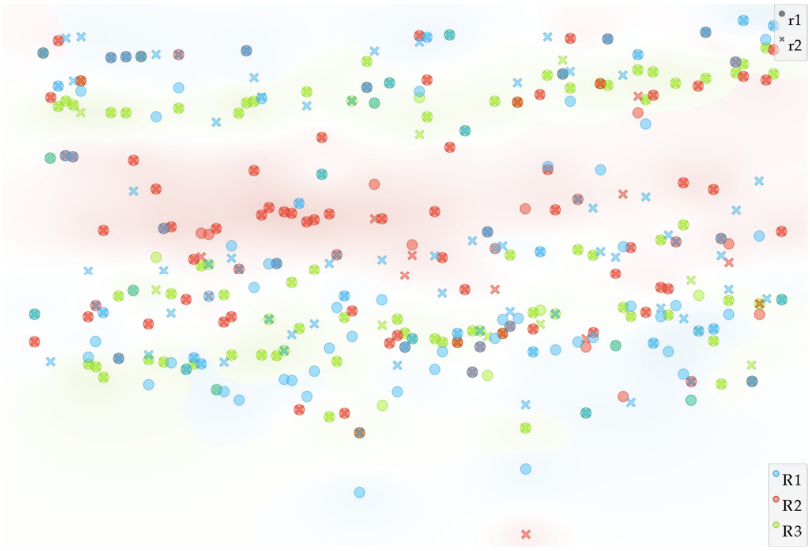

Figure 1. Results of multi-dimensional scaling concerning human rater agreement. Legend: R1—rater 1, R2—rater 2, R3—rater 3, r1—data from the first replication, r2—data from the second replication.

Figure 2, on the other hand, indicates a higher degree of disagreement when including the ratings from the DL model. While there were some cases of agreement between the ratings, many points in the RM are positioned well outside the ratings provided by the human experts, indicating the degree of disagreement in relation to the ground truth data.

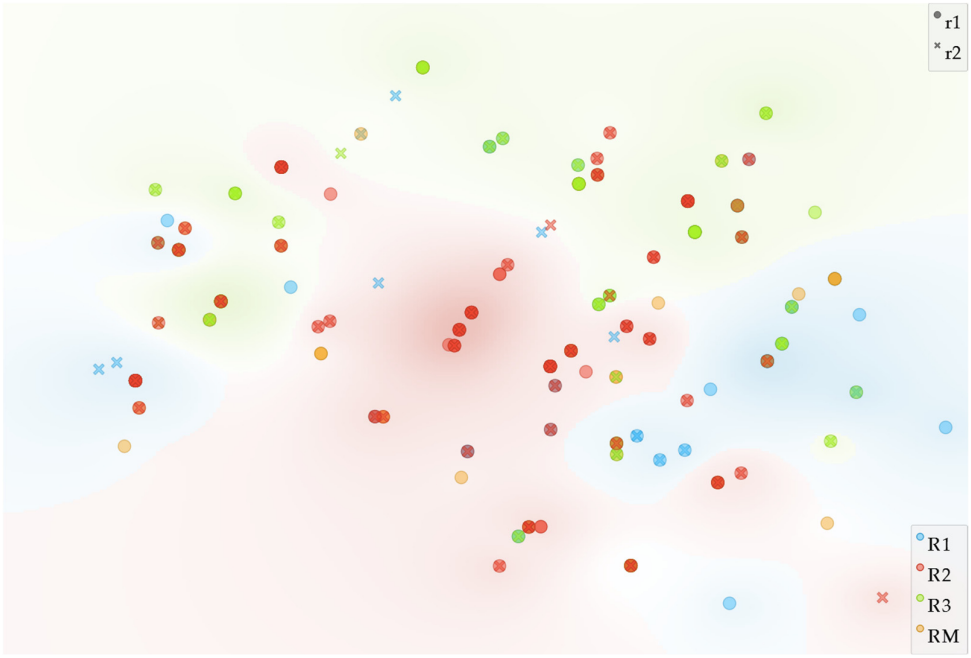

Figure 2. Results of multi-dimensional scaling concerning human raters and model agreement. Legend: R1—rater 1, R2—rater 2, R3—rater 3, RM—rating of the deep learning model, r1—data from the first replication, r2—data from the second replication. Note: for RM a single rating was used.

3.2. Intra-Rater Agreement

Table 3 shows the intra-rater agreement levels for three human raters (R1, R2, and R3) over two rounds of assessments (r1 and r2). The results indicate variability in the levels of intra-rater agreement between the two rounds of ratings (r1 and r2) among the three human raters, with observed agreement ranging from 0.61 to 1.00. This highlights a range of consensus, while the expected agreement by chance varied from 0.25 to 0.84. Cohen's kappa statistic, which adjusts for chance, ranges from 0.48 to 1.00, indicating moderate to almost perfect levels of agreement, and the percentage agreement spans from 61% to 100%. Most intra-rater agreements in this study are classified as moderate to almost perfect, with two instances reaching substantial levels [73], which in turn indicate a moderate to high level of consistency in ratings for the same rater. However, there were instances where the level of agreement showed that for the same image, decision-making regarding the correct posture changed after the second replication, as illustrated, for instance, by the first rater's case.

As such, moderate agreement was observed in cases like BR1r1 (ratings on the back posture made by rater 1 in replication 1) vs. BR1r2 (ratings on the back posture made by rater 1 in replication 2), accounting for 69% (k = 0.56) and ACR1r1 vs. ACR1r2 (61%, k = 0.48, where AC stands for the rating of action category), with a lower agreement on the action category data likely coming from a different evaluation of the back and legs' posture by R1. On the other hand, almost perfect agreement was found in cases such as AR2r1 vs. AR2r2 (100%, k = 1.00, where A stands for the rating on arm posture). The observed agreement showed strong intra-rater reliability overall, while the expected agreement by chance showed variability, contributing to differences in Cohen's kappa values.

Table 3. Results of intra-rater reliability for the three human raters.

Compared Datasets		# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
BR1r1	BR1r2	100	0.69	0.29	0.56	69	Moderate agreement
AR1r1	AR1r2	100	0.93	0.71	0.76	93	Substantial agreement
LR1r1	LR1r2	100	0.68	0.26	0.57	68	Moderate agreement
FR1r1	FR1r2	100	0.90	0.62	0.74	90	Substantial agreement
ACR1r1	ACR1r2	100	0.61	0.25	0.48	61	Moderate agreement
BR2r1	BR2r2	100	0.97	0.33	0.96	97	Almost perfect agreement
AR2r1	AR2r2	100	1.00	0.73	1.00	100	Almost perfect agreement
LR2r1	LR2r2	97	0.99	0.25	0.99	99	Almost perfect agreement
FR2r1	FR2r2	100	0.95	0.51	0.90	95	Almost perfect agreement
ACR2r1	ACR2r2	97	0.95	0.26	0.93	95	Almost perfect agreement
BR3r1	BR3r2	100	0.96	0.39	0.93	96	Almost perfect agreement
AR3r1	AR3r2	100	0.98	0.84	0.88	98	Almost perfect agreement
LR3r1	LR3r2	100	0.99	0.32	0.99	99	Almost perfect agreement
FR3r1	FR3r2	100	0.98	0.48	0.96	98	Almost perfect agreement
ACR3r1	ACR3r2	100	0.96	0.32	0.94	96	Almost perfect agreement

Note: *Po* denotes observed agreement; *Pe* denotes expected agreement by chance; *k* denotes Cohen's kappa statistic, B denotes the posture of the back, A denotes the posture of the arms, L denotes the posture of the legs, F denotes the level of force exertion, AC denotes the action category. The full abbreviations were composed by using the type of feature under assessment (B, A, L, or AC, Table 2) and the datasets presented in Table 1.

The intra-rater reliabilities observed in earlier studies that assessed the reliability of OWAS observations, and which reported generally high intra-rater reliability, provide strong support for the findings of this study. Karhu et al. [41] reported intra-rater reliability ranging from 70% to 100%, while de Bruijn et al. [50] reported reliability ranging from 83% to 100%, depending on the body parts assessed. Additionally, De Bruijn et al. [50] reported Cohen's kappa values above 0.6 in all comparisons when observers were adequately trained and adhered to clear guidelines. Thus, the high levels of agreement observed in this study suggest that raters likely followed well-defined criteria and possessed the necessary expertise. However, task complexity can affect reliability [50], which is reflected in the moderate agreement noted in some instances in this study, indicating that certain postures may have been more subjective or difficult to rate consistently. These results carry significant implications for the study; while the high levels of agreement demonstrate that

the rating process was robust, the moderate agreement in specific areas underscores the need for further refinement. Enhancing training or clarifying the rating criteria could help mitigate these inconsistencies and improve reliability in future assessments by humans, or machine learning models could be used to get around the reliability problem.

3.3. Pair-Based Inter-Rater Agreement

Table 4 presents the inter-rater reliability for three human raters (R1, R2, and R3) across the two rounds of assessments (r1 and r2). The results indicate a wide variability in levels of agreement among the raters, with observed agreement ranging from 0.32 to 0.92, highlighting a different spectrum of consensus. The expected agreement by chance ranged from 0.21 to 0.79. Cohen's kappa statistic, which adjusts for chance agreement, was found between 0.02 and 0.64, indicating levels of agreement ranging from slight to substantial. The percentage agreement spanned from 32% to 92%. However, most degrees of agreement in this study are classified as slight to moderate, although three instances reach substantial levels of agreement. For instance, comparisons such as BR2r2 vs. BR3r2 showed slight agreement (32%, k = 0.02), while AR1r2 vs. AR3r2 exhibited substantial agreement (92%, k = 0.63). However, some pairs, like LR1r1 vs. LR3r1, displayed only moderate agreement (64%, k = 0.52). These variations highlight differences in the raters' consistency, which may be influenced by factors such as task complexity, rater expertise, clarity of assessment criteria [50,53], and rater bias [72]. Studies indicate that if two or more raters are accurately observing the same postures, their assessments should be identical; any discrepancies in their reports are assumed to reflect the individual biases or characteristics of the raters [72].

Table 4. Results of inter-rater reliability among the three human raters.

Compared Datasets		# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
BR1r1	BR2r1	100	0.46	0.24	0.29	46	Fair agreement
BR1r1	BR3r1	100	0.62	0.36	0.41	62	Moderate agreement
BR2r1	BR3r1	100	0.34	0.29	0.07	34	Slight agreement
AR1r1	AR2r1	100	0.91	0.70	0.70	91	Substantial agreement
AR1r1	AR3r1	100	0.89	0.75	0.56	89	Moderate agreement
AR2r1	AR3r1	100	0.88	0.78	0.46	88	Moderate agreement
LR1r1	LR2r1	97	0.57	0.21	0.45	57	Moderate agreement
LR1r1	LR3r1	100	0.64	0.26	0.52	64	Moderate agreement
LR2r1	LR3r1	100	0.60	0.25	0.46	60	Moderate agreement
FR1r1	FR2r1	100	0.74	0.52	0.46	74	Moderate agreement
FR1r1	FR3r1	100	0.70	0.53	0.37	70	Fair agreement

Table 4. *Cont.*

Compared Datasets		# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
FR2r1	FR3r1	100	0.72	0.48	0.46	72	Moderate agreement
ACR1r1	ACR2r1	100	0.54	0.24	0.40	54	Fair agreement
ACR1r1	ACR3r1	100	0.52	0.27	0.34	52	Fair agreement
ACR2r1	ACR3r1	97	0.40	0.23	0.22	40	Fair agreement
BR1r2	BR2r2	100	0.58	0.28	0.41	58	Moderate agreement
BR1r2	BR3r2	100	0.41	0.30	0.15	41	Slight agreement
BR2r2	BR3r2	100	0.32	0.30	0.02	32	Slight agreement
AR1r2	AR2r2	100	0.90	0.73	0.62	90	Substantial agreement
AR1r2	AR3r2	100	0.92	0.79	0.63	92	Substantial agreement
AR2r2	AR3r2	100	0.86	0.78	0.37	86	Fair agreement
LR1r2	LR2r2	100	0.56	0.24	0.42	56	Moderate agreement
LR1r2	LR3r2	100	0.75	0.31	0.64	75	Substantial agreement
LR2r2	LR3r2	100	0.58	0.25	0.44	58	Moderate agreement
FR1r2	FR2r2	100	0.79	0.55	0.53	79	Moderate agreement
FR1r2	FR3r2	100	0.73	0.55	0.40	73	Fair agreement
FR2r2	FR3r2	100	0.75	0.48	0.52	75	Moderate agreement
ACR1r2	ACR2r2	100	0.56	0.25	0.42	56	Moderate agreement
ACR1r2	ACR3r2	100	0.41	0.25	0.22	41	Fair agreement
ACR2r2	ACR3r2	100	0.40	0.23	0.22	40	Fair agreement

Note: *Po* denotes observed agreement; *Pe* denotes expected agreement by chance; *k* denotes Cohen's kappa statistic, B denotes the posture of the back, A denotes the posture of the arms, L denotes the posture of the legs, F denotes the level of force exertion, AC denotes the action category. The full abbreviations were composed by using the type of feature under assessment (B, A, L, or AC, Table 2) and the datasets presented in Table 1.

The results of this study closely align with those of earlier studies carried out in real-world work settings. For instance, Karhu et al. [41] reported inter-observer reliability of 93%, while Heinsalmi [83] reported a 90% agreement on overall working posture. Similarly, the study by Lins et al. [53] found high inter-rater agreement exceeding 98% (k = 0.98) for arm postures, while leg posture classification showed slightly lower agreement levels, ranging from 66% to 97% (k = 0.85), and indicated that reliability is affected by the raters' familiarity with the method and the complexity of the analyzed postures. In this study, the moderate to substantial agreement observed in many comparisons suggests that the raters had a reasonable understanding of the assessment criteria, while the instances of slight agreement may point to challenges in consistently interpreting or applying these criteria [50,53]. Furthermore, De Bruijn et al. [50] observed that clear guidelines, well-defined criteria,

task complexity, and adequate training were key to achieving high reliability and can influence inter-rater reliability in OWAS observations. The results in this study support this observation, as the substantial agreement noted in some comparisons indicates adherence to clear guidelines, while the slight agreement in other instances underscores the need for further refinement of the criteria or additional support for the raters.

3.4. Pair-Based Agreement to the Ground Truth Data

The pair-based agreement results between the ratings of the DL model (RM) and those of the human raters (R1, R2, and R3) are displayed in Table 5. The findings indicate varying levels of agreement among the human and DL model ratings, with observed agreement ranging from 0.30 to 0.85. This demonstrates a spectrum of agreement, while the expected agreement by chance varied from 0.24 to 0.84. The Cohen's kappa statistic ranged from −0.03 to 0.34, indicating levels of agreement from poor to fair. Additionally, the percentage agreement spanned from 30% to 85%. Most agreements in this study are classified as slight to fair, with five instances categorized as poor. These findings reveal challenges in achieving consistency between the human raters and the ratings of the DL model across all assessed categories. The trained DL model showed fair agreement with human raters in some categories, such as FR3r2 vs. FRM (63%, $k = 0.34$), reflecting its potential to replicate human-like assessments when the human rater understands the movement in the assessed images. However, in other comparisons, like AR1r1 vs. ARM (75%, $k = -0.03$), poor agreement was observed, highlighting challenges in achieving consistency with human evaluations. The variation in agreement levels can be attributed to the DL model's reliance on learned visual features [63], which may not always align with the human raters' interpretation of complex or subtle posture variations. The ratings provided by the used model, a convolutional neural network adapted for posture analysis [59], are based on data-driven features learned during its training [60,84]. It identifies patterns in visual input to classify postures in accordance with the training data provided [59,60,84]. Unlike the DL model, the human raters used standardized guidelines for their ratings, which means that they may miss context such as movement. Despite the consistency of guidelines, differences in agreement levels suggest that while the DL model provides a stable reference, it may struggle to align with subjective human interpretations [35], especially in complex or nuanced classifications.

Table 5. Results of pair-based agreement between the human raters and the deep learning model.

Ratings Under Comparison		# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
BR1r1	BRM	100	0.43	0.34	0.13	43	Slight agreement
BR1r2	BRM	100	0.34	0.30	0.06	34	Slight agreement
BR2r1	BRM	100	0.32	0.30	0.03	32	Slight agreement
BR2r2	BRM	100	0.30	0.30	0.00	30	Poor agreement
BR3r1	BRM	100	0.57	0.37	0.32	57	Fair agreement
BR3r2	BRM	100	0.57	0.38	0.31	57	Fair agreement
AR1r1	ARM	100	0.75	0.76	−0.03	75	Poor agreement
AR1r2	ARM	100	0.79	0.79	−0.02	79	Poor agreement
AR2r1	ARM	100	0.78	0.78	−0.02	78	Poor agreement
AR2r2	ARM	100	0.78	0.78	−0.02	78	Poor agreement

Table 5. *Cont.*

Ratings Under Comparison		# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
AR3r1	ARM	100	0.85	0.84	0.04	85	Slight agreement
AR3r2	ARM	100	0.85	0.84	0.04	85	Slight agreement
LR1r1	LRM	100	0.38	0.24	0.18	38	Slight agreement
LR1r2	LRM	100	0.46	0.28	0.25	46	Fair agreement
LR2r1	LRM	97	0.44	0.25	0.26	44	Fair agreement
LR2r2	LRM	100	0.43	0.24	0.25	43	Fair agreement
LR3r1	LRM	100	0.50	0.29	0.29	50	Fair agreement
LR3r2	LRM	100	0.49	0.30	0.28	49	Fair agreement
FR1r1	FRM	100	0.60	0.47	0.24	60	Fair agreement
FR1r2	FRM	100	0.59	0.49	0.20	59	Slight agreement
FR2R1	FRM	100	0.53	0.44	0.16	53	Slight agreement
FR2r2	FRM	100	0.56	0.44	0.21	56	Fair agreement
FR3r1	FRM	100	0.61	0.44	0.31	61	Fair agreement
FR3r2	FRM	100	0.63	0.44	0.34	63	Fair agreement
ACR1r1	ACRM	100	0.32	0.26	0.08	32	Slight agreement
ACR1r2	ACRM	100	0.38	0.25	0.18	38	Slight agreement
ACR2r1	ACRM	97	0.35	0.24	0.15	35	Slight agreement
ACR2r2	ACRM	100	0.36	0.24	0.16	36	Slight agreement
ACR3r1	ACRM	100	0.50	0.29	0.29	50	Fair agreement
ACR3r2	ACRM	100	0.51	0.30	0.30	51	Fair agreement

Note: *Po* denotes observed agreement; *Pe* denotes expected agreement by chance; *k* denotes Cohen's kappa statistic, B denotes the posture of the back, A denotes the posture of the arms, L denotes the posture of the legs, F denotes the level of force exertion, AC denotes the action category. The full abbreviations were composed by using the type of feature under assessment (B, A, L, or AC, Table 2) and the datasets presented in Table 1.

Lins et al. [53] found that inter-rater agreement using the OWAS method was high for arm postures (k = 0.98) but lower for leg postures (k = 0.85). This highlights the inherent challenges in accurately classifying certain postures, particularly when variations are subtle. Similarly, Widyanti [75] emphasized the importance of training in ensuring consistent assessments, which may explain the variability observed among human raters in this study. While the human raters adhered to guidelines, ambiguities in posture categories could have introduced inconsistencies. De Bruijn et al. [50] emphasized the role of task complexity and guideline clarity in reliability studies. On the other hand, the DL model, as the reference, may excel in straightforward classifications but also may face some limitations in cases requiring more interpretive judgment. These findings suggest that some of the observed discrepancies could stem from differences in how the model and human raters interpret subtle features of certain postures. Therefore, the DL model serves as a reliable reference and can benefit from upcoming training data, which could enhance its ability to capture subtle posture variations. On the human side, providing additional training focused on ambiguous or complex cases, alongside improved guidelines, could help align human assessments more closely with ground truth predictions.

3.5. Overall Agreement to the Ground Truth Data

Table 6 shows the inter-rater reliability among the three human raters and the ratings of the DL model. The results indicate variability in the levels of agreement among the three human raters and the DL model, with observed agreement ranging from 0.49 to 0.89. This highlights a spectrum of consensus, while the expected agreement by chance varied from 0.23 to 0.79. The Fleiss' kappa statistic, which adjusts for chance, ranged from 0.26 to 0.49, indicating fair to moderate levels of agreement, and the percentage agreement spanned from 49% to 89%. Most agreements in this study are classified as fair, with two instances reaching moderate levels [73].

Table 6. Results of overall agreement among the three human raters and the ResNet-50 model.

Ratings Under Comparison				# Ratings	*Po*	*Pe*	*k*	%Agreement	Interpretation of Kappa
BR1R1	BR2R1	BR3R1	BRM	100	0.53	0.34	0.28	53	Fair agreement
AR1R1	AR2R1	AR3R1	ARM	100	0.88	0.77	0.49	88	Moderate agreement
LR1R1	LR2R1	LR3R1	LRM	97	0.52	0.23	0.37	52	Fair agreement
FR1R1	FR2R1	FR3R1	FRM	100	0.66	0.47	0.37	66	Fair agreement
ACR1R1	ACR2R1	ACR2R1	ACRM	97	0.52	0.26	0.35	52	Fair agreement
BR1R2	BR2R2	BR3R2	BRM	100	0.49	0.31	0.26	49	Fair agreement
AR1R2	AR2R2	AR3R2	ARM	100	0.89	0.79	0.47	89	Moderate agreement
LR1R2	LR2R2	LR3R2	LRM	100	0.53	0.25	0.38	53	Fair agreement
FR1R2	FR2R2	FR3R2	FRM	100	0.68	0.47	0.37	68	Fair agreement
ACR1R2	ACR2R2	ACR2R2	ACRM	100	0.51	0.27	0.33	51	Fair agreement

Note: *Po* denotes observed agreement; *Pe* denotes expected agreement by chance; *k* denotes Fleiss's kappa statistic, B denotes the posture of the back, A denotes the posture of the arms, L denotes the posture of the legs, F denotes the level of force exertion, AC denotes the action category. The full abbreviations were composed by using the type of feature under assessment (B, A, L, or AC, Table 2) and the datasets presented in Table 1.

The findings show a notable correspondence between the DL model's ratings (considered the ground truth) and the assessments by human raters. This alignment can be attributed to several factors, including the robust training of the DL model on a comprehensive dataset tailored to the task [59], which enabled it to make accurate predictions consistent with the assessment criteria used by human raters. When human ratings corresponded closely with those of the DL model, it suggested that both parties recognized similar characteristics in the data. On the other hand, the DL model provided a consistent benchmark for comparison, reinforcing its effectiveness in capturing the complexities of the postural classification task [59,84,85]. This agreement indicates that human raters applied consistent judgment criteria that aligned well with the DL model's training parameters. However, discrepancies between raters and the DL model may stem from differences in inter-human judgments or challenges in interpreting borderline cases, which the DL model

processed in a more objective manner, highlighting the inherent subjective interpretations of postural deviations by human raters [86]. This lack of consistency on the part of human raters might also arise from factors such as varying experience levels and confusion in terms of perception among raters [53,87].

Comparing these results with other studies reveals that the agreement achieved aligns with similar contexts where variability in human judgment and task complexity are critical factors [75]. For example, in a previous study by Widyanti [75] involving inter-rater reliability of OWAS using experts and new raters, percentage agreement ranged from 31.40% to 75%, while Fleiss' kappa ranged from 0.20 to 0.53, indicating fair to moderate levels of agreement, respectively. The characteristics of the task and the specific performance of the DL model in managing the dataset likely influenced these outcomes, suggesting that differences in interpretation, especially for borderline cases, could explain some discrepancies between the raters and the DL model. These findings have significant implications for the study's context, enhancing confidence in the model's reliability as a reference tool. By analyzing cases of divergence in ratings, researchers can refine both the model and the criteria for human assessments, thereby improving overall consistency and bridging the gap between algorithmic and human decision-making. The insights gained from this study emphasize the model's potential application in similar contexts where standardized and replicable ground truth references are essential.

3.6. Time Consumption

Ratings by the DL model took, on average, about one second per image. This highlights the significantly higher speed of machine ratings, which, on average, were approximately 19 to 53 times faster than those provided by human experts. This speed comparison was applicable to the computer architecture used in this study and to the sequence of computations performed, which included sequential image prediction, display, and storage. It is evident that for large datasets, the time required to make predictions on images without displaying them will be much lower, depending on the specific computer architecture employed.

Table 7 presents the results of the statistical comparison tests, highlighting three important findings. First, there were significant differences in time consumption at both the intra- and inter-rater levels during the assessment. An exception was noted for the third rater, who was more consistent in terms of time requirements to rate the images and who also utilized his expertise to train the DL model. Likely, this could be related to a greater familiarity with the images used, since they were selected from the dataset used to train the model, and the procedures used for assessment.

Second, there was a varying degree of time resources utilized, with increasing efficiency observed in the order of R3, R1, and R2, pointing out an inconsistency in terms of time resources when human experts carry on the rating tasks, resting in their different abilities to approach the problem in terms of speed. Lastly, with one exception, the raters demonstrated improved time resource utilization in the second replication compared to the first, which may be related to some degree of familiarization with that dataset and with the protocol used for making the ratings.

However, since this familiarization encompasses both the procedure used and the dataset itself, it is highly unlikely that the same trends in resource utilization will be maintained when approaching a new dataset. The dataset employed in this study consisted of 100 images, while real-world applications may involve much larger datasets, potentially leading to intellectual fatigue for human raters. This suggests that the observed performance in this study may not be replicated with new images from broader datasets, thus highlighting the effectiveness of machine learning models in addressing such tasks.

Table 7. Results of comparison tests for time consumption data.

Variables Under Comparison	Median Values (s)	Results of Normality Test [1]	Results of Comparison Test [2]
TR1r1-TR1r2	30.0–24.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR2r1-TR2r2	52.5–44.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR3r1-TR3r2	19.0–20.0	No, $p < 0.001$-No, $p < 0.001$	No, $p = 0.608$
TR1r1-TR2r1	30.0–52.5	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR1r1-TR3r1	30.0–19.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR2r1-TR3r1	52.5–19.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR1r2-TR2r2	24.0–44.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$
TR1r2-TR3r2	30.0–20.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p = 0.003$
TR2r2-TR3r2	44.0–20.0	No, $p < 0.001$-No, $p < 0.001$	Yes, $p < 0.001$

Note: [1]—According to Shapiro–Wilk test; [2]—significant differences according to Mann–Whitney two-tailed nonparametric test, T stands for the time consumption dataset.

The significantly improved speed and consistency of the deep learning (DL)-based OWAS assessment shown in this study opens numerous practical applications in real-world forest harvesting environments. For instance, integrating this technology into mobile applications could equip field supervisors with a quick and objective tool for spot-checking postures and identifying immediate ergonomic risks, thereby supplementing traditional observational methods [54]. Additionally, the possibility of automated analysis of video footage from stationary cameras, drones, or even body-worn devices presents an opportunity for extensive, longitudinal ergonomic risk surveillance across various operations [84]. This approach would facilitate the detection of high-risk patterns or tasks over longer durations and entire work teams, surpassing the limitations of traditional snapshot assessments previously noted [35,46]. However, to realize this potential, several practical challenges must be addressed, including the computational demands for real-time analysis on mobile or edge devices, ensuring model robustness against the variable environmental conditions commonly encountered in forestry—such as fluctuating lighting, precipitation, and obstructions from vegetation or equipment—which can affect computer vision performance [62,85]. Furthermore, continuous model maintenance and domain-specific fine-tuning are necessary to uphold accuracy as operational practices, tools, and worker demographics evolve [61,64]. Tackling these technical hurdles through further research and development will be essential for converting the demonstrated potential of DL-based postural assessment into widely used tools for enhancing occupational safety and health management in the challenging forestry sector.

4. Conclusions

This study shows that deep learning (DL) models present significant advantages for conducting OWAS-based postural assessments in manual and partly mechanized forest operations, offering remarkable speed enhancements (19 to 53 times faster on average) compared to traditional human-rater methods while achieving comparable levels of reliability. The findings showed that while human raters exhibited moderate to almost perfect intra-rater reliability (Cohen's kappa = 0.48–1.00), confirming individual consistency, their inter-rater agreement was considerably lower, ranging from slight to substantial (Cohen's kappa = 0.02–0.64). This discrepancy underscores the inherent subjectivity and variability present in human postural assessments, even among experts using a standardized method. Comparisons against the DL model, utilized as a consistent benchmark, revealed poor

to fair pairwise agreement between individual human raters and the model (Cohen's kappa = −0.03–0.34), yet achieved fair to moderate overall agreement when considering all human ratings collectively against the model (Fleiss' kappa = 0.28–0.49). These results suggest that while the DL model effectively captures general postural trends recognized collectively by humans, specific interpretations of individual postures can still diverge significantly between the automated system and individual expert assessments. Consequently, the DL model serves not only as a highly resource-efficient alternative, drastically reducing assessment time, but also as a stable reference point for evaluating OWAS assessments, effectively mitigating the challenges associated with human rater variability. Nonetheless, areas for improvement persist, particularly in enhancing the alignment between machine outputs and nuanced human interpretations for complex or borderline postures. Future research should prioritize the refinement of DL model parameters, the expansion of comprehensively annotated datasets reflecting diverse operational conditions, and the enhancement of training protocols for human raters to improve classification consistency. By addressing these aspects, the transformative potential of DL in revolutionizing postural assessment methods can be fully realized, paving the way for advancements essential to enhancing the occupational safety, operational efficiency, and overall sustainability of the forestry sector.

Author Contributions: Conceptualization, N.K. and S.A.B.; data curation, G.O.F.; formal analysis, G.O.F., M.V.M. and N.K.; funding acquisition, S.A.B.; investigation, G.O.F., M.V.M. and T.K.; methodology, G.O.F., M.V.M., N.K., T.K. and S.A.B.; project administration, S.A.B.; resources, G.O.F., M.V.M., N.K., T.K. and S.A.B.; software, G.O.F.; supervision, S.A.B.; validation, G.O.F., M.V.M., N.K., T.K. and S.A.B.; visualization, S.A.B.; writing—original draft, G.O.F., M.V.M., N.K., T.K. and S.A.B.; writing—review & editing, S.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by two grants of the Romanian Ministry of Education and Research, CNCS—UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2023-0141, and project number PN-IV-P8-8.1-PRE-HE-ORG-2024-0186, within PNCDI IV. Part of the study was funded by National Research Council of Thailand (NRCT) and Kasetsart University: contract number N42A670571. The APC was waived.

Data Availability Statement: Image data supporting this study may be made available upon a reasonable request to the first author of the study.

Acknowledgments: The authors are grateful to the Department of Forest Engineering, Forest Management Planning and Terrestrial Measurements for providing part of the infrastructure required to carry on this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Beims, R.F.; Arredondo, R.; Sosa Carrero, D.J.; Yuan, Z.; Li, H.; Shui, H.; Zhang, Y.; Leitch, M.; Xu, C.C. Functionalized Wood as Bio-Based Advanced Materials: Properties, Applications, and Challenges. *Renew. Sustain. Energy Rev.* **2022**, *157*, 112074. [CrossRef]
2. Jiang, F.; Li, T.; Li, Y.; Zhang, Y.; Gong, A.; Dai, J.; Hitz, E.; Luo, W.; Hu, L. Wood-Based Nanotechnologies toward Sustainability. *Adv. Mater.* **2018**, *30*, 1703453. [CrossRef] [PubMed]
3. Braga, C.I.; Petrea, S.; Zaharia, A.; Cucu, A.B.; Serban, T.; Ienasoiu, G.; Radu, G.R. Assessing the Greenhouse Gas Mitigation Potential of Harvested Wood Products in Romania and Their Contribution to Achieving Climate Neutrality. *Sustainability* **2025**, *17*, 640. [CrossRef]
4. Do, T.T.H.; Ly, T.B.T.; Hoang, N.T. A new integrated circular economy index and a combined method for optimization of wood production chain considering carbon neutrality. *Chemosphere* **2023**, *311*, 137029. [CrossRef]
5. Sudheshwar, A.; Vogel, K.; Nyström, G.; Malinverno, N.; Arnaudo, M.; Camacho, C.E.G.; Beloin-Saint-Pierre, D.; Hischier, R.; Som, C. Unraveling the climate neutrality of wood derivatives and biopolymers. *RSC Sustain.* **2024**, *2*, 1487–1497. [CrossRef]

6. Jarre, M.; Petit-Boix, A.; Priefer, C.; Meyer, R.; Leipold, S. Transforming the bio-based sector towards a circular economy—What can we learn from wood cascading? *For. Policy Econ.* **2020**, *110*, 101872. [CrossRef]
7. Hassegawa, M.; Brusselen, J.; Cramm, M.; Verkerk, P.J. Wood-based products in the circular bioeconomy: Status and opportunities towards environmental sustainability. *Land* **2022**, *11*, 2131. [CrossRef]
8. Nishiguchi, S.; Tabata, T. Assessment of social, economic, and environmental aspects of woody biomass energy utilization: Direct burning and wood pellets. *Renew. Sustain. Energy Rev.* **2016**, *57*, 1279–1286. [CrossRef]
9. Kropivšek, J.; Zupančič, A. Development of competencies in the Slovenian wood-industry. *Dyn. Relat. Manag. J.* **2016**, *5*, 3–20. [CrossRef]
10. Klein, D.; Kies, U.; Schulte, A. Regional employment trends of wood-based industries in Germany's forest cluster: A comparative shift-share analysis of post-reunification development. *Eur. J. For. Res.* **2009**, *128*, 205–219. [CrossRef]
11. Pang, S.; H'ng, P.; Chai, L.; Lee, S.; Paridah, M.T. Value added productivity performance of the Peninsular Malaysian wood sawmilling industry. *BioResources* **2015**, *10*, 7324–7338. [CrossRef]
12. Temu, B.J.; Monela, G.C.; Darr, D.; Abdallah, J.M.; Pretzsch, J. Forest sector contribution to the National Economy: Example wood products value chains originating from Iringa region, Tanzania. *For. Policy Econ.* **2024**, *164*, 103246. [CrossRef]
13. Michaud, G.; Jolley, G.J. Economic contribution of Ohio's wood industry cluster: Identifying opportunities in the Appalachian region. *Rev. Reg. Stud.* **2019**, *49*, 149–171. [CrossRef]
14. Heinimann, H.R. Forest operations engineering and management—The ways behind and ahead of a scientific discipline. *Croat. J. For. Eng.* **2007**, *28*, 107–121.
15. Marchi, E.; Picchio, R.; Spinelli, R.; Verani, S.; Venanzi, R.; Certini, G. Environmental impact assessment of different logging methods in pine forests thinning. *Ecol. Eng.* **2014**, *70*, 429–436. [CrossRef]
16. Szewczyk, G.; Spinelli, R.; Magagnotti, N.; Tylek, P.; Sowa, J.M.; Rudy, P.; Gaj-Gielarowiec, D. The mental workload of harvester operators working in steep terrain conditions. *Silva Fenn.* **2020**, *54*, 10355. [CrossRef]
17. Passicot, P.; Murphy, G.E. Effect of work schedule design on productivity of mechanized harvesting operations in Chile. *N. Z. J. For. Sci.* **2013**, *43*, 2. [CrossRef]
18. Moskalik, T.; Borz, S.A.; Dvořák, J.; Ferencik, M.; Glushkov, S.; Muiste, P.; Lazdiņš, A.; Styranivsky, O. Timber harvesting methods in Eastern European countries: A review. *Croat. J. For. Eng.* **2017**, *38*, 231–241.
19. Gerasimov, Y.; Sokolov, A. Ergonomic evaluation and comparison of wood harvesting systems in Northwest Russia. *Appl. Ergon.* **2014**, *45*, 318–338. [CrossRef]
20. Barbosa, R.P.; Fiedler, N.C.; Carmo, F.C.A.; Minette, L.J.; Silva, E.N. Analysis of posture in semi-mechanized forest harvesting in steep areas. *Rev. Árvore* **2014**, *38*, 733–738. [CrossRef]
21. Häggström, C.; Lindroos, O. Human, technology, organization and environment—A human factors perspective on performance in forest harvesting. *Int. J. For. Eng.* **2016**, *43*, 2. [CrossRef]
22. Grzywiński, W.; Wandycz, A.; Tomczak, A.; Jelonek, T. The prevalence of self-reported musculoskeletal symptoms among loggers in Poland. *Int. J. Ind. Ergon.* **2016**, *52*, 12–17. [CrossRef]
23. Calvo, A. Musculoskeletal disorders (MSD) risks in forestry: A case study to propose an analysis method. *Agric. Eng. Int.* **2009**, *11*, 1–9.
24. Cheţa, M.; Marcu, M.V.; Borz, S.A. Workload, exposure to noise, and risk of musculoskeletal disorders: A case study of motor-manual tree feeling and processing in poplar clear cuts. *Forests* **2018**, *9*, 300. [CrossRef]
25. Gómez-Galán, M.; Pérez-Alonso, J.; Callejón-Ferre, Á.J.; López-Martínez, J. Musculoskeletal disorders: OWAS review. *Ind. Health* **2017**, *55*, 314–337. [CrossRef]
26. Bevan, S. Economic Impact of Musculoskeletal Disorders (MSDs) on Work in Europe. *Best Pract. Res. Clin. Rheumatol.* **2015**, *29*, 356–373. [CrossRef]
27. Oh, I.H.; Yoon, S.J.; Seo, H.Y.; Kim, E.J.; Kim, Y.A. The economic burden of musculoskeletal disease in Korea: A cross-sectional study. *BMC Musculoskelet. Disord.* **2011**, *12*, 157. [CrossRef]
28. Borz, S.A.; Talagai, N.; Cheţa, M.; Chiriloiu, D.; Gavilanes Montoya, A.V.; Castillo Vizuete, D.D.; Marcu, M.V. Physical strain, exposure to noise and postural assessment in motor-manual felling of willow short rotation coppice: Results of a preliminary study. *Croat. J. For. Eng.* **2019**, *40*, 377–388. [CrossRef]
29. Pheasant, S.; Haslegrave, C.M. *Bodyspace: Anthropometry, Ergonomics and the Design of Work*, 3rd ed.; Taylor & Francis: Abingdon, UK, 2006.
30. Viviani, C.; Arezes, P.M.; Braganca, S.; Molenbroek, J.; Dianat, I.; Castellucci, H.I. Accuracy, precision and reliability in anthropometric surveys for ergonomics purposes in adult working populations: A literature review. *Int. J. Ind. Ergon.* **2018**, *65*, 1–16. [CrossRef]
31. Corella Justavino, F.; Jimenez Ramirez, R.; Meza Perez, N.; Borz, S.A. The use of OWAS in forest operations postural assessment: Advantages and limitations. *Bull. Transilv. Univ. Bras. Ser. II For. Wood Ind. Agric. Food Eng.* **2015**, *8*, 7–16.

32. Neitzel, R.; Yost, M. Task-based assessment of occupational vibration and noise exposure in forestry workers. *Aiha J.* **2002**, *63*, 617–627. [CrossRef]
33. Yongan, W.; Baojun, J. Effects of low temperature on operation efficiency of tree-felling by chainsaw in North China. *J. For. Res.* **1998**, *9*, 57–58. [CrossRef]
34. Li, G.; Buckle, P. Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods. *Ergonomics* **1999**, *42*, 674–695. [CrossRef] [PubMed]
35. David, G.C. Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders. *Occup. Med.* **2005**, *55*, 190–199. [CrossRef]
36. Kee, D. Systematic comparison of OWAS, RULA, and REBA based on a literature review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 595. [CrossRef]
37. Lopes, E.D.S.; Britto, P.C.; Rodrigues, C.K. Postural discomfort in manual operations of forest planting. *Floresta Ambient.* **2018**, *26*, 20170030. [CrossRef]
38. Denbeigh, K.; Slot, T.R.; Dumas, G.A. Wrist postures and forces in tree planters during three tree unloading conditions. *Ergonomics* **2013**, *56*, 1599–1607. [CrossRef]
39. Vosniak, J.; Lopes, E.D.S.; Fiedler, N.C.; Alves, R.T.; Venâncio, D.L. Demanded physical effort and posture in semi-mechanical hole-digging activity at forestry plantation. *Sci. For./For. Sci.* **2010**, *33*, 589–598.
40. Zanuttini, R.; Cielo, P.; Poncino, D. The OWAS method. Preliminary results for the evaluation of the risk of work-related musculoskeletal disorders (WMSD) in the forestry sector in Italy. *For. Riv. Selvic. Ecol. For.* **2005**, *2*, 242–255. [CrossRef]
41. Karhu, O.; Kansi, P.; Kuorinka, I. Correcting working postures in industry: A practical method for analysis. *Appl. Ergon.* **1977**, *8*, 199–201. [CrossRef]
42. Takala, E.P.; Pehkonen, I.; Forsman, M.; Hansson, G.Å.; Mathiassen, S.E.; Neumann, W.P.; Sjøgaard, G.; Veiersted, K.B.; Westgaard, R.H.; Winkel, J. Systematic evaluation of observational methods assessing biomechanical exposures at work. *Scand. J. Work Environ. Health* **2010**, *36*, 3–24. [CrossRef] [PubMed]
43. Helander, M. *A Guide to Human Factors and Ergonomics*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2006.
44. Burdorf, A.; Derksen, J.; Naaktgeboren, B.; Riel, M. Measurement of trunk bending during work by direct observation and continuous measurement. *Appl. Ergon.* **1992**, *23*, 263–267. [CrossRef] [PubMed]
45. Borz, S.A.; Castro Perez, S.N. Effect of the sampling strategy on the accuracy of postural classification: An example from motor-manual tree felling and processing. *Rev. Pădurilor* **2020**, *135*, 19–41.
46. Brandl, C.; Mertens, A.; Schlick, C.M. Effect of sampling interval on the reliability of ergonomic analysis using the Ovako Working Posture Analysing System (OWAS). *Int. J. Ind. Ergon.* **2017**, *57*, 68–73. [CrossRef]
47. Beek, A.J.; Mathiassen, S.E.; Windhorst, J.; Burdorf, A. An evaluation of methods assessing the physical demands of manual lifting in scaffolding. *Appl. Ergon.* **2005**, *36*, 213–222. [CrossRef]
48. Kee, D.; Karwowski, W. A comparison of three observational techniques for assessing postural loads in industry. *Int. J. Occup. Saf. Ergon.* **2007**, *13*, 3–14. [CrossRef]
49. Micheletti Cremasco, M.; Giustetto, A.; Caffaro, F.; Colantoni, A.; Cavallo, E.; Grigolato, S. Risk assessment for musculoskeletal disorders in forestry: A comparison between RULA and REBA in the manual feeding of a wood-chipper. *Int. J. Environ. Res. Public Health* **2019**, *16*, 793. [CrossRef]
50. De Bruijn, I.; Engels, J.A.; Van Der Gulden, J.W. A simple method to evaluate the reliability of OWAS observations. *Appl. Ergon.* **1998**, *29*, 281–283. [CrossRef]
51. Mattila, M.; Karwowski, W.; Vilkki, M. Analysis of working postures in hammering tasks on building construction sites using the computerized OWAS method. *Appl. Ergon.* **1993**, *24*, 405–412. [CrossRef]
52. Kivi, P.; Mattila, M. Analysis and improvement of work postures in the building industry: Application of the computerised OWAS method. *Appl. Ergon.* **1991**, *22*, 43–48. [CrossRef]
53. Lins, C.; Fudickar, S.; Hein, A. OWAS inter-rater reliability. *Appl. Ergon.* **2021**, *95*, 103357. [CrossRef] [PubMed]
54. Fığlalı, N.; Cihan, A.; Esen, H.; Fığlalı, A.; Çeşmeci, D.; Güllü, M.K.; Yılmaz, M.K. Image processing-aided working posture analysis: I-OWAS. *Comput. Ind. Eng.* **2015**, *85*, 384–394. [CrossRef]
55. Wahyudi, M.A.; Dania, W.A.; Silalahi, R.L. Work posture analysis of manual material handling using OWAS method. *Agric. Agric. Sci. Procedia* **2015**, *3*, 195–199. [CrossRef]
56. Miedema, M.C.; Douwes, M.; Dul, J. Recommended maximum holding times for prevention of discomfort of static standing postures. *Int. J. Ind. Ergon.* **1997**, *19*, 9–18. [CrossRef]
57. Gaskin, J.E. An ergonomic evaluation of two motor-manual delimbing techniques. *Int. J. Ind. Ergon.* **1990**, *5*, 211–218. [CrossRef]
58. Landekić, M.; Bačić, M.; Bakarić, M.; Šporčić, M.; Pandur, Z. Working posture and the center of mass assessment while starting a chainsaw: A case study among forestry workers in Croatia. *Forests* **2023**, *14*, 395. [CrossRef]
59. Forkuo, G.O.; Borz, S.A. Development and evaluation of automated postural classification models in forest operations using deep learning-based computer vision. *SSRN Preprint* **2024**. [CrossRef]

60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
61. Klie, J.C.; Castilho, R.E.; Gurevych, I. Analyzing dataset annotation quality management in the wild. *Comput. Ling.* **2024**, *50*, 817–866. [CrossRef]
62. Yogarajan, V.; Dobbie, G.; Pistotti, T.; Bensemann, J.; Knowles, K. Challenges in annotating datasets to quantify bias in under-represented society. *arXiv* **2023**, arXiv:2309.08624.
63. Mascarenhas, S.; Agarwal, M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification. In Proceedings of the International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 19–21 November 2021; pp. 96–99. [CrossRef]
64. Siddharth, T. Fine-Tuning ResNet50 Pretrained on ImageNet for CIFAR-10. 2023. Available online: https://sidthoviti.com/fine-tuning-resnet50-pretrained-on-imagenet-for-cifar-10/ (accessed on 12 March 2025).
65. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1–9. [CrossRef]
66. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [CrossRef]
67. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856. [CrossRef]
68. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
69. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382. [CrossRef]
70. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
71. Viera, A.J.; Garrett, J.M. Understanding interobserver agreement: The kappa statistic. *Fam. Med.* **2005**, *37*, 360–363. [PubMed]
72. DeVellis, R.F. Inter-Rater Reliability. In *Encyclopedia of Social Measurement*; Kimberly, K.-L., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; pp. 317–322. ISBN 9780123693983. [CrossRef]
73. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 59–174. [CrossRef]
74. Sim, J.; Wright, C.C. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys. Ther.* **2005**, *85*, 257–268. [CrossRef]
75. Widyanti, A. Validity and inter-rater reliability of postural analysis among new raters. *Malays. J. Public Health Med.* **2020**, *1*, 161–166. [CrossRef]
76. Fleiss, J.L.; Levin, B.; Paik, M.C. The measurement of interrater agreement. *Stat. Methods Rates Proportions* **1981**, *2*, 22–23.
77. Gwet, K.L. *Handbook of Inter-Rater Reliability*, 4th ed.; Advanced Analytics LLC: Wayne, IN, USA, 2014; ISBN 978-0970806284.
78. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data mining toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
79. Borg, I.; Groenen, P.J. *Modern Multidimensional Scaling: Theory and Aplications*; Springer Science Bussines Media: Berlin/Heidelberg, Germany, 2007; ISBN 100387251502.
80. Saeed, N.; Nam, H.; Haq, M.I.U.; Muhammad Saqib, D.B. A survey on multidimensional scaling. *ACM Comput. Surv.* **2018**, *51*, 1–25. [CrossRef]
81. JetBrains s.r.o. PyCharm Community Edition: The IDE for Pure Python Development. 2025. Available online: https://www.jetbrains.com/pycharm/download/?section=windows (accessed on 3 March 2025).
82. Zaiontz, C. Real Statistics Using Excel. 2025. Available online: https://real-statistics.com/ (accessed on 5 March 2025).
83. Heinsalmi, P. Method to Measure Working Posture Loads at Working Sites (OWAS). In *Ergonomics of Working Postures*; CRC Press: Boca Raton, FL, USA, 1986; pp. 100–104. [CrossRef]
84. Liu, B.; Yu, L.; Che, C.; Lin, Q.; Hu, H.; Zhao, X. Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms. *arXiv* **2023**, arXiv:2312.12872. [CrossRef]
85. Lee, J.; Kim, T.Y.; Beak, S.; Moon, Y.; Jeong, J. Real-time pose estimation based on ResNet-50 for rapid safety prevention and accident detection for field workers. *Electronics* **2023**, *12*, 3513. [CrossRef]

86. Forkuo, G.O.; Borz, S.A.; Bilici, E. Approaching full accuracy by deep learning and computer vision in OWAS postural classification: An example on how computer generated body keypoints can improve deep learning based on conventional 2D data. *SSRN Preprint SSRN-5037016* **2024**. [CrossRef]
87. Eliasson, K.; Palm, P.; Nyman, T.; Forsman, M. Inter-and intra-observer reliability of risk assessment of repetitive work without an explicit method. *Appl. Ergon.* **2017**, *62*, 1–8. [CrossRef]