Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: www.tandfonline.com/journals/terg20

# Automatic estimation of Hand Activity Level from upper-limb trajectories: a probabilistic regression framework

Ting-Hung Lin, Yu Hen Hu & Robert Radwin

**To cite this article:** Ting-Hung Lin, Yu Hen Hu & Robert Radwin (14 Aug 2025): Automatic estimation of Hand Activity Level from upper-limb trajectories: a probabilistic regression framework, Ergonomics, DOI: 10.1080/00140139.2025.2543047

**To link to this article:** https://doi.org/10.1080/00140139.2025.2543047

ERGONOMICS
https://doi.org/10.1080/00140139.2025.2543047

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE



# Automatic estimation of Hand Activity Level from upper-limb trajectories: a probabilistic regression framework

Ting-Hung Lin[a], Yu Hen Hu[a] and Robert Radwin[b]

[a]Department of Electrical & Computer Engineering, University of Wisconsin-Madison, Madison, WI, USA; [b]Department of Industrial & Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA

**ABSTRACT**

Accurate measurement of Hand Activity Level (HAL) is crucial for evaluating musculoskeletal injury risk in repetitive hand-intensive work. Manual HAL assessments are often subjective and impractical for large-scale or continuous monitoring. This study presents a probabilistic regression framework that leverages video-based upper-limb pose trajectories to automatically estimate HAL scores while providing associated confidence measures. By enabling ergonomic risk assessment with quantified uncertainty, the proposed method delivers objective and reliable HAL predictions. Experimental results demonstrate strong in-domain performance (Root Mean Square Error [RMSE] = 0.24, Mean Absolute Error [MAE] = 0.17) and robust cross-domain generalisation (RMSE = 0.74, MAE = 0.54), highlighting both the accuracy and transferability of the framework.

**Practitioner Summary:** This work presents a probabilistic regression framework for estimating Hand Activity Level (HAL) from video-based pose trajectories, providing both HAL predictions and associated confidence. The framework generalises across domains, enabling practitioners to rapidly assess HAL scores while identifying low-confidence cases that may require further review in diverse workplace settings.

**ARTICLE HISTORY**
Received 4 May 2025
Accepted 21 July 2025

**KEYWORDS**
Hand Activity Level; probabilistic regression; computer vision; workplace safety

## 1. Introduction

Hand-intensive work is a major contributor to musculoskeletal disorders (MSDs), particularly affecting the wrists and forearms of workers engaged in repetitive motion tasks (Barr, Barbe, and Clark 2004). A key metric to evaluate the risks associated is the Hand Activity Level (HAL), which quantifies the repetitiveness and intensity of hand use by considering factors such as exertions and recovery time (Latko et al. 1997). Given its ability to measure exposure to repetitive hand exertions, HAL is widely used in ergonomic assessments to identify the risk of MSDs (Burt et al. 2013; Garg et al. 2012; Kapellusch et al. 2014; Yung et al. 2019).

Recognising its importance, the American Conference of Governmental Industrial Hygienists (ACGIH®) incorporated HAL into Threshold Limit Value (TLV®) guidelines to help prevent hand and wrist injuries in occupations that require intensive manual labour. Musculoskeletal disorders remain a significant occupational health concern due to their high prevalence and economic impact. According to the U.S. Bureau of Labour Statistics, there were 502,380 workplace musculoskeletal disorders resulting in at least one day away from work, with an annualised incidence rate of 25.3 per 10,000 full-time equivalent workers (Employer-Reported Workplace Injuries and Illnesses, 2021-2022 2023).

Accurately assessing and mitigating hand activity risks is essential for workplace safety. Traditional methods, which rely on subjective evaluations by expert ergonomists, are often inconsistent and labour-intensive. Previous research has documented inter- observer variability in manual HAL ratings (Spielholz et al. 2008) and demonstrated that observational methods generally offer less reliability than objective measurement approaches (Dahlgren et al. 2023). These findings underscore the importance of supplementing observational methods with objective measurements that enhance both the consistency and accuracy of HAL assessments.

To address these challenges, quantitative models for HAL estimation have been introduced. For example, Radwin et al. (2015) proposed a parameterised equation that estimates HAL based on the frequency of exertion F and the percent duty cycle D, as defined in Equation (1). Despite providing a standardised estimation method, this approach requires precise manual identification of exertions and cycle times, limiting its

**CONTACT** Robert Radwin rradwin@wisc.edu Department of Industrial & Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

applicability to large-scale assessments. Subsequent studies sought to refine this methodology by incorporating speed-based re-parameterization techniques that estimate HAL from motion-derived features, ultimately reducing dependence on manually counted exertion frequencies (Akkas et al. 2015). Nevertheless, despite these enhancements in objectivity, the requirement for frame-by-frame annotation to extract key input variables from video recordings remains a significant barrier to widespread implementation in ergonomic evaluations.

$$\mathrm{HAL} = 6.56 \cdot ln(D) \cdot \frac{F^{1.31}}{1+\left(3.18 \cdot F^{1.31}\right)} \quad (1)$$

Building on these foundations, prior efforts have explored the use of computer vision to facilitate HAL estimation. Early approaches, such as template-matching algorithms, were developed to track hand motion and estimate frequency and duty cycle (Chen et al. 2013). Subsequently, machine learning techniques that incorporate motion- derived features, such as decision trees and k-nearest neighbour classifiers, demonstrated potential for HAL prediction (Akkas et al. 2016). However, these approaches often required manual initialisation, such as selecting the region of interest at the start of the analysis or manually identifying the first cycle as a reference sample, limiting their ability to scale across diverse workplace settings. Thus, despite promising advances, an automated pipeline capable of handling variations in camera viewpoints, motion complexity and environmental conditions remains an unmet need.

To address these challenges, we introduce a fully automated probabilistic regression framework for estimating HAL directly from video. Our framework leverages upper- limb pose estimation and multi-object tracking to extract joint trajectories, which are processed to predict HAL and its associated confidence score. This framework enables accurate and reliable ergonomic assessment across diverse workplace scenarios.

## 2. Methods

We present a probabilistic regression framework for automatic estimation of HAL from video, which operates on upper-limb pose trajectories and predicts HAL scores with explicit uncertainty quantification. An overview of the framework is shown in Figure 1.

The framework consists of two main stages: preprocess and probabilistic regression. In the preprocessing stage, raw video frames are transformed into structured joint trajectories through upper-limb pose estimation, multi-object tracking and sliding- window segmentation, producing temporally consistent input sequences suitable for downstream modelling. In the probabilistic regression stage, each trajectory segment is temporally encoded and passed to a regression model that jointly predicts the mean and variance of a

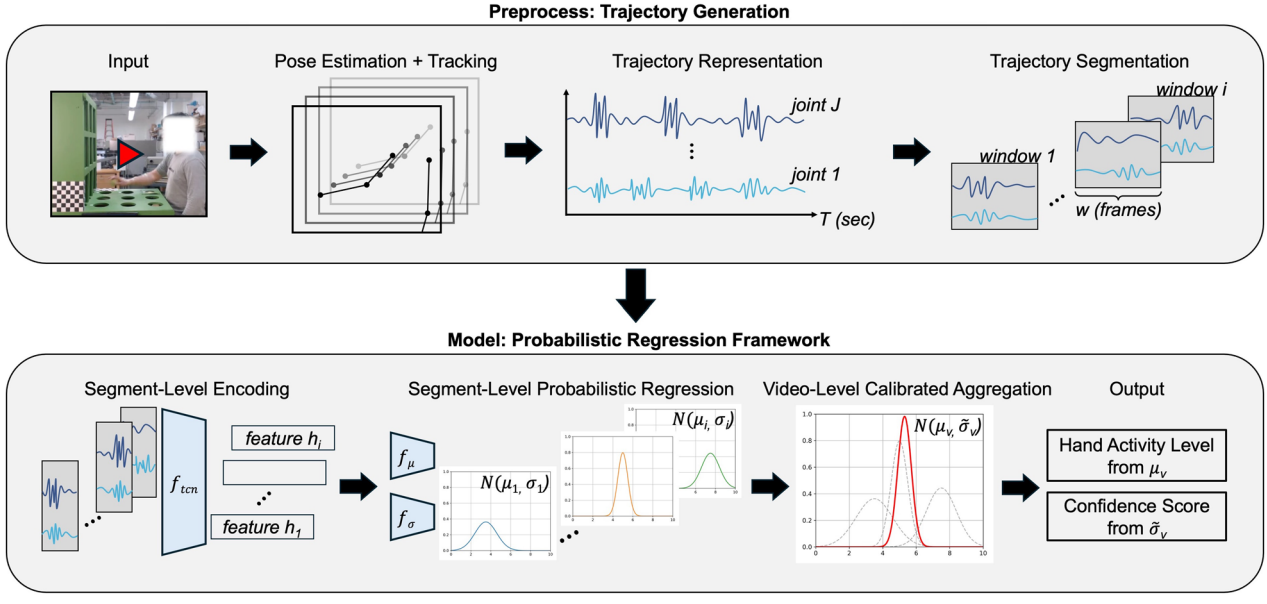


Figure 1. Overview of the proposed probabilistic regression framework for automatic HAL estimation. The framework consists of two main stages: a pre-process stage, which extracts overlapping windows of upper-limb pose trajectories from video and a probabilistic regression stage, in which each segment is temporally encoded, a Gaussian distribution over HAL is predicted per segment, and these distributions are fused into a final video-level HAL estimate with an associated confidence score.

Gaussian distribution over HAL, treating each segment as a probabilistic sample. Segment-level Gaussian predictions are then aggregated using a calibrated fusion scheme to produce a robust video-level HAL estimate, along with an interpretable confidence score.

Detailed descriptions of each component of the framework are provided in the following subsections.

### 2.1. Pre-process: trajectory generation

The preprocessing stage transforms raw video into structured joint trajectories through three main steps: pose estimation, tracking and trajectory segmentation. Pose estimation detects key joint positions in each frame, tracking maintains consistent joint identities across frames and segmentation divides the resulting trajectories into over- lapping temporal windows suitable for downstream modelling.

#### 2.1.1. Pose estimation

We use YOLOv10-Pose by Wang et al. (2024) for frame-wise upper-limb pose estimation, chosen for its balance of accuracy and efficiency. This model identifies 2D coordinates for key joints in each video frame. The modular design allows for future upgrades to more advanced pose estimators as needed.

#### 2.1.2. Tracking

To ensure temporal consistency, we apply BoT-SORT-ReID by Aharon, Orfaig, and Bobrovsky (2022) for multi-object tracking. This algorithm combines motion and appearance cues to reliably associate detected joints across frames, even in challenging or cluttered scenes.

#### 2.1.3. Trajectory segmentation

To leverage the repetitive nature of motions in HAL analysis, we segment the tracked trajectories into overlapping fixed-length windows. Each window is designed to span multiple motion cycles, ensuring that each segment captures representative local dynamics. This segmentation approach enables efficient batch processing of videos with variable durations while providing consistent input dimensions for model training. Let $X \in \mathbb{R}^{T\times J\times 2}$ represent the complete trajectory, where $T$ is the number of frames and $J$ is the number of tracked joints. Using a sliding window of length $w$ and stride $s$, we extract trajectory segments as follows:

$$\mathbf{X}_i = \mathbf{X}\left[s\cdot i : s\cdot i + w\right], \quad i = 0,\ldots,n_v - 1, \tag{2}$$

where $i$ indexes each extracted trajectory segment, and the number of segments for a video is given by $n_v = \left\lfloor \frac{T-w}{s} \right\rfloor$. Each trajectory segment is assigned the corresponding video-level HAL label $y_v$ under the assumption of piecewise stationarity. This assumption is well-suited to the repetitive, cyclical motions characteristic of HAL assessment tasks, where local segment dynamics are expected to reflect the overall activity. However, it is important to acknowledge that this assumption may be less accurate for segments near the beginning or end of a video, where transitional frames or incomplete motion cycles can introduce label noise. To mitigate this, we employ window sizes that span multiple motion cycles, thereby reducing the influence of non-representative segments. The effects of window length $w$ and stride $s$ on model performance are evaluated in Section 3.4.

### 2.2. Model: probabilistic regression

The original HAL rating scale is defined over discrete integer values from 0 to 10. However, the regression relationship between frequency, duty cycle and HAL in Equation (1) is continuous. This motivates framing HAL estimation as a regression problem, enabling finer-grained predictions beyond discrete categories. Such a continuous approach is valuable in practice, where task characteristics may not align with predefined HAL bins.

Given the variability in human motion and the inherent noise in video-derived pose data, deterministic models may yield overconfident or unreliable predictions. To mitigate this, we adopt a probabilistic framework that explicitly captures heteroscedastic uncertainty—uncertainty that varies across individual samples (Kendall and Gal 2017). This capability is particularly important in real-world environments, where the ambiguity of specific sequences may differ substantially across subjects and settings. By modelling uncertainty, the system generates confidence-aware predictions that are intended to be more robust and interpretable in diverse applications.

The probabilistic regression framework consists of four main steps: segment-level trajectory temporal encoding, segment-level probabilistic regression, video-level calibrated aggregation and interpretable confidence scoring. Temporal encoding extracts latent features from each trajectory segment, probabilistic regression predicts the mean and variance of a Gaussian distribution over HAL for each segment, calibrated aggregation combines segment-level predictions into a video-level estimate with associated

uncertainty, and confidence scoring maps the predictive distribution to an interpretable confidence metric.

#### 2.2.1. Segment-level trajectory temporal encoding

To capture temporal motion patterns within each trajectory segment, we employ a temporal convolutional network (TCN) as the temporal feature extraction backbone (Lea et al. 2016). TCNs leverage dilated convolutions to efficiently model dependencies in sequential data while enabling parallel computation and stable gradient propagation, making them well-suited for temporal sequence modelling.

Each trajectory segment $\mathbf{X}_i$ is transformed into a latent feature vector $\boldsymbol{h}_i$ of dimension $d$ using the temporal encoding function $f_{\text{tcn}}(\cdot)$:

$$\boldsymbol{h}_i = f_{\text{tcn}}(\cdot), \quad \boldsymbol{h}_i \in \mathbb{R}^d. \tag{3}$$

#### 2.2.2. Segment-Level probabilistic regression

We model each segment-level HAL label $y_i$ as a draw from a Gaussian distribution whose parameters are predicted from the input segment, following the approach introduced by Nix and Weigend (1994):

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), \tag{4}$$

where both the mean $\mu_i$ and variance $\sigma_i^2$ are predicted from the latent feature vector $\boldsymbol{h}_i$ extracted from the input segment $\mathbf{X}_i$.

Specifically, the mean and variance are computed as:

$$\mu_i = f_\mu(\boldsymbol{h}_i), \tag{5}$$

$$\sigma_i^2 = f_\sigma(\boldsymbol{h}_i), \tag{6}$$

where $f_\mu(\cdot)$ and $f_\sigma(\cdot)$ are independent fully connected neural network heads applied to the shared latent feature $\boldsymbol{h}_i$. The variance head is constrained to produce non- negative outputs, and the mean head is restricted to the valid HAL range. Additional architectural details are provided in Section 2.3.4.

The model is trained by minimising the average negative log-likelihood (NLL) over $n$ training segments:

$$\mathcal{L}_{NLL} = \frac{1}{n}\sum_{i=0}^{n-1}\left[\frac{(y_i-\mu_i)^2}{2\sigma_i^2} + \frac{1}{2}\log\sigma_i^2\right] + \frac{1}{2}\log(2\pi) \tag{7}$$

Minimising $\mathcal{L}_{NLL}$ jointly updates the TCN backbone and the $\mu-$ and $\sigma-$ heads so that $\mu_i$ best fits the observed HAL scores and $\sigma_i^2$ accurately quantifies the model's predictive uncertainty.

A larger $\sigma_i^2$ indicates greater uncertainty—ie the model considers a wider range of HAL values plausible—whereas a smaller $\sigma_i^2$ reflects higher confidence in the point estimate. No explicit uncertainty labels are required: the probabilistic formulation enables the network to learn both accurate predictions and well-calibrated uncertainty estimates directly from the data.

#### 2.2.3. Video-level calibrated aggregation

To obtain a robust video-level HAL estimate from the $n_v$ segment-level predictions, we aggregate these predictions while explicitly accounting for their associated uncertainties. A naïve approach, such as simple averaging, would treat all segments equally, disregarding differences in prediction confidence. Instead, we employ a precision-weighted aggregation scheme, which assigns greater influence on segments with lower predicted variance, thereby prioritising more reliable predictions.

For each segment $i$, precision is defined as $\tau_i = 1/\sigma_i^2$. The aggregated video-level mean and variance are computed as:

$$\mu_v = \frac{\sum_{i=1}^{n_v}\tau_i\mu_i}{\sum_{i=1}^{n_v}\tau_i} \tag{8}$$

$$\sigma_v^2 = \frac{1}{\sum_{i=1}^{n_v}\tau_i} \tag{9}$$

While precision-weighted aggregation is theoretically optimal for independent predictions, our use of overlapping windows introduces statistical dependencies that cause $\sigma_v^2$ to underestimate the true video-level uncertainty. To correct for this, we introduce a temperature scaling parameter $\lambda \geq 1$ to calibrate the aggregated variance:

$$\tilde{\sigma}_v^2 = \lambda^2\sigma_v^2 \tag{10}$$

With the aggregated mean $\mu_v$ and the calibrated variance $\tilde{\sigma}_v^2$, the final video-level predictive distribution is defined as a Gaussian:

$$y_v \sim \mathcal{N}\left(\mu_v, \tilde{\sigma}_v^2\right). \tag{11}$$

This calibrated distribution provides a statistically robust foundation for the interpretable confidence score detailed in the following section. A quantitative evaluation of this calibration is presented in Section 3.2.

#### 2.2.4. Interpretable confidence score mapping

To enhance the practical utility of the model's output, we map the video-level predictive distribution from Equation (11) to an interpretable confidence score. This score quantifies the probability that a prediction is sufficiently accurate, which we define as the true HAL value lying within a tolerance of $\epsilon = 1$ HAL unit from the predicted mean. This tolerance is consistent with established ergonomic literature, where deviations within ±1 unit are considered acceptable (Radwin et al. 2023).

The probability of the true HAL falling within this tolerance is derived from the standard normal cumulative distribution function (CDF), denoted by $\Phi(\cdot)$, as follow:

$$\begin{aligned}\Pr\left(|y_v-\mu_v|\le 1\right) &= \Pr\left(\mu_v-1\le y_v\le \mu_v+1\right)\\ &= \Pr\left(\frac{-1}{\tilde{\sigma}_v}\le\frac{y_v-\mu_v}{\tilde{\sigma}_v}\le\frac{1}{\tilde{\sigma}_v}\right)\\ &= \Phi\left(\frac{1}{\tilde{\sigma}_v}\right)\ \ \Phi\left(-\frac{1}{\tilde{\sigma}_v}\right)\\ &= \Phi\left(\frac{1}{\tilde{\sigma}_v}\right)-\left[1-\Phi\left(\frac{1}{\tilde{\sigma}_v}\right)\right]\\ &= 2\Phi\left(\frac{1}{\tilde{\sigma}_v}\right)-1\end{aligned} \tag{12}$$

Finally, to provide an intuitive metric for practitioners, we express this probability as a percentage-based confidence score:

$$\text{Confidence Score}\,(\%) = \left[2\Phi\left(\frac{1}{\tilde{\sigma}_v}\right)-1\right]\times 100. \tag{13}$$

This score provides a statistically grounded, interpretable measure of the model's certainty that its prediction is within an acceptable range, supporting reliable decision- making in ergonomic assessments.

### 2.3. Experimental setup and evaluation

This section provides a detailed overview of the experimental setup, including datasets, data augmentation strategies, model configuration and evaluation metrics.

#### 2.3.1. Datasets

We evaluate our approach using two video datasets: Dataset A and Dataset B. Both datasets involve cyclical load transfer tasks—including reaching, grasping, moving and releasing objects (see Figure 2). Ground truth HAL labels are obtained by annotating each video frame as either exertion or non-exertion. From these annotations, we compute the average cycle time, average exertion time and the number of exertions per video. Duty cycle, frequency and HAL are then derived according to the definitions provided in Radwin et al. (2015). The datasets are described as follows:

*Dataset A.* Based on Radwin (2011), Dataset A consists of 144 videos from 20 subjects performing a repetitive bottle loading task on a rotating table. HAL scores are summarised as Mean ± SD [Min-Max]: 1.6 ± 1.5 [0.1-5.1]. Video durations range from 70 to 230 seconds at 30 fps. This dataset is used exclusively for in-domain training and evaluation. We reserve 20% of the subjects (4 subjects, 24 videos) as a held-out test set, while the remaining 80% are used for training and hyperparameter tuning via 5-fold group-by-subject cross-validation, ensuring that

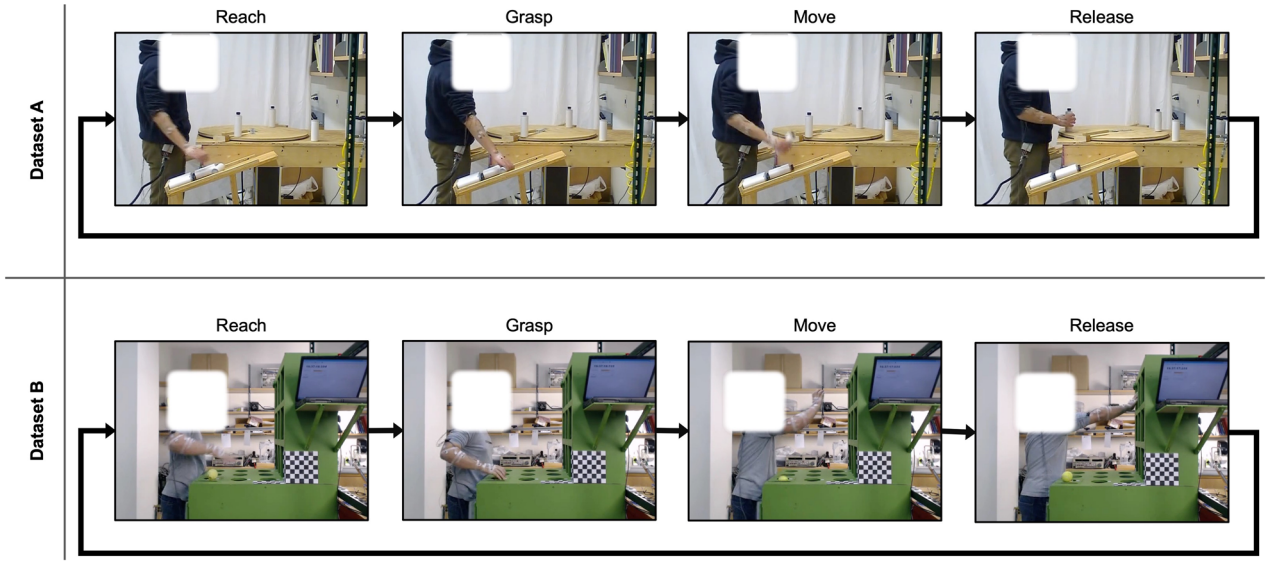


Figure 2. Representative frames from Datasets A and B illustrating the four-phase load transfer cycle. Dataset A (top row) features a bottle loading task, while Dataset B (bottom row) involves a shelf system with tennis balls. Each dataset captures the cyclical nature of the tasks, which includes the phases of Reach, Grasp, Move and Release.

all videos from each subject appear in only one split to prevent data leakage.

*Dataset B.* Based on Akkas et al. (2016), Dataset B contains 122 videos from 19 subjects performing similar load transfer tasks by repetitively moving tennis balls within a shelf system. HAL scores are summarised as Mean ± SD [Min-Max]: 4.6 ± 1.1 [1.1-6.0]. Video durations range from 20 to 120 seconds at 30 fps. All videos in Dataset B are reserved for cross-domain testing to assess generalisation.

### 2.3.2. Data augmentation

To address limitations in the video datasets—such as fixed camera positions, a pre- dominance of right-handed subjects and a limited range of HAL labels—we apply both temporal and spatial augmentation techniques to enrich the data and improve generalisability.

To increase representation of higher HAL values, we perform selective temporal augmentation by downsampling trajectory data during non-exertion periods. This approach preserves the integrity of exertion phases while reducing the overall cycle du- ration, thereby increasing the duty cycle and frequency. As a result, augmented samples tend to exhibit higher HAL values. Each video undergoes three randomised temporal augmentations, diversifying Dataset A to 576 videos with HAL labels summarised as Mean ± SD [Min-Max]: 3.7 ± 2.0 [0.1-6.4].

To mitigate biases from fixed viewpoints and handedness, we apply spatial augmentations to the pose trajectories. These include random rotations, random scaling and horizontal mirroring to simulate left-handed movements and varied camera perspectives.

### 2.3.3. Evaluation metrics

We evaluate the model using both point-estimate accuracy metrics and interval-based uncertainty calibration metrics. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) quantify the regression accuracy; RMSE penalises large errors more heavily, while MAE represents the average magnitude of error. The Expected Calibration Error (ECE) measures the discrepancy between nominal and empirical coverage probabilities, following the regression calibration procedure proposed by Kuleshov, Fenner, and Ermon (2018).

Let $m$ denote the number of test videos, $y_v^{(i)}$ the ground-truth HAL score for video $i$ and $\mu_v^{(i)}$ the predicted video-level mean from Equation (8).

*Root Mean Square Error (RMSE).*

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_v^{(i)} - \mu_v^{(i)}\right)^2} \tag{14}$$

*Mean Absolute Error (MAE).*

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_v^{(i)} - \mu_v^{(i)}\right| \tag{15}$$

#### Expected calibration Error (ECE)

$$\mathrm{ECE} = \frac{1}{K}\sum_{k=1}^{K}\left|\hat{c}(\gamma_k) - \gamma_k\right| \tag{16}$$

The ECE quantifies the average absolute difference between nominal coverage probabilities, $\gamma_k$ and their corresponding empirical coverage estimates, $\hat{c}(\gamma_k)$. These are evaluated over a set of $K = 19$ probability levels, where $\gamma_k \in \{0.05, 0.1, \ldots, 0.95\}$. The empirical coverage is defined as:

$$\hat{c}(\gamma_k) = \frac{1}{m}\sum_{i=1}^{m} 1\left\{ y_v^{(i)} \in \left[ \mu_v^{(i)} - z_{\left(\frac{1+\gamma_k}{2}\right)}\tilde{\sigma}_v^{(i)}, \mu_v^{(i)} + z_{\left(\frac{1+\gamma_k}{2}\right)}\tilde{\sigma}_v^{(i)} \right] \right\}, \tag{17}$$

where $\tilde{\sigma}_v^{(i)}$ is the calibrated standard deviation from Equation (10) and $z_{\left(\frac{1+\gamma_k}{2}\right)}$ is the standard-normal critical value defining a central $\gamma_k$ two-sided interval.

### 2.3.4. Model configuration and hyperparameters

The final model configuration was determined through 5-fold cross-validation on Dataset A, with hyperparameters selected to optimise the evaluation metrics defined in Section 2.3.3. The final configuration is as follows:

- **Trajectory Segmentation:** A sliding window with size $w = 450$ frames and stride $s = 45$ was used, as defined in Equation (2).
- **Input:** Each segment is a $w \times 4$ tensor representing the x/y offsets from the shoulder to the elbow and wrist joints. Uniform undersampling was applied to balance the HAL label distribution during training.
- **TCN Backbone:** Three 1D convolutional layers with channels [16, 64, 256], a kernel size of 8 and a dilation factor of 2 were used to produce a $d = 256$ latent feature vector, as described in Equation (3).
- **Prediction Heads**: Two parallel fully connected heads, each with a hidden size of 8, predict the segment-level distribution parameters. The mean head ($\mu_i$) uses a scaled sigmoid activation function to constrain its output between 0 and 10. The variance head predicts the log-variance

($\log \sigma_i^2$), which is converted to $\sigma_i^2$ via an exponential function to ensure positivity.

- **Activation and Regularisation:** A ReLU activation function was used after each layer, followed by a dropout rate of 0.2.
- **Optimisation:** The model was trained using the Adam optimiser with a learning rate of $1\times10^{-3}$ and a batch size of 256.
- **Uncertainty Calibration:** Temperature scaling ($\lambda = 4$) was used for variance calibration in Equation (10).

The final model comprises approximately 0.75 million trainable parameters, resulting in a model size of about 3 MB.

## 3. Results

This section presents a comprehensive evaluation of the proposed framework's performance, calibration and robustness. We first report primary regression metrics for both in-domain (Dataset A) and cross-domain (Dataset B) test sets, providing quantitative comparisons against baseline methods. To complement these metrics, we provide qualitative visualisations where each prediction is plotted against its ground truth value and colour-coded by its associated confidence score. We then analyse the statistical reliability of the model through a detailed model calibration assessment. Finally, to validate our methodological choices and assess their impact, we conduct an ablation study on core model components and a sensitivity analysis on key temporal parameters.

Table 1. Video-level HAL estimation accuracy for in-domain and cross-domain testing.

| | In-domain | | Cross-domain | |
|---|---|---|---|---|
| Method | RMSE↓ | MAE↓ | RMSE↓ | MAE↓ |
| Chen et al. (2013) | 1.80 | 1.36 | 2.42 | 2.33 |
| Akkas et al. (2016) | 1.63 | 1.21 | 1.87 | 1.73 |
| This work | 0.24 | 0.17 | 0.74 | 0.54 |

### 3.1. Model performance

We evaluated the proposed framework on in-domain and cross-domain test sets, as described in Section 2.3.1. The video-level HAL estimation results are summarised in Table 1, which compares our method with two baseline approaches: Chen et al. (2013) and Akkas et al. (2016). Our method achieved substantially lower RMSE and MAE across both datasets, indicating strong in-domain accuracy and robust cross-domain generalisation.

The relationship between predicted and true HAL values for both in-domain and cross-domain test sets is visualised in Figure 3. In the in-domain scenario, predictions are tightly clustered along the diagonal, indicating strong agreement with ground truth. In contrast, cross-domain predictions display increased dispersion, highlighting the challenges associated with domain shift. Each point is colour-coded by the interpretable confidence score described in Equation (13), offering insight into the model's uncertainty for individual predictions.

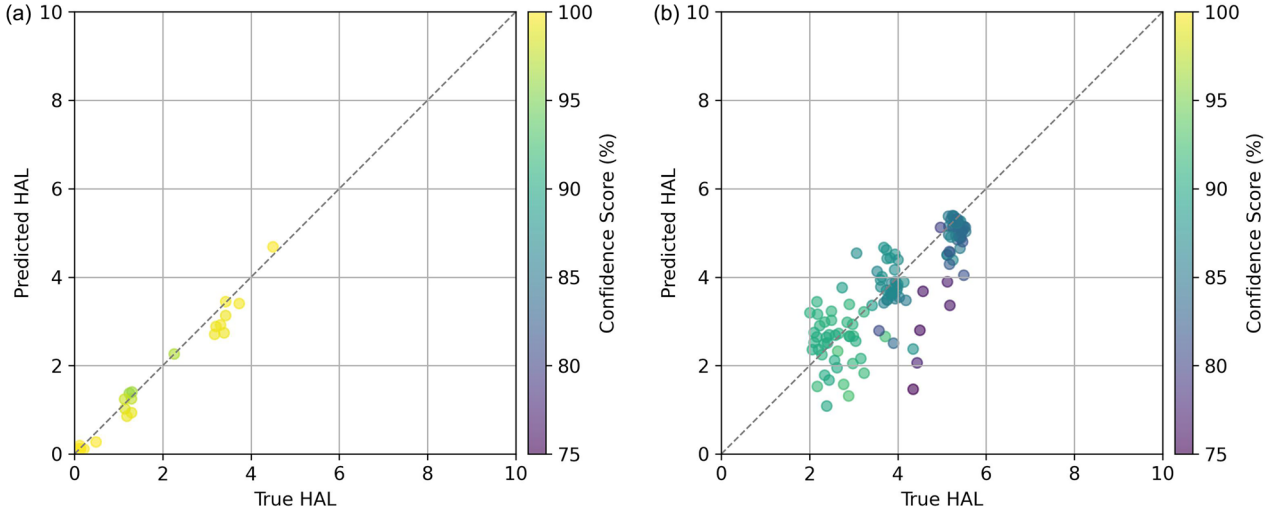


Figure 3. Prediction accuracy for in-domain (a) and cross-domain (b) testing. Predicted HAL values are plotted against ground truth, with the diagonal line indicating perfect agreement. Points are colour-coded by confidence score. In (a), predictions cluster tightly along the diagonal, indicating high in-domain accuracy. In (b), increased dispersion reflects the effect of domain shift.

### 3.2. Model calibration

We quantitatively assessed model calibration using the ECE metric, which measures the alignment between nominal and empirical coverage. Empirical coverage, as defined in Equation (17), was computed as the proportion of samples whose true HAL values fell within the predictive intervals constructed from the calibrated video-level variance $\tilde{\sigma}_v^2$. For comparison, we also report results using predictive intervals based on the uncalibrated video-level variance $\sigma_v^2$. Lower ECE values indicate improved calibration and more reliable uncertainty quantification.

For in-domain evaluation, ECE decreased from 0.31 (uncalibrated) to 0.04 (calibrated). In the cross-domain setting, ECE decreased from 0.36 (uncalibrated) to 0.03 (calibrated). These results demonstrate that calibration substantially improves the reliability of uncertainty estimates in both testing domains.

Calibration performance for three types of predictive intervals is compared in Figure 4: segment-level ($\sigma_l$), uncalibrated video-level ($\sigma_v$) and calibrated video-level ($\tilde{\sigma}_v$). The segment-level predictions, produced by joint training with NLL, are reasonably well-calibrated but show some underconfidence in the mid-range probabilities, especially in the in-domain case. Aggregating these predictions to the video level introduces systematic overconfidence, as indicated by the uncalibrated video-level curve falling below the diagonal. Applying temperature scaling as described in Section 2.2.3 to obtain calibrated video-level predictions ($\tilde{\sigma}_v$) corrects this overconfidence and aligns the reliability curve more closely with the ideal diagonal, confirming that calibration effectively addresses the miscalibration introduced during aggregation.

### 3.3. Ablation study

We conducted an ablation study to evaluate the contribution of two key components in the proposed framework: (1) *probabilistic regression*, which provides input-dependent uncertainty estimates via Gaussian distributions and (2) *precision-weighted aggregation*, which incorporates predicted uncertainty when combining segment-level predictions into a video-level estimate. These were compared against a deterministic baseline that produces point estimates without uncertainty modelling, trained using mean squared error loss and aggregated with simple mean. Incorporating probabilistic regression leads to improved performance over the deterministic baseline, and further gains are achieved with precision-weighted aggregation. The combination of both approaches yields the lowest prediction error on the in-domain dataset, as shown in Table 2.

### 3.4. Sensitivity analysis

We performed a sensitivity analysis on the temporal segmentation parameters by varying the window size ($w$)

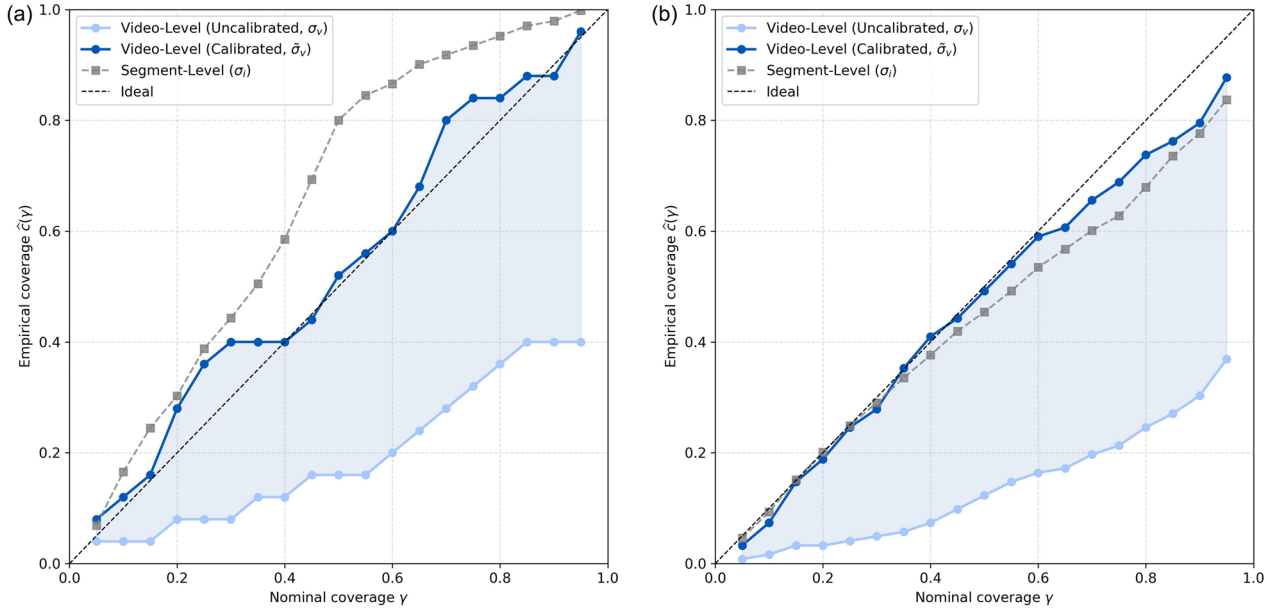


Figure 4. Reliability diagrams for in-domain (a) and cross-domain (b) testing. Empirical coverage $\hat{c}(\gamma_k)$ is plotted against nominal coverage ($\gamma$), with the diagonal indicating perfect calibration. Calibrated video-level predictions closely follow the diagonal, demonstrating improved reliability over the uncalibrated baseline. Shaded areas highlight the reduction in miscalibration achieved by temperature scaling.

Table 2. Ablation study evaluating the effects of probabilistic regression and precision- weighted aggregation on in-domain performance.

| | Model | Aggregation | RMSE | MAE |
|---|---|---|---|---|
| (1) | Deterministic | Mean | 0.25 | 0.18 |
| (2) | Probabilistic | Mean | 0.18 | 0.13 |
| (3) | Probabilistic | Precision-weighted | **0.13** | **0.09** |

Bolded values are the best performing model.

Table 3. Sensitivity analysis on window size *w* on in-domain performance.

| Window size (*w*) | RMSE | MAE |
|---|---|---|
| 150 | 0.36 | 0.34 |
| 300 | 0.28 | 0.25 |
| 450 | **0.20** | **0.15** |
| 600 | 0.35 | 0.30 |

Bolded values are the best performing model.

Table 4. Sensitivity analysis on window stride *s* on in-domain performance.

| Stride (*s*) | RMSE | MAE |
|---|---|---|
| 15 | **0.19** | **0.13** |
| 45 | 0.20 | 0.15 |
| 90 | 0.71 | 0.61 |
| 150 | 0.74 | 0.54 |
| 300 | 0.88 | 0.75 |
| 450 | 1.00 | 0.67 |

Bolded values are the best performing model.

and stride (*s*), while keeping all other settings fixed as described in Section 2.3.4 Model Configuration and Hyperparameters. Model performance was evaluated on the held-out test set of in-domain Dataset A. As summarised in Table 3 and Table 4, increasing the window size generally improves performance by providing more temporal context, though excessively large windows may reduce accuracy. Similarly, smaller strides, which increase overlap between segments and the number of training samples, tend to yield lower prediction error, highlighting the importance of balancing temporal context and sample diversity for optimal model accuracy.

## 4. Discussion

The goal of this study was to develop an objective and scalable method for estimating HAL directly from video, using pose trajectories as an intermediate representation. Our probabilistic HAL regression framework integrates pose estimation, multi-object tracking, temporal encoding and probabilistic regression to estimate HAL values while providing associated confidence measures. The model demonstrated strong performance in both in-domain and cross-domain evaluations, signifying substantial progress towards automated HAL estimation from video.

Within the broader research context, previous studies have consistently highlighted issues related to subjective variability and limited scalability in manual HAL assessments (Dahlgren et al. 2023; Spielholz et al. 2008). Our work builds upon and extends prior computer vision-based approaches, many of which depended on manual annotations or predefined regions of interest (Akkas et al. 2015, 2016; Chen et al. 2013). By preprocessing video into pose trajectories and integrating probabilistic regression with explicit uncertainty quantification, our framework not only achieves accurate HAL estimation but also enables uncertainty-informed ergonomic decision-making. This approach addresses critical gaps in existing methodologies and represents a significant step towards objective and reliable ergonomic assessment.

Beyond accuracy, the proposed framework is computationally efficient and well suited for real-world deployment. The HAL regression model operates with low latency and minimal memory requirements, with average inference time per video comparable to the time required for processing a single video frame during pose estimation. Furthermore, the modular design of the framework ensures that the HAL regressor remains compatible with future improvements in pose estimation models without requiring retraining, promoting scalability and long-term adaptability in practical applications.

The cross-domain evaluation revealed a modest decline in performance compared to in-domain testing. Although both datasets involve cyclical load-transfer tasks, they differ in several subtle but influential ways, including camera perspective, workspace geometry, object mass and movement tempo. These differences alter the model's input joint-trajectory patterns, likely contributing to the observed decrease in cross- domain performance. Nevertheless, the resulting errors remain well within the ±1 HAL tolerance commonly accepted as functionally equivalent to ground truth (Radwin et al. 2023), suggesting that the practical impact of the domain shift is limited. This level of generalisation is encouraging, given that some performance degradation is expected when transferring between datasets with differing task dynamics.

While these results demonstrate the feasibility and potential utility of the proposed framework for automated HAL estimation, several challenges remain. These limitations can be broadly categorised into issues related to data, modelling and interpretability.

### 4.1. Data limitations

Our evaluation was conducted using datasets collected in controlled laboratory environments, where tasks were highly uniform, camera perspectives were fixed and

movements followed constrained, repeatable patterns. These conditions inherently limit the generalisability of our findings to real-world industrial settings, which typically involve substantially greater variability in task execution, workspace configurations, lighting, occlusions and worker behaviour. Although cross-domain testing provided some evidence of model transferability, it does not fully capture the complexity and unpredictability of field environments. Moreover, the ground truth HAL values in our datasets did not cover the full 0 to 10 range. Even with temporal data augmentation, training examples for HAL values above 6.5 remained scarce, potentially limiting the model's reliability in high-load scenarios. Another limitation arises from the label assignment process. We assumed piecewise stationarity, meaning that local dynamics within each trajectory segment adequately represent the overall video-level HAL value. However, this assumption may not consistently hold, particularly near the start and end of recordings when transitional frames or incomplete motion cycles occur. These segments may not accurately reflect the assigned label, introducing label noise during training. While our precision-weighted aggregation scheme helps mitigate this by down-weighting low-confidence segments in the final video-level prediction, the impact of label noise remains a concern.

### 4.2. Modelling limitations

The current framework employs a multi-stage pipeline comprising pose estimation, multi-object tracking and HAL regression. While this modular structure facilitates the integration of specialised components at each stage, it also introduces the potential for error propagation across the pipeline. For example, inaccuracies in pose estimation can cascade through tracking and ultimately impact the HAL regression output. Additionally, although TCNs were selected for their faster training, improved stability and empirically superior performance over recurrent models in sequence tasks (Bai, Kolter, and Koltun 2018), their ability to capture long-range temporal dependencies is limited compared to more recent attention-based architectures, such as transformers, which have demonstrated state-of-the-art performance by effectively modelling global temporal structures.

### 4.3. Interpretability limitations

In our framework, HAL is directly predicted from pose trajectories, inherently disconnecting the output from the underlying biomechanical and kinematic features—such as hand speed, force and posture—that theoretically define the HAL construct. Although the model jointly predicts HAL and its associated uncertainty, and we have calibrated these uncertainty estimates to ensure their statistical reliability, the uncertainty scores do not provide direct insight into the specific motion characteristics—such as occlusions, rapid joint movements, or postural deviations—that contribute to prediction uncertainty or potential errors. This lack of interpretability may constrain practitioners' ability to diagnose prediction failures or tailor ergonomic interventions based on the model's output.

Future work should prioritise large-scale field validation and the expansion of datasets to include higher HAL values and greater task variability. Incorporating probabilistic label assignment at the segment level or transitioning to frame-level supervision may help mitigate label noise. From a modelling perspective, exploring transformer-based or hybrid architectures could further enhance the framework's temporal modelling capabilities. Finally, improving interpretability by explicitly linking HAL predictions and uncertainty estimates to specific kinematic features could provide actionable insights for ergonomic assessment and intervention.

## 5. Conclusion

In this work, we presented a probabilistic regression framework for estimating HAL from video-derived upper-limb trajectories. By combining advanced pose estimation, object tracking, temporal encoding and probabilistic modelling, our approach offers a scalable and automated alternative to manual ergonomic assessments. The method achieved accurate HAL predictions with meaningful confidence estimates, demonstrating its potential for reliable vision-based ergonomic risk assessment. Future work should validate the framework in diverse real-world settings and enhance model interpretability to support broader adoption.

## Author contributions

All listed authors meet the ICMJE criteria for authorship. Ting-Hung Lin, Yu Hen Hu and Robert Radwin made substantial contributions to the conception and design of the study. Ting-Hung Lin performed data analysis. Ting-Hung Lin, Yu Hen Hu and Robert Radwin drafted the manuscript. Yu Hen Hu and Robert Radwin critically revised the manuscript for important intellectual content. All authors approved the final version to be published and agree to be accountable for all aspects of the work.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This publication was supported in part by the Grant Number, T42 OH008455, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services. This research is also sponsored, in part by ErgoFactor, LLC.

## Data availability statement

The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research supporting data is not available.

## References

Aharon, N., R. Orfaig, and B.-Z. Bobrovsky. 2022. "BoT-SORT: Robust Associations Multi-Pedestrian Tracking." doi:10.48550/arXiv.2206.14651.

Akkas, O., D. P. Azari, C.-H. E. Chen, Y. H. Hu, S. S. Ulin, T. J. Armstrong, D. Rempel, and R. G. Radwin. 2015. "A Hand Speed and Duty Cycle Equation for Estimating the ACGIH Hand Activity Level Rating." *Ergonomics* 58 (2): 184–194. doi:10.1080/00140139.2014.966155.

Akkas, O., C.-H. Lee, Y. H. Hu, T. Y. Yen, and R. G. Radwin. 2016. "Measuring Elemental Time and Duty Cycle Using Automated Video Processing." *Ergonomics* 59 (11): 1514–1525. doi:10.1080/00140139.2016.1146347.

Bai, S., J. Z. Kolter, and V. Koltun. 2018. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling." doi:10.48550/arXiv.1803.01271.

Barr, A. E., M. F. Barbe, and B. D. Clark. 2004. "Work-Related Musculoskeletal Disorders of the Hand and Wrist: Epidemiology, Pathophysiology, and Sensorimotor Changes." *The Journal of Orthopaedic and Sports Physical Therapy* 34 (10): 610–627. doi:10.2519/jospt.2004.34.10.610.

Burt, S., J. A. Deddens, K. Crombie, Y. Jin, S. Wurzelbacher, and J. Ramsey. 2013. "A Prospective Study of Carpal Tunnel Syndrome: Workplace and Individual Risk Factors." *Occupational and Environmental Medicine* 70 (8): 568–574. doi:10.1136/oemed-2012-101287.

Chen, C.-H., Y. H. Hu, T. Y. Yen, and R. G. Radwin. 2013. "Automated Video Exposure Assessment of Repetitive Hand Activity Level for a Load Transfer Task." *Human Factors* 55 (2): 298–308. doi:10.1177/0018720812458121.

Dahlgren, G., P. Liv, F. Öhberg, L. Slunga Järvholm, M. Forsman, and B. Rehn. 2023. "Correlations between Ratings and Technical Measurements in Hand-Intensive Work." *Bioengineering* 10 (7): 867. doi:10.3390/bioengineering10070867.

Garg, A., J. Kapellusch, K. Hegmann, J. Wertsch, A. Merryweather, G. Deckow-Schaefer, and E. Malloy. 2012. "The Strain Index (SI) and Threshold Limit Value (TLV) for Hand Activity Level (HAL): Risk of Carpal Tunnel Syndrome (CTS) in a Prospective Cohort." *Ergonomics* 55 (4): 396–414. doi:10.1080/00140139.2011.644328.

Kapellusch, J. M., F. E. Gerr, E. J. Malloy, A. Garg, C. Harris-Adamson, S. S. Bao, S. E. Burt, A. M. Dale, E. A. Eisen, B. A. Evanoff, K. T. Hegmann, B. A. Silverstein, M. S. Theise, and D. M. Rempel. 2014. "Exposure–Response Relationships for the ACGIH Threshold Limit Value for Hand-Activity Level: Results from a Pooled Data Study of Carpal Tunnel Syndrome." *Scandinavian Journal of Work, Environment & Health* 40 (6): 610–620. doi:10.5271/sjweh.3456.

Kendall, A., and Y. Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

Kuleshov, V., N. Fenner, and S. Ermon. 2018. "Accurate Uncertainties for Deep Learning Using Calibrated Regression." In *International conference on machine learning* 2796–2804. PMLR. doi:10.48550/arXiv.1807.00263.

Latko, W. A., T. J. Armstrong, J. A. Foulke, G. D. Herrin, R. A. Rabourn, and S. S. Ulin. 1997. "Development and Evaluation of an Observational Method for Assessing Repetition in Hand Tasks." *American Industrial Hygiene Association Journal* 58 (4): 278–285. doi:10.1080/15428119791012793.

Lea, C., R. Vidal, A. Reiter, and G. D. Hager. 2016. "Temporal Convolutional Networks: A Unified Approach to Action Segmentation." In *European conference on computer vision* 47–54. Cham: Springer International Publishing. doi:10.48550/arXiv.1608.08242.

Nix, D. A., and A. S. Weigend. 1994. "Estimating the Mean and Variance of the Target Probability Distribution." *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, *1*, 55–60. doi:10.1109/ICNN.1994.374138.

Radwin, R. G. 2011. "Automated Video Exposure Assessment of Repetitive Motion." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 995–996. doi:10.1177/1071181311551207.

Radwin, R. G., D. P. Azari, M. J. Lindstrom, S. S. Ulin, T. J. Armstrong, and D. Rempel. 2015. "A Frequency-Duty Cycle Equation for the ACGIH Hand Activity Level." *Ergonomics* 58 (2): 173–183. doi:10.1080/00140139.2014.966154.

Radwin, R. G., Y. H. Hu, O. Akkas, S. Bao, C. Harris-Adamson, J.-H. Lin, A. R. Meyers, and D. Rempel. 2023. "Comparison of the Observer, Single-Frame Video and Computer Vision Hand Activity Levels." *Ergonomics* 66 (8): 1132–1141. doi:10.1080/00140139.2022.2136407.

Spielholz, P., S. Bao, N. Howard, B. Silverstein, J. Fan, C. Smith, and C. Salazar. 2008. "Reliability and Validity Assessment of the Hand Activity Level Threshold Limit Value and Strain Index Using Expert Ratings of Mono-Task Jobs." *Journal of Occupational and Environmental Hygiene* 5 (4): 250–257. doi:10.1080/15459620801922211.

U.S. Bureau of Labor Statistics. 2023. "Employer-Reported Workplace Injuries and Illnesses, 2021-2022." U.S. Bureau of Labor Statistics. Accessed 4/21/25. https://www.bls.gov/news.release/archives/osh_11082023.pdf

Wang, A., H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. 2024. "YOLOv10: Real-Time End-to-End Object Detection." *Advances in Neural Information Processing Systems*, 37: 107984–108011. doi:10.48550/arXiv.2405.14458.

Yung, M., A. M. Dale, J. Kapellusch, S. Bao, C. Harris-Adamson, A. R. Meyers, K. T. Hegmann, D. Rempel, and B. A. Evanoff. 2019. "Modeling the Effect of the 2018 Revised ACGIH® Hand Activity Threshold Limit Value® (TLV) at Reducing Risk for Carpal Tunnel Syndrome." *Journal of Occupational and Environmental Hygiene* 16 (9): 628–633. doi:10.1080/15459624.2019.1640366.