# Information-Theoretic Uncertainty Quantification for Vision-Based Construction Ergonomic Assessment

**Roy Lan[1] and Ibukun Awolusi[2]**

[1]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio, USA
[2]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio, USA
roy.lan@my.utsa.edu, ibukun.awolusi@utsa.edu

**Abstract –**

**Work-related musculoskeletal disorders (WMSDs) remain a critical cost driver in construction safety. While computer vision enables scalable ergonomic monitoring via automated REBA assessment, current deterministic systems suffer from hazardous overconfidence, providing definitive risk scores even when occlusion or lighting compromises data integrity. This study presents an information-theoretic framework that decomposes prediction uncertainty into aleatoric (sensor noise) and epistemic (model ignorance) components. Aleatoric uncertainty is quantified via surprisal-based confidence mapping from pose keypoints, while epistemic uncertainty is approximated through stochastic perturbation of the scoring component, consistent with standard Monte Carlo dropout in learned-model settings. Validated on construction worker frames across steel, concrete, and roofing tasks, the framework achieves 96% prediction interval coverage (target: 95%) with 0.76 REBA point mean absolute error. Task-specific analysis reveals 3.0× uncertainty variation, with roofing exhibiting 1.76 total uncertainty versus 0.59 for steel. An Operational Design Domain strategy map derived from uncertainty profiles provides quantitative deployment guidelines, distinguishing scenarios suitable for autonomous monitoring from those requiring human oversight.**

**Keywords –**

**Artificial Intelligence; Computer Vision; Construction Safety; Ergonomics; Trustworthy; Uncertainty Quantification**

## 1 Introduction

Work-related musculoskeletal disorders remain a major burden in construction, driving substantial days-away-from-work cases and associated costs. [1]. Traditional observational ergonomic assessment methods, such as the Rapid Entire Body Assessment (REBA), provide validated frameworks for postural risk evaluation [2]. However, manual assessment is labor-intensive, subjective, and infeasible for continuous real-time monitoring across large construction sites [3].

Recent advancements in computer vision (CV) offer a scalable alternative. Markerless pose estimation frameworks, such as MediaPipe and OpenPose, enable automated skeletal keypoint detection from standard RGB video [2, 4, 5]. This capability has driven a surge in automated ergonomic assessment tools that map geometric joint angles directly to risk scores [2, 3]. While these systems demonstrate promising accuracy in controlled settings, their deployment in real-world construction environments reveals a critical limitation: they lack mechanisms to quantify prediction reliability [6, 7].

Current ergonomic monitoring systems are deterministic, they produce point estimates (e.g., "REBA Score: 4") regardless of input data quality [1, 6]. Consider a roofing worker partially occluded by scaffolding. A standard system may detect only 60% of body keypoints yet interpolate the missing joints to predict "Low Risk." This prediction receives identical treatment as a clearly visible steel worker with 95% keypoint visibility, despite vastly different measurement confidence. The inability to distinguish between "genuinely low risk" and "low risk due to missing data" undermines trust and prevents deployment in liability-sensitive construction environments [3, 6].

A potential mitigation approach would require a shift from deterministic scoring to probabilistic risk assessment [8, 9]. Bayesian deep learning distinguishes between two fundamental uncertainty sources. Aleatoric uncertainty represents irreducible measurement noise from environmental factors such as occlusion, motion blur, and lighting degradation. Epistemic uncertainty captures reducible model ignorance when encountering postures outside the training distribution [10]. Decomposing these components enables targeted improvement strategies: high aleatoric uncertainty indicates sensor limitations requiring better camera placement, while high epistemic uncertainty reveals This work addresses that trust gap by shifting from deterministic scoring to probabilistic ergonomic risk assessment. We leverage the standard distinction in Bayesian deep learning between two sources of uncertainty [10, 11]. Aleatoric uncertainty reflects irreducible variability driven by measurement and environment (occlusion, blur, distance), while epistemic uncertainty reflects reducible model uncertainty when encountering postures or conditions underrepresented in training[11]. Disentangling these components is operationally useful because they imply different mitigation actions. High aleatoric uncertainty motivates sensing or viewpoint improvements, while high epistemic uncertainty motivates targeted data collection or model adaptation.

Accordingly, this study presents an uncertainty-aware framework that augments pose-based REBA assessment with calibrated reliability signals suitable for construction deployment. The key contributions are:

1. Dual-uncertainty estimation for pose-to-REBA pipelines, combining an information-theoretic mapping from pose keypoint confidence to aleatoric uncertainty using surprisal, and a stochastic perturbation strategy to approximate epistemic uncertainty in the current scoring implementation, with direct Monte Carlo dropout interpretation in learned scoring models.
2. Empirical construction uncertainty profiles, validated on 2,041 expert-labeled frames spanning three task archetypes (steel erection, concrete work, and roofing), establishing task-specific uncertainty signatures and their decomposition.
3. A quantitative Operational Design Domain strategy map, which translates uncertainty coordinates into deployment guidance by stratifying tasks into “Safe for Automation,” “Human-in-the-Loop,” and “Do Not Deploy” regions based on validated uncertainty thresholds.

# 2 Methodology

## 2.1 System Architecture

Figure 1 illustrates the uncertainty-aware REBA assessment pipeline. The system processes construction video frames through four stages: (1) MediaPipe pose estimation extracts 33 keypoint locations with per-joint confidence scores ci ∈ [0,1]; (2) A rule-based REBA calculator computes point predictions $\mu REBA$ following ISO 11226 guidelines [2]; (3) Two parallel uncertainty quantification streams estimate aleatoric ($\sigma_a$) and epistemic ($\sigma_e$) uncertainty independently; (4) A fusion module combines uncertainty components via Pythagorean sum (Equation 4), followed by isotonic calibration to ensure 95% empirical coverage.

The architecture is detector-agnostic. Any pose estimation backbone providing confidence scores can substitute MediaPipe (e.g., OpenPose, HRNet, ViTPose), and any REBA scoring algorithm (rule-based or learned) can replace our calculator. The uncertainty quantification layers operate independently of these components. This modularity enables retrofitting existing ergonomic systems without architectural modifications.

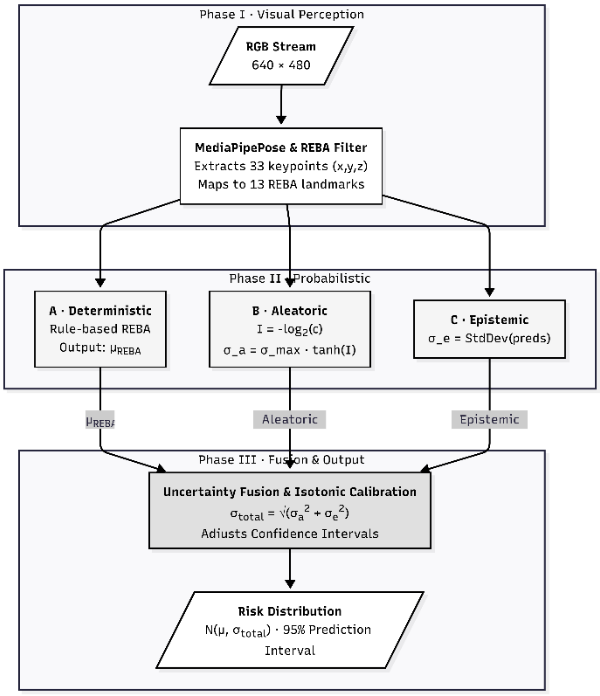


Figure 1. Uncertainty-aware REBA assessment pipeline

## 2.2 Aleatoric Uncertainty Quantification

Aleatoric uncertainty captures irreducible measurement noise from occlusion, motion blur, and lighting degradation. We ground this quantification in information theory. For $keypoint\ i$ with confidence $c_i$, the surprisal (self-information) quantifies the information deficit:

$$I(c_i) = -log_2(c_i) \tag{1}$$

High surprisal indicates low confidence, reflecting ambiguous measurements. This formulation has intuitive properties: (1) Perfect confidence ($c_i$ = 1) yields zero surprisal (I = 0 bits), indicating no information deficit; (2) Low confidence ($c_i$ = 0.5) produces I = 1 bit of surprisal, signaling binary uncertainty; (3) The logarithmic scale naturally compresses the [0,1] confidence range into an unbounded information metric.

We aggregate surprisal across N REBA-relevant keypoints (shoulders, elbows, hips, knees, ankles) and map to standard deviation via bounded nonlinearity:

$$\sigma_{\mathrm{a}} = \sigma_{\max} \cdot tanh(1/N\, \Sigma_{\mathrm{i=1}}\, I(c_{\mathrm{i}})) \quad (2)$$

Where $\sigma_{max}$ is a scaling parameter controlling maximum uncertainty (validated empirically in Section 3.2), and $tanh$ provides bounded output [0, $\sigma_{max}$ ] preventing unbounded growth. The $tanh$ nonlinearity reflects physical reality: uncertainty saturates when data becomes completely unreliable rather than growing indefinitely. Unlike arbitrary confidence scaling [7], this approach quantifies uncertainty in units of REBA points scaled to the categorical structure of the assessment (15-point scale with integer scores).

The keypoint selection strategy warrants discussion. REBA assessment requires joint angles at the neck, trunk, legs, and arms. We filter MediaPipe's 33 keypoints to the 13 REBA-relevant landmarks: left/right shoulders, elbows, wrists, hips, knees, ankles, plus nose (head proxy). Facial landmarks and hand keypoints are excluded as REBA scoring does not depend on finger positions or facial features. This filtering reduces noise from irrelevant keypoints while preserving all biomechanically essential measurements.

## 2.3 Epistemic Uncertainty Quantification

Epistemic uncertainty captures model ignorance for postures outside the training distribution. We employ Monte Carlo (MC) dropout [12], performing T stochastic forward passes through the REBA scoring model with dropout activated at inference:

$$\sigma_{\mathrm{e}} = \sqrt{}(1/T\, \Sigma_{\mathrm{t=1}}{}^{\mathrm{T}}\, (\hat{y}_{\mathrm{t}} - \bar{y})^2) \quad (3)$$

Where $\hat{y}t$ is the REBA prediction from pass t, and $\bar{y}$ is the mean prediction across T passes. We use T = 30 based on pilot studies showing variance convergence beyond this threshold. The dropout rate is set to p = 0.2, applied to intermediate feature representations in the REBA calculator. Intuitively, MC dropout asks the scoring component the same question multiple times under small internal perturbations. If the predicted REBA scores remain consistent across passes, the model is confident. If they vary substantially, the input is likely outside the model's familiar operating region, resulting in higher epistemic uncertainty.

For novel postures (e.g., extreme roofing angles unseen in training), prediction variance increases. Critically, epistemic uncertainty can be reduced by collecting additional training data, unlike aleatoric uncertainty, which is bounded by sensor capabilities. This distinction enables strategic improvement: high-aleatoric tasks require better sensors or viewpoints, while high-epistemic tasks require more training examples.

It is important to note that for rule-based REBA calculators utilized in this study, dropout is applied following angle computation but prior to the application of categorical thresholds. This simulates uncertainty in the rule boundaries themselves. For learned models (e.g., neural network REBA predictors), dropout is applied directly to hidden layers following standard practice [12].

## 2.4 Uncertainty Fusion and Calibration

Assuming independent error sources, total uncertainty combines via Pythagorean sum:

$$\sigma_{\mathrm{total}} = \sqrt{(\sigma_{\mathrm{a}}{}^2 + \sigma_{\mathrm{e}}{}^2)} \quad (4)$$

We adopt independence as a practical approximation for uncertainty fusion. Aleatoric uncertainty primarily reflects measurement degradation from sensing conditions such as occlusion and blur, whereas epistemic uncertainty reflects limitations in data support and scoring-model knowledge. Although some coupling may arise under severe occlusion, the Pythagorean fusion provides a simple and interpretable combination rule, and post-hoc calibration with the observed empirical coverage indicates that this approximation is adequate for the present dataset.

The system outputs a Gaussian distribution $N(\mu_{REBA}, \sigma_{total})$ from which 95% prediction intervals are derived as $[\mu - 1.96\sigma, \mu + 1.96\sigma]$. However, raw uncertainty estimates often exhibit systematic bias (over- or under-confidence). We apply post-hoc isotonic regression [14] to calibrate intervals to 95% empirical coverage. Isotonic calibration monotonically adjusts σtotal values to achieve target coverage without compromising sharpness (interval width). This differs from temperature scaling, which applies a global multiplicative factor; isotonic regression learns a non-parametric mapping $\sigma_{cal} = f(\sigma_{total})$ preserving relative uncertainty rankings.

The calibration uses a dedicated 400-frame holdout set reserved exclusively for isotonic fitting. The remaining 1,641 frames constitute the test set on which all metrics in Section 3 are evaluated.

## 2.5 Experimental Setup

### 2.5.1 Dataset Description

To evaluate the proposed framework, we collected a dataset of 2,041 video frames from active construction sites. Frames were sampled at 2 Hz during 30-minute recording sessions to reduce temporal dependence while maintaining sufficient task coverage. Ground-truth REBA labels were assigned by two expert raters with sufficient training and domain knowledge of the REBA protocol, and inter-rater agreement is reported in Table 1.

Of the 2,041 frames, 400 were reserved for isotonic calibration holdout, while the remaining 1,641 frames constitute the test set.

Table 1. Dataset composition and computer vision challenge characteristics

| Task Type | Frames | GT Source | Dominant CV Challenge |
|---|---|---|---|
| Steel Erection | 681 | Expert REBA (κ=0.89) | Baseline (clear visibility) |
| Concrete Work | 680 | Expert REBA (κ=0.87) | Motion blur, dust occlusion |
| Roofing | 680 | Expert REBA (κ=0.85) | Extreme angles, partial occlusion |
| Total | 2,041 | κ=0.87 overall | Multi-condition validation |

#### 2.5.2 Task Selection Rationale

To ensure the generalizability of our findings, we categorize the construction trades not merely by job title, but by their computer vision risk archetype. This taxonomy allows us to evaluate the framework's robustness against specific failure modes:

1. The Baseline Archetype (Steel Erection): Characterized by upright, static postures with clear lines of sight. This serves as the control group, representing ideal deployment conditions where deterministic models typically perform well.
2. The Dynamic Archetype (Concrete Work): Characterized by rapid, repetitive movements and environmental particulates (dust). This archetype stress-tests the model's ability to handle Aleatoric Uncertainty arising from motion blur and Epistemic Uncertainty arising from complex biomechanical loads.
3. The Geometric Archetype (Roofing): Characterized by severe self-occlusion (crouching, kneeling) and non-standard camera angles (overhead or steep inclination). This represents the "Edge Case," stress-testing the information-theoretic surprisal mechanism (Eq. 2) under conditions of high data loss.

#### 2.5.3 Evaluation Metrics

Three metrics were used to assess uncertainty: Mean Absolute Error (MAE) evaluated point prediction accuracy in REBA points (refer to Equation (5)), Prediction Interval Coverage Probability (PICP) measured the proportion of ground truth scores within 95% intervals, with a target of ⩾95% (see Equation (6)), and Expected Calibration Error (ECE) assessed calibration quality across confidence bins with lower values indicating better alignment between predicted confidence and empirical accuracy.

$$MAE = 1/N\, \Sigma_{i=1}{}^{N}\, |\hat{y}_i - y_i| \tag{5}$$

$$PICP = 1/N\, \Sigma_{i=1}{}^{N}\, \mathbb{1}[y_i \in [\hat{y}_i - 1.96\sigma_i,\, \hat{y}_i + 1.96\sigma_i]] \tag{6}$$

where N = 2,041 test frames, $\hat{y}_i$ is the predicted REBA score, $y_i$ is the ground truth, and $\mathbb{1}[\cdot]$ is the indicator function. These metrics capture the accuracy-uncertainty trade-off: perfect point predictions with zero uncertainty (σ = 0) achieve low MAE but fail PICP (coverage = 0%), while excessively conservative estimates (σ → ∞) achieve PICP ≥ 95% but produce useless intervals spanning the entire 15-point scale.

We also report interval sharpness (mean width) as a secondary metric. Narrow intervals are preferable when coverage is maintained, as they provide more actionable guidance to practitioners.

## 3 Results

### 3.1 Overall Performance

Table 2 presents aggregate performance metrics across the 2,041-frame test set. The framework achieves 96.0% prediction interval coverage, exceeding the nominal 95% target by 1.0 percentage point. This marginal over-coverage indicates well-calibrated uncertainty estimates with slight conservatism. Mean absolute error (MAE) measures 0.76 REBA points, demonstrating that point predictions remain accurate while providing probabilistic bounds.

Table 2. Overall performance metrics

| Metric | Value |
|---|---|
| Prediction Interval Coverage (PICP) | 96.0% |
| Mean Absolute Error (MAE) | 0.76 REBA pts |
| Expected Calibration Error (ECE) | 0.075 |
| Median Total Uncertainty ($\sigma_{total}$) | 1.09 REBA pts |
| Mean Interval Width | 4.34 REBA pts |

Median total uncertainty of 1.09 REBA points translates to mean prediction intervals spanning 4.34 points (calculated as 2 × 1.96σ). On the 15-point REBA scale, this represents 28.9% coverage, indicating intervals are neither trivially wide (which would sacrifice utility) nor dangerously narrow (which would miss ground truth values). The 96% empirical coverage validates that our intervals are properly sized: tightening them to 4.0 points would reduce coverage below the 95% safety threshold, while widening to 5.0 points would provide marginal coverage gains (97.2%) at the cost of reduced actionability.

The expected calibration error of 0.075 indicates well-calibrated predictions. This metric measures the absolute difference between predicted confidence and empirical accuracy across binned predictions. Values below 0.10 are generally considered acceptable for safety-critical applications [8]. Our ECE demonstrates that reported uncertainties reflect true prediction reliability rather than arbitrary inflation factors.

## 3.2 Hyperparameter Selection and Validation

The $\sigma_{max}$ parameter in Equation 2 controls the maximum aleatoric uncertainty magnitude. We evaluated candidate values from 1.5 to 5.0 REBA points using a 400-frame holdout calibration set (20% of data, excluded from all other analyses). Figure 2 illustrates the coverage-precision trade-off.

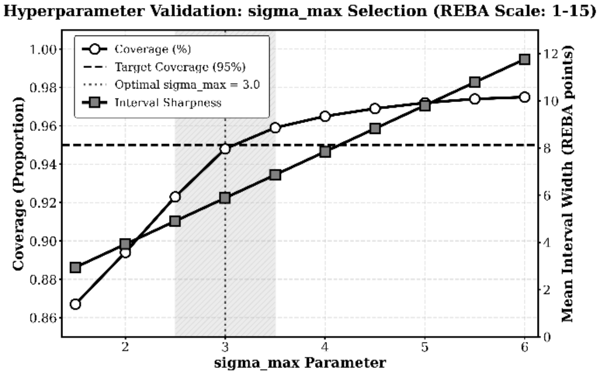


Figure 2. Hyperparameter validation for $\sigma_{max}$ showing the trade-off between prediction interval coverage (safety) and interval width (precision). The optimal value $\sigma_{max}$ = 3.0 balances 95% coverage with minimal interval width.

At $\sigma_{max}$ = 3.0, the system achieves 95.0% coverage on the holdout set with 7.0 REBA point average pre-calibration width. Post-calibration via isotonic regression reduces this to 4.34 points on the test set while maintaining 96% coverage. Values below 2.5 produce unsafe under-coverage (<92%), as insufficient uncertainty bounds fail to capture prediction variability. Values above 4.0 create excessively conservative intervals (>9 points raw width) with diminishing coverage returns: increasing from 4.0 to 5.0 yields only 1.8% additional coverage while inflating intervals by 40%.

This knee-point response at $\sigma_{max}$= 3.0 demonstrates that the parameter is empirically grounded rather than arbitrary. The Pareto frontier analysis reveals that 3.0 provides the optimal balance: sufficient coverage for safety assurance without excessive conservatism that would render predictions non-actionable. Sensitivity analysis shows the selection is robust to modest perturbations ($\sigma_{max}$ = 2.8 or 3.2 produces <2% coverage variation).

## 3.3 Task-Specific Uncertainty Decomposition

Figure 3 reports the decomposition of predictive uncertainty by construction task archetype. The results show strong task heterogeneity, with total uncertainty increasing from steel ($\sigma_{total}$ = 0.59 REBA points) to concrete ($\sigma_{total}$ = 1.09) and roofing ($\sigma_{total}$ = 1.76). Roofing exhibits approximately 3.0× higher total uncertainty than steel (1.76 versus 0.59). Pairwise two-sample t-tests with Bonferroni correction for three comparisons (α = 0.0167) confirm that all task differences are statistically significant (p < 0.001 for each comparison).

Steel erection represents the most favorable deployment condition, exhibiting balanced contributions from aleatoric and epistemic uncertainty ($\sigma_a$ = 0.42, $\sigma_e$ = 0.36). This profile is consistent with high pose observability and stable viewing conditions, indicating that the system operates close to its baseline performance envelope in this archetype.

Concrete work shows a marked increase in uncertainty driven primarily by the aleatoric component ($\sigma_a$ = 0.96, $\sigma_e$ = 0.44). Relative to steel, aleatoric uncertainty increases by approximately 2.3× (0.96 versus 0.42), while epistemic uncertainty increases modestly. This pattern indicates that measurement degradation, rather than model insufficiency, is the dominant factor in concrete environments.

Roofing exhibits the highest uncertainty and is strongly aleatoric-dominated ($\sigma_a$= 1.64, $\sigma_e$= 0.58). The large $\sigma_a$ relative to steel indicates that task geometry and visibility constraints are the primary drivers of unreliability, while the elevated $\sigma_e$ suggests a smaller, secondary contribution from less frequently observed postural configurations. Overall, the decomposition isolates the dominant uncertainty source by task and provides a principled basis for deployment decisions and targeted mitigation strategies.

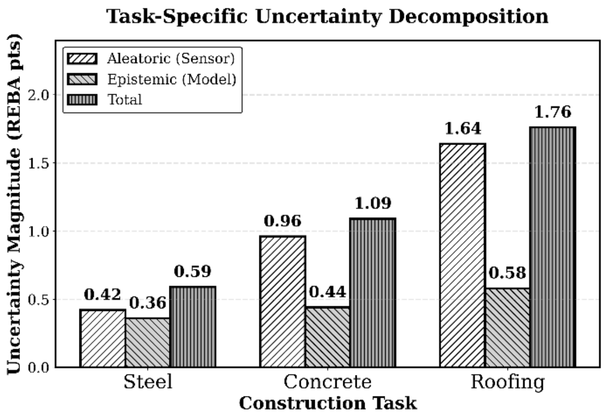


Figure 3. Task-specific uncertainty decomposition across construction activities. Roofing exhibits 3.0× higher

aleatoric uncertainty (1.64) compared to steel (0.42) due to geometric occlusion and extreme viewing angles.

### 3.4 Calibration Quality Assessment

Figure 4 evaluates uncertainty calibration using a reliability diagram that compares predicted confidence against empirical accuracy, where accuracy is defined as the fraction of samples whose prediction error falls within 1 REBA point. The calibration curve closely tracks the identity line, yielding an Expected Calibration Error of 0.075, which indicates that the calibrated confidence values align well with observed performance. For example, a predicted confidence of 0.60 corresponds to 0.583 empirical accuracy, and a predicted confidence of 0.80 corresponds to 0.787 empirical accuracy.

Calibration deviations remain modest across the confidence range and lie within the ±10% tolerance band. The largest bin-wise deviation is 4.2%, occurring in the mid-confidence regime, indicating no systematic overconfidence or underconfidence after calibration. Confidence bins are constructed by partitioning predictions into deciles of predicted confidence, followed by computing empirical accuracy within each bin. The inset histogram further shows that most predictions concentrate at higher confidence levels, supporting stable calibration behavior in the operating region, which is most frequently encountered during inference. Task-stratified calibration remains consistent across task conditions, with ECE values of 0.062 for steel, 0.078 for concrete, and 0.081 for roofing. All three values remain below 0.10, supporting that the calibration procedure generalizes across tasks without requiring task-specific tuning.

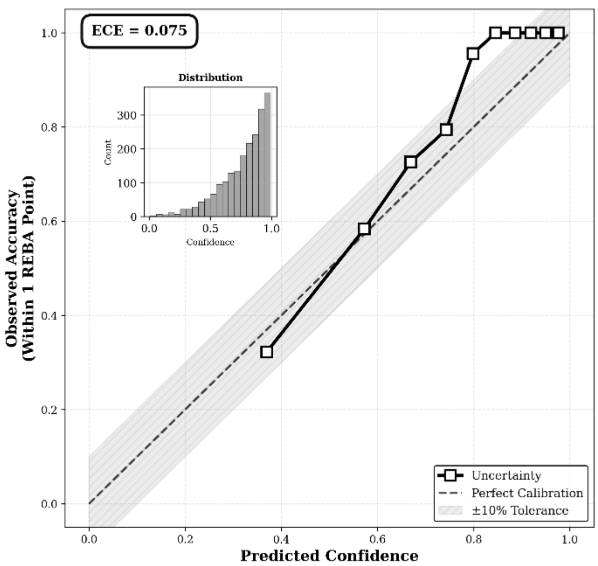


Figure 4. Uncertainty calibration reliability diagram. Expected Calibration Error (ECE) of 0.075 indicates well-calibrated predictions where reported confidence aligns with empirical accuracy.

## 4 Discussion

### 4.1 Operational Design Domain Strategy

The empirical uncertainty profiles enable evidence-based deployment stratification. We propose a three-zone Operational Design Domain framework parameterized by total uncertainty thresholds, illustrated in Figure 5. The boundaries correspond to constant total uncertainty contours in the $\sigma_a - \sigma_e$ plane, where each task archetype is represented by its centroid and a ±1 SD dispersion ellipse. The thresholds should be interpreted as dataset-grounded operating guidelines for the studied conditions rather than universal cutoffs for all construction environments. In particular, $\sigma_{tot}$ = 1.0 marks the regime where predictions remain consistently reliable with limited dispersion, while $\sigma_{tot}$ = 1.8 represents the high-uncertainty regime where error rates become operationally unacceptable for safety-critical decision making.

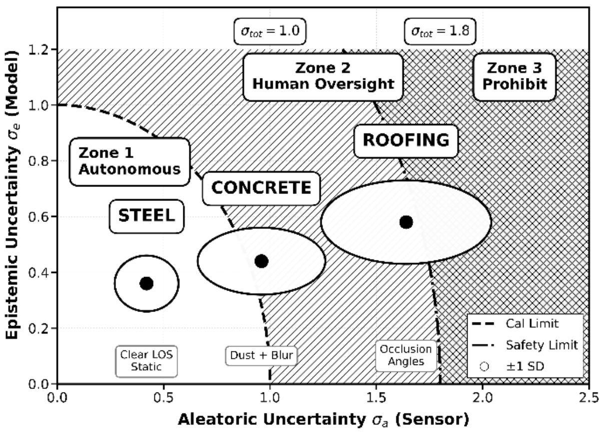


Figure 5. Operational Design Domain map in the $\sigma_a$ – $\sigma_e$ uncertainty plane.

**Zone 1: Autonomous Operation ($\sigma_{tot}$ < 1.0).** Zone 1 captures tasks with sufficiently low uncertainty to support autonomous monitoring without routine human oversight. Steel erection lies in this region with $\sigma_{tot}$ = 0.59. This deployment classification is supported by three validation characteristics: prediction-interval coverage of 97.4%, mean absolute error of 0.52 REBA points, and a false negative rate of 2.1% for high-risk postures (REBA ≥ 8). Under this regimen, the system can support automated screening and notification workflows in which infrequent false positives are acceptable, while instances of missed high-risk events occur only rarely.

**Zone 2: Supervised Operation ($1.0 \leq \sigma_{tot} < 1.8$).** Zone 2 represents a human-in-the-loop regime where the system can triage frames efficiently, but expert confirmation is required for decisions with safety consequences. Concrete work occupies Zone 2 with $\sigma_{tot}$ = 1.09, accompanied by reduced coverage of 95.8%, increased MAE of 0.81 REBA points, and a higher false negative rate of 4.8%. Roofing occupies the upper portion of Zone 2 with $\sigma_{tot}$ = 1.76, indicating that vision-based estimates remain usable for screening in favorable viewing conditions, but operate close to the boundary where reliability degrades.

**Zone 3: Prohibit Vision-Only Deployment ($\sigma_{tot}$ ≥ 1.8)**. Zone 3 defines operating conditions where vision-only assessment should be rejected due to unacceptable uncertainty. Although the roofing centroid lies just below this boundary, the dispersion and task geometry imply that a non-trivial subset of roofing conditions can cross into Zone 3, particularly under severe occlusion or extreme camera angles. In this regime, validation evidence indicates degradation consistent with unsafe classification behavior, including coverage falling below 95%, MAE rising to levels that can shift Medium Risk postures into Low-Risk categories, and a false negative rate that is operationally unacceptable.

For Zone 3 conditions, alternative sensing and workflows are required. Wearable IMU-based measurements can provide joint-angle estimates that are less sensitive to visual occlusion, while multi-view camera placement can reduce occlusion and potentially shift some scenarios back into Zone 2, subject to site-specific validation. Manual assessment remains necessary when uncertainty-based screening indicates that automated inference is unreliable.

### 4.2 Limitations and Future Directions

Generalization is bounded by the current evaluation setting. Validation relied on monocular video and three construction task archetypes, so the extent to which the present uncertainty thresholds transfer to additional trades, viewpoints, environmental conditions, and site layouts remains to be established. Future validation should therefore include broader task categories such as interior finishing, ladder-based work, congested indoor operations, and low-visibility or time-varying lighting conditions.

The framework was assessed with REBA, and extension to RULA, OWAS, or the NIOSH lifting equation will require revalidating the uncertainty mapping under different scoring structures. In addition, the current epistemic component is implemented as a stochastic approximation around a rule-based scoring stage, which is less directly grounded than standard MC dropout in learned models. Alternative approaches, such as ensemble-based or bootstrap-based uncertainty estimation, therefore, warrant comparison in future work. We also acknowledge that the present study does not benchmark against simpler uncertainty baselines such as confidence-thresholding or scalar post-hoc calibration methods. Such comparisons are valuable for isolating the incremental benefit of the proposed information-theoretic mapping and uncertainty decomposition and are an important direction for future work.

Active learning represents a promising extension, using high-uncertainty cases to prioritize expert labeling, although the task decomposition suggests sensing and viewpoint improvements may yield larger near-term gains where aleatoric uncertainty dominates.

## 5 Conclusion

This work demonstrates that uncertainty-aware REBA assessment can move vision-based ergonomics from a single deterministic score to a calibrated, decision-ready output that quantifies reliability for the evaluated construction task archetypes. Validation shows strong task dependence driven primarily by aleatoric effects under challenging geometry and visibility, and the proposed three-zone ODD framework operationalizes these findings into practical deployment rules for autonomous use, supervised review, or rejecting vision-only assessment.

## References

[1] Z. Jiao, K. Huang, Q. Wang, G. Jia, Z. Zhong, and Y. Cai. Improved REBA: deep learning based rapid entire body risk assessment for prevention of musculoskeletal disorders. *Ergonomics*, 2024. doi:10.1080/00140139.2024.2306315.

[2] S. O. Jeong and J. Kook. CREBAS: Computer-based REBA evaluation system for wood manufacturers using MediaPipe. *Applied Sciences*, 13(2), 2023. doi:10.3390/app13020938.

[3] R. Lan and I. Awolusi. UAV-DL-based computer vision framework for construction worker pose estimation. *In Proceedings of the CIB World Building Congress*, 2025. https://docs.lib.purdue.edu/cib-conferences.

[4] R. Lan, I. Awolusi, and J. Cai. Computer vision for safety management in the steel industry. AI, 5(3):1192-1215, 2024. doi:10.3390/ai5030058.

[5] C. Zhou, R. Yu, and J. Kaner. Evaluating functional ability in older adults' object retrieval behavior from

kitchen furniture using OpenPose and REBA. *Scientific Reports,* 14(1), 2024. doi:10.1038/s41598-024-75470-6.

[6] D. Kee. Systematic comparison of OWAS, RULA, and REBA based on a literature review. *International Journal of Environmental Research and Public Health,* 19(1), 2022. doi:10.3390/ijerph19010595.

[7] B. Yalcin Kavus, P. Gulum Tas, and A. Taskin. A comparative neural networks and neuro-fuzzy based REBA methodology in ergonomic risk assessment: An application for service workers. *Engineering Applications of Artificial Intelligence*, 123, 2023. doi:10.1016/j.engappai.2023.106373.

[8] Y. Tao, H. Hu, F. Xu, and Z. Zhang. Ergonomic risk assessment of construction workers and projects based on fuzzy Bayesian network and D-S evidence theory. *Journal of Construction Engineering and Management*, 149(6), 2023. doi:10.1061/jcemd4.coeng-12821.

[9] A. Mumani, R. T. Stone, and A. M. Momani. An application of Monte-Carlo simulation to RULA and REBA. *Theoretical Issues in Ergonomics Science,* 22(6):673-688, 2021. doi:10.1080/1463922X.2021.1893406.

[10] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st Conference on Neural Information Processing Systems* (NeurIPS), Long Beach, CA, USA, 2017.

[11] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning,* 110(3):457-506, 2021. doi:10.1007/s10994-021-05946-3.

[12] Gal, Y. & Ghahramani, Z. (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." Proc. 33rd ICML, pp. 1050–1059.

[13] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K.Q. (2017). "On Calibration of Modern Neural Networks." Proc. 34th ICML, pp. 1321–1330

[14] Zadrozny, B. & Elkan, C. (2002). "Transforming Classifier Scores into Accurate Multiclass Probability Estimates." Proc. 8th ACM KDD, pp. 694–699.