Knowledge Gaps and Methodology for Real-Time Deep Learning-Based Ergonomic Posture Evaluation

M. Roshni Thanka[1,3], Prof. Abeer Aljohani[2]
[1] Lincoln University College, 47301, Petaling Jaya, Selangor, Darul Eshan, Malaysia
[2] Applied College, Taibah University, Madinah, Saudi Arabia
[3] Karunya Institute of Technology and Sciences, Coimabore-641114, Tamil Nadu, India
Email ID : roshni@karunya.edu, aahjohani@taibahu.edu.sa

Abstract Work-related musculoskeletal disorders (MSDs) have become one of the more pressing concerns in occupational health, particularly among those whose jobs revolve around sustained computer use. While tools like RULA and REBA have served the ergonomics community well, their dependence on trained observers and periodic assessment windows means they routinely miss the gradual postural deterioration that accumulates during a typical workday. This paper charts a path toward addressing that limitation. Drawing on a structured review of the existing literature, five gaps are pinpointed the absence of systems that adapt in real time to individual users, the disconnect between AI-generated labels and clinically grounded risk categories, an underinvestment in temporal modelling of prolonged strain, a scarcity of datasets that reflect genuine office environments, and the persistent mismatch between model complexity and deployment-ready hardware. Against each gap, a corresponding methodological response is proposed: a pipeline built around YOLOv8n-Pose for keypoint detection, an LSTM or Temporal Transformer module for sequence analysis, Elastic Weight Consolidation for continual user adaptation, and an output scoring layer calibrated to RULA and REBA thresholds. Initial backbone testing returned 93% PCK@0.2 at 28 frames per second, which is taken as a feasibility signal.

Keywords: ergonomic assessment; posture monitoring; deep learning; YOLOv8n-Pose; musculoskeletal disorders; continual learning; RULA; REBA

Introduction

The clinical picture for work-related upper limb disorders is one of slow accumulation rather than sudden onset. Cervical spondylosis, rotator cuff tendinopathy, carpal tunnel syndrome these develop through the repeated compounding of postures that seem tolerable in any given moment but become damaging when sustained across months and years of desk work. The World Health Organization records musculoskeletal conditions as affecting over 1.7 billion people globally [1], with sustained computer use identified as a significant occupational contributor. Ergonomists have addressed this with structured assessment tools: RULA [2] scores upper limb and neck posture against action level thresholds; REBA [3] extends coverage to the whole body. Both are well-validated and genuinely useful. Both also require a trained assessor to be physically present at the moment of scoring, producing a record of a posture at a point in time rather than of how that posture has evolved through a working day.

The question this research is organised around answering is whether computer vision and deep learning can be combined to close that temporal gap in a way that is practically deployable. Tracking a skeleton from a webcam in real time is now routine; sequence architectures capable of reading meaning into how that skeleton changes over extended periods are well-developed. Whether they can be assembled into something accurate enough to be clinically trusted, efficient enough to run on ordinary office hardware, and personalised enough to account for individual variation in body proportions and working style is still open.

Related Work

The literature relevant here divides loosely into two communities working on adjacent problems without fully converging. Ergonomics researchers have used computer vision instrumentally, primarily as a means of extracting joint angles that would otherwise require a measuring instrument and a trained observer. Kim et al. [4] fed OpenPose joint angle estimates into a rules-based REBA scoring module and found that, compared against motion-capture reference data, agreement was close enough to be practically meaningful. The limitation the authors themselves note is that the assessment is still static a posture scored at a moment with no account taken of what the worker was doing in the hours beforehand. Singhtaun et al. [5] took a comparable approach using MediaPipe Pose with real-time RULA scoring, confirming that the pipeline operates within reach of standard hardware. The same limitation holds: the system has no capacity to learn anything about the individual using it or to track what that individual was doing earlier in the session.

The computer vision community has advanced pose estimation considerably, though without occupational health as a design objective. Cao et al. [6] established OpenPose as a general-purpose real-time framework; the Part Affinity Field architecture it introduced shaped a generation of subsequent work. Dong and Du [7] recently attached a context coordinate attention mechanism to YOLOv8-Pose and recovered roughly 2.8 percentage points of average precision on MS COCO 2017, a meaningful gain on a backbone already valued for its accuracy-to-speed ratio. Samkari et al. [8] surveyed the field systematically and noted that the distance between what pose estimation achieves under controlled evaluation conditions and what has reached functioning occupational health applications is large and poorly documented. That finding matches the experience of searching for a deployable system for the present research.

Spanning both communities is a persistent gap around adaptation and temporal duration. Vaswani et al. [9] provided the Transformer architecture now standard for sequence tasks where long-range dependencies matter — precisely the property needed to detect strain developing across an afternoon rather than in a single badly-held posture. Kirkpatrick et al. [10] addressed how to update a neural model on new data without losing what it had already learned; Elastic Weight Consolidation offers a principled mechanism, and the fact that it does not require retaining raw footage from previous sessions has direct value in any workplace setting where storing identifiable video is a compliance concern. Neither contribution was motivated by ergonomics, but both address problems the ergonomics literature has not resolved. Table 1 maps prior work against the dimensions the proposed framework is designed to cover.

Table 1. Prior work mapped against dimensions relevant to the proposed framework

Review	Approach	Real-time	Risk scoring	Principal limitation
Kim et al. [4], 2021	OpenPose + REBA rules	Partial	Rule-based	Static assessment; no temporal or adaptive component
Singhtaun et al. [5], 2024	MediaPipe + RULA/REBA	Yes	Rule-based	No user-specific learning; single-frame analysis only
Cao et al. [6], 2021	OpenPose (PAF)	Yes	None	General pose estimation; no ergonomic integration

Dong & Du [7], 2024	CCAM-YOLOv8x-Pose	Yes	None	Accuracy-focused; no scoring, no adaptation, no session modelling
Samkari et al. [8], 2023	Systematic review	N/A	N/A	Review only; gap between lab results and clinical deployment highlighted
This work	YOLOv8n-Pose + LSTM/Transformer + EWC + RULA/REBA scoring	Yes	DL-integrated, clinically grounded	First framework unifying real-time operation, temporal modelling, user adaptation, and interpretable clinical output

Key Contribution

The gap the review above describes is not a gap in any single capability but in the absence of a system that brings several capabilities together for the specific purpose of continuous occupational health monitoring. Kim et al. [4] demonstrate that clinical scoring can follow from pose estimation. Dong and Du [7] move the accuracy-efficiency boundary. Kirkpatrick et al. [10] show how adaptation without data retention is possible. What the literature does not contain is a framework that combines these things in a single pipeline intended for deployment in real office environments. The present work proposes exactly that: keypoint detection at speeds viable on machines that already exist in most workplaces; a temporal module that treats posture as a sequence evolving over the session rather than a collection of independent frames; a continual learning component that adjusts to individual workers without demanding retraining and without retaining their video; and output expressed as Low, Medium, or High risk in the terms RULA and REBA assessors would use. The contributions at this stage of the programme are: formal documentation of five knowledge gaps from the literature; a design architecture where each gap maps to a deliberate component decision; a data collection plan that takes ecological validity seriously; a five-experiment evaluation protocol with pre-specified success thresholds; and backbone feasibility results that justify proceeding to the full programme.

Method, Experiments and Preliminary Results

3.1 Knowledge Gaps

Five gaps appeared consistently regardless of which model or dataset a given study used. They are stated here concisely because the methodology is organised around addressing each one.

- **Real-time personalisation is absent:** existing systems are either real-time or adaptive to individual users, but not both simultaneously. Adding a new user to a deployed system typically requires retraining from scratch, which makes individual calibration impractical at any meaningful scale [4,5].
- **Clinical risk vocabulary is missing from outputs:** classifiers produce probability scores or posture category labels. Translating those into the action levels a RULA or REBA assessment would produce requires an explicit bridging mechanism that virtually no published pipeline includes [8].
- **Sessions are not modelled as sessions:** the great majority of published work treats each video frame as an independent data point. A mildly awkward posture sustained for three hours produces physiologically different outcomes than the same posture sustained for three minutes, and methods that ignore this distinction cannot detect the difference.

- **Benchmark datasets do not represent the target environment:** COCO, MPII, and Human3.6M are valuable general-purpose collections, but none was designed to reflect the posture patterns of someone at a computer workstation for a full day.
- **High-performing models are not deployable on standard hardware:** the detectors with the strongest benchmark results require GPU resources that ordinary office machines do not have. The accuracy-efficiency trade-off has not been resolved for this specific application [7].

3.2 Proposed Framework

The proposed pipeline processes webcam video through eight stages. Preprocessing applies normalisation and noise reduction. YOLOv8n-Pose [11] then extracts 17 COCO-standard skeletal keypoints at 28 frames per second — selected because it meets the deployment efficiency requirement while offering sufficient geometric accuracy for the downstream angle computations. Joint angles for the neck, trunk, upper arms, and wrists are derived from those keypoints. An LSTM or Temporal Transformer [9] processes the resulting angle time series over a sliding window of between 30 and 300 frames, the window width depending on what temporal granularity the strain pattern of interest requires. The temporal module connects to an Elastic Weight Consolidation layer [10] that absorbs new user-specific data without overwriting what the model learned from earlier users, and without retaining any raw video; the absence of retained footage is a structural feature of the architecture rather than a policy add-on. A scoring layer converts model outputs to Low, Medium, or High risk using joint angle thresholds derived from RULA [2] and REBA [3] standards, presented through a lightweight desktop interface.

3.3 Data Strategy

Collection proceeds in three phases. Pre-training draws on COCO-Pose, MPII, and Human3.6M to provide broad initialisation before ergonomics-specific fine-tuning. The second phase collects a dedicated dataset from at least 50 participants working in genuine office, home-office, and academic settings, in sessions of three or more hours per participant, annotated by trained assessors with RULA and REBA risk labels; joint angle ground truth is computed directly from keypoint coordinates. Ethics approval is in place before any recruitment; participation is voluntary; no facial identification is performed at any point; raw video is deleted immediately after keypoint extraction. The third stream is collected during deployment: per-user sequences supporting EWC adaptation are never retained beyond the extraction step.

3.4 Experimental Plan

The validation programme has five components. The first compares YOLOv8n-Pose against HRNet-W32 on PCK@0.2 and Object Keypoint Similarity across both public and custom datasets, establishing where the accuracy-efficiency trade-off sits within this domain rather than the general one. The second compares the LSTM and Temporal Transformer variants on posture sequence classification across sessions of varying length, evaluated by accuracy, F1, and AUC. The third assesses continual learning: EWC against iCaRL, using backward transfer and forward transfer metrics on multi-session data streams. The fourth — most consequential for practical adoption — measures agreement between model risk outputs and independent expert RULA and REBA assessments using Cohen's Kappa with a pre-committed acceptance threshold of $\kappa \geq 0.70$. The fifth records end-to-end latency, throughput, model size, and resource draw on standard office hardware against a 25 FPS target.

3.5 Preliminary Results

The YOLOv8n-Pose backbone was evaluated on COCO-Pose before the main programme begins, as a technical feasibility check. It returned PCK@0.2 = 0.93, OKS = 0.90, 28 FPS, 35 ms per frame, a 6.2 MB model file, and 62% GPU utilisation on a standard desktop. These figures are not findings in the sense the research ultimately requires they confirm that the underlying premise is sound and that the full experimental programme is worth running.

Discussions

There is a simpler version of this research that would have been much quicker to complete: select a strong off-the-shelf pose estimator, add a REBA scoring layer, and present the result as an ergonomic monitoring system. Kim et al. [4] and Singhtaun et al. [5] have each done something along those lines, producing demonstrations that have genuine value as proofs of concept. The reason the present work takes a more involved path is that the simpler approach leaves the three most demanding problems temporal extent, per-user adaptation, and clinical interpretability exactly as they were. Making those three requirements part of the design from the outset, rather than additions to be addressed later, is what shaped most of the decisions described in the preceding sections.

The EWC choice warrants a few additional sentences because continual learning does not appear often in ergonomics research. The obvious alternative, standard fine-tuning, overwrites earlier model parameters when new data arrives. That is not a problem when training happens once, but it is a serious limitation in a system intended to become progressively more calibrated to each individual worker as it accumulates observations. EWC's Fisher information penalty preserves the parameters important for earlier users while permitting updates where new data warrants them. The practical consequence that no raw session footage from prior users needs to be stored is not incidental: in a workplace context, retaining identifiable video of workers carries real compliance costs, and building around that requirement architecturally is preferable to managing it as a data governance policy.

The outcome of Experiment 4, the clinical validation, will determine whether the framework's claimed clinical relevance is substantiated or needs to be qualified. A Cohen's Kappa of 0.70 or above against trained RULA and REBA assessors represents substantial agreement by conventional standards, and reaching it would constitute reasonable evidence that the system's output is meaningful to occupational health practitioners. The threshold was specified before data collection begins, which is the correct moment to make such decisions.

Conclusions

This paper presents the contribution of an ongoing post-doctoral research programme. The following conclusions are drawn:

- The primary research problem is the absence of a real-time, personalised, interpretable, and deployable ergonomic posture risk prediction system for computer-aided workers has been formally established and five specific knowledge gaps have been documented through systematic literature review.
- A novel deep learning framework integrating YOLOv8n-Pose, LSTM/Temporal Transformer, EWC-based continual learning, and RULA/REBA-inspired risk scoring has been designed, with each component choice explicitly justified against the identified gaps and evaluated against alternatives.
- Preliminary baseline evaluation confirms the feasibility of the approach: the YOLOv8n-Pose backbone achieves 93% PCK@0.2, 0.90 OKS, 28 FPS, and a model size of 6.2 MB on standard

hardware — meeting the deployment performance targets established in the research objectives.

- Limitations at this stage include the absence of the full custom real-world dataset and complete experimental validation. The incremental learning module, temporal modelling comparison, and clinical risk scoring validation remain to be executed.

References

[1] World Health Organization, "Musculoskeletal conditions," WHO Fact Sheet, 2023. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

[2] L. McAtamney and E. N. Corlett, "RULA: a survey method for the investigation of work-related upper limb disorders," Applied Ergonomics, vol. 24, no. 2, pp. 91–99, 1993. https://doi.org/10.1016/0003-6870(93)90080-S

[3] S. Hignett and L. McAtamney, "Rapid entire body assessment (REBA)," Applied Ergonomics, vol. 31, no. 2, pp. 201–205, 2000. https://doi.org/10.1016/S0003-6870(99)00039-3

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

[5] C. Singhtaun, S. Natsupakpong, and P. Lorprasertkul, "Ergonomic risk assessment using human pose estimation with MediaPipe Pose," in Proc. 2024 7th Artificial Intelligence and Cloud Computing Conference, 2024. https://doi.org/10.1145/3719384.3719453

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: real-time multi-person 2D pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021. https://doi.org/10.1109/TPAMI.2019.2929257

[7] C. Dong and G. Du, "An enhanced real-time human pose estimation method based on modified YOLOv8 framework," Scientific Reports, vol. 14, p. 7834, 2024. https://doi.org/10.1038/s41598-024-58146-z

[8] E. Samkari, M. Arif, M. Alghamdi, and M. A. Al Ghamdi, "Human pose estimation using deep learning: a systematic literature review," Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1612–1659, 2023. https://doi.org/10.3390/make5040081

[9] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. https://doi.org/10.48550/arXiv.1706.03762

[10] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences, vol. 114, no. 13, pp. 3521–3526, 2017. https://doi.org/10.1073/pnas.1611835114

[11] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. https://doi.org/10.5281/zenodo.7347926