Does Sensor Calibration Matter for Surgical Skill Assessment?

Jason Au[a] and Matthew S. Holden[a]

[a]Carleton University, 1125 Colonel By Dr, Ottawa, ON K1S 5B6, Canada
[1]Correspondance: jasonau@cmail.carleton.ca

ABSTRACT

Accurate skill assessment in surgical training relies on precise sensor-based tracking, yet the extent to which calibration accuracy is required for reliable machine learning–based evaluation remains unclear. This study examines the sensitivity of kinematic skill assessment models to sensor calibration errors in ultrasound-guided needle insertion. Using data from 38 novices and 5 experts, we simulate 7 controlled perturbation scenarios (e.g, human-scale translations, phantom-scale misalignments, and random and constant calibration drift) to assess robustness to spatial error. Temporal Convolutional Networks (TCNs) and ROCKET, a lightweight time-series classification method, are evaluated for predicting normalized skill scores under perturbed coordinate transformations. Contrary to expectations, large spatial perturbations (e.g. translations up to 1500–2720 mm and full 360° rotation) do not degrade performance and in some cases they improve prediction accuracy (Cohen's *d* = -0.42, *p* ¡ 0.05). These results suggest that models primarily learn relative kinematic patterns rather than relying on absolute spatial alignment. The best-performing approach, ROCKET, achieves a trial-level mean absolute error of 1.93 for out-of-plane insertions, comparable to the inter-expert variability MAE of 1.95. Together, these findings challenge the assumption that frequent sensor recalibration is essential for robust skill assessment and highlight relative motion features as invariant descriptors of procedural skill.

Keywords: ultrasound-guided needle insertion, sensor tracking and calibration, medical simulation and training, ROCKET time-series classifier, Temporal Convolutional Network (TCN), calibration robustness

1. INTRODUCTION

Accurate skill assessment in surgical training increasingly relies on sensor-based tracking systems that capture procedural kinematics. Wearable and pose-tracking technologies, including optical and electromagnetic sensors, provide rich motion data that can be used to quantify surgical performance.[1] However, many of these systems rely on non-sterilizable components and require frequent recalibration, raising concerns about workflow disruption and whether strict calibration accuracy is necessary for reliable skill evaluation.

Machine learning methods are widely used to analyze time series data derived from surgical motion. Early approaches employed distance-based techniques such as Dynamic Time Warping, while more recent work has shifted toward feature-based classifiers and deep learning architectures, including support vector machines, random forests, long short-term memory networks, and one-dimensional convolutional neural networks.[2] These models have demonstrated strong performance in capturing temporal structure in kinematic data and have been applied successfully to surgical skill assessment tasks.[3]

In image-guided procedures such as ultrasound-guided needle insertion, spatial alignment between tracked tools and imaging data is typically enforced through calibration.[4] Calibration enables consistent coordinate transformations across sensors and imaging frames, supporting trajectory reconstruction and synchronized analysis. While essential for image-to-space registration, the requirement for precise calibration in downstream tasks, like machine learning–based skill assessment, remains unclear.

Prior work in related domains suggests that learning-based models may be robust to sensor inaccuracies. Data augmentation strategies such as Gaussian noise injection have been shown to improve generalization and stability in neural networks,[5] while simulated sensor drift has been used to train models resilient to measurement error in continuous glucose monitoring systems.[6] These findings raise the possibility that surgical skill assessment models may similarly tolerate (or even benefit from) calibration perturbations.

Medical Imaging 2026: Image-Guided Procedures, Robotic Interventions, and Modeling, edited by Maryam E. Rettmann, Pierre Jannin, Proc. of SPIE Vol. 13927, 1392728
2026 Published by SPIE · 1605-7422 · doi: 10.1117/12.3084063

This study investigates whether frequent sensor calibration is essential for reliable machine learning–based skill assessment in ultrasound-guided needle insertion. By introducing controlled perturbations that simulate calibration drift and spatial misalignment, we evaluate the sensitivity of temporal models to tracking error and examine whether skill-related kinematic patterns remain predictive under degraded calibration conditions.

2. DATA

This study uses an ultrasound-guided needle insertion dataset introduced by Xia et al.,[4] collected using the Perk Tutor training platform. The dataset includes recordings from 38 novice medical students and 5 expert physicians performing procedures using two standard ultrasound guidance techniques: in-plane (IP) and out-of-plane (OOP). These approaches are routinely used in clinical practice and present distinct visualization and motor-control challenges relevant to skill assessment.

Novice participants were randomly assigned to one of the two technique groups. The IP group comprised 20 novices who completed 120 procedures, while the OOP group included 19 novices who performed 114 procedures. Expert physicians completed a total of 15 procedures (8 IP, 7 OOP) to serve as performance benchmarks. For a subset of trials, three expert preceptors independently rated video recordings using an OSATS-like rubric, yielding total scores ranging from 5 to 25.[4]

Pre-processing

Following prior work,[7] each procedure was segmented into overlapping windows of 60 frames using window slicing. Samples were randomly resampled to mitigate class imbalance between novice and expert trials.

The dataset includes static transformations generated using the PLUS toolkit, including ImageToProbe, NeedleTipToNeedle, and ReferenceToVessel. These were combined with dynamic transformations (NeedleToReference and ProbeToReference) to construct composite representations NeedleTipToVessel and ImageToVessel, which align needle motion and ultrasound imaging within a common anatomical reference frame.

Expert ratings were available only for baseline and final trials and were therefore restricted to these conditions for regression analysis. Ratings were based on five criteria: time in motion, instrument handling, flow of procedure, bimanual dexterity, and overall performance, each scored on a 5-point scale. Total scores were normalized to the $[0, 1]$ range using:

$$\text{Normalized Score} = \frac{x - 5}{20}, \tag{1}$$

where x denotes the summed OSATS score.[8]

3. METHODOLOGY

Experiments

We conducted 7 perturbation experiments, each of which represents a different type and magnitude of translation and rotation error. Experiment 0 (not part of the perturbation count) acts as the control condition, using the original calibrated transformations without added noise, and provided the reference point for statistical comparisons. The remaining experiments were designed to evaluate model robustness to different conditions. Perturbation magnitudes were selected to include realistic clinical variability (i.e., minor tracker reattachment and calibration drift) and worst-case deployment scenarios (i.e., phantom-to-human transfer), enabling evaluation across practical and extreme operating conditions. The magnitudes and distributions of perturbations used in the experiments are detailed in Table 1.

Table 1: Perturbation experiments, rationale, and error distributions.

Exp.	Transformations	Rationale	Translation	Rotation
1	ReferenceToVessel	Simulates a domain shift from phantom to human body due to size differences, altering the coordinate frame.	Uniform: 1500–2720 mm (torso width)	Uniform: 0–360°
2	ReferenceToVessel	Simulates inconsistent reference tracker placement on the ultrasound simulator, introducing spatial variability across the length of the phantom.[9]	Uniform: 50–150 mm (phantom size)	Uniform: 0–360°
3	ReferenceToVessel	Models small calibration drifts or misalignments of the reference tracker.	Normal: $\mu = 0$, $\sigma =$ 10 mm	Normal: $\mu = 0°$, $\sigma \approx 5°$ (0–15°)
4	NeedleTipToNeedle	Captures minor pivot calibration variations or reattachment without recalibration.	Normal: $\mu = 0$, $\sigma =$ 10 mm	Normal: $\mu = 0°$, $\sigma \approx 5°$ (0–15°)
5	ImageToProbe	Captures small ultrasound probe calibration errors or reattachment without recalibration.	Normal: $\mu = 0$, $\sigma =$ 10 mm	Normal: $\mu = 0°$, $\sigma \approx 5°$ (0–15°)
6	NeedleTipToNeedle	Simulates systematic error from a miscalibrated needle reused across sessions.	Fixed per trial: sampled from Normal, $\sigma =$ 10 mm	Fixed per trial: sampled from Normal (0–15°)
7	ImageToProbe	Simulates systematic ultrasound calibration error reused across sessions.	Fixed per trial: sampled from Normal, $\sigma =$ 10 mm	Fixed per trial: sampled from Normal (0–15°)

Coordinate Systems and Calibrations

Accurate tracking in ultrasound systems relies on sensors attached to the anatomy (Reference sensor), the ultrasound probe (Probe sensor), and any other tools (e.g., Needle sensor). A pose tracking system dynamically captures transformations between sensors (e.g., NeedleToReference, ProbeToReference). Two main calibrations must be computed to spatially synchronize needle and ultrasound probe motion: needle and probe calibration. Needle calibration enables localization of the needle tip. This was achieved through the Algebraic One-Step method for pivot calibration[10] and spin calibration using axial rotation of the needle.[4] Ultrasound probe calibration enables tracking the ultrasound image plane relative to other objects in the scene. and the external tracking system. This was performed using the Tracked Pointer Method detailed by Welch et al.[11]

Furthermore, registration is required to determine the position of a reference sensor mounted to the anatomy relative to a pre-operative image. In this case, we manually registered the reference sensor to a model of a vessel. Because the reference was rigidly fixed to the phantom in the dataset and the pre-operative image is not used for skill assessment, the exact value of the ReferenceToVessel transform is inconsequential.

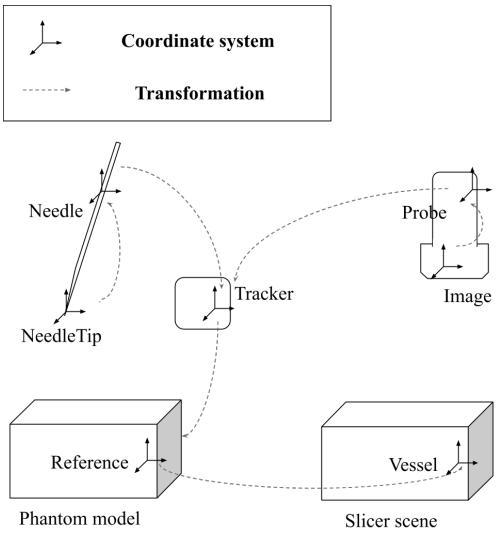

Figure 1: A schematic of the coordinate systems and transformations involved in the ultrasound-guided procedure workflow. Inspired by Ungi et al.[12]

The transformation from the needle tip coordinate system to the vessel coordinate system is obtained by composing three rigid transformations. First, the needle tip is expressed in the needle coordinate system using the needle tip–to–needle calibration transform, denoted $T^{\mathrm{NT}\to\mathrm{N}}$. This is followed by the dynamically tracked transformation from the needle coordinate system to the reference coordinate system, $T^{\mathrm{N}\to\mathrm{Ref}}$. Finally, the reference coordinate system is mapped into the vessel coordinate system using the reference-to-vessel registration transform, $T^{\mathrm{Ref}\to\mathrm{V}}$. The resulting composite transformation is given by

$$T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{N}\to\mathrm{Ref}} \cdot T^{\mathrm{NT}\to\mathrm{N}} \tag{2}$$

Similarly, the transformation from the ultrasound image coordinate system to the vessel coordinate system is obtained by composing three rigid transformations. The ultrasound image is first expressed in the probe coordinate system using the image-to-probe calibration transform, denoted $T^{\mathrm{I}\to\mathrm{P}}$. The probe coordinate system is then mapped to the reference coordinate system via the dynamically tracked probe-to-reference transformation, $T^{\mathrm{P}\to\mathrm{Ref}}$. Finally, the reference coordinate system is mapped into the vessel coordinate system using the reference-to-vessel registration transform, $T^{\mathrm{Ref}\to\mathrm{V}}$. The resulting composite transformation is given by

$$T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{P}\to\mathrm{Ref}} \cdot T^{\mathrm{I}\to\mathrm{P}} \tag{3}$$

We constructed a perturbation matrix P that incorporates both rotational and translational noise. Rotational perturbations were generated by first sampling a random axis vector, with each component drawn from a uniform distribution and subsequently normalized to define a valid rotation axis. A rotation angle was then sampled from a normal distribution, with the magnitude of the perturbation controlled by the specified standard deviation. The resulting axis–angle representation was converted into a rotation matrix. Translational perturbations were generated by sampling a three-dimensional noise vector, with identical distributions applied independently along each spatial dimension. The rotational and translational components were finally combined into a single homogeneous transformation matrix P.

In each experiment, a single perturbation matrix P was applied to one transformation within the kinematic chain, with the placement of the perturbation determined by the calibration or reference component being simulated. We adopt the convention that left-multiplication applies a perturbation in the parent coordinate frame, while right-multiplication applies a perturbation in the local coordinate frame of the object being perturbed.

For reference-level perturbations, the perturbation matrix was applied via left-multiplication to the reference-to-vessel transformation. Because this transformation appears in both the needle-to-vessel and image-to-vessel chains, the perturbation affects both modalities simultaneously. The perturbed needle and image transformations are therefore given by

$$T^{\mathrm{NT}\to\mathrm{V}}_{\mathrm{perturbed}} = P^{\mathrm{Ref}} \cdot T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{N}\to\mathrm{Ref}} \cdot T^{\mathrm{NT}\to\mathrm{N}} \quad T^{\mathrm{I}\to\mathrm{V}}_{\mathrm{perturbed}} = P^{\mathrm{Ref}} \cdot T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{P}\to\mathrm{Ref}} \cdot T^{\mathrm{I}\to\mathrm{P}} \tag{4}$$

For needle calibration perturbations, the perturbation matrix was applied via right-multiplication to the needle tip–to–needle calibration transform, thereby perturbing the local needle frame without affecting the reference or vessel frames. The resulting perturbed needle-to-vessel transformation is given by

$$T^{\mathrm{NT}\to\mathrm{V}}_{\mathrm{perturbed}} = T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{N}\to\mathrm{Ref}} \cdot \left(T^{\mathrm{NT}\to\mathrm{N}} \cdot P^{\mathrm{N}}\right) \tag{5}$$

Similarly, for image calibration perturbations, the perturbation matrix was applied via right-multiplication to the image-to-probe calibration transform, perturbing the local image frame only. The perturbed image-to-vessel transformation is therefore given by

$$T^{\mathrm{I}\to\mathrm{V}}_{\mathrm{perturbed}} = T^{\mathrm{Ref}\to\mathrm{V}} \cdot T^{\mathrm{P}\to\mathrm{Ref}} \cdot \left(T^{\mathrm{I}\to\mathrm{P}} \cdot P^{\mathrm{I}}\right) \tag{6}$$

In Experiments 0 through 3, both needle-to-vessel and image-to-vessel transformations were available for each trial. These representations were combined to form a single multivariate input by horizontally concatenating their corresponding feature matrices. Let $T_1 \in \mathbb{R}^{n\times d_1}$ denote the needle-to-vessel transformation sequence and $T_2 \in \mathbb{R}^{n\times d_2}$ denote the image-to-vessel transformation sequence, where n is the number of time steps and d_1 and d_2 are the respective feature dimensions. The combined input representation is given by

$$T_{\mathrm{combined}} = [T_1 \parallel T_2] \in \mathbb{R}^{n\times(d_1+d_2)} \tag{7}$$

where $\parallel$ denotes horizontal concatenation along the feature dimension.

Model Architecture

We implemented two machine learning methods for skill assessment. First, we implemented a Temporal Convolutional Network (TCN) following Liu and Holden 2.[7] Their work showed the TCN outperformed LSTM architectures across various data representations. In addition to the TCN, we evaluated ROCKET (RandOm Convolutional KErnel Transform),[13] a near state-of-the-art approach on general time series tasks. ROCKET employs thousands of randomly initialized 1D convolutional kernels followed by a linear classifier, providing an efficient and powerful baseline without deep learning.

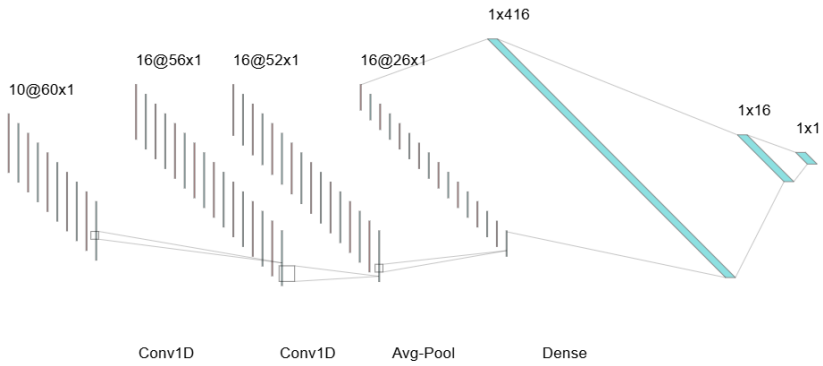

Figure 2: TCN architecture, inspired by Liu and Holden.[7] Top values indicate feature dimensions; bottom labels indicate layer types. The input is shown as 10@60, consistent with our 10 input features and 60-frame window slicing.

We used MiniRocketMultivariate,[13] with 5,000 convolutional kernels and fixed seed 18 for reproducibility. Similar to the TCN setup, extracted features were normalized and passed to a ridge regression model with regularization strengths selected via cross-validation (`alphas` = {0.1, 1.0, 10.0}).

We encode transformations using 6 values for rotation (vectorized rotation matrix), 3 values for translation, and 1 timestamp value, resulting in a 10-dimensional feature vector per frame. We sample 60 frame clips, yielding an input size of 10×60 (features $\times$ time steps). Clips were randomly resampled to balance novice and expert data. Ratings were linearly rescaled from the range $[5, 25]$ to the range $[0, 1]$ for regression.

We compared the performance of TCN and ROCKET on unperturbed data (Experiment 0) to identify the most suitable model for subsequent robustness experiments. We measure performance using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both models were trained to minimize MSE, with the TCN leveraging early stopping (patience = 5), dropout (0.8), weight regularization (0.01), and a learning rate of 0.0001.

Experimental Design

Model evaluation is performed using leave-one-user-out five-fold cross-validation. Folds were stratified by skill level, ensuring data from one expert is in each fold. This provides a robust measure of model generalization to unseen users. Each experiment was repeated 20 times to ensure reliably of results across different random perturbations.

The model reports both Clip MAE and Trial MAE to assess performance at different granularities. Clip MAE evaluates prediction performance on a 60 frame clip, capturing short-term performance. Trial MAE aggregates predictions within each full trial and compares the resulting average to the corresponding ground truth OSATS score, providing a high-level measure of performance.

To assess model sensitivity to perturbations, we performed statistical analyses to compare performance under perturbed and unperturbed conditions using effect size (Cohen's d) and a one-sided independent samples *t-test* for non-inferiority (with margin of 0.5).

Effect size was quantified using Cohen's d computed relative to the unperturbed baseline:

$$d = \frac{\mu_{\text{pert}} - \mu_{\text{base}}}{s_p}, \quad s_p = \sqrt{\frac{(n_{\text{pert}} - 1)s_{\text{pert}}^2 + (n_{\text{base}} - 1)s_{\text{base}}^2}{n_{\text{pert}} + n_{\text{base}} - 2}}, \tag{8}$$

where μ and s denote the mean and standard deviation of the metric across runs, and s_p is the pooled standard deviation.

4. RESULTS

We first compared the two model architectures (i.e., TCN and ROCKET) on the unperturbed data in experiment 0. Evaluation was performed using multiple metrics (Clip MAE, Clip RMSE, Trial MAE, Trial RMSE) for both approaches (Table 2).

Table 2: Experiment 0: Comparison of TCN and ROCKET performance (mean ± SD across 5 folds).

Model	**IP Clip MAE**	**IP Clip RMSE**	**IP Trial MAE**	**IP Trial RMSE**	**OOP Clip MAE**	**OOP Clip RMSE**	**OOP Trial MAE**	**OOP Trial RMSE**
TCN	4.17 ± 1.59	4.76 ± 1.63	3.48 ± 0.95	4.21 ± 1.08	2.35 ± 0.87	3.05 ± 1.10	4.93 ± 0.90	6.86 ± 1.49
ROCKET	2.81 ± 0.34	3.18 ± 0.34	2.13 ± 0.18	2.60 ± 0.20	1.23 ± 0.19	1.61 ± 0.19	1.93 ± 0.19	2.76 ± 0.29

ROCKET ($\alpha = 1.0$) consistently outperformed the TCN across all evaluated metrics, particularly at the clip level. As a result, ROCKET was selected as the primary model for subsequent perturbation experiments. The

distributions of MAE and RMSE across experimental conditions are shown in Figs. 3 and 4, which summarize variability and outliers over $n = 20$ repeated runs per condition.

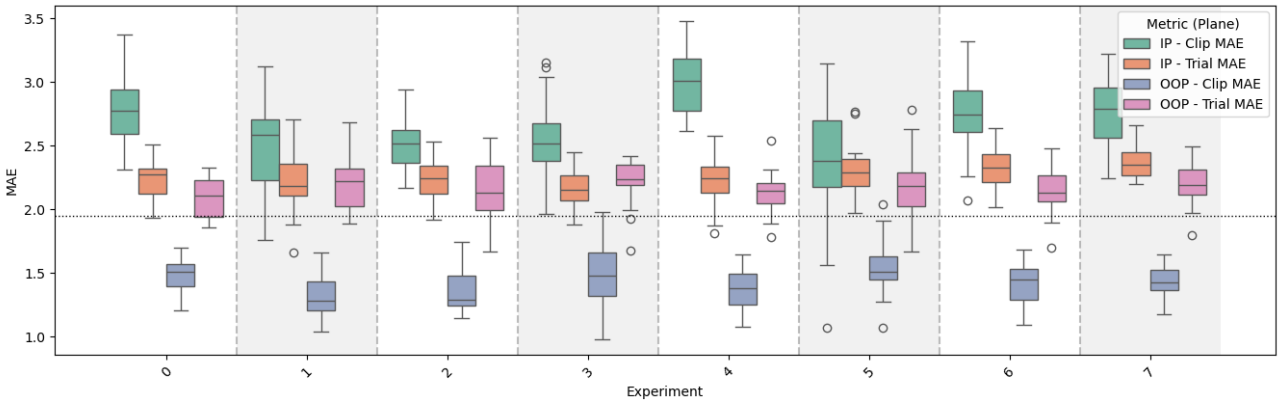

Figure 3: Box-and-whisker plot of MAE values across perturbation experiments ($n = 20$ runs per condition).

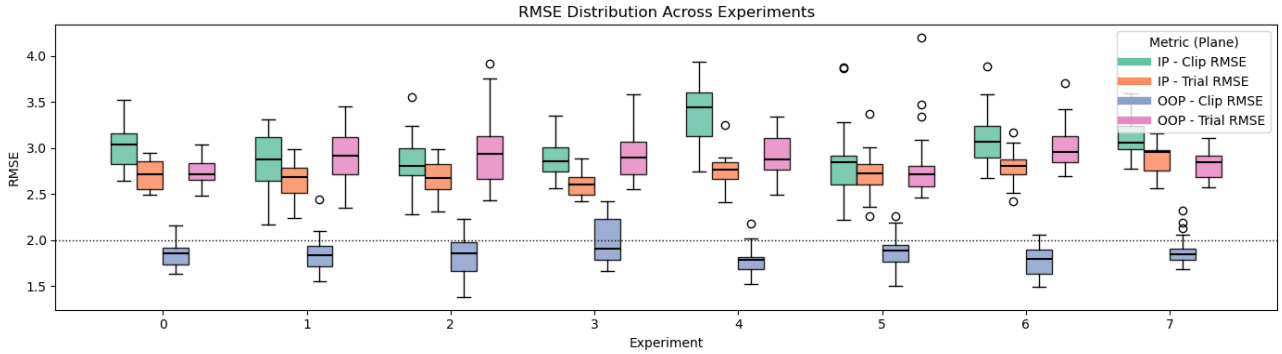

Figure 4: Box-and-whisker plot of RMSE values across perturbation experiments ($n = 20$ runs per condition).

Cohen's d was computed across all perturbation experiments and evaluation metrics ($n = 56$ comparisons). Effect sizes ranged from –4.35 to 5.99 (mean = 0.47, SD = 2.85). Positive values of d indicate higher error under perturbation (worse performance), while negative values indicate lower error (improvement) relative to baseline.

To assess robustness, one-sided non-inferiority tests were performed for all experimental conditions relative to the unperturbed baseline. All resulting p-values were below the predefined 0.05 threshold, indicating that none of the perturbations led to a statistically significant degradation in performance.

5. DISCUSSION

Across experiments, ROCKET achieved trial-level errors comparable to inter-expert variability under unperturbed conditions, indicating that automated assessment can approach human agreement when calibration is stable. Performance differences between clip-level and trial-level evaluation were expected, as errors accumulate over longer temporal aggregation windows. Out-of-plane insertions yielded lower error overall, consistent with prior findings that reduced spatial ambiguity improves model performance.[7]

Although some perturbation conditions yielded large absolute values of Cohen's d, these cases were associated with very low run-to-run variance, which can substantially inflate effect size estimates even when absolute differences in MAE or RMSE are modest. This explains why large $|d|$ values are not visually associated with strong separation between distributions in Figs. 3 and 4. Consistent with the non-inferiority analysis, these large effect sizes do not reflect meaningful performance degradation.

Across perturbation experiments, model performance was largely preserved and in several cases modestly improved relative to the unperturbed baseline. This behavior contradicts the initial hypothesis that calibration error would degrade skill prediction. Instead, the results suggest that transformation noise may function as a form of implicit regularization, encouraging the model to rely on stable, skill-relevant temporal motion patterns rather than absolute spatial alignment. Similar robustness effects have been reported in prior work on noise injection for improving generalization in deep learning models.[5]

6. LIMITATIONS AND FUTURE WORK

This study is a simulation analysis of sensor calibration error. Perturbations were synthetically applied to recorded transformations and may not fully capture the complexity, temporal structure, or failure modes of real-world sensor miscalibration. Additionally, results are specific to ultrasound-guided needle insertion and may not generalize to other procedures or sensor configurations without further validation.

Future studies should evaluate these findings under real sensor miscalibration and across a broader range of procedures to confirm their clinical relevance. If validated in real-world settings, this robustness could reduce the need for frequent recalibration, lowering setup burden and improving the practicality of sensor-based skill assessment in simulation-based training and image-guided interventions.

7. CONCLUSION

This study demonstrated that machine learning–based surgical skill assessment can remain accurate under substantial sensor calibration perturbations. Contrary to the initial hypothesis, both large and small spatial errors did not degrade performance and, in some cases, were associated with modest improvements. These results suggest that temporal models may primarily encode relative motion patterns rather than relying on precise absolute spatial alignment.

ACKNOWLEDGMENTS

This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2020-05582. This research was enabled in part by support provided by https://www.computeontario.ca and the Digital Research Alliance of Canada https://alliancecan.ca.

REFERENCES

[1] Armitage, H., "The right moves: Tracking surgeons' motions to understand success." *Stanford Medicine: Data Sciences*, Issue 1 (April 2020). Available: https://stanmed.stanford.edu/analyzing-motion-data-pinpoint-surgical-success/.

[2] Foumani, N. M., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., and Salehi, M., "Deep learning for time series classification and extrinsic regression: A current survey," *ACM Computing Surveys* **1**(1), 47 (2023).

[3] Kulik, D., Bell, C. R., and Holden, M. S., "Fast skill assessment from kinematics data using convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery* , 1–7 (2023).

[4] Xia, S., Keri, Z., Holden, M. S., Hisey, R., Lia, H., Ungi, T., Mitchell, C. H., and Fichtinger, G., "A learning curve analysis of ultrasound-guided in-plane and out-of-plane vascular access training with perk tutor," in [*Proc. SPIE Medical Imaging*], **10576**, 1057625 (2018).

[5] Camuto, A., Willetts, M., Şimşekli, U., Roberts, S., and Holmes, C., "Explicit regularisation in gaussian noise injections," in [*Advances in Neural Information Processing Systems (NeurIPS)*], (2020).

[6] Drecogna, S., Misitano, A., Cabras, L., Pisano, A., Anedda, F., and Moncada, G., "Continuous glucose monitoring (cgm) accuracy enhancement using synthetic datasets and machine learning models," *Biomedinformatics* **4**(2), 1519–1530 (2024).

[7] Liu, R. and Holden, M. S., "Kinematics data representations for skills assessment in ultrasound-guided needle insertion," in [*Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*], 183–190, Springer International Publishing (2020).

[8] Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., Murakami, S., Saeki, S., Mukaida, H., and Takiyama, W., "Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room," *Surgery Today* **43**, 271–275 (Mar. 2013).

[9] GTSimulators.com, "Branched 2 vessel ultrasound training block." Accessed April 10, 2025.

[10] Yaniv, Z., "Which pivot calibration?," in [*Proc. SPIE Medical Imaging*], **9415**, 94153U (2015).

[11] Welch, M., Andrea, J., Ungi, T., and Fichtinger, G., "Freehand ultrasound calibration: phantom versus tracked pointer," in [*Proc. SPIE Medical Imaging*], **8671**, 86711C (2013).

[12] Ungi, T., Sargent, D., Moult, E., Lasso, A., Pinter, C., McGraw, R. C., and Fichtinger, G., "Perk tutor: An open-source training platform for ultrasound-guided needle insertions," *IEEE Transactions on Biomedical Engineering* **59**(12), 3475–3481 (2012).

[13] Dempster, A., Petitjean, F., and Webb, G. I., "Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery* **34**(5), 1454–1495 (2020).