Machine Learning Methods for Automated Technical Skills Assessment with Instructional Feedback in Ultrasound-Guided Interventions

Matthew S. Holden[1]*, Sean Xia[1], Hillary Lia[1], Zsuzsanna Keri[1], Colin Bell[2], Lindsey Patterson[3], Tamas Ungi[1], Gabor Fichtinger[1]

[1]Laboratory for Percutaneous Surgery, School of Computing, Queen's University, Kingston, ON, Canada

[2]Department of Emergency Medicine, School of Medicine, Queen's University, Kingston, ON, Canada

[3]Department of Anesthesiology and Perioperative Medicine, School of Medicine, Queen's University, Kingston, ON, Canada

*Email: 72mh@queensu.ca

Abstract

Objective: Currently, there is a worldwide shift towards competency-based medical education. This necessitates the use of automated skills assessment methods during self-guided interventions training. Making assessment methods that are transparent and configurable will allow assessment to be interpreted into instructional feedback. The purpose of this work is to develop and validate skills assessment methods in ultrasound-guided interventions that are transparent and configurable.

Methods: We implemented a method based upon decision trees and a method based upon fuzzy inference systems for technical skills assessment. Subsequently, we validated these methods for their ability to predict scores of operators on a 25-point global rating scale in ultrasound-guided needle insertions and their ability to provide useful feedback for training.

Results: Decision tree and fuzzy rule-based assessment performed comparably to state-of-the-art assessment methods. They produced median errors (on a 25-point scale) of 1.7 and 1.8 for in-plane insertions and 1.5 and 3.0 for out-of-plane insertions, respectively. In addition, these methods provided feedback that was useful for trainee learning. Decision tree assessment produced feedback with median usefulness 7 out of 7; fuzzy rule-based assessment produced feedback with median usefulness 6 out of 7.

Conclusion: Transparent and configurable assessment methods are comparable to the state-of-the-art and, in addition, can provide useful feedback. This demonstrates their value in self-guided interventions training curricula.

Keywords: ultrasound-guided needle insertion, simulation-based training, medical education, objective skill assessment

Introduction

Globally, skills training for medical interventions is transitioning from a time-based model to a competency-based model. Under the old time-based model, trainees would practice an intervention for a fixed amount of time, at which point they would be deemed competent and graduate, or they would be deemed incompetent and have to undertake significant remediation. Under the new competency-based model, trainees practice until they achieve a predefined competency benchmark. This scheme allows each trainee to practice the precise amount of time they need to achieve competency. The drawback of this method is that trainees' competency needs to be continually monitored.

Expert-based methods for skills assessment include checklists, global rating scales, and entrustments scores. Checklists are application-specific rubrics which assess whether the operator performs each step in the intervention correctly [1]. Global rating scales (GRS) offer application-independent assessment of interventions across several different facets [2]. Entrustment scores assess to what degree a supervisor trusts the trainee to complete each face of the intervention [3]. While these methods provide reliable assessment, in particular when combined [4], they rely on experts. With increasing medical class sizes and demands on expert time, it is not feasible to implement expert-based assessment on a wide scale. Instead, skill assessment should be automated.

Automated skills assessment can be applied to many different interventions (e.g. laparoscopy, open surgery, needle insertion, etc.), and can use data from many sources (e.g. instrument tracking, video, surgeon status, patient monitors) [5], [6]. Perhaps the most common method for automated skills assessment is metrics-based assessment. Under this paradigm, clinical experts specify what aspects of the intervention are relevant to operator skill. Subsequently, these can be implemented into a set of performance metrics: quantities that are understandable to trainees and clinicians and can be readily computed from measurable data. From these performance metrics, an overall skill level can be derived using pattern recognition or machine learning approaches.

Metrics-based overall skills assessment was initially addressed as an optimization problem, where each metric is treated as a cost and the most skillful operator is the one who best minimized the weighted sum of costs [7], [8]. Since, pattern recognition approaches have been used to achieve improved reliability in assessment. Chmarra et al. showed that linear discriminant analysis reliably distinguishes novices from intermediates from experts in laparoscopic training tasks [9]. Likewise, Allen et al. showed that support vector machines outperform cost-based approaches for skill classification in laparoscopic training tasks [10]. Oropesa et al. also demonstrated that support vector machines outperform linear discriminant analysis and adaptive-neuro fuzzy inference systems for laparoscopic training tasks [11]. Ahmidi et al. use support vector machines for skill classification for several difference types of performance metrics in septoplasty [12]. Fard et al. contrasted support vector machines with k-nearest neighbors and logistic regression for identifying novices and experts in robotic suturing tasks on real patients [13]. Kramer et al. have suggested learning vector quantization and self-organizing maps for assessment in simulated vascular surgery [14]. Neural network-based approaches have seen some success [15]. Fuzzy pattern recognition approaches have also gained some traction, including rule-based methods [16], [17] and adaptive fuzzy inference systems [18].

In consultation with clinical experts, we suggest two criteria which metrics-based skills assessment methods should meet in order to be clinically useful: transparency and configurability. A machine learning approach is considered transparent if both the model is easy to interpret and the principal of the method is easily understood [19], [20]. A machine learning approach is considered configurable if it has parameters which can be configured to improve performance based on domain-knowledge from a domain expert (Chiticariu et al. consider this a component of transparency [20]). In interventional skills assessment, transparency allows both the supervisor and trainee to understand why the trainee received a particular score and to interpret their results into actionable strategies to improve performance. Configurability allows the expert to adjust the assessment to their particular training scenario or to emphasize particular skills.

Of course, there are other methods for interventional skills assessment that are not based on performance metrics. In particular, temporal modelling [21], process monitoring [22], and end-to-end deep learning approaches [23] have shown some promise for skills assessment. Unfortunately, these methods do not provide adequate transparency to allow trainees and supervisors to interpret results into actionable feedback to improve performance. Crowd-sourcing can also provide accurate skills assessment and is effectively automated [24], but cannot provide immediate feedback.

The objective of this work is to develop and validate methods for overall skills assessment in percutaneous interventions. The methods should be transparent, configurable, and conducive to self-guided training. Subsequently, we evaluated (1) the accuracy of the proposed methods compared to state-of-the-art computer-assisted assessment and (2) the usefulness of the feedback provided by our proposed methods.

A preliminary version of this work has been reported [25].

Methods

Skills Assessment Algorithms

We aim to implement skills assessment algorithm which are transparent and configurable. Transparency and configurability are inherently subjective and fuzzy. In a review of common machine learning techniques, Kotsiantis identified three techniques as highly transparent: decision trees, naïve Bayes, and rule-based learners [19]. Naïve Bayes is more applicable for classification, and performs poorly for regression tasks [26]. This leaves decision trees and rule-based learners as transparent and configurable machine learning approaches for interventional skills assessment. Each skills assessment method takes a set of performance metrics as input (i.e. feature vector) and computes an overall skill level as output.

Decision Tree Assessment

For transparent and configurable assessment using decision trees, we use importance-aided decision trees [27] (Figure 1). This method is intended to incorporate domain-knowledge into decision trees, especially in lower levels of the decision tree when training data is limited.

In decision tree learning, a split is made based on the attribute and value which optimizes some measure of purity of each branch. In our case, for regression, we choose the within-branch variance as the attribute selection score. Following Al Iqbal et al., for an attribute x, we create a new attribute selection score S based on a linear combination of the within-branch variance score S_v and the attribute's weight W [27].

$$S(x) = (1 - \rho)S_v(x) + (\rho)W(x)$$

We select the attribute and split point which optimizes this new attribute selection score S. The coefficient in the linear combination ρ grows inversely with the proportion of remaining training samples in the branch [27]. The splitting is stopped once the within-branch variance decreases beyond a certain threshold. At this point, all training instances in the branch will effectively have the same skill level. We observe that in the case of equal attribute weights, this functions in the same way as a classical decision tree.

As identified by Kotsiantis, this method is transparent [19]. The user is presented with the traversal of the decision tree and the splitting criteria. As actionable feedback, we can identify the metrics associated with splits in the traversal where the branch center changed most. The feedback “well done” is provided when all splits in the traversal result in positive change in the branch center. This method is configurable in that the weights associated with each attribute can be adjusted. As demonstrated by Al Iqbal et al., incorporating this domain-knowledge into the decision tree can improve the accuracy of assessment [27].

Fuzzy Rule-Based Assessment

For rule-based assessment that is transparent and configurable, we use a set of fuzzy inference rules (Figure 1). In particular, we choose to use rules of the form: IF <metric> is <skill level> THEN operator is <skill level>. For example, IF elapsed time is expert THEN operator is expert. Such a rule is defined for each metric and skill level pair.

In practice, this requires us to define a membership function for each skill class and a membership function for each metric for each skill class. We define the skill class membership functions as symmetrical triangular functions on the range [0, 1] overlapping such that membership over all classes sums to one [16], [17]. The metric membership function for each skill class is computed empirically from the training data by Gaussian kernel density estimation, using the Silverman rule-of-thumb to estimate the bandwidth [28]. Importantly, each training instance may have membership in multiple skill classes and contribute with different weight to multiple metric membership functions.

We use clipping based on the membership in the input function to compute the output membership function for each rule. The set of fuzzy rules is combined and defuzzified by computing the mean of the maximum of the output membership functions.

This rule-based assessment method is transparent [19]. The user is presented with the rules that were applied and their strengths. As actionable feedback, we can identify the metrics for which the net influence of all associated fuzzy rules is the strongest. The feedback “well done” is provided when for each metric the net influence of all rules associated with that metric is positive. This method can be configured by allowing the weights associated with each rule to be adjusted or fuzzy rules to be added or removed. In particular, experts can add more sophisticated rules based on their domain-knowledge for improved accuracy.

Validation of Assessment Accuracy

We validated our assessment methods on both in-plane and out-of-plane needle insertions on a vascular access phantom (CAE Healthcare), following the setup used in Xia et al. [29]. We recorded 19 trainees and 5 experts performing in-plane insertions and 19 trainees and 5 experts performing out-of-plane insertions. Trainees were recorded at two points during a training curriculum (Figure 2). Experts were each recorded once. Operators used a Telemed MicrUs linear ultrasound probe (Telemed Medical Systems). We recorded videos of participants’ hands and tracked the needle and ultrasound probe. Tools were tracked with the Ascension trakStar (Northern Digital Inc.), and data was recorded using the PLUS Toolkit (www.plustoolkit.org) [30] and Perk Tutor (www.perktutor.org) [31]. We computed eight metrics for in-plane insertions and seven metrics for out-of-plane insertions (Table 1) [29]. These metrics were designed based on consultation with clinical experts, and they are intended to cover all relevant aspects of the ultrasound-guided needle insertion tasks.

As ground-truth assessment, we did not use participants' level of training. Instead, we recruited three clinical experts to assess participants' performance via anonymized hand motion videos using a previously validated global rating scale [32], [33]. The mean overall expert assessment provides a ground-truth skill level out of 25.

To determine the weight associated with each metric, we interviewed the same three clinical experts who provided ratings on the global rating scale, and we asked them to rate the importance of each metric for skills assessment on a seven-point Likert scale. We linearly scaled these ratings onto the interval [0, 1].

Subsequently, we validated the performance of our proposed assessment methods using leave-one-user-out cross-validation. We computed difference in the output of the proposed assessment methods with the mean expert rating. We then compared these results with the results achieved from several standard methods: (1) zero-rule regression (i.e. always guessing the mean scores), (2) linear regression, an empirically optimal version of the sum of z-scores method [8], (3) support vector machine regression, which has been shown to achieve state-of-the-art results in several assessment tasks [10]–[13], (4) nearest-neighbor regression with sequential forward feature selection, which achieves highly accurate assessment in suturing and knot tying [34], [35], (5) random regression forests, a generalization on decision tree regression. To compare the methods, we used a Friedmann test with pairwise Dunn's post hoc tests with Bonferroni correction (α=0.05). To determine whether our assessment methods are comparable to these other methods, we performed non-inferiority sign-rank tests (α=0.05) with the pooled standard deviation in expert ratings as the non-inferiority margin.

To determine the added value of expert-configured assessment, we used Bonferroni-corrected sign-rank tests (α=0.05) to compare unconfigured assessment with expert-configured assessment. We tested: (1) assessing the mean expert-assigned score using the mean expert configuration and (2) assessing each expert-assigned score using each expert's respective configuration.

Validation of Feedback Accuracy

To assess the quality of feedback provided by our methods, we mapped each metric to a plain-language description (Table 2). This was done in consultation with our clinical experts to ensure the vocabulary covers all possible feedback an expert might provide to trainees during in a typical training scenario. Subsequently, we asked one expert to review each trainee's post-training video (as this was identified by experts as the most useful stage for feedback). At the end of each video, we showed the expert all the different feedbacks and asked them to rate the usefulness of each one on a seven-point Likert scale (1 = strongly disagree that feedback was useful; 4 = neutral; 7 = strongly agree that feedback was useful).

We compared the usefulness of the feedback provided by the proposed methods with the usefulness of the kth most useful feedback by sign-rank test (α=0.05), for all k. We report the smallest k for which the predicted feedback is significantly more useful than the kth most useful feedback provided by the expert. This provides evidence of the usefulness of the proposed methods relative to expert feedback, without being skewed by the fact that experts found the majority of feedbacks to be useful. We also report confusion matrices for the truly most useful feedback compared to the predicted feedback.

Results

Assessment Accuracy

For ground-truth skill, the average measures intraclass correlation coefficient was 0.90 for the in-plane insertions and 0.93 for the out-of-plane insertions, indicating good reliability. For decision tree assessment and fuzzy rule-based assessment respectively, the median errors were 1.7 and 1.8 for in-plane insertions and 1.5 and 3.0 for out-of-plane insertions (Figure 3). Post hoc tests revealed decision tree assessment significantly outperformed all methods except support vector machine assessment (Table 3).

Decision tree assessment was non-inferior to all other assessment methods for both in-plane and out-of-plane insertions. Fuzzy rule-based assessment was non-inferior to all other assessment methods for in-plane insertions. For out-of-plane insertions, however, significance was not achieved. In fact, for out-of-plane insertions, fuzzy rule-based assessment was non-inferior to only zero-rule and nearest neighbor with sequential forward feature selection.

Reliability in the mean expert-defined weights was poor. The average measures intraclass correlation coefficient was 0.49 for in-plane insertions and 0.32 for out-of-plane insertions. When we used the expert-defined weights in the configurable assessment methods, the change in accuracy was insignificant (Figure 4).

Feedback Accuracy

The usefulness of the feedback was rated a median 7 out of 7 and a mean 5.8 out of 7 on a Likert scale for decision tree assessment. Likewise, the usefulness of the feedback was rated a median 6 out of 7 and a mean 5.3 out of 7 on a Likert scale for fuzzy rule-based assessment (Figure 5). Decision tree assessment produced useful feedback 74% of the time, and fuzzy rule-based assessment produced useful feedback 63% of the time (feedback rated 5, 6, or 7 out of 7 on a Likert scale). Furthermore, both methods produced significantly better than neutral feedback. Confusion matrices illustrate the most commonly misclassified feedback (Figure 6).

Compared to expert feedback, we found that for in-plane insertions, both decision tree assessment and fuzzy rule-based assessment produced significantly better feedback than the 5th best expert feedback. For out-of-plane insertions, decision tree assessment produced significantly better feedback than the 3rd best expert feedback, and fuzzy rule-based assessment produced significantly better feedback than the 5th best expert feedback. In all cases, the feedback produced by the proposed methods was better than the median expert feedback, but this was significant only for decision tree assessment in out-of-plane insertions.

Discussion

The results show that transparent and configurable assessment methods (1) perform comparably to state-of-the-art methods and (2) provide useful feedback for training. In particular, decision tree assessment performed the most accurately and provided the most useful feedback for our dataset. We did not observe a significant change in assessment accuracy when experts configured the proposed methods based on their domain knowledge [36], [37]. We believe the lack of significant improvement in the presented results may be due to our definition of ground-truth skill as a sum of global rating scale scores, without considering the importance of each aspect. Furthermore, the experts found all the metrics that were defined to be useful on average (rated as 5 or higher out of 7), and thus, the expert configurations are not substantially different from the default configuration.

We have identified our methods as transparent and identified methods such as support vector machines as opaque. While we have followed the work of Kotsiantis in identifying machine learning techniques as transparent [19], these classifications are inherently fuzzy. Although there is ongoing work in introspection in deep learning [38] allowing users to gain some understanding of how the deep model reached the result, it is unclear how well such methods will be accepted into practice [39].

One of the limitations of our feedback is the finite nature of our feedback vocabulary. While our feedback vocabulary was generated in consultation with experts to cover every aspect of the intervention, it does not allow feedback to be tailored to a particular trainee, as preceptors would do in practice. We observed that experts rated the top feedback as a median of 7 out of 7, indicating they agreed that the feedback from the vocabulary was indeed useful.

Another challenge of this work was determining the weights for each feature. We used a Likert scale to capture experts' opinion about the importance of each aspect of the intervention, and linearly scaled these responses to weights. But we observed that there was poor consistency between experts. This indicates that each expert may value different aspects of ultrasound-guided insertions. Our methods would allow the assessment to be tailored to each expert individually.

Although there are five experts and nineteen trainees for each of in-plane and out-of-plane insertions, we observe that ground-truth global rating scores cluster towards the higher end of the scale. This creates a problem of unbalanced regression and may affect the reported results.

We observed that in 8% and 29% of cases, "well done" was incorrectly predicted as the most useful feedback, for decision tree and fuzzy rule-based assessment, respectively. But the feedback "well done" may not be the most instructive for trainees. The proposed methods could be adapted to provide this feedback less frequently. For decision tree assessment, this could be achieved by requiring all splits in the traversal to have a change in branch center above a certain threshold. Analogously for fuzzy rule-based assessment, the could be achieved by requiring the net influence of all rules associated with each metric to be above a certain threshold. The threshold value could be tuned to optimize a sensitivity and specificity criterion.

We have shown that the proposed methods work effectively for skills assessment and feedback in both in-plane and out-of-plane ultrasound-guided needle insertions. Our setup has shown evidence for face and content validity [29]. Because the overwhelming majority of ultrasound-guided interventions use one of these approaches, we suggest the results will apply to most ultrasound-guided interventions. Recent work has shown it takes approximately 85 practice attempts to reach proficiency in ultrasound-guided needle insertions [40]. Our experts believe that the feedback provided by our system will be most applicable after 10 practice attempts when the trainee has fully understood the basics of the intervention.

Our results are consistent with other work demonstrating the utility of metrics-based assessment of interventional skills [7], [8]. In the context of ultrasound-guided needle insertions, we have shown that transparent and configurable methods are comparable to state-of-the-art methods for assessment but, in addition, can provide useful feedback.

In the future, we suggest further study into how the proposed methods perform in specific ultrasound-guided interventions (e.g. biopsy, epidural, central line) and how they may be extended to other types of interventions. It has been previously shown that generic performance metrics may not be equally applicable to all interventions [41] and application-specific metrics provide added value over generic metrics [42]. Thus, in order to extend these methods to other interventions, it is necessary to develop application-specific performance metrics and a feedback vocabulary in consultation with expert clinicians. These are the only places where we have infused application-specific knowledge into the proposed methods.

We also suggest a future longitudinal study examining the effect of providing the proposed computer-generated feedback on trainee learning. Such a study could better identify the added value of the proposed feedback methods over self-guided training without feedback. Prior work has shown that feedback through 3D visualization can improve ultrasound-guided interventions learning [43], but has not evaluated the impact of targeted feedback.

We make the proposed methods available to the community through Perk Tutor (www.perktutor.org) [31].

Conclusion

We have demonstrated that transparent and configurable skills assessment methods are comparably accurate to state-of-the-art methods. In contrast to state-of-the-art methods, however, transparent and configurable methods were shown to provide useful feedback for training. Importance-aided decision tree assessment provided the most accurate assessment with feedback.

Thus, transparent and configurable assessment methods can be adopted into practice to provide feedback without compromising accuracy. We have also demonstrated that they can be customized by experts to suit the particular application or emphasize particular skills.

We envision these methods could be employed in an ultrasound-guided interventions training curriculum. They would monitor trainee learning curves and provide automated instructions during self-directing learning. This would serve to supplement supervision and assessment from expert preceptors.

Acknowledgements

Matthew S. Holden is supported by the Link Foundation Fellowship in Modelling, Simulation, and Training. Gabor Fichtinger is supported as a Canada Research Chair in Computer-Integrated Surgery.

Compliance with Ethical Standards

Conflict of Interest

All authors declare that they have no conflict of interest.

Ethical Approval

All procedures in this study involving human participants were performed in accordance with the ethical standards of the institution, and were approved by the research ethics board at Queen's University. This study does not contain any procedures involving animals.

Informed Consent

All participation was voluntary, and written informed consent was obtained from all participants.

References

[1] R. M. Harden, M. Stevenson, W. W. Downie, and G. M. Wilson, "Assessment of clinical competence using objective structured examination.," *Br. Med. J.*, vol. 1, no. 5955, pp. 447–51, 1975.

[2] C. P. Winckel, R. K. Reznick, M. Frcsc, and R. Cohen, "Reliability and Construct Validity of a

Structured Technical Skills Assessment Form," *Am. J. Surg.*, vol. 167, pp. 423–427, 1994.

[3] W. T. Gofton, N. L. Dudek, T. J. Wood, F. Balaa, and S. J. Hamstra, "The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence," *Acad. Med.*, vol. 87, no. 10, pp. 1401–1407, 2012.

[4] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown, "Objective structured assessment of technical skill (OSATS) for surgical residents," *Br. J. Surg.*, vol. 84, no. 2, pp. 273–278, 1997.

[5] C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surg. Endosc.*, vol. 25, no. 2, pp. 356–366, Feb. 2011.

[6] S. S. Vedula, M. Ishii, and G. D. Hager, "Objective Assessment of Surgical Technical Skill and Competency in the Operating Room," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 301–325, Jun. 2017.

[7] S. A. Fraser, D. R. Klassen, L. S. Feldman, G. A. Ghitulescu, D. Stanbridge, and G. M. Fried, "Evaluating laparoscopic skills, setting the pass/fail score for the MISTELS system," *Surg. Endosc. Other Interv. Tech.*, vol. 17, no. 6, pp. 964–967, 2003.

[8] N. Stylopoulos, S. Cotin, S. K. K. Maithel, M. Ottensmeye, P. G. G. Jackson, R. S. S. Bardsley, P. F. F. Neumann, D. W. W. Rattner, S. L. L. Dawson, M. Ottensmeyer, P. G. G. Jackson, R. S. S. Bardsley, P. F. F. Neumann, D. W. W. Rattner, and S. L. L. Dawson, "Computer-enhanced laparoscopic training system (CELTS): bridging the gap," *Surg Endosc*, vol. 18, no. 5, pp. 782–789, May 2004.

[9] M. K. Chmarra, S. Klein, J. C. F. de Winter, F.-W. W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surg. Endosc. Other Interv. Tech.*, vol. 24, no. 5, pp. 1031–1039, May 2010.

[10] B. Allen, V. Nistor, E. Dutson, G. Carman, C. Lewis, and P. Faloutsos, "Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks," *Surg Endosc*, vol. 24, no. 1, pp. 170–178, Jan. 2010.

[11] I. Oropesa, P. Sánchez-González, M. K. Chmarra, P. Lamata, R. Pérez-Rodríguez, F. W. Jansen, J. Dankelman, and E. J. Gómez, "Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis," *Surg. Endosc. Other Interv. Tech.*, vol. 28, no. 2, pp. 657–670, Feb. 2014.

[12] N. Ahmidi, P. Poddar, J. D. Jones, S. S. Vedula, L. Ishii, G. D. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, no. 6, pp. 981–991, 2015.

[13] M. J. Fard, S. Ameri, R. Darin Ellis, R. B. Chinnam, A. K. Pandya, and M. D. Klein, "Automated robot-assisted surgical skill evaluation: Predictive analytics approach," *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1850, 2017.

[14] B. D. Kramer, D. P. Losey, M. K. O'Malley, and M. K. O'Malley, "SOM and LVQ Classification of Endovascular Surgeons Using Motion-Based Metrics," in *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, vol. 428, E. Merényi, M. J. Mendenhall, and P. O'Driscoll, Eds. Cham: Springer International Publishing, 2016, pp. 227–237.

[15] M. Uemura, M. Tomikawa, T. Miao, R. Souzaki, S. Ieiri, T. Akahoshi, A. K. Lefor, and M. Hashizume, "Feasibility of an AI-Based Measure of the Hand Motions of Expert and Novice Surgeons," *Comput. Math. Methods Med.*, 2018.

[16] I. Hajshirmohammadi and S. Payandeh, "Fuzzy set theory for performance evaluation in a surgical simulator," *Presence-Teleoperators Virtual Environ.*, vol. 16, no. 6, pp. 603–622,

2007.

[17] M. Riojas, C. Feng, A. Hamilton, and J. Rozenblit, "Knowledge elicitation for performance assessment in a computerized surgical training system," *Appl. Soft Comput. J.*, vol. 11, no. 4, pp. 3697–3708, 2011.

[18] J. Huang, S. Payandeh, P. Doris, and I. Hajshirmohammadi, "Fuzzy classification: towards evaluating performance on a surgical simulator.," *Stud. Health Technol. Inform.*, vol. 111, pp. 194–200, 2005.

[19] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[20] L. Chiticariu, Y. Li, and F. Reiss, "Transparent Machine Learning for Information Extraction: State-of-the-Art and the Future," in *Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 4–6.

[21] J. Rosen, J. D. Brown, L. Chang, M. Barreca, M. Sinanan, and B. Hannaford, "The BlueDRAGON - a system for measuring the kinematics and dynamics of minimally invasive surgical tools in-vivo," in *IEEE International Conference on Robotics and Automation*, 2002, vol. 2, pp. 1876–1881.

[22] G. Forestier, F. Lalys, L. Riffaud, B. Trelhu, and P. Jannin, "Classification of surgical processes using dynamic time warping," *J. Biomed. Inform.*, vol. 45, no. 2, pp. 255–264, 2012.

[23] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination," in *Computer Vision and Pattern Recognition*, 2018, p. in press.

[24] T. M. Kowalewski, B. Comstock, R. Sweet, C. Schaffhausen, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, J. Kaouk, B. Knudsen, J. Landman, B. Lee, B. F. Schwartz, E. McDougall, and T. S. Lendvay, "Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks," *J. Urol.*, vol. 195, no. 6, pp. 1859–1865, 2016.

[25] M. S. Holden, H. Lia, S. Xia, Z. Keri, T. Ungi, and G. Fichtinger, "Configurable Overall Skill Assessment in Ultrasound-Guided Needle Insertion," in *16th Annual Imaging Network Ontario Symposium (ImNO)*, 2018.

[26] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, "Technical note: Naive Bayes for regression," *Mach. Learn.*, vol. 41, no. 1, pp. 5–25, 2000.

[27] M. R. Al Iqbal, S. Rahman, S. I. Nabil, and I. U. A. Chowdhury, "Knowledge based decision tree construction with feature importance domain knowledge," in *2012 7th International Conference on Electrical and Computer Engineering*, 2012, pp. 659–662.

[28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, no. 1951. 1986.

[29] S. Xia, Z. Keri, M. S. Holden, R. Hisey, H. Lia, T. Ungi, C. H. Mitchell, and G. Fichtinger, "A learning curve analysis of ultrasound-guided in-plane and out-of-plane vascular access training with Perk Tutor," in *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2018, vol. 10576, p. 66.

[30] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "PLUS: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2527–2537, 2014.

[31] T. Ungi, D. Sargent, E. Moult, A. Lasso, C. Pinter, R. C. McGraw, and G. Fichtinger, "Perk Tutor: An Open-Source Training Platform for Ultrasound-Guided Needle Insertions," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3475–3481, Dec. 2012.

[32] K. Domuracki, A. Wong, L. Olivieri, and L. E. M. Grierson, "The impacts of observing flawed and flawless demonstrations on clinical skill learning," *Med. Educ.*, vol. 49, no. 2, pp. 186–

192, 2015.

[33] I. W. Y. Ma, N. Zalunardo, G. Pachev, T. Beran, M. Brown, R. Hatala, and K. McLaughlin, "Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation," *Adv. Heal. Sci. Educ.*, vol. 17, no. 4, pp. 457–470, 2012.

[34] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, "Automated Assessment of Surgical Skills Using Frequency Analysis," in *Medical Image Computing and Computer-Assisted Interventions - MICCAI 2015, Pt I*, 2015, vol. 9349, pp. 430–438.

[35] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, 2018.

[36] S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, and J. Herlocker, "Toward harnessing user feedback for machine learning," in *Proceedings of the 12th international conference on Intelligent user interfaces - IUI '07*, 2007, p. 82.

[37] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan, "EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers," *Learning*, pp. 1283–1292, 2009.

[38] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9908 LNCS, pp. 3–19.

[39] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Machine Stud.*, vol. 27, no. 5–6, pp. 527–539, Nov. 1987.

[40] R. McGraw, T. Chaplin, C. McKaigney, L. Rang, M. Jaeger, D. Redfearn, C. Davison, T. Ungi, M. Holden, C. Yeo, Z. Keri, and G. Fichtinger, "Development and Evaluation of a Simulation-based Curriculum for Ultrasound-guided Central Venous Catheterization," *CJEM*, pp. 1–9, May 2016.

[41] V. Datta, S. Mackay, M. Mandalia, and A. Darzi, "The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model," *J. Am. Coll. Surg.*, vol. 193, no. 5, pp. 479–485, 2001.

[42] M. S. Holden, Z. Keri, T. Ungi, and G. Fichtinger, "Overall Proficiency Assessment in Point-of-Care Ultrasound Interventions: The Stopwatch is not Enough," in *Imaging for Patient-Customized Simulations and Systems for Point-of-Care Ultrasound: International Workshops, BIVPCS 2017 and POCUS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings*, M. J. Cardoso, T. Arbel, J. M. R. S. Tavares, S. Aylward, S. Li, E. Boctor, G. Fichtinger, K. Cleary, B. Freeman, L. Kohli, D. Shipley Kane, M. Oetgen, and S. Pujol, Eds. Cham: Springer International Publishing, 2017, pp. 146–153.

[43] H. Lia, Z. Keri, M. S. Holden, V. Harish, C. H. Mitchell, T. Ungi, and G. Fichtinger, "Training with Perk Tutor improves ultrasound-guided in-plane needle insertion skill," in *SPIE Medical Imaging*, 2017, p. 101350T--101350T.

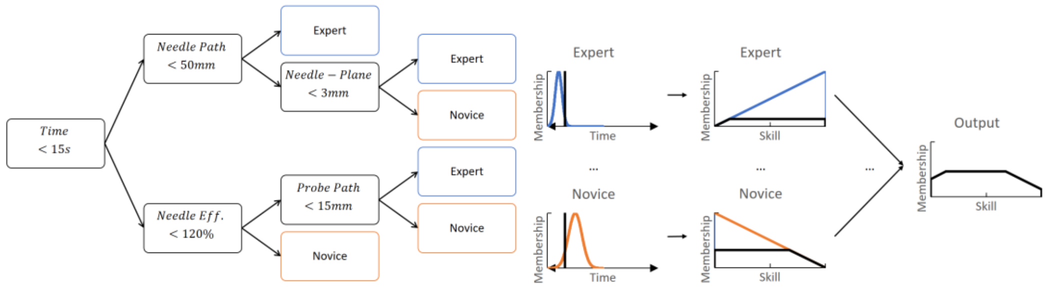

Figure 1. Illustration of importance-aided decision tree assessment (left) and fuzzy rule-based assessment (right) in ultrasound-guided needle insertion assessment using performance metrics.

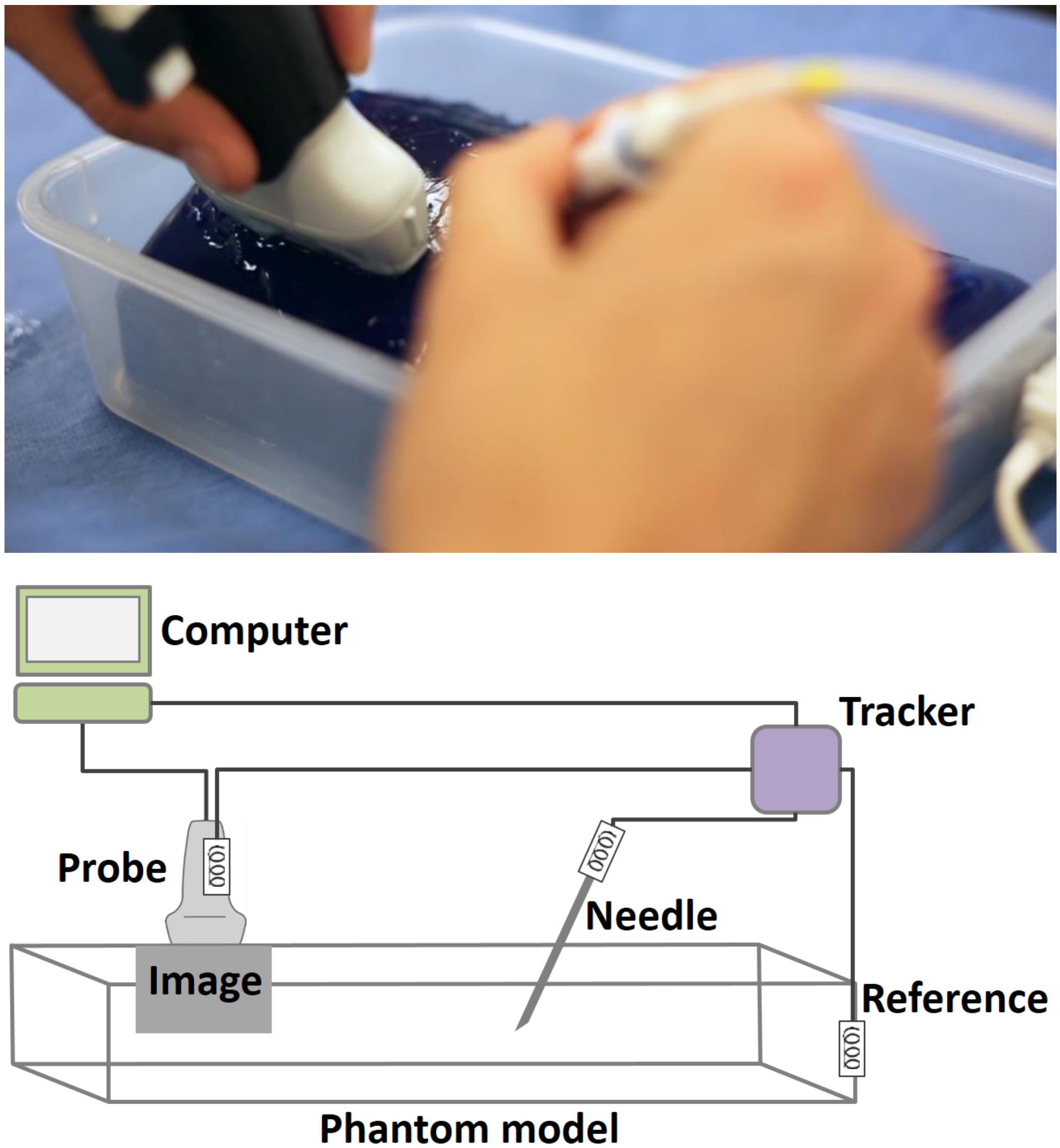

Figure 2. Photograph of a trainee participant performing an ultrasound-guided in-plane needle insertion (top) and schematic diagram of setup (bottom). The electromagnetic pose trackers used are attached to the base of the needle, base of the ultrasound probe, and exterior of the phantom.

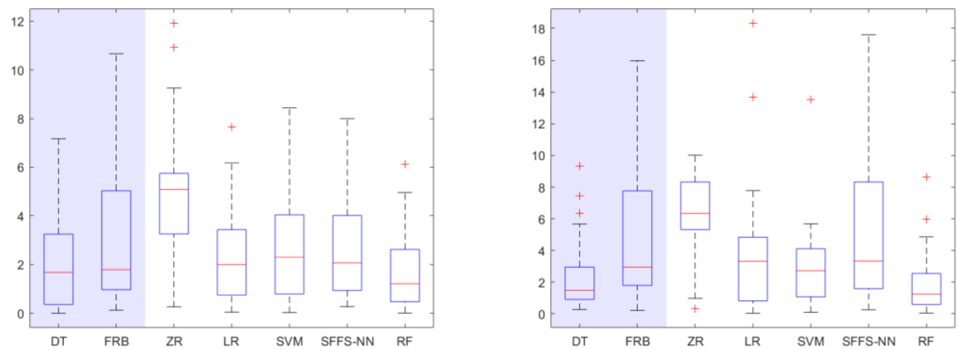

Figure 3. Error in assessment for the decision tree (DT), fuzzy rule-based (FRB), zero-rule (ZR), linear regression (LR), support vector machine (SVM), nearest neighbor with sequential forward feature selection (SFFS-NN), and random forest (RF) assessment methods for in-plane insertions (left) and out-of-plane insertions (right). Data from 24 users over 43 trials.

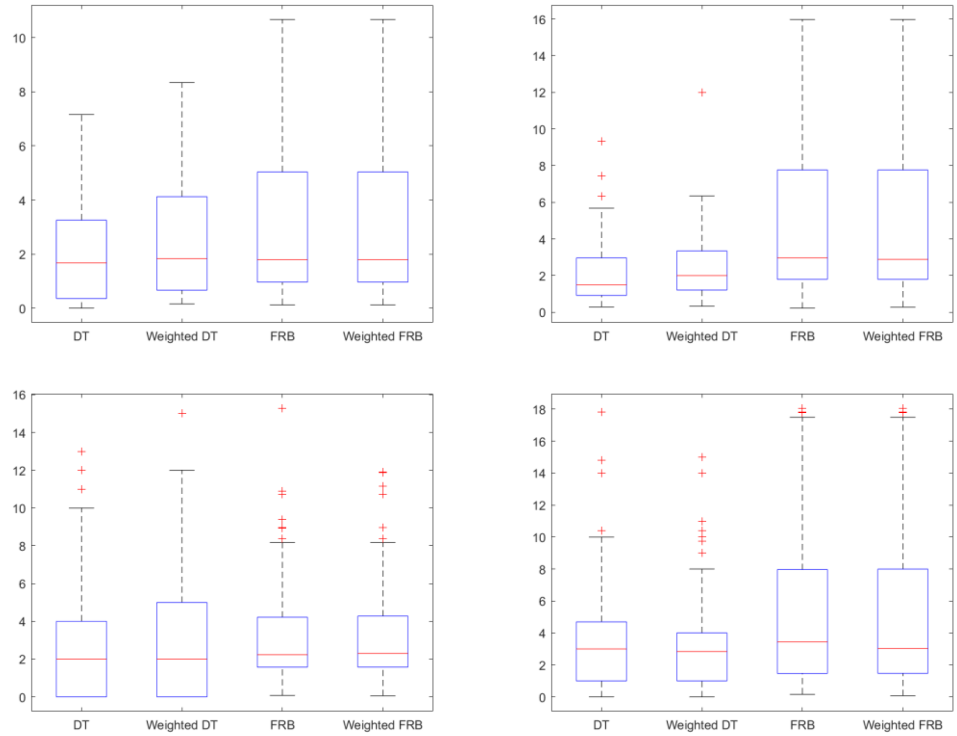

Figure 4. Error in assessment for decision tree (DT) and fuzzy rule-based (FRB) assessment methods with or without expert-defined weights for in-plane insertions (left) and out-of-plane insertions (right). Top row shows results from predicting the mean expert-assigned score using the mean expert configuration; bottom row shows results from predicting each individual expert-assigned score with each expert's respective configuration. Data from 24 users over 43 trials.

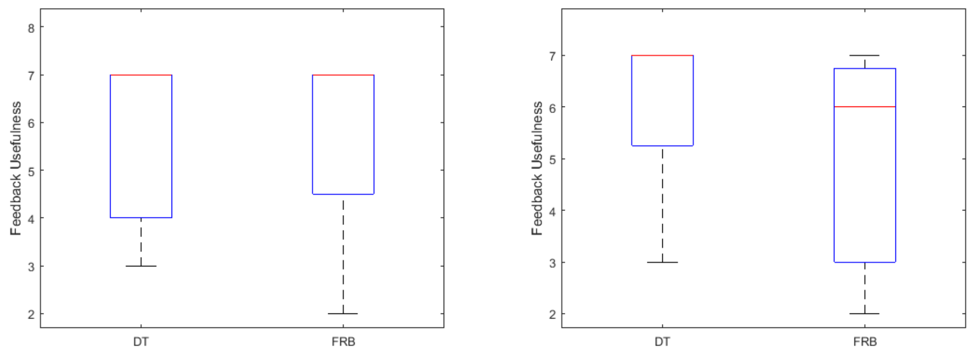

Figure 5. Usefulness of feedback produced by the decision tree (DT) and fuzzy rule-based (FRB) assessment methods for in-plane insertions (left) and out-of-plane insertions (right). Red line indicates median. Data from 24 users over 43 trials.

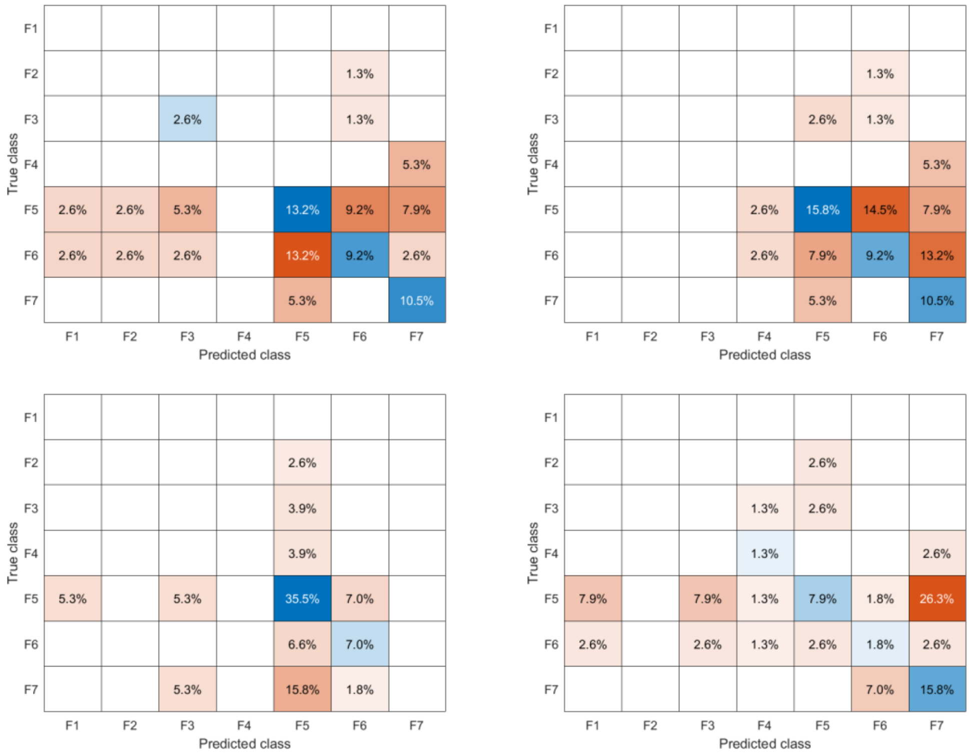

Figure 6. Confusion matrices illustrating errors in predicted feedback for in-plane (top) and out-of-plane (bottom) ultrasound-guided needle insertions using importance-aided decision tree (left) and fuzzy rule-based (right) assessment. Ties are distributed across all tied labels. Labels correspond to feedback vocabulary. Blue shading indicates correct predictions; red shading indicates incorrect predictions. Intensity of shading indicates higher concentration. Data from 19 users over 19 trials.

Table 1. Description of performance metrics for in-plane and out-of-plane insertions.

In-Plane Metrics	
Elapsed Time (s)	Total time from the start of the insertion to the end of the insertion
Needle path length (mm)	Total distance travelled by the tip of the needle
Probe path length (mm)	Total distance travelled by the foot of the ultrasound probe
Needle path efficiency (%)	Ratio of the needle's path length to the distance between the needle's start and end points
Average needle to ultrasound plane distance (mm)	Average orthogonal distance between the needle tip and the ultrasound plane
Maximum needle to ultrasound plane distance (mm)	Maximum orthogonal distance between the needle tip and the ultrasound plane
Average needle to ultrasound plane angle (°)	Average angle between the needle and the ultrasound plane
Maximum needle to ultrasound plane angle (°)	Maximum angle between the needle and the ultrasound plane
Out-of-Plane Metrics	
Elapsed Time (s)	Total time from the start of the insertion to the end of the insertion
Needle path length (mm)	Total distance travelled by the tip of the needle
Probe path length (mm)	Total distance travelled by the foot of the ultrasound probe
Needle path efficiency (%)	Ratio of the needle's path length to the distance between the needle's start and end points
Maximum distance needle is past ultrasound plane (mm)	Maximum orthogonal distance the needle tip travels past the ultrasound plane
Total time needle is past ultrasound plane (s)	Total time spent with the needle tip past the ultrasound plane
Average rotation from needle to ultrasound plane normal (°)	Average angle between the needle and the plane orthogonal to the ultrasound marked-unmarked vector

Table 2. Plain-language feedback associated with each performance metric for in-plane and out-of-plane insertions.

<table>
<tr><td colspan="2">In-Plane Metrics</td></tr>
<tr><td>Elapsed Time (s)</td><td>F1. Keep practicing with proper technique to improve your time efficiency.</td></tr>
<tr><td>Needle path length (mm)</td><td>F2. Look at the depth of your target, and try to estimate the correct angle of needle insertion.</td></tr>
<tr><td>Probe path length (mm)</td><td>F3. Get a longitudinal ultrasound image of the middle of the vessel and stabilize the probe using your hand/finger against the gel surface.</td></tr>
<tr><td>Needle path efficiency (%)</td><td>F4. Try to focus on a smooth, straight needle path while inserting the needle as close to the ultrasound plane as possible.</td></tr>
<tr><td>Average needle to ultrasound plane distance (mm)</td><td rowspan="2">F5. Start with the needle in the middle of the ultrasound probe and try to keep it aligned with the ultrasound plane during needle insertion.</td></tr>
<tr><td>Maximum needle to ultrasound plane distance (mm)</td></tr>
<tr><td>Average needle to ultrasound plane angle (°)</td><td rowspan="2">F6. Do not change the angle between the needle and the ultrasound plane during needle insertion. This will make sure that there is perfect alignment.</td></tr>
<tr><td>Maximum needle to ultrasound plane angle (°)</td></tr>
<tr><td></td><td>F7. Well done.</td></tr>
<tr><td colspan="2">Out-of-Plane Metrics</td></tr>
<tr><td>Elapsed Time (s)</td><td>F1. Keep practicing with proper technique to improve your time efficiency.</td></tr>
<tr><td>Needle path length (mm)</td><td>F2. Look at the depth of your target, and try to estimate the correct angle of needle insertion.</td></tr>
<tr><td>Probe path length (mm)</td><td>F3. Do not move the probe when advancing the needle. Advance the probe very slightly when the needle appears in the ultrasound image.</td></tr>
<tr><td>Needle path efficiency (%)</td><td>F4. Insert the needle in a straight, smooth path.</td></tr>
<tr><td>Maximum distance needle is past ultrasound plane (mm)</td><td rowspan="2">F5. Keep the ultrasound plane slightly ahead of the needle. If you see the needle tip on the screen, move the ultrasound slightly ahead until the needle disappears and then continue needle insertion until the needle appears again.</td></tr>
<tr><td>Total time needle is past ultrasound plane (s)</td></tr>
<tr><td>Average rotation from needle to ultrasound plane normal (°)</td><td>F6. Start with the target in the middle of the ultrasound screen, with the needle in the middle of the probe at 90° to the probe and 45° to the gel surface. Do not change this angle during the needle insertion.</td></tr>
<tr><td></td><td>F7. Well done.</td></tr>
</table>

Table 3. Results of post hoc testing for differences in decision tree (DT), fuzzy rule-based (FRB), zero-rule (ZR), linear regression (LR), support vector machine (SVM), nearest neighbor with sequential forward feature selection (SFFS-NN), and random forest (RF) assessment. Mean ranks indicates the mean rank of accuracy for the method when compared to the other methods. Significant indicates which methods were significantly different, and whether the method was more accurate (<) or less accurate (>).

In-Plane		
Assessment Method	Mean Rank	Significant
Decision Tree	3.09	<ZR
Fuzzy Rule-Based	4.40	>RF
Zero-Rule	5.74	>DT, LR, SVM, SFFS-NN, RF
Linear Regression	3.63	<ZR
Support Vector Machine	4.09	<ZR
SFFS-Nearest Neighbor	4.21	<ZR
Random Forest	2.84	<FRB, ZR
Out-of-Plane		
Assessment Method	Mean Rank	Significant
Decision Tree	3.19	<FRB, ZR, SFFS-NN
Fuzzy Rule-Based	4.63	>DT, SVM, RF
Zero-Rule	5.81	>DT, LR, SVM, RF
Linear Regression	3.88	<ZR
Support Vector Machine	3.21	<FRB, ZR
SFFS-Nearest Neighbor	4.60	>DT, RF
Random Forest	2.67	<FRB, ZR, SFFS-NN