FAST skill assessment from kinematics data using convolutional neural networks

Daniil Kulik[1], Colin R. Bell MD[2] and Matthew S. Holden PhD[1]

[1]School of Computer Science, Carleton University, 1125 Colonel By Dr, Ottawa, K1S 5B6, ON, Canada.
[2]Department of Emergency Medicine and Cumming School of Medicine, University of Calgary, 3330 Hospital Dr NW, Calgary, T2N 4N1, AB, Canada.

Contributing authors: Daniil.Kulik@carleton.ca; Colin.Bell@albertahealthservices.ca; Matthew.Holden@carleton.ca;

Abstract

Purpose: FAST is a point of care ultrasound (POCUS) study that evaluates for the presence free fluid, typically hemoperitoneum in trauma patients. FAST is an essential skill for Emergency Physicians. Thus, it requires objective evaluation tools that can reduce the necessity of direct observation for proficiency assessment. In this work, we use deep neural networks to automatically assess operators' FAST skills. **Methods:** We propose a deep convolutional neural network for FAST proficiency assessment based on motion data. Prior work has shown that operators demonstrate different domain-specific dexterity metrics that can distinguish novices, intermediates, and experts. Therefore, we augment our dataset with this domain knowledge and employ fine-tuning to improve the model's classification capabilities. Our model, however, does not require specific points-of-interest (POIs) to be defined for scanning. **Results:** The results show that the proposed deep convolutional neural network can classify FAST proficiency with 87.5% accuracy and 0.884, 0.886, 0.247 sensitivity for Novices, Intermediates and Experts, respectively. It demonstrates the potential of using kinematics data as an input in FAST skill assessment tasks. We also show that the proposed domain-specific features and region fine-tuning increase the model's classification accuracy and sensitivity.

Conclusion: Variations in probe motion at different learning stages can be derived from kinematics data. These variations can be used for automatic and objective skill assessment without prior identification of clinical POIs. The proposed approach can improve the quality and objectivity of FAST proficiency evaluation. Furthermore, skill assessment combining ultrasound images and kinematics data can provide a more rigorous and diversified evaluation than using ultrasound images alone.

Keywords: FAST Ultrasound, Skill assessment, Kinematics data, Surgical Data Science

1 Introduction

1.1 Motivation

Point-of-care ultrasound (POCUS) is the real-time diagnostic or interventional use of ultrasound by a clinician at the bedside, and is routinely performed in the emergency department. One of the common examinations is the Focused Assessment with Sonography in Trauma (FAST) [1]. FAST is a POCUS study that rapidly establishes the presence or absence of free fluid, typically hemoperitoneum, bile, urine or stomach contents, after a traumatic injury [2]. The protocol requires scanning four body regions: right upper quadrant (RUQ), left upper quadrant (LUQ), pericardial space (HEART), and pelvic area (PELVIC) [2]. POCUS imaging has many benefits over other radiographic imaging. It is portable, cost-effective and reduces patients' exposure to harmful radiation [3]. In particular illness presentations, including trauma, POCUS has been demonstrated to shorten lengths of stay at the hospital, streamline patient care and reduce complications [4].

FAST is a core POCUS competency that is included in medical undergraduate and postgraduate curricula [1] (Fig. 1). Thus, it warrants the employment of validated, objective assessment tools to ensure trainees' evaluation consistency [5, 6]. Most of the existing assessment tools rely on direct observation by experts and are subject to human error [5]. However, with the help of automated assessment, we can improve the education process by reducing the necessity of direct expert observation with increased skill evaluation accuracy and consistency [7, 8].

The main objective of this work is to propose a new approach to training a convolutional neural network (CNN) to improve its performance in skill evaluation. The model will train on kinematics data combined with domain-specific features with further region specific fine-tuning. The kinematics data was recorded from an ultrasound probe during a series of FAST procedures. Furthermore, this work investigates a different perspective on POCUS skill evaluation that focuses not on acquired images but on the operator's hand and probe motion. They can ensure the acquisition of high-quality images, which leads to optimal interpretation and clinical integration [9].

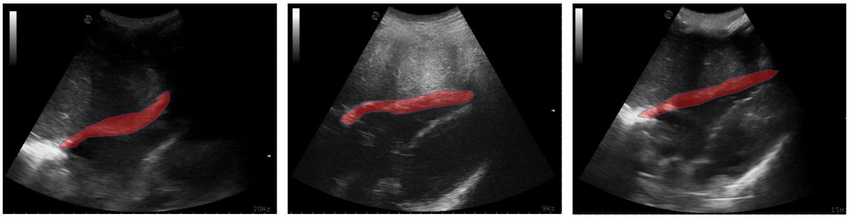

Fig. 1 Ultrasound images of the pericardial space. **Left:** (novice) right ventricle to septum incompletely visualized, cardiac apex obscured by bowel gas. **Middle:** (intermediate) good visualization of the pericardium and structures, more fatpad is visualized in the near screen. **Right:** (expert) pericardium visualized with intraventricular towards apex

1.2 Previous work

At present, the prevailing method of POCUS competency assessment is based on direct expert observation [5]. There are several FAST assessment tools with a comprehensive list of scoring metrics, which include probe motion and positioning [8, 10] though none are in widespread use [5, 7]. Therefore, kinematics data collected from the probe during FAST can be used to evaluate operators' proficiency and track their progression [11].

Previous work on FAST kinematics data analysis for proficiency evaluation is focused on summary statistics [7, 12–14]. [7, 13] demonstrate that trained operators require less time and shorter path length with less probe motion to complete FAST compared to operators with less experience.

Modern approaches to skill assessment from kinematics data employ deep learning models. Primarily, surgical skill assessment from time-series data with CNNs dominating the field [15–18]. CNNs are a type of deep learning model used in different domains, from time-series analysis to video recognition tasks [19]. One of the main advantages of CNNs is their ability to automatically learn hierarchical representations of the input data, which can lead to better performance than traditional machine learning models that rely on handcrafted features [19]. CNN-based models demonstrated high performance in skill assessment tasks on the open-source JIGSAW dataset [20]. Moreover, it was shown that CNN skill assessment correlates with existing expert-based proficiency assessment tools [21].

2 Methods

2.1 Dataset

Three groups of people (fourteen novices, fifteen intermediates, and three experts) with self-reported proficiency participated in the data collection [7]. The novices had limited FAST experience obtained during their dedicated curriculum. The intermediates completed the same curriculum with at least 50 supervised FAST diagnostics on patients. The experts were FAST instructors who completed additional POCUS training.

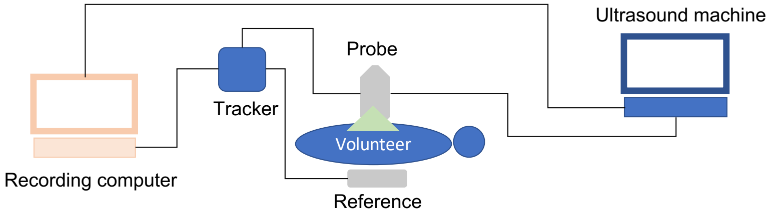

Fig. 2 Experiment setup for tracking and recording kinematics data

The data was collected in a simulated environment on a healthy volunteer with a sensor attached to the ultrasound probe, a reference sensor placed under the volunteer's back and a tracking device (see Fig. 2). The participants were asked to complete the FAST examination and scan all four regions of interest separately. The recorded data contains sequences of transformation matrices from the probe sensor to the reference sensor ($T_{probe \to ref}$), with a timestamp for each matrix. The motion data was recorded using the open-source software SlicerIGT with the PLUS software library extension [22]. To our knowledge, this is the only available dataset of this kind.

2.2 Data processing

To make the data robust to the reference position on the patient's back, we compute the transformations relative to the previous transformation as follows

$$T'_{probe \to ref_i} = T^{-1}_{probe \to ref_{i-1}} \cdot T_{probe \to ref_i} \tag{1}$$

Additionally, we propose the use of domain knowledge by including expert-defined summary statistics in the data as one of the contributions of this work. Previous studies show that operators with better skills show more efficient probe movements with less time to complete the procedure [7, 12–14]. Thus, we augment each transformation matrix with path length L_i defined as follows

$$L_i = \|x_i - x_{i-1}\|_2, \tag{2}$$

where x_i and x_{i-1} are the current and previous position of the probe sensor relative to the reference sensor. Also we add the relative time difference Δt_i as

$$\Delta t_i = t_i - t_{i-1}, \tag{3}$$

where t_i and t_{i-1} are the timestamps of the current transformation and the previous transformation. To further emphasize the importance of hand motion in the dataset we compute angular and linear speed. The angular speed A_i is computed as

$$A_i = \frac{\cos^{-1}\left(\frac{Trace\left(R_{probe \to ref_i} \cdot R^{-1}_{probe \to ref_{i-1}}\right) - 1}{2}\right)}{\Delta t_i}, \tag{4}$$

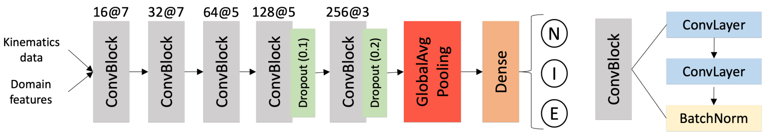

Fig. 3 The proposed network architecture. The top values show number of filters @ kernel size with dropout rate provided in the brackets

where $R_{probe \to ref_i}$ and $R^{-1}_{probe \to ref_{i-1}}$ are rotation matrices of $T'_{probe \to ref_i}$ and $T'^{-1}_{probe \to ref_{i-1}}$, respectively. Finally, we add linear speed S_i to the data as follows

$$S_i = \frac{L_i}{\Delta t_i}. \tag{5}$$

The original dataset consists of homogeneous transformation matrices

$$T'_{probe \to ref_i} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & d_x \\ r_{21} & r_{22} & r_{23} & d_y \\ r_{31} & r_{32} & r_{33} & d_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

where r_{ij} are rotation components and $d_{x,y,z}$ are translation components. We flatten the first three rows of the matrix, append the previously calculated domain-specific values and reshape the resulting matrices into vectors $\vec{v}_{probe \to ref_i} \in R^{16 \times 1}$ as follows

$$\vec{v}_{probe \to ref_i} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & d_x & \cdots & L_i & A_i & S_i & \Delta t_i \end{pmatrix}^T \tag{7}$$

We address the potential overfitting that is typical for CNNs by employing a window slicing approach to augment time series data as was proposed in [23]. Since the original dataset has only 32 training instances of variable length time-series, we use window slicing to increase the number of training samples and get fixed length time-series samples. Each sample has 70 slices and preserves the original class label.

2.3 Architecture

The proposed architecture (Fig. 3) is based on [15, 16] with kinematic input vectors augmented with the domain features as described in Eq. 7. The network has five convolutional blocks. Each block includes two convolutional layers and a batch normalization layer with the ReLU activation function. As proposed in [15], the last two layers in the network are a global average pooling layer followed by a fully connected layer with a softmax activation function. We apply cross-entropy loss to compute the probabilities for each class with batch size 16. To regularize the model's weights and reduce overfitting, we use L2 regularization with a factor 0.01 and dropout layers. The network uses Adam optimizer with a learning rate 0.001.

2.4 Training and testing

We propose several training and testing modifications to improve the model's generalization and performance. First, the model will follow a user-out 3-fold cross-validation protocol during training and testing to ensure better generalization on unseen data. Since the dataset contains a different number of classes, each fold has one expert sample, five intermediates and five novices (except one fold with four novices). Second, we introduce fine-tuning on each region, as it was shown that transfer learning helps improve performance in similar tasks [24, 25].

In the first stage, the model trains on the data from all four regions for 80 epochs. The best set of hyperparameters is chosen based on the classification accuracy that the model demonstrates on the validation set. In the second stage of training, the network is fine-tuned on each region separately for 20 epochs. We compute the performance as the average between the fine-tuned models for each region. This scheme provides the most reliable approach to evaluate the generalization of the network on unseen data.

2.5 Ablation studies

We ablate our model on two contributions that we propose in this work. Firstly, we test the added value that the incorporation of the domain-specific features brings to the model. We record the model's performance without features, using only one feature, and with all proposed features together. Secondly, we conduct an ablation study on fine-tuning to determine its effect on the model's performance.

3 Results

3.1 Classification results

Table 1 illustrates performance measures calculated after the fine-tuning stage for time-series slices. We show that classification accuracy and macro-sensitivity is consistent with prior work [24]. The proposed model achieves 80.1%, 94.9%, 93.6% and 81.3% accuracy for the RUQ, LUQ, HEART, and PELVIC areas, respectively. The model shows 0.672 macro-sensitivity for all scanned regions with experts and 0.885 with experts excluded.

Table 2 shows the performance comparison with other FAST proficiency assessment models. In Table 3, we demonstrate confusion matrices using majority vote on time-series slices. Namely, a participant's time-series is sliced into sub-sequences, and each sub-sequence receives a class label. The participant receives the mode of labels assigned to sub-sequences.

3.2 Ablation results

In Table 4, we compare against ablations of the proposed domain specific features. The highest accuracy and novice sensitivity scores are achieved by the

Table 1 Performance of the proposed model on region specific data computed for time-series. Novice (Nov.), Intermediate (Int.), Expert (Exp.)

Region	Accuracy, %	Sensitivity Nov.	Int.	Exp.	Macro
RUQ	80.1	0.752	0.869	0.416	0.679
LUQ	94.9	0.961	0.990	0.116	0.689
HEART	93.6	0.942	0.953	0.276	0.724
PELVIC	81.3	0.881	0.730	0.180	0.597
All	87.5	0.884	0.886	0.247	0.672

Table 2 Classification performance for different models in FAST proficiency assessment. Best results for automated methods are in **bold**

Baseline	Accuracy, %	Sensitivity Nov.	Int.	Exp.	Macro
Ultrasound video [24]	82.6	0.866	**0.967**	0.00	0.611
Kinematics (Random forest) [13]	70.0	0.696	0.767	**0.417**	0.627
Kinematics (ours)	**87.5**	**0.884**	0.886	0.247	**0.672**

Table 3 Confusion matrices for FAST skill level for different models on the same dataset. Best results are in **bold**

		Predicted: Kinematics (ours)			Ultrasound video [24]			Random forest [13]		
		Nov.	Int.	Exp.	Nov.	Int.	Exp.	Nov.	Int.	Exp.
True	Nov.	**54**	2	0	48.5	7.5	0	39	16	1
	Int.	5.5	54	0.5	2	**58**	0	10	46	4
	Exp.	6.5	3	2.5	9	4	0	0	7	**5**

model that includes only time as an extra feature in the transformation matrices. However, the model that includes all four domain features demonstrates the highest macro-sensitivity score.

We also tested the benefits of the region fine-tuning. Table 5 shows that after the fine-tuning, the model demonstrates better overall accuracy and sensitivity scores.

4 Discussion

This study demonstrates that a CNN model can accurately assess FAST skills from kinematics data. It provides a method of skill assessment when access to expert supervision is limited, which can reinforce video-based expert review with supplementary information from kinematics.

The proposed model shows that expert-defined points of interest on the patient's anatomy are not required for skills assessment. Furthermore, the model is insensitive to the position of the reference sensor on the patient's

Table 4 Accuracy and sensitivity scores of the network for different extra features. The sensitivity scores are computed for each class separately. Best results are in **bold**

Feature	Accuracy, %	Sensitivity Nov.	Int.	Exp.	Macro
No features	88.9	0.910	**0.913**	0.003	0.619
Time	**89.5**	**0.917**	0.883	0.170	0.657
Linear speed	87.2	0.883	0.877	0.163	0.641
Angular speed	88.8	0.896	0.900	0.102	0.633
Path length	86.6	0.886	0.851	0.149	0.629
All features	87.5	0.884	0.886	**0.247**	**0.672**

Table 5 Performance scores and confidence intervals. Left values - tuned model, right values - non-tuned model. Best are in **bold**

Region	Accuracy, %	Sensitivity Nov.	Int.	Exp.	Macro
RUQ	**80.1**/80.0	0.752/**0.764**	**0.869**/0.848	**0.416**/0.242	**0.679**/0.618
LUQ	**94.9**/90.9	**0.961**/0.902	**0.990**/0.984	0.116/**0.135**	**0.689**/0.674
HEART	93.6/**95.9**	0.942/**0.963**	0.953/**0.978**	**0.276**/0.030	**0.724**/0.657
PELVIC	**81.3**/80.1	**0.881**/0.866	**0.730**/0.721	**0.180**/0.000	**0.597**/0.529
All	**87.5**/86.7	**0.884**/0.874	**0.886**/0.883	**0.247**/0.102	**0.672**/0.620
Accuracy 95% CI for All, %			[87.3, 87.7]/[86.5, 86.9]		
Macro-Sensitivity 95% CI for All			[0.669, 0.675]/[0.617, 0.623]		

body due to the use of relative transformations. Thus, the model provides both a practical advancement and performance improvement over previous work, which uses a random forest classifier based on summary statistics and requires an expert to scan beforehand and manually delineate points-of-interest in order to perform a proficiency assessment [7, 13] (see Table 2 and 3).

The model achieves equivalent accuracy and macro-sensitivity while performing better in classifying experts compared to a CNN-based classifier which uses ultrasound video data from the same dataset [24] (Table 2 and 3). Our work demonstrates that correct probe motion during image acquisition provides essential information about FAST proficiency. It is consistent with prior work in POCUS competency evaluation, which shows the process of image acquisition as a critical facet of skill [11]. Therefore, using hand and probe motion trackers for skills assessment in FAST will have added value. Similarly, kinematics data with domain specific features was used successfully for skills assessment on a different dataset but the results are not directly comparable [14].

The use of automated assessment tools can bring several advantages to medical education. It can (1) reduce the necessity of direct observation and manual evaluation needed from instructors, (2) eliminate the biases from human interaction, and (3) facilitate the learning process for students by providing more opportunities to hone their skills in a self-guided low-stakes setting. Thus, it can improve learning outcomes in synergy with classical teaching practices. The proposed model shows promise as a tool for incorporation into POCUS training curricula as a standalone software or as a part of an existing automatic assessment.

5 Limitations

The primary limitation of this work is that the dataset was collected from a single healthy male model. The physiological differences and patient pathology might affect the image acquisition techniques used by the operators and, thus, the model's performance on unseen data. Additionally, this dataset only includes participants with typical movement patterns. However, the model has the potential to accurately evaluate the skills of participants with non-typical movement patterns, if such samples were included in the dataset.

The other limitation of this research is the size of the dataset. There are only 32 samples with three expert instances. Therefore, with the 3-fold cross-validation, the training set has one expert sample. The data augmentation technique we use to increase the number of training instances brings even more class imbalance to the data. As a result, most Experts are classified as Novices since the imbalance pushes the model to overfit Novices, who have the highest number of data slices in a fold. However, sensitivity for Novices and Intermediates (excluding Experts) is 0.885, which indicates that the network captures dexterity differences at a fine granularity. The differences between Novices and Intermediates are more subtle and harder to detect compared to the differences between Novices and Experts. Thus, we conjecture that given more expert data our model will increase its performance on the expert class. Our results are consistent with prior work (see Table 2), where the model achieves zero expert class sensitivity on the same dataset due to the class imbalance [24]. In contrast, a random forest classifier based on the summary statistics showed better expert sensitivity (see Table 2) [13].

We also acknowledge that our model may incorrectly classify a series of efficient probe motions that fail to capture clinically relevant images. However, in this work, we focus on image acquisition proficiency which is an important part of the POCUS skillset [11]. In addition, it is worth noting that the proposed network relies on ground-truth class labels assigned based on the number of completed training hours [7]. It may be beneficial to base these labels on an existing assessment tool, such as QUICk or UCAT [8, 10], to achieve better generalization and granularity.

6 Conclusion

In this study we establish that CNNs can distinguish trainees with different skill levels based on their probe motion during FAST. We demonstrate that (1) prior identification of points of clinical interest is not required for kinematics-based skill assessment, and (2) motion data could reinforce automatic POCUS skill assessment from images to provide more accurate and diversified proficiency evaluation. We demonstrated the improvement in the network's performance as the result of augmenting input vectors with expert-defined domain features and region fine-tuning. Therefore, a CNN-based automated assessment on kinematics data can become a useful tool for tracking learners' progress with high accuracy and objectivity.

We propose several ways to improve this study. In the future, more data augmentation techniques should be explored to overcome the lack of expert samples. Moreover, extending the model to predict UCAT score [10] can help improve feedback granularity [21]. In addition, an ensemble model with skill assessment from ultrasound video and kinematics data should be investigated to increase performance [26].

Acknowledgments. This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2020-05582 and the Carleton University I-CUREUS program. This research was enabled, in part, by support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

7 Declarations

Conflict of Interest. All authors declare that they have no conflict of interest.

Ethical Approval. This work involves secondary analysis of anonymous data and, thus, did not require ethics approval according to institutional policy.

Informed Consent. Informed consent was not required for this study.

References

[1] Ultrasound Guidelines: Emergency, Point-of-Care and Clinical Ultrasound Guidelines in Medicine. Annals of Emergency Medicine **69**(5), 27–54 (2017). https://doi.org/10.1016/j.annemergmed.2016.08.457

[2] Lagi, A., Marini, F.: In: Sarti, A., Lorini, F.L. (eds.) Focused Assessment with Sonography for Trauma, pp. 397–399. Springer, Milano (2012). https://doi.org/10.1007/978-88-470-2583-7_46

[3] Moore, C.L., Copel, J.A.: Point-of-care ultrasonography. New England Journal of Medicine **364**(8), 749–757 (2011)

[4] Melniker, L.A., Leibner, E., McKenney, M.G., Lopez, P., Briggs, W.M., Mancuso, C.A.: Randomized Controlled Clinical Trial of Point-of-Care, Limited Ultrasonography for Trauma in the Emergency Department: The First Sonography Outcomes Assessment Program Trial. Annals of

Emergency Medicine **48**(3), 227–235 (2006). https://doi.org/10.1016/j.annemergmed.2006.01.008

[5] Damewood, S.C., Leo, M., Bailitz, J., Gottlieb, M., Liu, R., Hoffmann, B., Gaspari, R.J.: Tools for Measuring Clinical Ultrasound Competency: Recommendations From the Ultrasound Competency Work Group. AEM Education and Training **4**, 106–112 (2020)

[6] Schmidt, J.N., Kendall, J., Smalley, C.: Competency assessment in senior emergency medicine residents for core ultrasound skills. Western Journal of Emergency Medicine **16**(6), 923–926 (2015). https://doi.org/10.5811/westjem.2015.9.28587

[7] Bell, C.R., McKaigney, C.J., Holden, M., Fichtinger, G., Rang, L.: Sonographic Accuracy as a Novel Tool for Point-of-care Ultrasound Competency Assessment. AEM education and training **1**(4), 316–324 (2017)

[8] Ziesmann, M.T., Park, J., Unger, B.J., Kirkpatrick, A.W., Vergis, A., Logsetty, S., Pham, C., Kirschner, D., Gillman, L.M.: Validation of the quality of ultrasound imaging and competence (QUICk) score as an objective assessment tool for the FAST examination. Journal of Trauma and Acute Care Surgery **78**(5), 1008–1013 (2015). https://doi.org/10.1097/TA.0000000000000639

[9] Bahner, D.P., Hughes, D., Royall, N.A.: I-AIM: a novel model for teaching and performing focused sonography. Journal of Ultrasound in Medicine **31**(2), 295–300 (2012)

[10] Bell, C., Hall, A.K., Wagner, N., Rang, L., Newbigging, J., McKaigney, C.: The Ultrasound Competency Assessment Tool (UCAT): Development and Evaluation of a Novel Competency-based Assessment Tool for Point-of-care Ultrasound. AEM Education and Training **5**(3), 1–12 (2021). https://doi.org/10.1002/aet2.10520

[11] Kumar, A., Kugler, J., Jensen, T.: Evaluation of Trainee Competency with Point-of-Care Ultrasonography (POCUS): a Conceptual Framework and Review of Existing Assessments. Journal of General Internal Medicine **34**(6), 1025–1031 (2019). https://doi.org/10.1007/s11606-019-04945-4

[12] Zago, M., Sforza, C., Mariani, D., Marconi, M., Biloslavo, A., La Greca, A., Kurihara, H., Casamassima, A., Bozzo, S., Caputo, F., Others: Educational impact of hand motion analysis in the evaluation of FAST examination skills. European Journal of Trauma and Emergency Surgery, 1–8 (2019)

[13] Holden, M.S., Ungi, T., McKaigney, C., Bell, C., Rang, L., Fichtinger,

G.: Objective evaluation of sonographic skill in focussed assessment with sonography for trauma examinations. In: CARS 2015—Computer Assisted Radiology and Surgery Proceedings of the 29th International Congress and Exhibition Barcelona, pp. 79–80 (2015)

[14] Ziesmann, M.T., Park, J., Unger, B., Kirkpatrick, A.W., Vergis, A., Pham, C., Kirschner, D., Logestty, S., Gillman, L.M.: Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. Journal of Trauma and Acute Care Surgery **79**(4), 631–637 (2015). https://doi.org/10.1097/TA.0000000000000813

[15] Wang, Z., Majewicz Fey, A.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International Journal of Computer Assisted Radiology and Surgery (2018) https://arxiv.org/abs/1806.05796. https://doi.org/10.1007/s11548-018-1860-1

[16] Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Evaluating Surgical Skills from Kinematic Data Using Convolutional Neural Networks. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2018). https://doi.org/10.1007/978-3-030-00937-3_25

[17] Castro, D., Pereira, D., Zanchettin, C., MacEdo, D., Bezerra, B.L.D.: Towards Optimizing Convolutional Neural Networks for Robotic Surgery Skill Evaluation. In: Proceedings of the International Joint Conference on Neural Networks (2019). https://doi.org/10.1109/IJCNN.2019.8852341

[18] Nguyen, X.A., Ljuhar, D., Pacilli, M., Nataraja, R.M., Chauhan, S.: Surgical skill levels: Classification and analysis using deep neural network model and motion signals. Computer Methods and Programs in Biomedicine (2019). https://doi.org/10.1016/j.cmpb.2019.05.008

[19] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

[20] Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. IEEE Transactions on Biomedical Engineering (2017). https://doi.org/10.1109/TBME.2016.2647680

[21] Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A.: Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery **14**(9), 1611–1617 (2019) https://arxiv.org/abs/1908.07319. https://doi.org/10.1007/s11548-019-02039-4

[22] Lasso, A., Heffter, T., Rankin, A., Pinter, C., Ungi, T., Fichtinger, G.: PLUS: Open-source toolkit for ultrasound-guided intervention systems. IEEE Transactions on Biomedical Engineering (2014). https://doi.org/10.1109/TBME.2014.2322864

[23] Guennec, A.L., Malinowski, S., Tavenard, R.: Data Augmentation for Time Series Classification using Convolutional Neural Networks. ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data (2016)

[24] Tyrrell, R.E., Holden, M.S.: Ultrasound video analysis for skill level assessment in FAST ultrasound. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization, 1–15 (2020). https://doi.org/10.1080/21681163.2020.1835549

[25] Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Transfer learning for time series classification. In: Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018 (2019). https://doi.org/10.1109/BigData.2018.8621990

[26] Webb, G.I., Zheng, Z.: Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. IEEE Transactions on Knowledge and Data Engineering **16**(8), 980–991 (2004). https://doi.org/10.1109/TKDE.2004.29