Expert-quality Dataset Labeling via Gamified Crowdsourcing on Point-of-Care Lung Ultrasound Data

Nicole M. Duggan, MD*
Department of Emergency Medicine
Brigham and Women's Hospital
Boston, MA, 02115
nmduggan@bwh.harvard.edu

Mike Jin, PhD*
Centaur Labs
Boston, MA, 02116
mike@centaurlabs.com

Maria Alejandra Duran Mendicuti, MD
Department of Radiology
Brigham and Women's Hospital
Boston, MA, 02115
mduranmendicuti@bwh.harvard.edu

Stephen Hallisey, MD
Department of Emergency Medicine
Brigham and Women's Hospital
Boston, MA, 02115
shallisey@bwh.harvard.edu

Denie Bernier, RDMS
Department of Emergency Medicine
Brigham and Women's Hospital
Boston, MA, 02115
Dbernier1@bwh.harvard.edu

Lauren A. Selame, MD
Department of Emergency Medicine
Brigham and Women's Hospital
Boston, MA, 02115
lselame@bwh.harvard.edu

Ameneh Asgari-Targhi, PhD
Department of Radiology
Brigham and Women's Hospital
Boston, MA, 02115
aasgaritarghi@bwh.harvard.edu

Chanel E. Fischetti, MD
Department of Emergency Medicine
Brigham and Women's Hospital
Boston, MA, 02115
cfischetti@bwh.harvard.edu

Ruben Lucassen, MS
Department of Biomedical Engineering
Eindhoven University of Technology
The Netherlands
r.t.lucassen@tue.nl

Anthony E. Samir, MD, MPH
Department of Radiology
Massachusetts General Hospital
Boston, MA, 02114
asamir@mgh.harvard.edu

Erik Duhaime, PhD+
Centaur Labs
Boston, MA, 02116
erik@centaurlabs.com

Tina Kapur, PhD+
Department of Radiology
Brigham and Women's Hospital
Boston, MA, 02115
tkapur@bwh.harvard.edu

*These authors contributed equally to primary authorship
+These authors contributed equally to senior authorship

Abstract

Machine learning tools can aid medical imaging data interpretation. Building such tools requires labeled training datasets. We tested whether a gamified crowdsourcing approach can produce clinical expert-quality lung ultrasound clip labels. 2,384 lung ultrasound clips were retrospectively collected. Six lung ultrasound experts classified 393 of these clips as having no B-lines, one or more discrete B-lines, or confluent B-lines to create two sets of reference standard labels: a training and test set. Sets trained users on a gamified crowdsourcing platform, and compared concordance of the resulting crowd labels to the concordance of individual experts to reference standards, respectively. 99,238 crowdsourced opinions were collected from 426 unique users over 8 days. Mean labeling concordance of individual experts relative to the reference standard was 85.0% ± 2.0 (SEM), compared to 87.9% crowdsourced label concordance (p=0.15). Scalable, high-quality labeling approaches such as crowdsourcing may streamline training dataset creation for machine learning model development.

Keywords: Machine learning, artificial intelligence, ultrasound, POCUS, crowdsourcing

URI: https://hdl.handle.net/10125/106853
978-0-9981331-7-1

1. Introduction

Machine learning (ML) models applied to medical image analysis can improve medical diagnostic accuracy and streamline healthcare processes (Tschandl et al., 2020). Widespread ML tool development is limited by the need for large-scale labeled datasets for model training (Gulshan et al., 2016; Esteva et al., 2017; Lee at al., 2019; Malone et al., 2004; Pesapane et al., 2018; Rajpurkar et al., 2017). These labeled datasets are time- and labor-intensive to produce, and as such are often costly.

Crowdsourcing is the practice of collecting large numbers of unskilled opinions to complete a given task. Crowdsourcing can produce more accurate interpretations than those from a single individual, and has been shown to improve efficiency, lower costs, and offer high-quality in repetitive task completion (Grote et al., 2019; Hautz et al., 2015; Surowiecki, 2005). Crowdsourcing for biomedical image labeling is often made difficult by the complexity of the tasks and need to ensure label quality control. Despite this, crowdsourcing has been successful in some healthcare-related tasks (Heim et al., 2018; Meakin et al., 2019; Monu et al., 2020). Combining crowdsourcing with gamification, the persuasive system design which uses game-like tasks to engage participants competitively for rewards, can encourage crowd participation and improve performance accuracy (Foncubierta Rodriguez et al., 2012; Von Ahn et al., 2013; Kattan et al., 2016; Kentley et al., 2023; Kurvers et al., 2016; Wolf et al., 2015).

Point-of-care ultrasound (POCUS) is a medical imaging technique used at patients' bedside to make accurate, real-time diagnoses (Lichtenstein et al., 2015; ACEP Policy Statement, 2017). Unfortunately, to use POCUS accurately for clinical care, extensive user training in image acquisition and interpretation is required (Blehar et al., 2015). As such, ML models which automate POCUS image interpretation hold exceptional potential clinical value. Here, we examined whether a gamified crowdsourcing approach can classify lung POCUS clips at comparable accuracy to trained ultrasound experts.

2. Methods

2.1. Study Design and Setting

This was a prospective analysis performed using lung POCUS clips retrospectively collected between March 1st, 2020 and February 28th, 2022 from an academic tertiary care center emergency department. This study was approved by the local institutional review board.

2.2. Dataset Curation

In total, 2,391 POCUS clips were downloaded from the hospital electronic medical record and storage system and were de-identified using a software package (Smith B, 2022). Clips were randomly divided by patient into two sets: dataset A (102 patients, 1,271 clips) and dataset B (101 patients, 1,120 clips). 200 random clips from dataset A were selected as a crowd training set, and 200 random clips from dataset B were selected as a test set to evaluate crowd label quality. 5 training set clips and 2 test set clips were excluded for being flagged by at least one expert as not containing lung (Figure 1A).

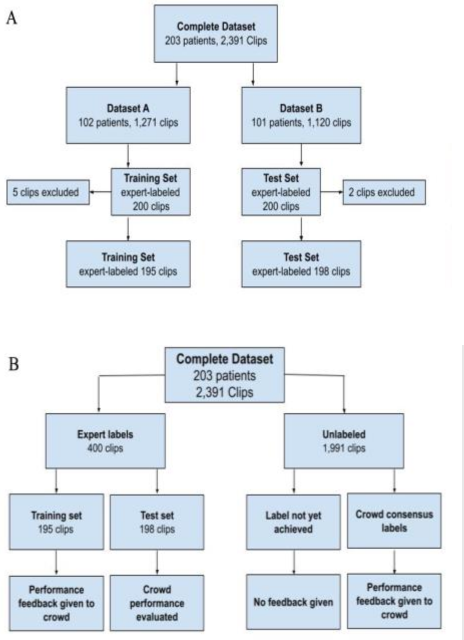

Figure 1. Dataset workflow. A. The complete dataset is divided into training and test sets. B. Labeling schema for the complete dataset.

2.3. Task Definition and Reference Standards

On lung POCUS, B-lines are vertical, hyperechoic, dynamic structures which indicate the presence of pulmonary congestion. Their presence, quantity, and thickness (discrete vs confluent) correlate with severity of several pathological conditions (Pang et al., 2021). Expert and crowd users were asked to classify B-lines on lung POCUS clips into either: a) no B-lines, b) one or more discrete B-lines, or c) confluent B-lines (Figure 2). Clips were classified based on the highest B-line severity present throughout the clip. Discrete B-lines were defined as hyperechoic lines originating from the pleural line, demonstrated sliding with the pleura, and extending to the bottom of the sonographic field. Confluent B-lines were defined as hyperechoic sections originating from the pleural line, demonstrated sliding with the pleura, and had thickness along the pleura beyond that of discrete B-lines (Lichtenstein et al., 2008; Lichtenstein et al., 2020).

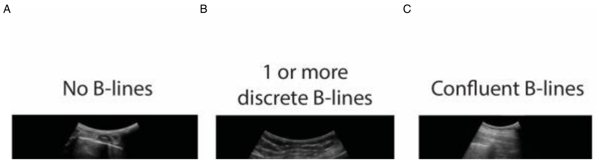

Figure 2. B-line classifications from lung point-of-care ultrasound clips. A. No B-lines. B. One or more discrete B-lines. C. Confluent B-lines.

Six experts with advanced training in lung POCUS (four ultrasound fellowship-trained emergency medicine physicians; one emergency radiologist, and one Registered Diagnostic Medical Sonographer) provided independent classification opinions for all training and test set clips via DiagnosUs (Centaur Labs, Boston, MA), an iOS application where users compete in medical data labeling contests to win cash prizes based on their labeling accuracy. Reference standard labels on training and test set clips were assigned using expert consensus – the majority rule of the six experts' opinions, with ties broken randomly.

2.4. Gamified Crowdsourcing

Crowd opinions were collected using DiagnosUs via gamified contests. Crowd users could include anyone in the general public with access to the iOS-based application. There were no criteria regarding level of expertise to participate, and users voluntarily participated in labeling contests based on interest and the potential to earn rewards.

Crowd users were trained by optional tutorial cases with accompanying layperson explanations which they could access at any time, and, after submitting an opinion, immediate feedback via revelation of the current label (i.e., the reference standard label on training set clips, or the current crowd consensus label if one exists). In the DiagnosUs app, clips were shown to users in random order, and any test clip could be shown to a user multiple times. Crowd consensus labels were assigned for all clips using the majority rule of the top crowd labelers' opinions, and were assigned to any of the initially unlabeled clips once sufficient agreement was reached amongst the top crowd labelers for that clip (Figure 1B). Top labelers were identified via performance monitoring based on their trailing accuracy on clips with labels.

2.5. Assessing Crowd Consensus Label Quality

Crowd label quality was assessed by comparing how well crowd consensus labels and individual expert opinions matched the reference standard labels on the test set. Both of these were measured by concordance, the percent of matching labels across the test set clips.

2.6. Statistical Analyses

Analysis was performed with Python 3.10 (Pedregosa et al. 2011). Paired t-tests were used to assess differences between crowd concordance and average individual expert concordance. All mean calculations are reported as mean ± standard error of mean (SE). Significance was set at $p < 0.05$.

3. Results

3.1. Dataset Characteristics

Patients in the lung POCUS database had a mean age of 60.0 years (standard deviation 19.0), 105 (51.7%) were female; 43 (21.2%) were Hispanic, 42 (20.7%) were Black, and 114 (56.1%) were White. From the ED, 64% of patients were admitted to the hospital floor, 28.6% were discharged home, 6.4% were admitted to the intensive care unit, 1% went directly to the operating room, 4.4% had an alternate disposition and 0% expired in the ED (Table 1). Over the test clips the reference standard label distribution was 70% no B-lines, 18% discrete B-lines, and 12% confluent B-lines.

3.2. Collecting Opinions

Experts spent an average of 1.7 hours (minimum 0.9 hours, maximum 2.5 hours) submitting opinions for training and test clips (3.9 opinions per minute on average). On the training clips, the reference standard label distribution (based on the experts' majority opinion) was 58% no B-lines, 29% discrete B-lines, and 13% confluent B-lines.

Table 1. Patient characteristics

Characteristic	Subjects, N = 203 [no. (%)]
Sex	
Male	98 (48.3)
Female	105 (51.7)
Ethnicity	
Hispanic or Latino	43 (21.2)
Not Hispanic or Latino	160 (78.8)
Race	
American Indian or Alaskan Native	0 (0)
Asian	9 (4.4)
Native Hawaiian or Pacific Islander	0 (0)
Black or African American	42 (20.7)
Caucasian	114 (56.2)
Other	38 (18.7)
Emergency Department Disposition	
Discharge home	57 (28.6)
Floor admission	122 (64.0)
Intensive care unit admission	13 (6.4)
Operating room	2 (1.0)
Other	9 (4.4)

Overall, 99,238 crowdsourced opinions were collected from 426 users across all clips. The number of users contributing an opinion to each test set clip ranged from 28 to 48. Of these, 34,363 opinions from 114 unique users contributed to crowd consensus labels based on quality thresholds. The live contest was launched over 138 hours with a mean acquisition rate of 12.0 opinions per minute. The total cash prize payout throughout the entire competition was 1,100 USD. The maximum prize earned by an individual user was 25 USD.

3.3. Label Concordance with Reference Standard

The six experts' concordances on the test clips relative to the reference standard were 77.2%, 81.3%, 84.8%, 87.3%, 88.4%, and 90.9%, with a mean of 85.0 ± 2.0. Comparatively, the crowd concordance on these clips was 87.9% relative to the reference standard (p=0.15) (Figure 3). For clips designated by the reference standard as having no B-lines, experts had an average concordance with the reference standard of 91.5% ± 2.3, compared to a crowd concordance of 99.3% (p<0.001). For cases with discrete B-lines, experts had an average concordance of 63.9% ± 13.2 compared to the crowd concordance of 50% (p=0.088). For cases with confluent B-lines, expert average concordance (79.2% ± 6.5) and crowd concordance were both 79.2% (p=1.0). .2 for individual experts relative to leave-one-out reference standard. Calculated from randomly sampled subsets of collected opinions, 7 quality-filtered opinions were sufficient to achieve near the maximum crowd accuracy (Figure 4).

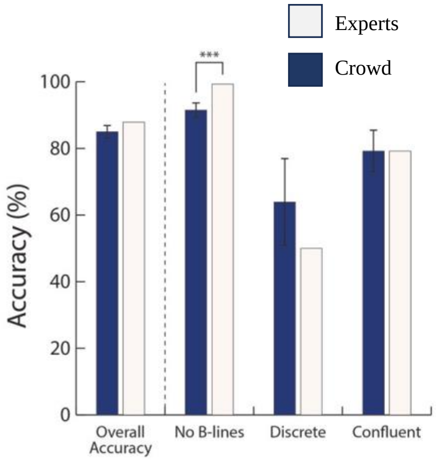

Figure 3. Expert compared to crowd opinion concordance with reference standard.

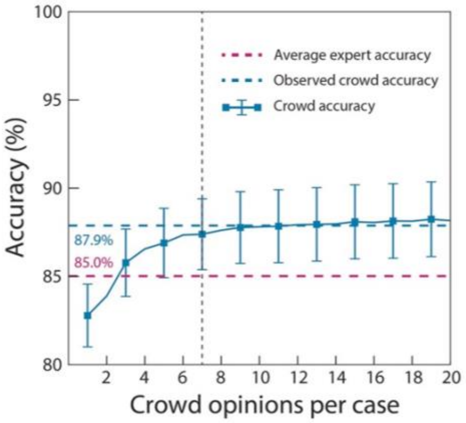

Figure 4. The number of crowd opinions needed to maximize crowd-consensus accuracy. Solid blue line indicates estimated crowd-consensus accuracy as dependent on the number of crowd opinions collected. Vertical dotted line indicates the crowd opinions sufficient to achieve the observed crowd accuracy. Error bars indicate standard error of the mean.

4. Discussion

Our data suggest that gamified crowdsourcing can produce expert-quality labels for B-line classification on lung POCUS clips. There was considerable variability in individual expert accuracies for classifying B-lines. Given that medical imaging data interpretation is often complex, it may be that the variability is explained by a proportion of clips having inherent ambiguity that even experts disagree on. This is consistent with previously published work that shows expert interrater agreement for identifying B-lines is imperfect (Gravel et al., 2020; Herraiz et al., 2023; Sustic et al., 2022; Nowak et al., 2010). Inherent ambiguity is likely a theme that extends across medical imaging data beyond lung POCUS and may represent a challenge for imaging database labeling overall.

Variability in expert concordance may also be attributable to variable expert baseline skill. Consistent with existing literature which uses medical experts to label POCUS images, our experts all had either fellowship-level training or advanced certification in interpreting lung ultrasound as well as years of clinical experience (Moore et al., 2022; Pare et al., 2022; van Sloun et al., 2020). Ground truth labeling for training POCUS-based ML models is commonly derived from a small handful of experts (typically 1-5 individuals). Thus, this work combining opinions from 6 experts to form our reference standard is consistent with accepted practices. Currently there is no established method for defining ground truth in B-line identification beyond expert opinion. Given the recent widespread adoption of lung POCUS globally and the recognized utility of B-lines as a clinical disease marker, our work highlights the critical need for clarifying how ground truth interpretation of lung POCUS and POCUS overall is defined.

Comparing individual expert opinions against the consensus of all six experts likely inflates expert accuracy estimates since each expert’s opinion is influencing the consensus label. Despite this, crowd opinions achieved the same accuracy as experts. Computing consensus labels using a “leave-one-out” consensus, or a consensus that excludes the original opinion from the expert who is being assessed may offer a more accurate picture of crowd versus individual expert accuracy. Given that we expect consensus labels derived from all six experts as in the present work inflates expert accuracy, we anticipate that the crowd may demonstrate similar if not higher accuracy than experts if a “leave-one-out” consensus were used. Overall, our approach suggest that favorable crowd performance compared to experts is a true effect.

Ultimately we aim to identify strategies for scalable and accurate medical data labeling. Our findings highlight a possible triage approach to dataset labeling using gamified crowdsourcing. Crowd opinion for clips with a high degree of crowd agreement would be accepted as truth, and expert review would only be necessary for cases where crowd agreement drops below a certain threshold. This approach could significantly decrease the proportion of clips requiring expert review and optimize both time and cost associated with current expert-based dataset labeling approaches.

5. Limitations

Our dataset had an oversampling of clips with no lung pathology with more than 50% of clips in both the training and test datasets containing no B-lines. Since the crowd demonstrated higher concordance with reference standard than individual experts on clips with no B-lines, but performed worse than experts on classifying discrete B-lines, it is possible that the crowd may be less skilled at identifying subtle diagnostic findings, but we are not adequately seeing this trend due to dataset bias. Further investigation with larger more varied training and test sets is warranted.

While the iOS application used here openly available, users with medical background may be more likely to engage in labeling contests. We do have a general understanding of the proportion of our users with medical background, however the precise demographic breakdown of prior medical experiences was poorly defined. It is possible that our crowd is not representative of the general population and may be more consistent with a population of semi-skilled labelers. The generalizability of our findings across variable crowd populations with clearly defined experience levels will need to be explored further.

While our findings show crowdsourcing may be a promising approach to streamlining lung POCUS data labeling this approach may not be generalizable to other medical imaging data labeling tasks. Next steps are to apply this approach to similar questions in lung POCUS data such as segmenting B-lines. This will help us understand more about the generalizability of our findings.

6. Conclusion

Our work demonstrates that gamified crowdsourcing can produce B-line classification labels that match consensus labels as well as individual experts themselves. Innovative and scalable approaches to generating high-quality labeled medical image databases such as crowdsourcing could help streamline future ML model development.

7. References

Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. Acad Emerg Med. 2015 May;22(5):574-82. doi: 10.1111/acem.12653.

Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb 2;542(7639):115-118. doi: 10.1038/nature21056.

Foncubierta Rodríguez, A. and H. Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. in Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia. 2012.

Gravel CA, Monuteaux MC, Levy JA, Miller AF, Vieira RL, Bachur RG. Interrater reliability of pediatric point-of-care lung ultrasound findings. Am J Emerg Med. 2020 Jan;38(1):1-6. doi: 10.1016/j.ajem.2019.01.047.

Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of Histological Image Labeling and Object Delineation by Medical Students. IEEE Trans Med Imaging. 2019 May;38(5):1284-1294. doi: 10.1109/TMI.2018.2883237.

Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402-2410. doi: 10.1001/jama.2016.17216.

Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. JAMA. 2015 Jan 20;313(3):303-4. doi: 10.1001/jama.2014.15770.

Heim E, Roß T, Seitel A, et al. Large-scale medical image annotation with crowd-powered algorithms. J Med Imaging (Bellingham). 2018 Jul;5(3):034002. doi: 10.1117/1.JMI.5.3.034002.

Herraiz JL, Freijo C, Camacho J, et al. Inter-Rater Variability in the Evaluation of Lung Ultrasound in Videos Acquired from COVID-19 Patients. *Applied Sciences*. 2023; 13(3):1321. https://doi.org/10.3390/app13031321

Kattan MW, O'Rourke C, Yu C, Chagin K. The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones. Med Decis Making. 2016 May;36(4):536-40. doi: 10.1177/0272989X15581615.

Kentley J, Weber J, Liopyris K, et al. Agreement Between Experts and an Untrained Crowd for Identifying Dermoscopic Features Using a Gamified App: Reader Feasibility Study. JMIR Med Inform. 2023 Jan 18;11:e38412. doi: 10.2196/38412.

Kurvers RH, Herzog SM, Hertwig R, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci U S A. 2016 Aug 2;113(31):8777-82. doi: 10.1073/pnas.1601827113.

Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng. 2019 Mar;3(3):173-182. doi: 10.1038/s41551-018-0324-9.

Lichtenstein DA, Mezière GA. Relevance of lung ultrasound in the diagnosis of acute respiratory failure: the BLUE protocol. Chest. 2008 Jul;134(1):117-25. doi: 10.1378/chest.07-2800. Epub 2008 Apr 10. Erratum in: Chest. 2013 Aug;144(2):721. doi: 10.1378/chest.07-2800.

Lichtenstein DA. BLUE-protocol and FALLS-protocol: two applications of lung ultrasound in the critically ill. Chest. 2015 Jun;147(6):1659-1670. doi: 10.1378/chest.14-1313.

Lichtenstein DA. Lung ultrasound for the cardiologist-a basic application: The B-profile of the Bedside Lung Ultrasound in Emergencies protocol for diagnosing haemodynamic pulmonary oedema. Arch Cardiovasc Dis. 2020 Aug-Sep;113(8-9):489-491. doi: 10.1016/j.acvd.2020.05.005.

Malone, T.W., The Future of Work–How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life, Harvard Business School Press. Boston, MA, 2004.

Meakin JR, Ames RM, Jeynes JCG, et al. The feasibility of using citizens to segment anatomy from medical images: Accuracy and motivation. PLoS One. 2019 Oct 10;14(10):e0222523. doi: 10.1371/journal.pone.0222523.

Monu J, Triplette M, Wood DE, et al. Evaluating Knowledge, Attitudes, and Beliefs About Lung Cancer Screening Using Crowdsourcing. Chest. 2020 Jul;158(1):386-392. doi: 10.1016/j.chest.2019.12.048.

Moore CL, Wang J, Battisti AJ, et al. Interobserver Agreement and Correlation of an Automated Algorithm for B-Line Identification and Quantification With Expert Sonologist Review in a Handheld Ultrasound Device. J Ultrasound Med. 2022 Oct;41(10):2487-2495. doi: 10.1002/jum.15935.

Nowak S, Ruger S. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. *Proc Int*

Conf Multimedia Info Retriev. 2010: 557 –5. doi: 10.1145/1743384.

Pang PS, Russell FM, Ehrman R, et al. Lung Ultrasound-Guided Emergency Department Management of Acute Heart Failure (BLUSHED-AHF): A Randomized Controlled Pilot Trial. JACC Heart Fail. 2021 Sep;9(9):638-648. doi: 10.1016/j.jchf.2021.05.008.

Pare JR, Gjesteby LA, Telfer BA, et al. Transfer Learning for Automated COVID-19 B-Line Classification in Lung Ultrasound. Annu Int Conf IEEE Eng Med Biol Soc. 2022 Jul;2022:1675-1681. doi: 10.1109/EMBC48229.2022.9871894.

Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12(2011); 2825-2830.

Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp. 2018 Oct 24;2(1):35. doi: 10.1186/s41747-018-0061-6.

Rajpurkar, P, Irvin, J., Zhu, K., et al., Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225, 2017.

Smith B. ClipDeidentifier. Core Ultrasound. URL: https://www.coreultrasound.com/clipdeidentifier/

Surowiecki, J., The Wisdom of Crowds. 2005, New York, NY: Knopf Doubleday Publishing Group.

Šustić, A., Mirošević, M., Szuldrzynski, K. *et al.* Inter-observer reliability for different point-of-care lung ultrasound findings in mechanically ventilated critically ill COVID-19 patients. *J Clin Monit Comput* 36, 279–281 (2022). doi: 10.1007/s10877-021-00726-9.

Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020 Aug;26(8):1229-1234. doi: 10.1038/s41591-020-0942-0.

van Sloun RJG, Demi L. Localizing B-Lines in Lung Ultrasonography by Weakly Supervised Deep Learning, In-Vivo Results. IEEE J Biomed Health Inform. 2020 Apr;24(4):957-964. doi: 10.1109/JBHI.2019.2936151. Epub 2019 Aug 19.

Von Ahn, L. Duolingo: learn a language for free while helping to translate the web. in Proceedings of the 2013 international conference on Intelligent user interfaces. 2013. ACM.

Wolf M, Krause J, Carney PA, Bogart A, Kurvers RH. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS One. 2015 Aug 12;10(8):e0134269. doi: 10.1371/journal.pone.0134269.