Matthew Jacqueline (Orcid ID: 0000-0003-4754-0322)
Day Thomas George (Orcid ID: 0000-0001-8391-7903)

Interaction between clinicians and artificial intelligence to detect fetal atrioventricular septal defects on ultrasound: how can we optimize collaborative performance?

T. G. Day[1,2], J. Matthew[1], S. F. Budd[1], L. Venturini[1], R. Wright[1], A. Farruggia[1], T. V. Vigneswaran[1,2], V. Zidere[1,2,3], J. V. Hajnal[1], R. Razavi[1,2], J. M. Simpson[1,2] and B. Kainz[1,4,5]

1. School of Biomedical Engineering and Imaging Sciences, Faculty of Life Sciences and Medicine, King's College London, London, UK
2. Department of Congenital Heart Disease, Evelina London Children's Healthcare, Guy's and St Thomas' NHS Foundation Trust, London, UK
3. Harris Birthright Centre, King's College London NHS Foundation Trust, London, UK
4. Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
5. Department of Computing, Faculty of Engineering, Imperial College London, London, UK

Corresponding author: Dr T. G. Day, School of Biomedical Engineering and Imaging Sciences, Faculty of Life Sciences and Medicine, King's College London, London, UK. e-mail: thomas.day@kcl.ac.uk

Running title: human-AI collaboration in detecting fetal AVSD

Keywords: artificial intelligence; machine learning; congenital heart disease; fetal medicine; fetal cardiology; atrioventricular septal defect

Contribution

What are the novel findings of this work?

Artificial intelligence (AI) can improve the performance of clinicians in detecting fetal AVSD on ultrasound, even if the AI performance is worse than the clinicians alone. When the AI was incorrect, this resulted in a deterioration in clinician performance. Giving additional information about AI model workings and model confidence did not improve overall performance.

What are the clinical implications of this work?

These results support the possibility of integrating AI in the clinical workflow of fetal ultrasound screening. Even if AI models alone do not reach expert level performance, they still have potential to improve overall collaborative human-machine performance. We have not identified a reliable method to mitigate against the risk of incorrect AI.

Abstract

Objectives

Artificial intelligence (AI) has shown promise in improving the performance of fetal ultrasound screening in detecting congenital heart disease (CHD). The effect of giving AI advice to human operators has not been studied in this context. Giving additional information about AI model workings, such as confidence scores for AI predictions, may be a way of improving performance further. Our aims were to investigate whether AI advice improved overall diagnostic accuracy (using a single CHD lesion as an exemplar), and to see what, if any, additional information given to clinicians optimized the overall performance of the clinician-AI team.

Methods

An AI model was trained to classify a single fetal CHD lesion (atrioventricular septal defect, AVSD), using a retrospective cohort of 121,130 cardiac four chamber images extracted from 173 ultrasound scan videos (98 with normal hearts, 75 with AVSD). A ResNet50 model architecture was used. Temperature scaling of model prediction probability was performed on a validation set, and gradient-weighted class activation maps (grad-CAMs) produced. Ten clinicians (two consultant fetal cardiologists, three trainees in pediatric cardiology, and five fetal cardiac sonographers) were recruited from a center of fetal cardiology to participate. Each participant was shown 2000 fetal four chamber images in a random order (1,000 normal and 1,000 AVSD). The dataset was comprised of 500 images, each shown in four conditions: 1) image alone without AI output; 2) image with binary AI classification; 3) image with AI model confidence; 4) image with gradient-weighted class activation map image overlays. The clinicians were asked to classify each image as normal or AVSD.

Results

20,000 image classifications were recorded from 10 clinicians. The AI model alone achieved an accuracy of 0.798 (95% CI 0.760 – 0.832), sensitivity of 0.868 (95% CI 0.834 – 0.902) and specificity of 0.728 (95% CI 0.702 – 0.754, and the clinicians without AI achieved an accuracy of 0.844 (95% CI 0.834

– 0.854), sensitivity of 0.827 (95% CI 0.795 – 0.858) and specificity of 0.861 (95% CI 0.828 – 0.895). Showing a binary (normal or AVSD) AI model output resulted in significant improvement in accuracy to 0.865 ($p < 0.001$). This effect was seen in both experienced and less experienced participants. Giving incorrect AI advice resulted in significant deterioration in overall accuracy from 0.761 to 0.693 ($p < 0.001$), which was driven by an increase in both type I and type II error by the clinicians. This effect was worsened by showing model confidence (accuracy 0.649, $p < 0.001$) or grad-CAM (accuracy 0.644, $p < 0.001$).

Conclusions

AI has the potential to improve performance when used in collaboration with clinicians, even if the model performance does not reach expert level. Giving additional information about model workings such as model confidence and class activation map image overlays did not improve overall performance, and actually worsened performance for images where the AI model was incorrect.

Introduction

Antenatal diagnosis of fetal congenital heart disease (CHD) is associated with improved morbidity and mortality after birth[1–3]. Many countries have instigated mid-trimester ultrasound screening to detect structural malformations such as CHD, but these do not achieve universal detection, with considerable regional variation. For atrioventricular septal defect (AVSD), in the UK the Fetal Anomaly Screening Program detects an estimated 69.2% of cases via the anomaly ultrasound scan, and 79.4% overall, including first trimester screening[4].

Artificial intelligence (AI) using convolutional neural networks (a form of deep learning) has proved to be a powerful tool in many medical imaging tasks. This includes promising performance in the automatic detection of fetal CHD using ultrasound[6].

Ultrasound is an operator-dependent modality, usually performed and interpreted at the same clinic visit. Despite promising performance, it is unlikely that AI would be used autonomously of the human ultrasound operator. This means that collaboration will be required between AI and clinicans[7]. Ideally, overall performance of the clinician-AI team would be better than either the clinician or the AI operating alone. However, this is not guaranteed. Recent work suggests that providing AI assistance to radiologists does not improve their performance in chest X-ray interpretation[8]. AI could worsen the performance of clinicians, if, for example, they choose to trust the AI when it was incorrect[7]. Providing more information about the model, for example model confidence, or the area of the image most influential for the classification, might be a means of mitigating against the risk of algorithm aversion (not trusting enough) or automation bias (trusting too much). These might help the operator appropriately calibrate their trust of the AI, meaning that they can make good decisions about when, and when not, to use the AI outputs[7].

Little work has been undertaken exploring the interaction between humans and AI in ultrasound interpretation. Our aims were twofold:

1. To investigate whether AI assistance given to clinicians increases overall collaborative performance in ultrasound disease classification, using fetal AVSD as an example lesion.
2. To investigate whether additional information about the AI model provided to the clinician impacts the overall collaborative clinician-AI performance.

Methods

Setting

The study was undertaken at a tertiary fetal cardiology referral center. This permitted access to an image archive with a large number of cardiac abnormalities and facilitated analysis of images by staff with specific training in fetal echocardiography.

Development of AI models

An AI model was trained to classify fetal ultrasound four-chamber cardiac images into normal and AVSD. Detailed information about the technical aspects of AI model development is shown in Appendix S1. The dataset was retrospectively acquired and consisted of 173 fetal ultrasound scans (98 with normal hearts and 75 with AVSD), all diagnosed antenatally. Multiple ultrasound videos containing four-chamber views were used per fetus. Ultrasound videos were manually labelled by image plane and quality, and only frames labelled as high-quality four-chamber views were used. The total dataset size was 121,130 images.

Human interaction with AI

To compile the dataset for the experiment investigating human and AI combined performance, the test and validation sets from the AI model development datasets were used (i.e., images not used for training the model). 500 four-chamber images from 36 fetuses (16 with AVSD and 20 with normal heart) were randomly selected for the experiment, split evenly between normal hearts and AVSD (250 images for each). Temperature scaling was undertaken to improve calibration of model confidence score, as previously described[9]. Gradient-weighted class activation maps (grad-CAMS) were generated for each image, using previously published techniques[10].

The 500 images (250 normal and 250 AVSD) were presented to each clinician in four different conditions which are outlined below, and shown diagrammatically in Figure 1. This meant that a total of 2000 images were classified as either normal or AVSD by each clinical reviewer.

- Condition 1: the plain unprocessed image, with no additional AI information.
- Condition 2: the image with the addition of a binary AI model classification ('normal' or 'AVSD').
- Condition 3: as in condition 2, but with the addition of the temperature-scaled model confidence, expressed as a percentage likelihood that the image represents a case of AVSD.
- Condition 4: as in condition 3, but with an additional grad-CAM image displayed adjacent to the plain image.

Clinicians (both medical and non-medical) were recruited from our tertiary fetal cardiology unit. Years of experience (both since qualification, and specifically in fetal ultrasound) and professional background (consultant fetal cardiologist, specialist trainee doctor in pediatric cardiology, or sonographer) were recorded. The 2000 images were displayed to each clinician in a random order, using a bespoke platform implemented in Python version 3.10. For each image the participant was asked to select normal or AVSD as the most likely diagnosis. Before starting they were informed of the prevalence of AVSD among the image dataset (i.e., 50%), and the accuracy of the AI model. They also received tuition on the meaning of the AI model output, the model confidence, and grad-CAM image. Accuracy was defined as number of correct classifications / number of images. 95% confidence intervals were calculated using the exact Clopper-Pearson method. Accuracy was compared between conditions using the paired McNemar test for proportions. A p-value of less than 0.05 was considered significant.

This work was undertaken as part of a research project which has approval from the East of Scotland Research Ethics Service, reference 20/ES/0005.

Results

Ten clinician participants were recruited (two consultant fetal cardiologists, five non-medical sonographers, and three doctors undertaking specialist higher training in pediatric cardiology). The consultant fetal cardiologists had 29 and 20 years of post-qualification medical experience, and 22 and 7 years of consultant-level experience in fetal cardiology. The sonographers had a mean post-qualification experience of 27.4 years (range 17-39 years), with a mean of 14.2 years' experience in fetal ultrasound (range 5-25 years). All sonographers had completed the Fetal Medicine Foundation online training course in Fetal Echocardiography[11]. The pediatric cardiology trainee doctors were all working at fellow / registrar level, had a mean post-qualification experience of 8 years (range 7-9 years), and a mean experience in pediatric cardiology (including postnatal echocardiography) of 2.7 years (range 1-4 years). All the trainee doctors had less than one year's experience in fetal cardiology. For analysis the operators were split into more experienced (the sonographers and fetal cardiology consultants), and less experienced (the trainee doctors). 20,000 human clinician classifications were recorded in total.

The accuracy of the AI model in diagnosing AVSD in this dataset was 0.798 (95% CI 0.760 – 0.832), with a sensitivity of 0.868 (95% CI 0.834 – 0.902) and specificity of 0.728 (95% CI 0.702 – 0.754). The overall clinician performance when shown images without AI assistance (condition 1) was an accuracy of 0.844 (95% CI 0.834 – 0.854), significantly better than the AI model ($p < 0.001$), and a sensitivity of 0.827 (95% CI 0.795 – 0.858) and specificity of 0.861 (95% CI 0.828 – 0.895). This superiority in accuracy was restricted to the more experienced participants (accuracy 0.873, $p < 0.001$), with the less experienced group showing a non-significant trend towards poorer performance compared to the AI model (accuracy 0.777, 95% CI 0.755 – 0.798, $p = 0.161$).

Figure 2 and Table 1 show the performance of the clinicians when shown the image in the different conditions. Showing a binary AI model prediction with the image (condition 2) improved overall performance compared to a plain image (condition 1), with an accuracy of 0.865 vs 0.844 ($p < 0.001$).

This effect was seen in both more experienced and less experienced operators. Giving additional information (either model confidence (condition 3), or grad-CAM (condition 4)) along with the image did not significantly improve performance compared to giving a binary AI model output. For the clinicians as a whole, giving model confidence (condition 3) resulted in a deterioration in overall performance compared to condition 2 (accuracy 0.850 vs 0.865 respectively, p = 0.002). For the more experienced group, both model confidence (condition 3) and grad-CAM (condition 4) resulted in a worse performance compared to condition 2 (accuracy 0.872 and 0.877 respectively vs 0.888, p = 0.001 and 0.029 respectively). For the less experienced group, giving model confidence or grad-CAM did not significantly change the accuracy compared to condition 2.

To investigate the effect of professional group in more detail, we further stratified the more experienced group into the fetal cardiology consultants and the sonographers so that the performance of fetal cardiology consultants, sonographers and trainees can be seen separately. The results of this analysis are shown in Table S1 and Figure S1. The patterns of performance seen with the different conditions of image are similar between the sonographers and consultants, with both groups significantly outperforming the AI model. Although the consultants' accuracy was higher when shown the binary AI model output (condition 2) compared to operating alone, this improvement did not reach statistical significance.

Figure 3 and Table 2 show the results stratified by whether the AI advice was correct or incorrect. The unassisted clinicians had higher performance on the images where the AI was correct compared to where the AI was incorrect (accuracy 0.865 vs 0.761, p < 0.001), probably reflecting the fact that these were easier images to classify (either because image quality was better, or because the pathological findings were more obvious). For the images where AI was correct, the addition of AI advice with a binary AI classification resulted in a significant improvement in accuracy compared to clinicians using the plain images (0.908 vs 0.865, p <0.001). Giving additional information (model confidence or grad-CAM) did not change performance compared to binary AI advice.

Giving incorrect AI advice resulted in a significant deterioration in clinician performance (accuracy 0.693 vs 0.761, $p < 0.001$). This effect was driven by an increase in both type I and type II error, (i.e., both false negatives and false positives) with a deterioration seen in both sensitivity and specificity when the AI was incorrect. The effect was worsened when additional information (model confidence or grad-CAM was given (0.649 and 0.644 respectively, both $p < 0.001$).

Discussion

We have demonstrated, for the first time in fetal ultrasound, that giving AI advice to clinicians results in a significant improvement in the overall diagnostic performance of the clinician-AI team, compared to either AI or clinicians operating alone. This improvement is seen in both experienced experts and less experienced trainee doctors, and is also seen even if the AI performance alone is inferior to clinicians. This supports the hypothesis that a human-in-the-loop AI system may have utility in improving fetal ultrasound screening for structural malformations, even if neither the human sonographer nor AI model achieves 100% accuracy.

The relative performance of AI alone compared to clinicians varied by the experience of the human operator. In this study, the performance of experienced fetal cardiology clinicians was independently superior to AI, but the performance of less experienced operators was similar to AI alone. Whether this potential benefit would be magnified in a less specialist setting remains unknown. Work in other imaging modalities has shown variable results, with some finding AI does not improve expert diagnostic performance, and others findings a significant improvement[8,12]. We have also shown that giving incorrect AI model outputs worsened clinician performance, in keeping with work in other imaging modailites[13]. This effect was driven by causing an increase in both false negative and false positive decisions by the clinicians, resulting in a decrease in both sensitivity and specificity.

In previous work we have discussed the issue of trust calibration[7]. The AI model used in this work did not have 100% sensitivity or specificity, meaning that it was incorrect for an important number of images. If we had a model that operated perfectly, trust calibration would not be required, as clinicians would simply trust the model output every time. However, we feel that this situation is unlikely to be reality in the near future, especially for AVSD detection given how subtle this lesion can be (making it difficult for both humans and AI to detect). For this reason, we feel trust calibration will be important for any clinical use of AI in this context.

Trust calibration is the concept of human operators calibrating their trust according to AI performance, meaning that they trust the AI model output when it is correct, but distrust it (and therefore appropriately overrule it) when it is wrong. We postulated that giving additional information to the human operator (such as the confidence of the AI model, or a grad-CAM showing which area of the image has been most influential in the model decision) may improve their trust calibration. Our findings do not support this hypothesis, in contrast to work in other imaging modalities[14]. When the AI model was correct, giving additional information did not make the clinicians more likely to trust it compared to just simply displaying the overall model diagnosis as 'normal' or 'AVSD'. When the AI model was incorrect, giving additional information counterintuitively made the clinicians *more* likely to inappropriately trust the AI, resulting in significantly worse performance.

The causes of this are not completely clear. We have shown that for the images where the AI was incorrect, the clinicians also found these difficult to classify, with an overall lower diagnostic accuracy compared to the other images. In such situations, the clinicians may be more likely to rely on the AI diagnosis, even though it is wrong. It is possible that giving additional information simply adds credence to the AI decision, making it seem more trustworthy even though it should be distrusted, so-called "automation bias"[15].

Output prediction scores of deep neural networks, if interpreted as likelihood point estimates, are known to be overly confident, meaning it becomes difficult for the human user to interpret these and use them to help calibrate trust. Temperature scaling, as we have done here, only partially ameliorates this, and as we have shown the model confidence outputs were not helpful to the clinicians when they were judging whether to trust or distrust the AI. How to improve the relationship between AI model confidence and likelihood of correctness is an area of active research, and more accurate metrics may become much more useful to human operators. However, even if more accurate, rather than real-value probabilities, alternative methods of communicating these metrics may be more effective (for

example, a general indicator that the AI should be 'trusted', 'not trusted', or 'approached with caution').

Similarly, we have shown that our class activation map image overlays were not useful to the clinicians. A grad-CAM image is a relatively simplistic graphical display of the final layer of a highly complex neural network, and does not necessarily reflect the workings of the multiple hidden layers of the network. Additionally, the area of the image most influential in the final model decision is not necessarily helpful in deciding if the model is correct. For some images the grad-CAM of an incorrect AI model output clearly shows that an area outside the heart is being highlighted, meaning that the clinician may use this to decide to overrule the AI. However, in many images this was not the case, with appropriate highlight of the cardiac crux. In these cases, this may give false reassurance leading to inappropriate trust of the AI by the clinician.

There are several limitations to the current work. Firstly, the setting was a retrospective review of still images, which is different to how these clinicians would work in clinical practice, i.e., real-time interpretation of moving ultrasound video. Diagnosing AVSD from still images is not the same task as making an overall diagnosis of AVSD in a fetus, and this probably accounts for the lower accuracy seen in this task than expected if they were assessing on a per-fetus basis, and limits the generalizability of this work to clinical practice. It is far more difficult to diagnose any malformation from a single image compared to a complete ultrasound examination, and we acknowledge that this will have influenced human performance in this study. The psychological effect of forcing the clinicians to make a binary decision based on a single image is very different to how clinicians operate in either a screening or specialist setting, where much richer data would be available, either from other images or other clinical information. We still feel that assessing interaction between human clinicians and AI in this context is informative and, but further work is planned to develop these AI techniques so that the AI models can be run in real time, and allow simultaneous image acquisition and interpretation with AI assistance, to allow a closer representation of true clinical workflow.

Secondly, all cases included in this study were diagnosed antenatally, so may not be representative of the cases that are most important to focus on, i.e. those cases that are currently missed. In addition, as in most studies in this field, the available dataset is relatively small meaning that the images used in this study were taken from a limited number of fetuses, with multiple images taken from the same fetus. This is important, as some AVSDs are far easier to detect than others, and it is possible the cases chosen for this study are not representative of the overall population of fetuses with AVSD. For more 'obvious' AVSDs it would be easier for the clinician to overrule the AI if it was incorrect, so the nature of trust calibration would be altered. Future prospective clinical trials based in screening units will be an important way of assessing these issues, as it is likely that the AI model performance will deteriorate with this type of clinical translation, in part because the types of cases encountered at a population screening level may be different to a specialist level, causing a degree of covariate shift.

Thirdly, we used only a single fetal CHD lesion, AVSD. This was selected as it has a relatively poor antenatal detection rate compared to other major congenital heart diseases[16]. Also, it is diagnosable from a single plane (the four-chamber view), and although long-term outcomes are good following surgery[17], due to its extremely high association with chromosomal disorders antenatal detection is of great importance. Our group and others are working on broadening AI classification models to cover the entire CHD spectrum, and further work will be required to investigate how applicable the results of the present study are to disparate fetal conditions.

Finally, the clinicians were recruited from a tertiary referral center for pediatric and fetal cardiology. This means that their performance specific to CHD may be superior to sonographers operating at a screening level. However, we have shown that our findings hold true even when restricted to a far less experienced group. Despite this, is it possible that impact of AI assistance in the setting of screening ultrasound (where the vast majority of examinations will be normal), may be completely different to when used in a specialist setting. Involvement of screening level sonographers in future studies will be

important to examine the generalizability of this work when we consider exactly how AI might be integrated into the actual clinical workflow of a national screening program.

Conclusion

We have shown that human clinicians and AI working together to diagnose fetal CHD on ultrasound can have superior performance compared to either clinicians or AI operating alone. This is supportive of the idea that AI might be of real clinical value in this context, even if the AI does not reach expert level performance. However, giving incorrect AI advice results in a deterioration in clinician performance. We examined the utility of providing additional information about the AI model to the clinicians to mitigate against this, but conversely found that this resulted in a further deterioration in accuracy. Further work is required to investigate methods of avoiding this potentially dangerous phenomenon, if clinical integration of AI is being considered.

Acknowledgments

TD and JM are supported by NIHR Doctoral Fellowships (NIHR301448 and NIHR300555 respectively). This work was supported by the Wellcome Trust [IEH Award, 102431], by core funding from the Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z], by The AI Centre for Value Based Healthcare [OLS Award, 106231], and by core funding from Innovate UK Office of Life Sciences. BK received funding from the European Research Council (project MIA-NORMAL 101083647). The funding bodies had no influence in the study design, data collection, analysis, interpretation, preparation of the manuscript, or decision to submit or publication.

References

1. Mahle WT, Clancy RR, McGaurn SP, Goin JE, Clark BJ. Impact of prenatal diagnosis on survival and early neurologic morbidity in neonates with the hypoplastic left heart syndrome. *Pediatrics*. 2001;107(6):1277-1282.

2. Calderon J, Angeard N, Moutier S, Plumet M-H, Jambaqué I, Bonnet D. Impact of prenatal diagnosis on neurocognitive outcomes in children with transposition of the great arteries. *J Pediatr*. 2012;161(1):94-98.

3. Holland BJ, Myers JA, Woods CR. Prenatal diagnosis of critical congenital heart disease reduces risk of death from cardiovascular compromise prior to planned neonatal cardiac surgery: A meta-analysis. *Ultrasound Obstet Gynecol*. 2015;45(6):631-638.

4. Aldridge N, Pandya P, Rankin J, Miller N, Broughan J, Permalloo N, McHugh A, Stevens S. Detection rates of a national fetal anomaly screening program: a national cohort study. *Br J Obstet Gynaecol*. Published online 2022.

5. Day TG, Kainz B, Hajnal J, Razavi R, Simpson JM. Artificial intelligence, fetal echocardiography, and congenital heart disease. *Prenat Diagn*. 2021;41(6):733-742.

6. Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat Med*. 2021;27(5):882-891.

7. Day TG, Matthew J, Budd S, Hajnal J V, Simpson JM, Razavi R, Kainz B. Sonographer interaction with artificial intelligence: collaboration or conflict? *Ultrasound Obstet Gynecol*. 2023;62:167-174.

8. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. *National Bureau of Economic Research* 2023.

9. Temperature scaling. Accessed August 23, 2023. https://github.com/gpleiss/temperature_scaling

10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis*. 2020;128(2):336-359.

11. Fetal Medicine Foundation. https://fetalmedicine.org/education/fetal-echocardiography-1

12. Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, Rajpurkar P, Amrhein T, Gupta R, Halabi S, Langlotz C, Lo E, Mammarappallil J, Mariano AJ, Riley G, Seekins J, Shen L, Zucker E, Lungren M. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digit Med*. 2019;2(1).

13. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag J V., Colak E, Ghassemi M. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit Med*. 2021;4(1).

14. Gaube S, Suresh H, Raue M, Lermer E, Koch TK, Hudecek MFC, Ackery AD, Grover SC, Coughlin JF, Frey D, Kitamura FC, Ghassemi M, Colak E. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep*. 2023;13(1):1-11.

15. Wickens CD, Clegg BA, Vieane AZ, Sebok AL. Complacency and Automation Bias in the Use of Imperfect Automation. *Hum Factors*. 2015;57(5):728-739.

16. *National Congenital Heart Disease Audit, Summary Report*. The National Institute for Cardiovascular Outcomes Research; 2021.

17. Ginde S, Lam J, Hill GD, Cohen S, Woods RK, Mitchell ME, Tweddell JS, Earing MG. Long-term outcomes after surgical repair of complete atrioventricular septal defect. *J Thorac Cardiovasc Surg*. 2015;150(2):369-374.

18. Labelbox. Accessed July 27, 2023. https://labelbox.com

19. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* ; 2016:770-778.

Figure legends

Figure 1: examples of the four conditions in which each image was displayed to the clinician participants. A: condition 1, plain ultrasound image, without AI output; B: condition 2, image with binary model output; C: condition 3, image with model output and model confidence, expressed a probability of an AVSD diagnosis; D: condition 4, image with model output, model confidence score, and gradient-weighted class activation map (grad-CAM). In the grad-CAM, red and yellow colors indicate a greater relative influence of those pixels to the final model output, and green and blue color indicate a lesser relative influence.

Figure 2: diagnostic performance accuracy, stratified by level of experience of the human operator. More experienced: fetal cardiology consultants and sonographers; less experienced: pediatric cardiology trainee doctors; AVSD: atrioventricular septal defect; grad-CAM: gradient-weighted class activation map; AI: artificial intelligence. Error bars represent one standard error. * = $p < 0.05$.

Figure 3: diagnostic performance accuracy of clinicians, stratified by correctness of AI model output. AVSD: atrioventricular septal defect; grad-CAM: gradient-weighted class activation map; AI: artificial intelligence. Error bars represent one standard error. * = $p < 0.05$.

Tables

Table 1: AVSD classification accuracy, by image type and clinicians' experience.

	All clinicians				**More experienced clinicians**				**Less experienced clinicians**			
	Accuracy (95% CI)	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*P value*	*Accuracy (95% CI)*	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*P value*	*Accuracy (95% CI)*	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*P value*
Condition 1 *(image with no AI information)*	0.844 (0.834–0.854)	0.827 (0.795–0.858)	0.861 (0.828–0.895)		0.873 (0.861–0.883)	0.868 (0.828–0.908)	0.877 (0.836–0.918)		0.777 (0.755–0.798)	0.731 (0.682–0.779)	0.824 (0.766–0.882)	
Condition 2 *(image with AI classification)*	0.865 (0.855–0.874)	0.866 (0.832–0.899)	0.864 (0.830–0.897)	**<0.001** †	0.888 (0.877–0.898)	0.893 (0.852–0.935)	0.883 (0.836–0.918)	**0.004** †	0.810 (0.789–0.830)	0.801 (0.746–0.857)	0.819 (0.762–0.876)	**0.005** †
Condition 3 *(image with AI confidence)*	0.850 (0.839–0.859)	0.847 (0.815–0.880)	0.852 (0.819–0.885)	0.292 † **0.002** ‡	0.872 (0.860–0.883)	0.878 (0.838–0.919)	0.866 (0.842–0.924)	0.917 † **0.001** ‡	0.797 (0.776–0.817)	0.775 (0.722–0.828)	0.820 (0.763–0.877)	0.103 † 0.262 ‡
Condition 4 *(image with grad-CAM)*	0.858 (0.848–0.868)	0.845 (0.813–0.878)	0.871 (0.837–0.905)	**0.011** † 0.180 ‡	0.877 (0.866–0.888)	0.874 (0.833–0.914)	0.880 (0.839–0.921)	0.455 † **0.029** ‡	0.814 (0.793–0.833)	0.779 (0.725–0.832)	0.849 (0.790–0.909)	**0.003** † 0.722 ‡

† comparison of accuracy with condition 1 (plain image without AI support); ‡ comparison of accuracy with condition 2 (assistance with binary AI output)

Table 2: AVSD classification accuracy by clinicians, stratified by correctness of AI model.

	Images with correct AI				**Images with incorrect AI**			
	Accuracy (95% CI)	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*P value*	*Accuracy (95% CI)*	*Sensitivity (95% CI)*	*Specificity (95% CI)*	*P value*
Condition 1 (image with no AI information)	0.865 (0.854–0.875)	0.843 (0.808–0.878)	0.891 (0.851–0.932)		0.761 (0.734– 0.787)	0.721 (0.649– 0.793)	0.781 (0.725– 0.837)	
Condition 2 (image with AI classification)	0.908 (0.899–0.917)	0.898 (0.861–0.936)	0.920 (0.878–0.962)	**<0.001** †	0.693 (0.664– 0.721)	0.652 (0.592– 0.711)	0.713 (0.664– 0.762)	**<0.001** †
Condition 3 (image with AI confidence)	0.901 (0.891–0.910)	0.887 (0.850–0.924)	0.917 (0.875–0.959)	**<0.001** † 0.137 ‡	0.649 (0.618– 0.678)	0.588 (0.543– 0.633)	0.678 (0.633– 0.723)	**<0.001** † **0.001** ‡
Condition 4 (image with grad-CAM)	0.907 (0.898–0.916)	0.886 (0.849–0.923)	0.932 (0.890–0.975)	**<0.001** † 0.840 ‡	0.644 (0.634– 0.693)	0.579 (0.536– 0.622)	0.706 (0.658– 0.754)	**<0.001** † **0.046** ‡

† comparison of accuracy with condition 1 (plain image without AI support); ‡ comparison of accuracy with condition 2 (assistance with binary AI output)

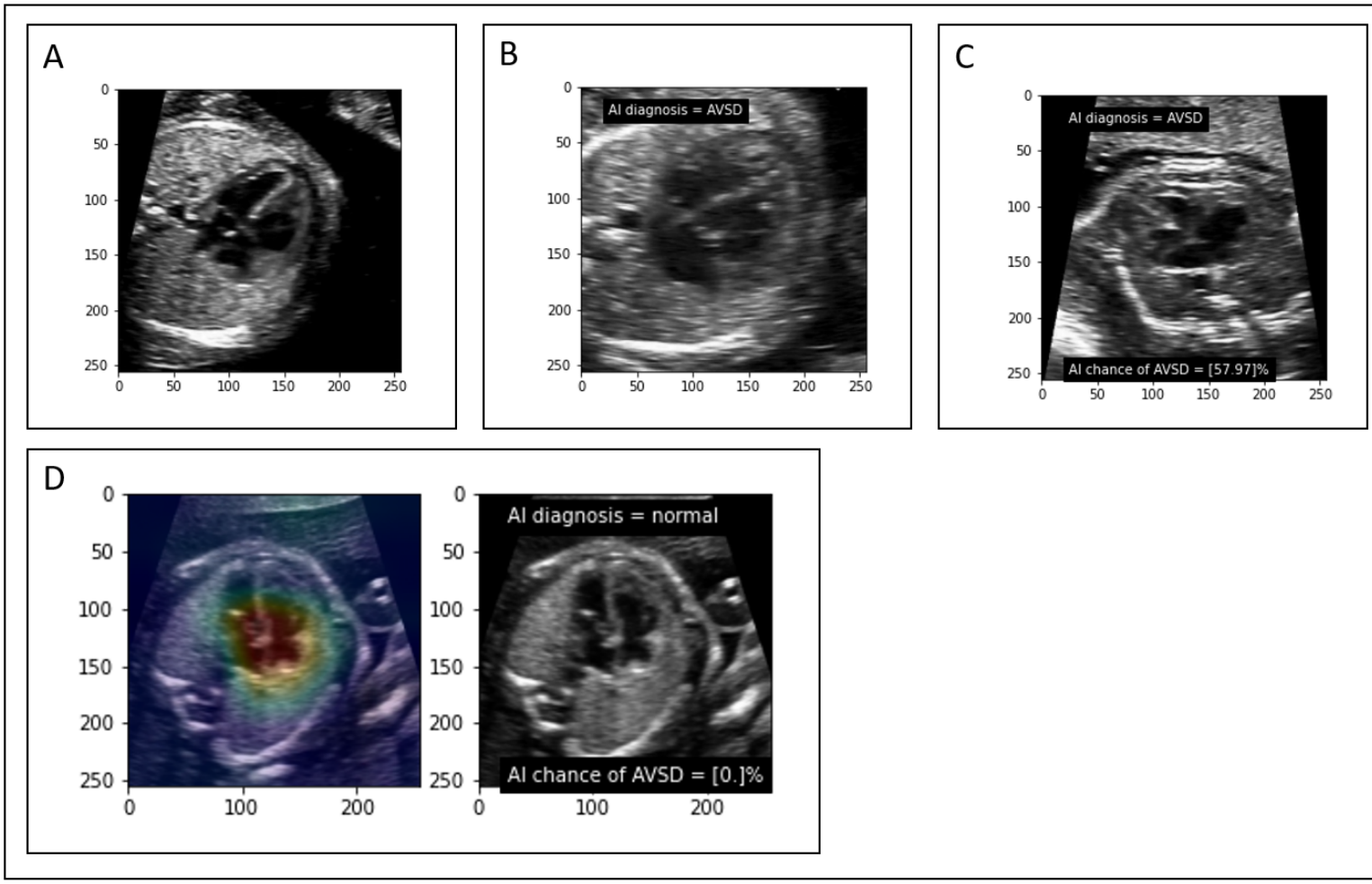

fig1.PNG

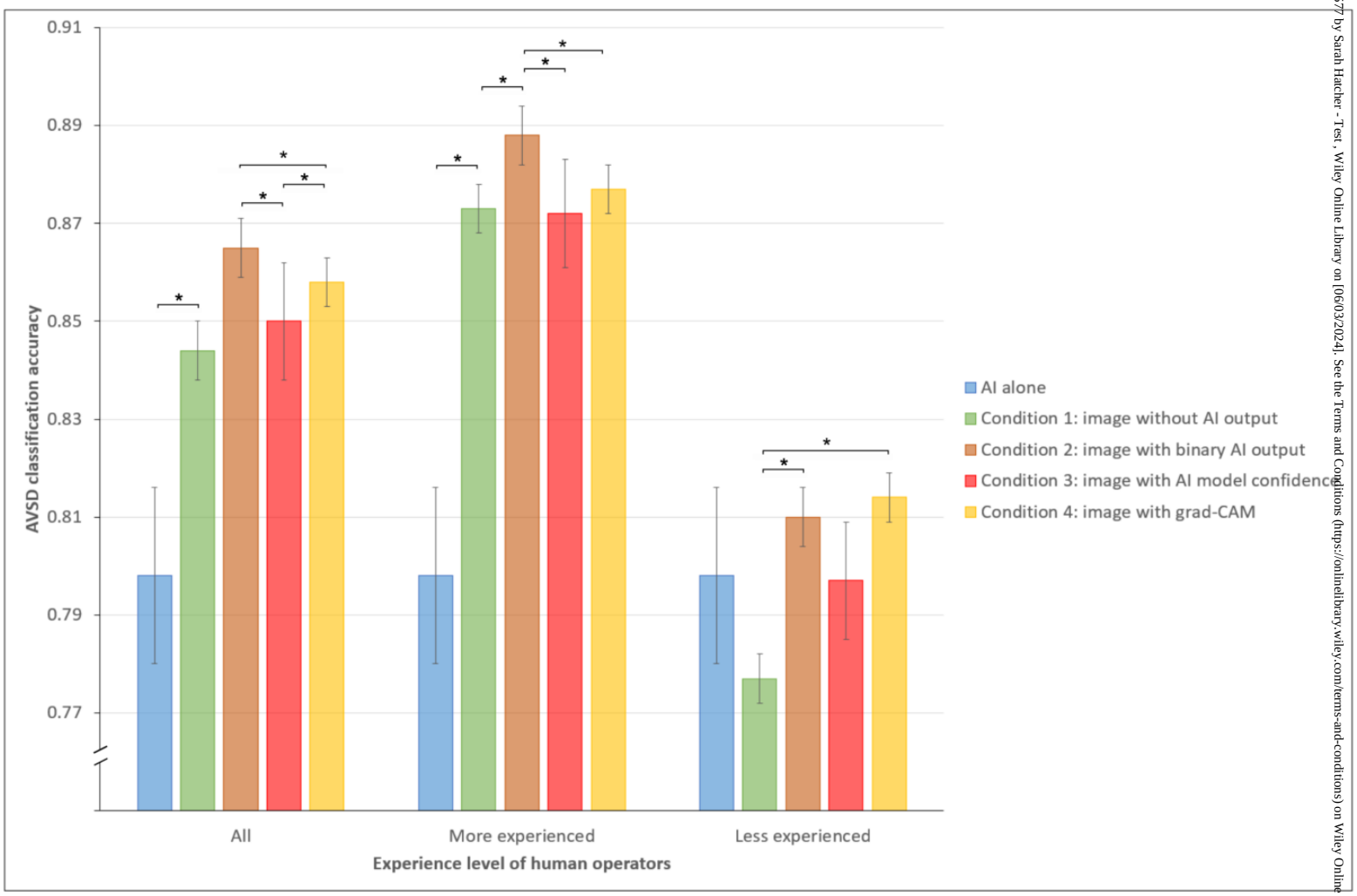

fig2with sig.png

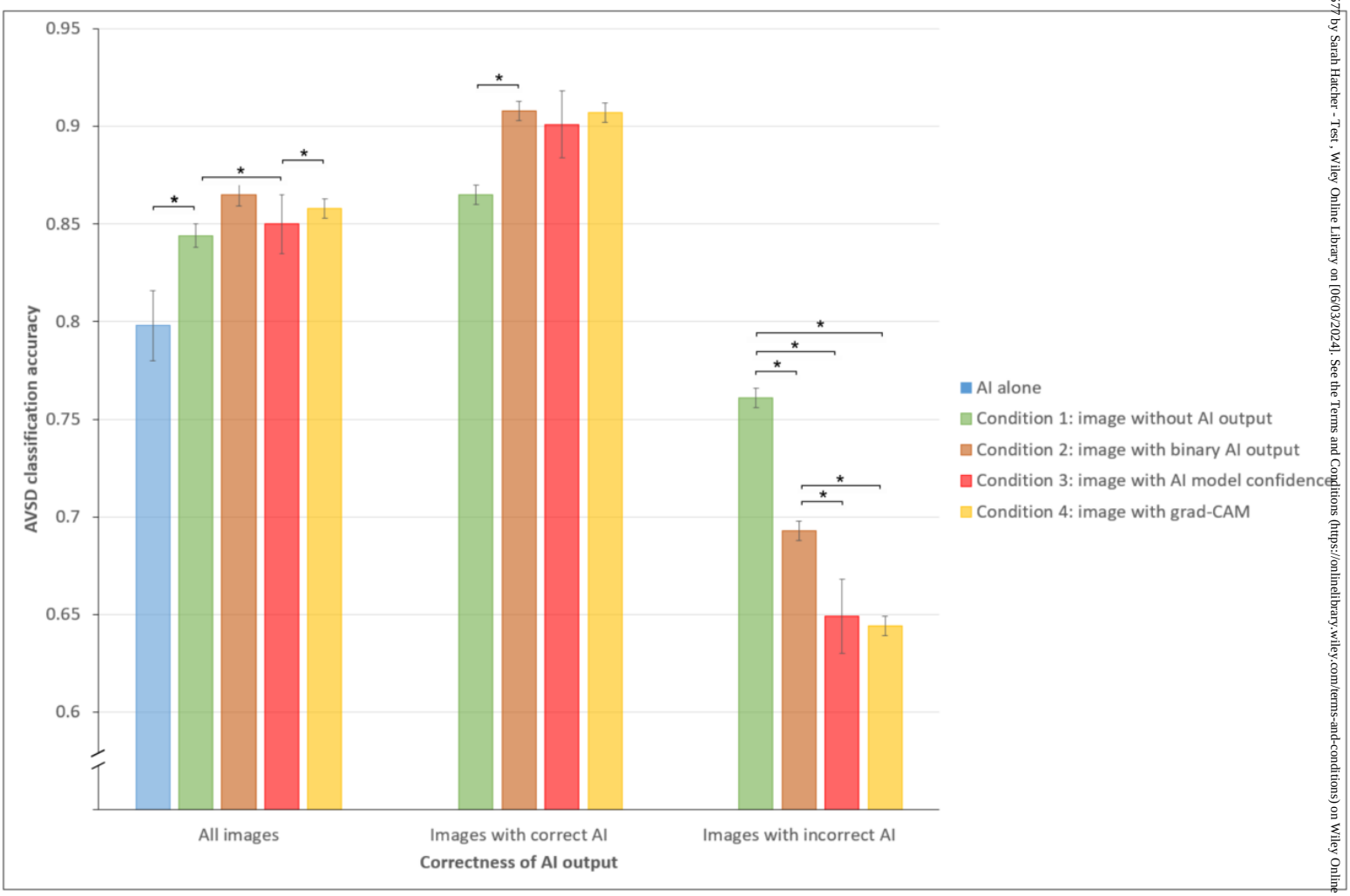

fig3with sig.PNG