Abdominal Circumference Operator-agnostic UltraSound measurement in Low-Income Countries using Artificial Intelligence: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Abdominal Circumference Operator-agnostic UltraSound measurement in Low-Income Countries using Artificial Intelligence

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

ACOUSLIC-AI

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Fetal growth restriction (FGR), affecting up to 10% of pregnancies, is a critical factor contributing to perinatal morbidity and mortality (1-3). Strongly linked to stillbirths, FGR can also lead to preterm labor, posing risks to the mother (4,5). This condition often results from an impediment to the fetus' genetic growth potential due to various maternal, fetal, and placental factors (6). Measurements of the fetal abdominal circumference (AC) as seen on prenatal ultrasound are a key aspect of monitoring fetal growth. When smaller than expected, these measurements can be indicative of FGR, a condition linked to approximately 60% of fetal deaths (4). FGR diagnosis relies on repeated measurements of either the fetal abdominal circumference (AC), the expected fetal weight, or both. These measurements must be taken at least twice, with a minimum interval of two weeks between them for a reliable diagnosis (7). Additionally, an AC measurement that falls below the third percentile is, by itself, sufficient to diagnose FGR (7-9). However, the routine practice of biometric obstetric ultrasounds, crucial for AC measurements, is limited in low-resource settings due to the high cost of sonography equipment and the scarcity of trained sonographers.
The use of low-cost ultrasound devices and standardized blind-sweep protocols has been proposed for novice operators to acquire obstetric data in these settings (10-12). Blind-sweep acquisition protocols are characterized by operators performing scans without viewing the ultrasound images. These protocols yield sequences of 2D ultrasound frames that are captured as the ultrasound probe follows specific trajectories across the gravid abdomen. Unlike traditional clinical sonography, where experienced sonographers search for the standard plane to conduct biometry measurements, blind-sweep data poses a distinct set of challenges. The quality of the image data is limited and may not contain the precise standard planes conventionally used for measurements (13).

Addressing these limitations, a growing body of literature focuses on the use of artificial intelligence (AI) to automate prenatal assessment tasks on free-hand ultrasound sequences acquired following standardized protocols, bypassing the need for expert sonographic interpretation. Such tasks include fetal biometry measurements (13,14), gestational age estimation (13,15-17) and pregnancy risk detection (14,15,18-22). These AI solutions have the potential to be embedded into mobile devices, offering a complete, offline, low-cost, and portable solution suitable for resource-limited settings, as demonstrated in (15,21).
Unlike previous challenges that focused on ultrasound imaging data acquired in clinical settings, this is the first challenge to propose the use of blind-sweep data for fetal biometry tasks. The goal is to develop and benchmark AI models for the automated measurement of fetal abdominal circumference on this specific data type, with the aim to broaden the accessibility of prenatal care in areas with limited resources. A similar task was explored in (23), where the authors propose an AI model to perform fetal biometry measurements including AC on automatically selected standard planes. This model specifically worked on data captured by expert sonographers. Unlike a standardized blind-sweep protocol, the sonographers in this study were directed to ensure that each imaging sequence they acquired included the standard planes necessary for accurate measurements.
Participants in this challenge will develop AI models to estimate AC in blind sweep 2D prenatal abdominal ultrasound sequences, acquired by novice operators in five African peripheral healthcare units and one European hospital. The models must identify the optimal frame for measurement and accurately segment the fetal abdomen within that frame. They must provide the identified frame and the corresponding segmentation mask, which will be used to precisely measure the fetal abdominal circumference. The models will be evaluated against expert estimates derived from blind-sweep data. This challenge represents a first step into FGR detection in low-resource settings. Its main aim is to accurately estimate AC from blind sweep data acquired by novice operators. These estimates could eventually be used to detect FGR, though FGR detection is beyond the scope of the challenge itself. Our end goal is to create effective AI applications for ultrasound imaging that will help improve the care provided to pregnant women and neonates in these regions.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

bdominal circumference, prenatal ultrasound, low-income countries, novice operators, artificial intelligence

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

N/A

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect 20 teams to participate, based on previous challenges on similar topics: Fetal Brain Tissue Annotation and Segmentation Challenge, (feta-2021, 21 teams) and Fetal Tissue Annotation and Segmentation Challenge (feta-2022, 16 teams).

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

The publication policy for this challenge is as follows:

- Authorship in the final challenge paper: Up to three members of each participating team, specifically those who played a significant role in the algorithm's design, will be acknowledged as co-authors in the final challenge paper.
- Independent publication of results: Participants are free to publish their algorithms and findings independently once the challenge concludes and the embargo period is over (see below). However, we require that any such independent publications reference both the summary paper of the challenge and the data publication paper.
- Embargo period: Although teams are permitted to publish their results independently, there is an embargo period of 6 months to ensure all analyses and computations outlined in this proposal can be conducted. This means that for the first six months following the end of the challenge, the challenge organizers retain exclusive rights to publish the collective results, after which teams may proceed with their individual publications. However, we acknowledge the value of early research dissemination and, therefore, permit the submission of arXiv preprints during the embargo period.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

The challenge will be hosted on the online platform grand-challenge.org. Participants will download the public training data from zenodo and train their solutions using their own compute power. They will be able to submit and evaluate their algorithms as docker containers. Inference will be run offline on grand-challenge against the Development and Testing phases.

TASK 1: Segmentation of fetal abdomen on the optimal frame of 2D prenatal ultrasound blind-sweep sequences.

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Fetal growth restriction (FGR), affecting up to 10% of pregnancies, is a critical factor contributing to perinatal morbidity and mortality (1-3). Strongly linked to stillbirths, FGR can also lead to preterm labor, posing risks to the mother (4,5). This condition often results from an impediment to the fetus' genetic growth potential due to various maternal, fetal, and placental factors (6). Measurements of the fetal abdominal circumference (AC) as seen on prenatal ultrasound are a key aspect of monitoring fetal growth. When smaller than expected, these measurements can be indicative of FGR, a condition linked to approximately 60% of fetal deaths (4). FGR diagnosis relies on repeated measurements of either the fetal abdominal circumference (AC), the expected fetal weight, or both. These measurements must be taken at least twice, with a minimum interval of two weeks between them for a reliable diagnosis (7). Additionally, an AC measurement that falls below the third percentile is, by itself, sufficient to diagnose FGR (7-9). However, the routine practice of biometric obstetric ultrasounds, crucial for AC measurements, is limited in low-resource settings due to the high cost of sonography equipment and the scarcity of trained sonographers.

The use of low-cost ultrasound devices and standardized blind-sweep protocols has been proposed for novice operators to acquire obstetric data in these settings (10-12). Blind-sweep acquisition protocols are characterized by operators performing scans without viewing the ultrasound images. These protocols yield sequences of 2D ultrasound frames that are captured as the ultrasound probe follows specific trajectories across the gravid abdomen. Unlike traditional clinical sonography, where experienced sonographers search for the standard plane to conduct biometry measurements, blind-sweep data poses a distinct set of challenges. The quality of the image data is limited and may not contain the precise standard planes conventionally used for measurements (13). Addressing these limitations, a growing body of literature focuses on the use of artificial intelligence (AI) to automate prenatal assessment tasks on free-hand ultrasound sequences acquired following standardized protocols, bypassing the need for expert sonographic interpretation. Such tasks include fetal biometry measurements (13,14), gestational age estimation (13,15-17) and pregnancy risk detection (14,15,18-22). These AI solutions have the potential to be embedded into mobile devices, offering a complete, offline, low-cost, and portable solution suitable for resource-limited settings, as demonstrated in (15,21).

Unlike previous challenges that focused on ultrasound imaging data acquired in clinical settings, this is the first challenge to propose the use of blind-sweep data for fetal biometry tasks. The goal is to develop and benchmark AI models for the automated measurement of fetal abdominal circumference on this specific data type, with the aim to broaden the accessibility of prenatal care in areas with limited resources. A similar task was explored in (23), where the authors propose an AI model to perform fetal biometry measurements including AC on automatically selected standard planes. This model specifically worked on data captured by expert sonographers. Unlike a standardized blind-sweep protocol, the sonographers in this study were directed to ensure that each imaging sequence they acquired included the standard planes necessary for accurate measurements.

Participants in this challenge will develop AI models to estimate AC in blind sweep 2D prenatal abdominal

ultrasound sequences, acquired by novice operators in five African peripheral healthcare units and one European hospital. The models must identify the optimal frame for measurement and accurately segment the fetal abdomen within that frame. They must provide the identified frame and the corresponding binary segmentation mask, which will be used to precisely measure the fetal abdominal circumference. The models will be evaluated against expert estimates derived from blind-sweep data. This challenge represents a first step into FGR detection in low-resource settings. Its main aim is to accurately estimate AC from blind sweep data acquired by novice operators. These estimates could eventually be used to detect FGR, though FGR detection is beyond the scope of the challenge itself. Our end goal is to create effective AI applications for ultrasound imaging that will help improve the care provided to pregnant women and neonates in these regions.

1. Gardosi J, Chang A, Kalyan B, Sahota D, Symonds EM. Customised antenatal growth charts. The Lancet. 1992 Feb 1;339(8788):283-7.
2. Bernstein IM, Horbar JD, Badger GJ, Ohlsson A, Golan A. Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. Am J Obstet Gynecol. 2000 Jan 1;182(1):198-206.
3. Unterscheider J, O'Donoghue K, Daly S, Geary MP, Kennelly MM, McAuliffe FM, et al. Fetal growth restriction and the risk of perinatal mortality-case studies from the multicentre PORTO study. BMC Pregnancy Childbirth. 2014 Feb 11;14(1):63.
4. Lawn JE, Blencowe H, Pattinson R, Cousens S, Kumar R, Ibiebele I, et al. Stillbirths: Where? When? Why? How to make the data count? The Lancet. 2011 Apr 23;377(9775):1448-63.
5. Lawn JE, Ohuma EO, Bradley E, Idueta LS, Hazel E, Okwaraji YB, et al. Small babies, big risks: global estimates of prevalence and mortality for vulnerable newborns to accelerate change and improve counting. The Lancet. 2023 May 20;401(10389):1707-19.
6. Society for Maternal-Fetal Medicine (SMFM). Electronic address: pubs@smfm.org, Martins JG, Biggio JR, Abuhamad A. Society for Maternal-Fetal Medicine Consult Series 52: Diagnosis and management of fetal growth restriction: (Replaces Clinical Guideline Number 3, April 2012). Am J Obstet Gynecol. 2020 Oct;223(4):B2-17.
7. van Scheltinga JAT, Scherjon SA, van Dillen J. Nederlandse Vereniging voor Obstetrie en Gynaecologie (NVOG). 2017 [cited 2024 Jan 17]. NVOG-Richtlijn Foetale Groeirestrictie (FGR). Available from: https://www.nvog.nl/wp-content/uploads/2017/12/Foetate-groeirestricie-FGR-15-09-2017.pdf
8. Gordijn SJ, Beune IM, Thilaganathan B, Papageorghiou A, Baschat AA, Baker PN, et al. Consensus definition of fetal growth restriction: a Delphi procedure. Ultrasound Obstet Gynecol. 2016;48(3):333-9.
9. Lees CC, Stampalija T, Baschat AA, Da Silva Costa F, Ferrazzi E, Figueras F, et al. ISUOG Practice Guidelines: diagnosis and management of small-for-gestational-age fetus and fetal growth restriction. Ultrasound Obstet Gynecol. 2020 Aug;56(2):298-312.
10. Abuhamad A, Zhao Y, Abuhamad S, Sinkovskaya E, Rao R, Kanaan C, et al. Standardized Six-Step Approach to the Performance of the Focused Basic Obstetric Ultrasound Examination. Am J Perinatol. 2016 Mar;02(1):90-8.
11. DeStigter KK, Morey GE, Garra BS, Rielly MR, Anderson ME, Kawooya MG, et al. Low-Cost Teleradiology for Rural Ultrasound. In: 2011 IEEE Global Humanitarian Technology Conference. 2011. p. 290-5.
12. Self A, Chen Q, Desiraju BK, Dhariwal S, Gleed A, Mishra D, et al. Developing Clinical Artificial Intelligence for Obstetric Ultrasound to Improve Access in Underserved Regions: Protocol for a Computer-Assisted Low-Cost Point-of-Care UltraSound (CALOPUS) Study. JMIR Res Protoc. 2022 Sep;11(9):e37374.
13. van den Heuvel T, Petros H, Santini S, de Korte C, van Ginneken B. AUTOMATED FETAL HEAD DETECTION AND CIRCUMFERENCE ESTIMATION FROM FREE-HAND ULTRASOUND SWEEPS USING DEEP LEARNING IN RESOURCE-LIMITED COUNTRIES. ULTRASOUND Med Biol. 2019 Mar;45(3):773-85.
14. Arroyo J, Marini TJ, Saavedra AC, Toscano M, Baran TM, Drennan K, et al. No sonographer, no radiologist: New

system for automatic prenatal detection of fetal biometry, fetal presentation, and placental location. PLOS ONE. 2022 Feb 9;17(2):e0262107.
15. Gomes RG, Vwalika B, Lee C, Willis A, Sieniek M, Price JT, et al. A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. Commun Med. 2022 Oct 11;2(1):1-9.
16. Lee C, Willis A, Chen C, Sieniek M, Watters A, Stetson B, et al. Development of a Machine Learning Model for Sonographic Assessment of Gestational Age. JAMA Netw OPEN. 2023 Jan 4;6(1):e2248685.
17. Pokaprakarn T, Prieto JC, Price JT, Kasaro MP, Sindano N, Shah HR, et al. AI Estimation of Gestational Age from Blind Ultrasound Sweeps in Low-Resource Settings. NEJM Evid. 2022 Apr 26;1(5):EVIDoa2100058.
18. Gleed AD, Chen Q, Jackman J, Mishra D, Chandramohan V, Self A, et al. Automatic Image Guidance for Assessment of Placenta Location in Ultrasound Video Sweeps. Ultrasound Med Biol. 2023 Jan 1;49(1):106-21.
19. Gleed AD, Mishra D, Chandramohan V, Fu Z, Self A, Bhatnagar S, et al. Towards Multi-Sweep Ultrasound Video Understanding: Application in Detection of Breech Position Using Statistical Priors. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) [Internet]. Cartagena, Colombia: IEEE; 2023 [cited 2023 Dec 29]. p. 1-5. Available from: https://ieeexplore.ieee.org/document/10230662/
20. Maraci MA, Bridge CP, Napolitano R, Papageorghiou A, Noble JA. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. Med Image Anal. 2017 Apr;37:22-36.
21. Schilpzand M, Neff C, van Dillen J, van Ginneken B, Heskes T, de Korte C, et al. AUTOMATIC PLACENTA LOCALIZATION FROM ULTRASOUND IMAGING IN A RESOURCE-LIMITED SETTING USING A PREDEFINED ULTRASOUND ACQUISITION PROTOCOL AND DEEP LEARNING. ULTRASOUND Med Biol. 2022 Apr;48(4):663-74.
22. Self A, Chen Q, Noble J a., Papageorghiou A t. OC10.03: Computer-assisted low-cost point of care ultrasound: an intelligent image analysis algorithm for diagnosis of malpresentation. Ultrasound Obstet Gynecol. 2020;56(S1):28-28.
23. Plotka S, Klasa A, Lisowska A, Seliga-Siwecka J, Lipa M, Trzcinski T, et al. Deep learning fetal ultrasound video model match human observers in biometric measurements. Phys Med Biol. 2022 Feb 21;67(4).

Keywords

List the primary keywords that characterize the task.

abdominal circumference, prenatal ultrasound, low-income countries, novice operators, artificial intelligence

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

María Sofía Sappia, Medical Imaging department, Radboudumc, the Netherlands;
Keelin Murphy, Medical Imaging department, Radboudumc, the Netherlands;
Chris de Korte , Medical Imaging department, Radboudumc, the Netherlands;
Jeroen van Dillen, Department of Obstetrics and Gynaecology, Radboudumc, the Netherlands; Bram van Ginneken, Medical Imaging department, Radboudumc, the Netherlands;
Julio Joaquín López González, Delft Imaging Systems, the Netherlands;
Frank Vijn, Delft Imaging Systems, the Netherlands;
Guido Geerts, Delft Imaging Systems, the Netherlands;
Enya Séguin, Delft Imaging Systems, the Netherlands

Note: we are seeking collaboration with an African sonographer.

b) Provide information on the primary contact person.

María Sofía Sappia, Medical Imaging department, Radboudumc, the Netherlands,
mariasofia.sappia@radboudumc.nl

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One-time event with fixed submission deadline.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

This challenge will be hosted during MICCAI2024 as part of a half-day thematic event on Abdominal analysis, together with two additional challenges:
24: CURVAS: Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation (Meritxell Riera-Marin)
142: AMOS-MM: Abdominal Multimodal Analysis Challenge (Yuanfeng Ji)
During this event, each challenge will be assigned a timeslot for each organizer to present an overview of the task(s) and dataset involved, and for top-performing teams to present their solutions. Additionally, an opening and closing session hosted by the organizers of all three challenges will be held. The proposed program for this event is as follows:
00:00 - 00:10: Opening session (all challenges)
00:10 - 01:10: ACOUSLIC-AI
01:10 - 01:30: Break
01:30 - 02:30: CURVAS
02:30 - 02:50: Break
02:50 - 03:50: AMOS-MM
03:50 - 04:00: Closing session (all challenges)
Please note that times are indicated in terms of duration and will be adjusted accordingly once the starting time of the event is known.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

Not available yet, but will be found on grand-challenge.org.

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

For this challenge, only automatic methods encapsulated in Docker containers will be allowed. Submissions involving semi-automated or interactive methods will not be considered. Furthermore, all Docker containers will be operated in an offline environment, meaning they won't have internet access for downloading or uploading resources. Therefore, participants must ensure that all required resources, such as pre-trained AI model weights, are included within the Docker containers before submission.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Participants are allowed to use pre-trained AI models from computer vision or medical imaging datasets (like ImageNet, Medical Segmentation Decathlon), as well as other datasets beyond the ones provided in this challenge. However, this is conditional on the data and/or models being published under a permissive license. Moreover, for each submission, participants must explicitly declare the source and intended use-case of such data or models.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Individuals affiliated with any of the sponsoring or organizing bodies (i.e. Radboud University Medical Center, Delft Imaging Systems) are welcome to participate in the challenge. However, they will not be eligible for inclusion in the final ranking during the Testing Phase, nor will they be eligible to receive any prizes.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

We are seeking sponsorship to award prizes to the top 3 performing participants.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

During the Development Phase, the performance and rankings of all submitted AI algorithms will be made public via the challenge's live leaderboard.
After the final submission date of the Testing Phase, the corresponding leaderboard will be made publicly availabe in the challenge website. In addition, the top three performing AI algorithms and their authors will be added to the homepage of the challenge website. Additionally, these results will be formally presented at the MICCAI 2024 challenge session, where they will be announced. The top five performers will be given the opportunity to present their results during this session.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)

- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

The publication policy for this challenge is as follows:
- Authorship in the final challenge paper: Up to three members of each participating team, specifically those who played a significant role in the algorithm's design, will be acknowledged as co-authors in the final challenge paper.
- Independent publication of results: Participants are free to publish their algorithms and findings independently once the challenge concludes and the embargo period is over (see below). However, we require that any such independent publications reference both the summary paper of the challenge and the data publication paper.
- Embargo period: Although teams are permitted to publish their results independently, there is an embargo period of 6 months to ensure all analyses and computations outlined in this proposal can be conducted. This means that for the first six months following the end of the challenge, the challenge organizers retain exclusive rights to publish the collective results, after which teams may proceed with their individual publications. However, we acknowledge the value of early research dissemination and, therefore, permit the submission of arXiv preprints during the embargo period.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants are required to use their own computing resources for model training. For each submission, they can submit one AI model, packaged as a Docker container, on the grand-challenge platform. During the Development Phase, their performance will be assessed using the hidden validation and tuning cohort, and team rankings will be updated in real-time on a live, public leaderboard.
In the Testing Phase, teams are allowed to submit a single AI algorithm for assessment. This evaluation will be conducted on the hidden testing cohort. The algorithms' performance on this cohort will be used to establish the final rankings, which will determine the top 5 best-performing models.
During both phases (Development and Testing), the algorithms will be run on the grand-challenge.org platform.

We will provide participants with access to a code repository containing a baseline algorithm, as well as comprehensive instructions on how to package it as a Docker container. This will serve as a foundational template for participants to build upon and customize for their specific requirements.

General instructions for submitting AI models encapsulated in Docker containers to the Grand Challenge platform can be found in: https://grand-challenge.org/documentation/creating-an-algorithm-container/

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Development phase: this is a preliminary evaluation phase for participants to familiarize themselves with the submission system and test their algorithms on a small set of cases (50). During this phase, a maximum of 10 submissions per team is allowed. Performance metrics and live ranking will be provided, but will not be

considered for the final official ranking.
Testing phase: a one-time submission that is used to compute the final performance metrics and ranking.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Release of training data: 1st April 2024
Registration period: 1st April - 30th April 2024
Last submission: 15th July 2024
Results release: 8th October 2024

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The training set includes data acquired in three peripheral healthcare units (PHUs) in Sierra Leone. These data are under the ethical approval of the Office of the Sierra Leone Ethics and Scientific Review Committe, Government of Sierra Leone (Ref. Num.: 015/06/2022).
The test set contains data from two PHUs in Tanzania and from Radboumc, in the Netherlands. Data from Tanzania were approved by the Karatu District Council, Government of Tanzania (Ref. Num.: KDC/DED/M2/1/VOL.IV/26). Data from Radboudumc were approved by the local ethics committee CMO Arnhem-Nijmegen (Ref. Num.: 2020-6701).

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Data can be used and distributed under a CC BY-NC-SA license.

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation software will be made publicly available in the form of a github repository by the date the challenge is officially released.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

It is required that participants link their algorithms to a public repository with a version tag and an Apache 2.0 license or MIT license file. Should a participant prefer a different permissive open-source license, they should leave a message in the channel website's Forum.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Delft Imaging Systems is the sponsor and funder of this challenge. Because we opted for a non-commercial CC BY-NC-SA license, there are no conflicts of interest as all algorithms developed on this challenge will fall under the same license type and the data cannot be used for commercial enterprises.
Access to the test case labels will be restricted to Delft Imaging Systems employees, María Sofía Sappia (Radboudumc), Keelin Murphy (Radboudumc), Chase Neff (Radboudumc), Anette Beverdam (Radboudumc), Suzie Klaassen (Delft Imaging Systems), Liron Bundel (Thirona), possibly also an African sonographer (to be determined).

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Main fields of application: screening, research.

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Task categories: classification, segmentation, detection, localization, prediction.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Target cohort: women with confirmed or suspected pregnancy that come for an antenatal care visit in low-resource settings.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Data from two different cohorts are included in this challenge. For both cohorts, data was collected by novice operators using a low-cost handheld ultrasound probe and following a blind-sweep acquisition protocol.
African cohort: women with confirmed or suspected pregnancy receiving antenatal care at Peripheral Healthcare Units (PHUs) located in Sierra Leone and Tanzania.
Radboudumc cohort: pregnant women aged 18 years or older, fluent in either Dutch or English, who are undergoing routine ultrasound examinations at the Department of Obstetrics and Gynecology, Radboud University Medical Center, Nijmegen.
In both cohorts, only scans with sufficient image quality - as reported by expert analysts - were included. Data was restricted to scans corresponding to singletons with gestational ages in the range 20 - 32 (incl.) weeks. Cases from

the Radboudumc cohort that did not have an accompanying abdominal circumference measurement from a conventional clinical ultrasound made by an expert sonographer were excluded

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

2D B-mode transabdominal ultrasound using a low-cost handheld device and a blind-sweep acquisition protocol performed by a novice operator.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

- Abdomen annotation masks: Abdomen pixel masks derived from ellipse annotations are provided for both the public training data, as well as for the hidden validation and test sets. These annotations are of two types:
1. Optimal planes: annotation masks on frames considered to be optimal for the estimation of fetal abdominal circumference.
2. Suboptimal planes: annotation masks on frames considered to be suboptimal but still good enough to provide an estimation of the fetal abdominal circumference.
- Abdominal circumference measurements: each case is accompanied by a corresponding abdominal circumference measurement (in mm.)

b) ... to the patient in general (e.g. sex, medical history).

No patient information will be provided to the participants, as it is not necessary for solving the proposed task.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Abdomen of pregnant women intended to include full uterus and fetus within, shown in 2D B-mode transabdominal ultrasound blind sweep sequences.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The output target of the algorithm is a 2D binary mask of the fetal abdomen together with the corresponding frame, identified as the optimal for measurement, derived from scans acquired by novice operators following a standardized blind-sweep protocol. This output is evaluated based on two criteria: a) the selection of the frame (classification task) and b) the segmentation accuracy of the fetal abdomen within the selected frame (segmentation task). The masks provided should be fit for an ellipse fitting tool to measure its circumference during evaluation. This tool will be provided for reference in the publicly available evaluation code.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

The challenge involves analyzing a series of 2D ultrasound frames extracted from blind-sweep sequences acquired by novice operators. Participants are tasked with identifying the most suitable frame for measuring the fetal abdominal circumference. Along with selecting this optimal frame, participants must also provide the binary segmentation mask of the abdomen on the ultrasound image corresponding to the selected frame. This segmentation mask should be fit for an ellipse fitting tool to measure its circumference during evaluation.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

All blind-sweep imaging data were acquired using the MicrUs Pro-C60S (Telemed, Lithuania), a low-cost ultrasound probe.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Imaging data included in this challenge was acquired by novice users (1 hour training) with a
low cost portable probe (MicrUs Pro-C60S, Telemed, Lithuania) connected to a
smartphone. The users were blinded to the ultrasound images during acquisition
and performed free-hand sweeps according to instructions provided. The instructions were presented in the smartphone screen and guided them through the obstetric sweep protocol (OSP) proposed by Stigter et al., 2011. The OSP is a blind sweep acquisition protocol consisting of six sweeps over the gravid abdomen: three transverse sweeps (1-3) in caudocranial direction, and three sagittal sweeps (4-6) from participants' left to right.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

This study includes data acquired at:
- three peripheral healthcare units in Tonkolili District, in Sierra Leone;
- two peripheral healthcare units in Karatu District, Tanzania;
- Radboud University Medical Center (Radboudumc), the Netherlands.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Data used in this study were acquired by novice operators with one hour of training. While none of the operators in this study had experience in performing ultrasound scans, operators in the Radboudumc cohort were medical students, assumed to have prior knowledge about the technical background of ultrasound equipment and about practical ultrasound techniques.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases consist of a pair of 2D B-mode ultrasound sweeps and abdominal circumference annotations. Both the sweeps and annotations correspond to a series of 840 frames of shape 744x562 pixels and a fixed spacing of 0.28 mm/pixel. The annotations correspond to pixel masks of the abdomen on individual frames, and pertain to either of two categories: optimal and suboptimal planes for abdominal circumference measurement. Per frame, the annotation pixels adopt one of three values: pixel value 0 indicates no annotation (background), pixel value 1 indicates a mask drawn on an optimal plane, and pixel value 2 indicates a mask drawn on a suboptimal plane. Cases are also accompanied by the corresponding abdominal circumference reference value (in mm).

b) State the total number of training, validation and test cases.

This challenge is structured into three distinct data splits with the following characteristics and use cases:
- Public Training and Development Set (300 cases): This set is accessible to all participants for the purpose of developing AI models during the Development Phase. The data, released under a CC BY-NC-SA license, encompass cases from three Public Health Units (PHUs) in Sierra Leone.
- Hidden Validation and Tuning Set (50 cases): Designed to support a live, public leaderboard, this set aids in model selection and tuning during the Development Phase. It consists of previously unseen data from two PHUs in Tanzania.
-Hidden Test Set (250 cases): Utilized for benchmarking AI models at the conclusion of the Testing Phase, this set contains unseen testing data from two PHUs in Tanzania (150 cases) and from Radboudumc (100 cases).

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

In this challenge, a total of 600 cases were selected to develop, tune and benchmark AI models, including a small subset of 100 cases to benchmark these solutions against clinical standards. The 600 cases are distributed across four data splits to facilitate the development, tuning, and evaluation of the AI models. The Public Training and Development Set (300 cases) provides a foundational dataset for initial model development, the Hidden Validation and Tuning Set (50 cases) allows for adjusting and checking the models during development, and the Hidden Test Set (250 cases) enable a final assessment of the models' performance on unseen data from three independent centers.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

For all datasets and cohorts, incuding training, testing and validation, the following should be noted:
- Cases that did not contain any optimal or suboptimal frames for fetal abdominal circumference measurement were excluded. This was due to the insufficient number of such cases to support effective training of the algorithms.
- The included cases were restricted to gestational ages between 20 and 32 weeks. This range is the most commonly represented and was chosen based on the availability of cases. There were not enough cases outside of this range in both the African and Radboudumc cohorts to effectively train an AI model. Specifically, at Radboudumc, this situation occurs as most ultrasound appointments are scheduled at 20 and 32 weeks of gestation.
- Only singleton pregnancies were included in the study. This decision was based on the limited number of multiple gestation cases and the complexities in defining an optimal plane for measurement. In multiple gestations, variations in the sizes of fetuses' abdomens and the potential presence of more than one fetus in a single frame complicate the process of obtaining a single, definitive measurement.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

For both the training and validation sets, ellipse annotations were obtained from manual annotations performed by human readers on every initial and final frame where the corresponding structure (transverse plane of the abdomen) and type (optimal/suboptimal) were observed. Annotations on intermetiade frames were automatically generated using linear interpolation. In cases where the size or position of the structure changed greatly, additional manual annotations were provided to ensure accuracy in the interpolation process. All ellipse annotations were filled in to provide participants with pixel mask annotations rather than ellipse contours. Per frame, the annotation pixels adopt one of three values: pixel value 0 indicates no annotation (background), pixel value 1 indicates a mask drawn on an optimal plane, and pixel value 2 indicates a mask drawn on a suboptimal plane.

For all cases in the Public Training and Development Set (300 cases), annotations were performed by two readers with 20 hours of training in acquiring and analyzing blind-sweep ultrsaound data. Their experience extends over two years, with one reader dedicating a total of 120 hours and the other 300 hours to analyzing such data. Each reader annotated cases independently, with an approximate distribution of 50% cases each. For cases in the Hidden Validation and Tuning Set (50 cases) and in the Hidden Test Set (250 cases), the annotation process was performed by two readers with higher expertise. A radiologist in training with experience in ultrasound annotated all cases. Subsequently, these annotations were reviewed by a sonographer with 37 years of experience. Instances of disagreement were resolved via a consensus meeting.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The text provided to the annotators is reproduced here below:

Fetal structures to segment - definitions

Abdominal circumference standard plane

The abdominal circumference standard plane consists of a transverse section through the upper abdomen, revealing key fetal landmarks such as the fetal stomach, umbilical vein, and portal sinus. This view includes symmetrical fetal ribs, the junction of the umbilical vein, and the fetal stomach, while the kidneys and cord insertion should not be visible. There should be no maternal abdominal compression.

Abdominal circumference - optimal plane

Optimal planes refer to views of the abdomen that approximate the abdominal circumference standard plane as close as possible for a given case. The best planes that would be selected to provide an abdominal circumference measurement, even if not ideal.

Abdominal circumference - suboptimal planes

Suboptimal planes of the abdominal circumference represent secondary options for estimating abdominal circumference for a given case.

Note:

The selected optimal/suboptimal planes should be deemed good enough to provide an abdominal circumference estimate. To decide which plane is optimal/suboptimal for a given case, try asking yourself the following questions:

- Would you definitely choose this plane and consider it correct for measuring AC? Then this is an optimal plane.
- Would you only choose this plane for measuring AC if you couldn't find a better one? Then this is a suboptimal plane.

Annotation instructions

Both structures should only be annotated in transverse views. The annotation should be performed using the corresponding ellipse tool (one per structure). Make sure that the ellipse includes the whole structure. Bear in mind that it is not necessary to be pixel-specific.

The first and last frame where each structure type (optimal/suboptimal) is visible should be annotated. It is not necessary to annotate intermediate frames unless the size or position of the target structure in it changes with respect to previous ones. In those cases, a new ellipse should be drawn.

In case no plane corresponding to either case is present in the scan, the annotation question should be left blank.

If multiple planes of the same type (optimal/suboptimal) are present across different sweeps, all of them should be annotated.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Public Training and Development Set: annotations were performed by two readers with 20 hours of training in acquiring and analyzing blind-sweep ultrsaound data. Their experience extends over two years, with one reader dedicating a total of 120 hours and the other 300 hours to analyzing such data.

Hidden Validation and Tuning Set: annotations were performed by the more experienced reader in the training set (20 hours of training, 300 hours of experience analyzing blind-sweep data).

Hidden Test Set: a radiologist in training with experience in ultrasound annotated all cases. These annotations were reviewed by a sonographer with 37 years of experience.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

There were no multiple annotations in none of the data splits used for this challenge.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

No pre-processing was applied to the blind-sweep video sequences. For the ellipse annotations, the following steps were applied: the ellipses were filled in to create mask annotations and any pixels located outside the field of view were set to zero. Consecutive integer labels were assigned to each ellipse type: optimal (1) and suboptimal (2). Background was set to zero.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Possible sources of error related to inter- and intra-annotator variability in image annotation include the difficulty of providing pixel-perfect annotations, the ellipse tool's limitations in accurately delineating non-oval shapes, and errors in circumference estimation when the abdominal structure being measured falls outside the field of view. Additionally, subjective decisions in determining whether a plane is optimal or suboptimal, challenges in clearly delineating boundaries in poorly defined planes, and the selection of cross-sectional planes that deviate from the transverse plane further contribute to potential inaccuracies. All above mentioned error sources are also present in clinical practice in prenatal ultrasound as standard in high-income countries.

Other sources of error relate to the measurement of abdominal circumference, where the plane used for measurement is unlikely to be the precise standard plane that an expert sonographer would be able to acquire.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other relevant sources of error are present.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

To assess the performance of the algorithms in this challenge's task, we employ a combination of metrics that reflect accuracy, precision, and clinical relevance.
- Dice Similarity Coefficient (DSC): This metric evaluates the spatial accuracy of the algorithm's ability to segment the fetal abdomen. It measures the overlap between the algorithm-provided binary segmentation mask and the

ground truth pixel mask annotation.

- Weighted Frame Selection Score (WFSS): The WFSS is a custom metric designed to assess the algorithm's ability to classify the frames and select the most clinically appropriate frame for measurement. It can adopt three possible values: [0, 0.6, 1]. Given that the dataset includes annotations for both optimal and suboptimal planes, the WFSS assigns the highest weight (1) to frames containing optimal planes when they are correctly identified and selected by the algorithm. Specifically, if a frame containing an optimal plane is present and correctly identified, the algorithm receives a full score (1) for that case. If the algorithm selects a frame with a suboptimal plane when an optimal one is available, it receives an intermediate score (0.6) proportionate to the clinical relevance of suboptimal planes. If no optimal planes are present, the maximum score (1) is awarded for the correct identification of frames containing suboptimal planes. If the algorithm selects a frame that does not contain either an optimal or suboptimal plane in a case where at least one such plane is available, the algorithm receives the minimum score (0). The choice of 0.6 for suboptimal frames - slightly higher than the average of 0.5 - was made to underscore the relative importance of identifying a suboptimal frame over missing a frame entirely.
- Hausdorff Distance (HD): This metric evaluates the maximum discrepancy between the predicted and ground truth pixel structure boundaries.
- Mean Absolute Error (MAE): For the assessment of abdominal circumference, the Mean Absolute Error (MAE) offers an interpretable measure of accuracy. This metric will be used to evaluate the precision of the abdominal circumference measurements, which are computed using an ellipse fitting tool on the binary segmentation masks provided by the algorithms. The MAE reflects the average deviation of these derived measurements from the ground truth values, expressed in millimeters. A lower MAE signifies greater accuracy of the algorithm in segmenting the region of interest in a manner that leads to precise circumference calculations, even though the circumference measurement itself is not a direct output of the algorithm.

Rankings are computed by considering a composite score that integrates all the above metrics, with a predetermined weighting system that reflects their relative importance as determined by the assessment aims. The weighting system is designed to prioritize the clinical utility of the algorithm, ensuring that the highest-ranked algorithms are not only accurate but also align with clinical decision-making practices.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The chosen metrics for evaluating the algorithms in the context of this challenge were selected to address both the technical performance and the clinical applicability of the algorithm.

Dice Similarity Coefficient (DSC): DSC is a widely accepted metric for assessing the accuracy of spatial overlap in medical image segmentation. In the context of fetal abdominal circumference measurement, the DSC ensures that the algorithm's predicted masks align closely with the expert annotations, which is critical for reliable measurements.

Weighted Frame Selection Score (WFSS): The WFSS is a custom metric designed to assess the algorithm's ability to select the most clinically appropriate frame for measurement from a mixture of optimal and suboptimal planes in a case. This task is of great importance given the nature of the data used in this challenge regarding the acquisition protocol and the inexperience of the target operators. This task mirrors the critical decision-making that is required by human experts to report the most accurate possible clinical measurement on the available blind sweep data. In this context, the WFSS score pairs the algorithm performance in selecting a clinically appropriate frame for measurement with that of human experts.

Hausdorff Distance (HD): The HD provides an important measure of the geometric accuracy of the predicted ellipses

Mean Absolute Error (MAE): The MAE is a direct measure of the difference in abdominal circumference measurements between the algorithm's predictions (as derived from the output binary segmentation mask) and the ground truth. This metric is critical as it translates directly into the clinical utility of the algorithm. An algorithm with a low MAE indicates that it can provide measurements that are close to what an expert would obtain, which is the ultimate goal of the imaging task.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Composite Score:
The performance rank for all submitted algorithms is determined based on the following composite score:

Composite score = w_1 DSC + w_2 WFSS + w_3 (1 - HD_norm) + w_4 (1 - MAE_norm)

Metric Components:
DSC (Dice Similarity Coefficient): Measures the overlap between the predicted and ground truth pixel masks.
WFSS (Weighted Frame Selection Score): Reflects the clinical appropriateness of the selected frame, with a higher score for optimal planes.
HD_norm (Normalized Hausdorff Distance): Represents the maximum distance between the predicted and ground truth structure boundaries, normalized to a [0, 1] range where 0 represents the smallest (best) distance and 1 represents the largest (worst) distance. Since lower values reflect higher performance, we subtract HD_norm from 1.

HD_norm=(HD - HD_min)/(HD_max - HD_min)
Where:
HD_min is the smallest HD in the dataset.
HD_max is the largest HD in the dataset.

MAE_norm (Normalized Mean Absolute Error): Represents the accuracy of the abdominal circumference measurement. The MAE is normalized to a range of [0, 1] where 0 represents the best (most accurate) score and 1 represents the worst (least accurate) score. Since lower values reflect higher performance, we subtract MAE_norm from 1.

MAE_norm=(MAE - MAE_min)/(MAE_max - MAE_min)
Where:
MAE_min is the smallest MAE in the dataset (most accurate).
MAE_max is the largest MAE in the dataset (least accurate).

Weight Assignments: The weights for each metric sum to 1 and are assigned as follows:
w1 (DSC weight): 0.25
w2 (WFSS weight): 0.30
w3 (HD_norm weight): 0.10
w4 (MAE_norm weight): 0.35

The weight assignment reflects the prioritization of the accuracy of measurements (MAE_norm) as the most critical factor, emphasizing the importance of precise clinical measurement. This is followed closely by the clinical relevance of the frame selection (WFSS), which ensures the selection of the most appropriate planes for assessment. The accuracy of the segmentation overlap (DSC) is also given significant importance but is ranked slightly lower, focusing on the preciseness of the segmentation in the images. Lastly, the geometric accuracy of the boundary delineation (HD_norm) is considered, but with the least priority compared to the other factors. Additionally, we will consider conducting an analysis of the challenge ranking's robustness, specifically employing Kendall's tau as a statistical measure to ensure the reliability and consistency of our rankings.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Submissions with missing results will be disqualified, and not presented on any leaderboard.

c) Justify why the described ranking scheme(s) was/were used.

The ranking scheme and its corresponding weight assignment are designed to align with the primary objectives of accurately measuring fetal abdominal circumference on the most clinically relevant ultrasound frames. This approach prioritizes MAE_norm, underscoring the importance of precise clinical measurements. The substantial weight assigned to WFSS reflects the necessity of choosing the most appropriate frames for this measurement, given the varying image quality and diverse anatomical views of abdominal planes in the data. The DSC, while assigned a slightly lower weight, remains an essential metric, ensuring that segmentation accuracy is adequately considered, as it directly impacts the reliability of the measurements. Finally, HD_norm, with the least weight, addresses the geometric accuracy of the segmentation.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The following statistical methods will be employed in the challenge paper to compare the performance of the top three performing algorithms on the Testing Phase:

Comparison of Abdominal Circumference Measurements:
Objective: The aim is to determine whether the AC values predicted by the algorithm are statistically different from the ground truth values specified. The test is designed to verify whether the differences are statistically

different from zero. This comparison will assess the accuracy of each algorithm in estimating fetal abdominal circumference, based on the binary segmentation masks they provide. The statistical analysis will determine how closely each algorithm's derived measurements align with the reference AC values.
Method: A paired t-test is conducted to compare the AC measurements derived from the algorithm's output with the corresponding ground truth for each case. This test was chosen to assess whether the average difference between paired observations (algorithm vs. ground truth) is statistically different from zero.
Normality Check: Prior to conducting the t-test, the Shapiro-Wilk test is used to evaluate the normality of the differences between paired measurements. If the differences do not follow a normal distribution, a non-parametric alternative, the Wilcoxon Signed Rank Test, is utilized.
Significance Level: The threshold for statistical significance is set at $p < 0.05$.

b) Justify why the described statistical method(s) was/were used.

For abdominal circumference measurements, the paired t-test is employed due to its suitability in comparing the algorithms' derived measurements with the ground truth on a case-by-case basis. This test was selected to assess the mean difference between each pair of algorithm-generated and ground truth measurements. Before employing the t-test, the normality of data is verified using the Shapiro-Wilk test, a standard procedure for ensuring the data meets the t-test's normal distribution requirement. If normality is not observed, the Wilcoxon Signed Rank Test is used as a non-parametric alternative.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

In the forthcoming challenge paper, we plan to conduct further analyses on the results. In particular, we will investigate whether combining algorithms via ensembling significantly improves the performances. We will also investigate which is the most suitable light-weight model for implementation on mobile devices, given the limited hardware capabilities. Choosing a model that can run on a mobile device eliminates the need for more expensive hardware or connectivity to transfer data and results on a network. This is optimal for a low resource setting.
Additionally, we will assess the computational efficiency of the models, utilizing metrics such as frames per second (FPS), to ensure our solutions are not only effective but also practically deployable in various settings.
In this paper, we will also benchmark the performance of the top three performing methods against clinical standards, by comparing their fetal abdominal circumference (AC) measurement outputs with those obtained by expert sonographers in clinical practice.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

N/A