9-2022

Generalisability of deep learning models in low-resource imaging settings: A fetal ultrasound study in 5 African countries

Carla Sendra-Balcells

V´ıctor M. Campello

Jordina Torrents-Barrena

Yahya Ali Ahmed

Mustafa Elattar

See next page for additional authors

Authors

Carla Sendra-Balcells, V´ıctor M. Campello, Jordina Torrents-Barrena, Yahya Ali Ahmed, Mustafa Elattar, Benard Ohene Botwe, Pempho Nyangulu, William Stones, Sikolia Wanyonyi, and Marleen Temmerman

Generalisability of deep learning models in low-resource imaging settings: A fetal ultrasound study in 5 African countries

Carla Sendra-Balcells[a], Víctor M. Campello[a], Jordina Torrents-Barrena[b], Yahya Ali Ahmed[c], Mustafa Elattar[d,e], Benard Ohene Botwe[f], Pempho Nyangulu[g], William Stones[g], Mohammed Ammar[h], Lamya Nawal Benamer[i], Harriet Nalubega Kisembo[k], Senai Goitom Sereke[j], Sikolia Z. Wanyonyi[l], Marleen Temmerman[m], Kamil Mikolaj[n], Martin Grønnebæk Tolsgaard[n], Karim Lekadir[a]

[a]*Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[b]*HP Inc., Barcelona, Spain*
[c]*Obstetrics and Gynecology Dept., School of Medicine, Suez University, Egypt*
[d]*Medical Imaging and Image Processing, Center of Informatics Science, Nile University, Egypt*
[e]*Research & Development Division, Intixel, Egypt*
[f]*Dept. of Radiography, School of Biomedical & Allied Health Sciences, College of Health Sciences, University of Ghana, Ghana*
[g]*Kamuzu University of Health Sciences, Blantyre, Malawi*
[h]*Dept. of Electrical Engineering Systems & Laboratory of Engineering System and Telecommunication, University of M'Hamed Bougara Boumerdes, Algiers, Algeria*
[i]*Obstetrics and Gynecology Dept., School of Medicine, Algiers University, Algeria*
[j]*Dept. of Radiology and Radiotherapy, School of Medicine, Makerere University College of Health Sciences, Kampala, Uganda*
[k]*Dept. of Radiology, Mulago National Referral and Teaching Hospital, Kampala, Uganda*
[l]*Dept. of Obstetrics and Gynaecology, Aga Khan University Hospital, 3rd Parklands Avenue, Nairobi, Kenya.*
[m]*UNDP/UNFPA/UNICEF/WHO/The World Bank Special Programme of Research, Development and Research Training in Human Reproduction (HRP), Dept. of Reproductive Health and Research, World Health Organization, Geneva, Switzerland.*
[n]*Copenhagen Academy for Medical Education and Simulation and Dept. of Obstetrics, Rigshospitalet, Denmark*

Abstract

Most artificial intelligence (AI) research and innovations have concentrated in high-income countries, where imaging data, IT infrastructures and clinical expertise are plentiful. However, slower progress has been made in limited-resource environments where medical imaging is needed. For example, in Sub-Saharan Africa the rate of perinatal mortality is very high due to limited access to antenatal screening. In these countries, AI models could be implemented to help clinicians acquire fetal ultrasound planes for diagnosis of fetal abnormalities. So far, deep learning models have been proposed to identify standard fetal planes, but there is no evidence of their ability to generalise in centres with low resources, *i.e.* with limited access to high-end ultrasound equipment and ultrasound data. This work investigates for the first time different strategies to reduce the domain-shift effect arisen from a fetal plane classification model trained on one clinical centre with high-resource settings and transferred to a new centre with low-resource settings. To that end, a classifier trained with 1,792 patients from Spain is first evaluated on a new centre in Denmark in optimal conditions with 1,008 patients and is later optimised to reach the same performance in five African centres (Egypt, Algeria, Uganda, Ghana and Malawi) with 25 patients each. The results show that a transfer

---

learning approach for domain adaptation can be a solution to integrate small-size African samples with existing large-scale databases in developed countries. In particular, the model can be re-aligned and optimised to boost the performance on African populations by increasing the recall to 0.92 ± 0.04 and at the same time maintaining a high precision across centres. This framework shows promise for building new AI models generalisable across clinical centres with limited data acquired in challenging and heterogeneous conditions and calls for further research to develop new solutions for usability of AI in countries with less resources and, consequently, in higher need of clinical support.

1. Introduction

In the age of Big Data and digital health, a great excitement has been generated around the extraordinary opportunities that artificial intelligence (AI) may offer in tomorrow's healthcare. In particular, deep learning based methods have shown great promise for the analysis of complex biomedical data. In medical imaging, deep learning has already made an impact in a wide range of applications such as cardiology [1], cancer [2] or brain imaging [3] among many others [4, 5]. However, existing developments have been mostly focused on applications in high-resource settings, where there is a greater access to large clinical imaging datasets for training deep learning networks. Applications in environments with limited resources, such as countries across the African continent, have been scarce. Thus, there is a risk of increasing inequalities in global health due to the disparities in the application of AI in medical imaging.

Paradoxically, in places with important medical imaging needs the development of AI is at its lowest level. For example, in Sub-Saharan Africa, there is a very high level of neonatal mortality (*i.e.* 27 deaths per 1,000 births in 2019, compared to an average rate of 3.4 per 1,000 in the European Union [6]) and a very high burden of stillbirths [7], but the access to antenatal screening is limited, especially in rural Africa. A statistical study from the World Health Organisation highlights a critical shortage of trained obstetricians across African countries [8]. However, the diagnosis of fetal abnormalities in clinical practice requires clinical expertise in order to acquire and quantify ultrasound images of the fetus.

At the same time, the literature is abundant with AI developments and deep learning implementations to facilitate the tasks of fetal ultrasound scanning, quantification and diagnosis [9, 10, 11, 12, 13, 14]. For example, several techniques have been developed to help clinicians acquire fetal ultrasound planes that are adequate for further quantification, *i.e.* that cover specific structures such as the fetal head, fetal femur, fetal abdomen and fetal thorax. Baumgartner et al. [15] were the first to identify and localise 13 fetal standard planes in real time during 2D ultrasound mid-pregnancy examinations using convolutional neural networks (CNNs). Later, similarly but in a retrospective manner, Maraci et al. [16] automated the task of detecting the fetal presentation and heartbeat from a pre-defined free-hand ultrasound video with support vector machines and a conditional random field model instead of CNNs, due to the limited number of samples available for their study. Alternative approaches involved obtaining standard planes from 3D ultrasound (US) volumes to learn the mapping between the 2D plane and the transformation required to move towards the standard plane within the 3D volume [17]. More recently, Burgos-Artizzu et al. [18] evaluated the maturity of state-of-the-art CNNs to automatically classify 2D maternal fetal US and released a large open-source dataset to promote further research on the matter. While some AI tools are being implemented to recognise multiple views in fetal imaging [19], others focus their work on specific anatomical structures [20, 21, 22].

2

Despite the big amount of research studies, the proposed techniques have been trained and/or tested with fetal ultrasound datasets from high-income countries, such as the UK [15, 16, 17], China [21], United States [19] or Spain [18, 20]. There is no evidence of the ability of these deep learning models to generalise in centres with low resources, *i.e.* with lower image quality and limited number of data to calibrate the models. To our knowledge, only one study assessed the applicability of a fetal ultrasound plane classifier across countries, but exclusively in high-resource imaging settings: a classifier was trained in the UK and tested successfully in Denmark [23].

This paper presents the very first study investigating the transferability of AI models trained in high-income settings and their applicability to process images acquired in low-income settings. Specifically, we build fetal ultrasound classifiers using large-scale datasets from Spain and assess their performance when applied to fetal ultrasounds acquired in five different African countries with differences in scanners, populations and contexts. Furthermore, transfer learning, an effective strategy to adapt a pre-trained model to a new domain [24], is implemented and assessed under real world conditions, *i.e.* based on very small size samples of new images from the local sites, to assess the resources and efforts needed to re-calibrate the original model to obtain high-performance in each local clinical centre. The main contributions of this work are summarised as follows:

- The collection and annotation of a new multi-centre fetal ultrasound dataset comprised of 120 patients from five African countries, namely Algeria, Egypt, Malawi, Uganda and Ghana. This dataset is the most diverse ultrasound dataset acquired to our knowledge and represents real clinical scenarios across low-resource imaging settings. As a result, the ultrasound appearance, both globally and locally, exhibits marked differences as clearly illustrated in Figure 1.

- Application and evaluation of a fetal plane classification model trained in high-income settings within a diversity of clinical centres in low-resource imaging settings. To that end, a deep learning model is built using a large and diverse European dataset from Spain (12,400 images from two hospitals) [18] and externally evaluated in two different settings: 1) with fetal ultrasounds from four centres located in another European country, namely Denmark [23], to assess generalisability under similar settings and resources; and 2) with the fetal ultrasound images acquired from five African countries to assess generalisability under different conditions and contexts.

- Implementation of adaptive learning approaches to optimise the scalability of the classifiers in settings with distinct imaging conditions and lower image quality. This includes the estimation of the minimal number of samples required for calibrating the models to reach maximal performance for standard fetal plane classification in each of the local African centres. As a result, realistic strategies are derived for the scalability and applicability of existing deep learning models across new centres and imaging settings with minimal efforts.

2. Methods

2.1. Datasets

The data used in this study and described in Table 1, consist of seven datasets acquired in distinct clinical centres. The first is a publicly available dataset from two centres in Spain [18]. The second has been acquired in four different centres in Denmark. Finally, the last five datasets contain US samples from five African countries (Malawi, Algeria, Uganda, Ghana and Egypt). The subjects have been scanned using a range of scanner US vendors by GE Medical Systems, Siemens, Edan Instruments, Shenzhen Mindray Bio-Medical Electronics and Aloka. Moreover, the frequency range of the curved transducer used during the acquisition ranges from 3 to 7.5 MHz and the screening
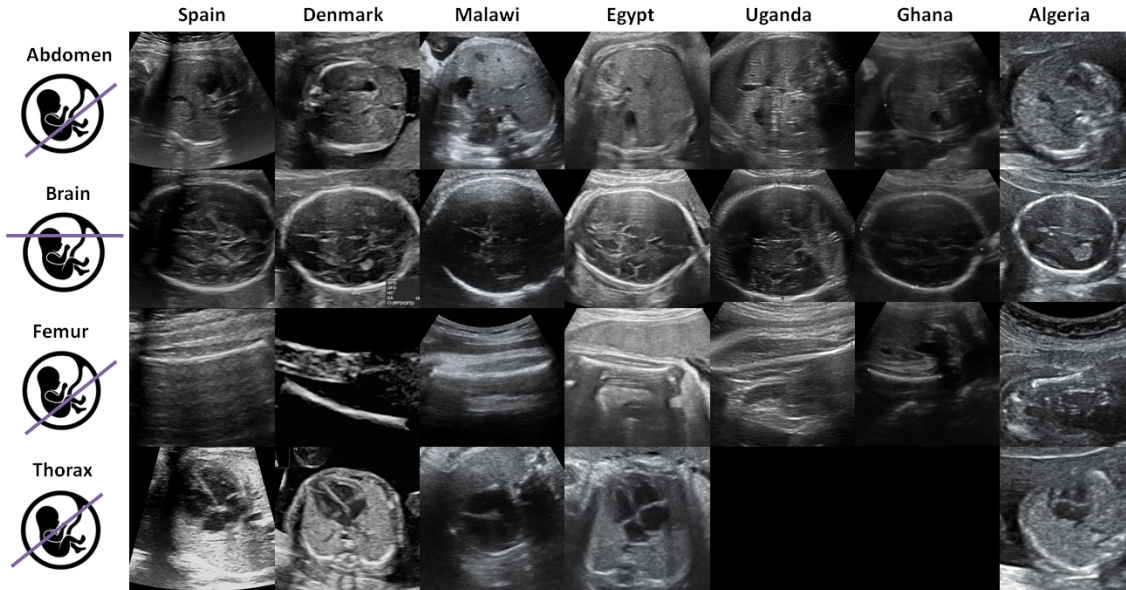
Figure 1: Image examples of the maternal-fetal US categories from our multi-centre dataset.

was performed during 2nd and/or 3rd trimester of the pregnancy period. Finally, clinical experts or research technicians who received extensive training in US classification were asked to classify all images into femur, thorax, head and abdomen. The other less common planes were classified into the 'other' category. However, the five African datasets lack the 'other' category, and two of them do not contain samples of the thorax, representing a real clinical scenario in low-resource settings. Details about the number of samples used in each case are summarised in Table 2.

2.1.1. European datasets

The first large and open-source dataset with 1,792 patients was collected at BCNatal [18], an institution with two associated hospitals (Hospital Clínic and Hospital Sant Joan de Déu, Barcelona, Spain). All pregnant women attending for routine pregnancy screening during the 2nd and 3rd trimesters in between October 2018 and April 2019 were included in the study. Images were acquired using six different US machines by different operators with similar experience. The US machines used were three Voluson E6, one Voluson S8, one Voluson S10 (GE Medical Systems) and one Aloka. Images were taken using a curved transducer with a frequency range between 3 and 7.5 MHz. Operators were instructed to avoid using any type of post-processing or artifacts such as smoothing, noise, pointers or calipers when possible.

The second big dataset of this study was composed by 2nd and 3rd trimester US scans obtained from four fetal medicine centres in Denmark (Copenhagen University Hospital Rigshospitalet, Hvidovre Hospital, Herlev Hospital and Nordsjællands Hospital Hillerød). The scans were completed in the period in between 2009 and 2017 and were obtained using three GE machines: Voluson E7, Voluson E8 and Voluson E10. Images were taken using a curved transducer with a frequency range between 3 and 7.5 MHz and no post-processing or artifacts were applied.

Table 1: Details about the seven datasets considered in the study.

Country	Vendors	Type of transducer	Freq range (MHz)	Name of clinical centre	Trimester pregnancy
Spain	Voluson E6, Voluson S8 and Voluson S10 (GE Medical Systems, Zipf, Austria), and Aloka (Aloka CO., LTD.)	Curved transducer	3 to 7.5	Hospital Clínic and Hospital Sant Joan de Déu	2nd and 3rd
Denmark	Voluson E6, Voluson S8 and Voluson S10 (GE Medical Systems, Zipf, Austria)	Curved transducer	3 to 7.5	Copenhagen University Hospital Rigshospitalet, Hvidovre Hospital, Herlev Hospital and Nordsjællands Hospital Hillerød.	2nd and 3rd
Malawi	Mindray DC-N2 (Shenzhen Mindray Bio-Medical Electronics Co., Ltd, China/Germany)	Curved transducer	3.5	Queen Elizabeth Central Hospital	2nd and 3rd
Egypt	Voluson P8 (GE Medical Systems, Zipf, Austria)	Curved transducer	7	Sayedaty Center	2nd
Uganda	ACUSON X600 (Siemens)	Curved transducer	3 to 7.5	Mulago National Referral Hospital	3rd
Ghana	EDAN DUS 60 (Edan Instruments, Inc., Shenzhen, China)	Curved transducer	3.5 to 5	KBTH Polyclinic (Accra)	2nd and 3rd
Algeria	Voluson S8 (GE Medical Systems, Zipf, Austria)	Curved transducer	3 to 7.5	EPH Kouba and Clinique Des Lilas	2nd and 3rd

2.1.2. African datasets

Five African datasets were collected specifically for this study between November 2021 and February 2022. Each dataset included 25 patients and the images were not processed after the acquisition.

The first African dataset was acquired at Queen Elizabeth Central Hospital in Malawi using a Mindary DC-N2 US machine. The 2nd and 3rd trimester US samples were obtained between December and February using a curved transducer with a 3.5 MHz frequency.

The second small-size dataset was obtained at Sayedaty centre in Egypt using a Voluson P8 (GE), a GE Voluson series ultrasound with less image quality. The 2nd trimester ultrasound scans were acquired between November and December using a curved transducer with a 7 MHz frequency.

The third African dataset consists of images from pregnant women during the 3rd trimester from the Mulago National Referral Hospital in Uganda. The US images were obtained in December 2021 with a Siemens US machine and using a curved transducer with a frequency range between 3 and 7.5 MHz.

The fourth dataset was acquired with an EDAN DUS 60 US machine and comprises samples from women in their 2nd or 3rd trimester pregnancy period. The images were acquired at the KBTH Polyclinic centre, Accra, Ghana, in December. Operators used a curved transducer with a frequency range between 3.5 and 5 MHz.

The last dataset was acquired between November 2021 and December 2021 at the EPH Kouba and Clinique Des Lilas centres in Algeria and using the same acquisition as in the European datasets, but with a lower quality GE machine.

2.1.3. Data preparation and pre-processing

All patient data was fully anonymised by removing the header in the original image and was later stored in PNG (Portable Network Graphic) format without compression to avoid quality degradation. Since US images do not contain color information, images were stored as grayscale bitmaps. Images

Table 2: Number of subjects for each of the four datasets used during the training, validation and testing phases.

Country	Abdomen	Brain	Femur	Thorax	Other	TOTAL	Patients
Spain	711	1781	1040	1718	4213	9463	1792
Denmark	771	635	844	291	0	2541	1008
Malawi	25	25	25	25	0	100	25
Egypt	25	25	25	25	0	100	25
Uganda	25	25	25	0	0	75	25
Ghana	25	25	25	0	0	75	25
Algeria	25	25	25	25	0	100	25

with visual artifacts were flagged by the clinician or technician and excluded from the analysis. We cropped the region containing most of the field of view but excluding the vendor logo and ultrasound control indicators. Finally, a min-max normalization was used after resizing the image to 224x224 to keep the same intensity range in images from the same dataset.

2.2. Model architecture and implementation details

Based on the model benchmark presented by Burgos et al. [18] for fetal plane classification, the best performing method, a Densenet-169 [25], was implemented in this study. The Spanish dataset described previously was used as the training set for the aforementioned benchmark, achieving a 0.94 average class precision. The Adam optimiser was for training the model with a constant learning rate of 10^{-3} and first and second moment of 0.9 and 0.999, respectively. At each iteration a batch of size 24 was used to calculate the loss function (cross-entropy loss) and optimise the network parameters. To improve the learning of the model, data augmentation was implemented during training. At each batch, images were randomly flipped, cropped between 0–20%, translated from 0–10 pixels and randomly rotated up to an angle of 15 degrees. The code was developed in PyTorch [26] and the hardware used during training and inference consisted of a 8 GB NVIDIA GeForce RTX 2080 Ti GPU. The different datasets were split into 50% for training and 50% for testing because of the small sample size of the African datasets. During training, 20% of the studies were kept for validation for those datsets with a large sample size case, while this amount was increased to 50% in the small sample size case. The same class distribution as in the Spanish dataset was enforced in the remaining datasets (40% 'other' and 15% for each of the other planes: femur, thorax, abdomen and head) and the missing categories of the African datasets ('other' or thorax) were filled with samples from the Spanish dataset.

2.3. Ablation study

Several strategies were considered to study the model transferability from high to low-resource imaging settings. We classified them in single-centre and multi-centre. The first approach uses a single centre in high-resource settings and the second approach includes also samples from the target centre in low-resource settings.

1. **Single-centre**: We train the neural network with the Spanish dataset using a different amount of samples (n = 125, 250, 500, 1000, 2000, 4000). Then, we evaluate the classification performance on the new datasets to test the generalisability of the classification model either on the Danish dataset, acquired in high-resource settings, or on the African datasets acquired in low-resource settings.
2. **Multi-centre**:

Combination: We train the neural network directly using two samples with varying sizes from two different datasets, the Spanish dataset ($n = 0, 125, 250, 500, 1000, 2000, 4000$) and each of the African datasets separately ($n = 25$). With this, we investigate the effect of the sample size of the first centre on the performance of the model in the second centre.

Transfer learning: We fine-tune the last 4 layers of a pre-trained neural network with the few US images from the new clinical site in Africa. To that end, we pre-train the neural network with the Spanish dataset using again a different amount of samples ($n = 0$, 125, 250, 500, 1000, 2000, 4000), being $n = 0$ the case in which a model pre-trained on the public ImageNet database for object detection and image classification is used. Then, the model is fine-tuned with the available US samples from each of the African centres. Additionally, we estimate the minimum number of patients ($p = 2, 4, 6, 8, 10, 12$) needed from the African centre during the fine-tuning to obtain the desired level of performance, compared to the case in which no pre-trained model is used.

2.4. Performance evaluation

The average accuracy metric is used for monitoring the evolution of the model training. Thus, the model that reaches the maximum accuracy on the validation data and the minimum validation loss is saved as the final model and is later assessed against the testing set. For all experiments and results, confusion matrices are calculated and saved. The average recall, a metric that did not influence the training of the model, is used to assess all the experiments. In this way, we can also consider the false negatives of each class.

3. Results

3.1. Generalisation of the fetal plane classifier to high or low-resource settings using a single-centre approach

In the baseline experiment we evaluated the generalisation of the model trained with the Spanish dataset for different amount of samples. Figure 2 shows that a good performance is achieved on the Danish dataset (average recall of 90.3%) when a sufficiently large amount of samples is used ($n > 500$). Although the Spanish model has a good generalisation ability in other European dataset, the African clinical centres with low-resource settings manifest an inconsistent pattern and lower performance. Even in the case where the images are acquired using a similar protocol to that of the European datasets, such as Algeria (Table 1), the model shows a low performance (average recall of 71.2%) and a similar behaviour as compared to the other African datasets.

3.2. Transferability of the fetal plane classifier from high to low-resource settings using a multi-centre approach: combination or transfer learning

The first approach tested to improve the generalisability of the model from high to low-resource settings consisted in the combination of the Spanish dataset with the small-size datasets from each African centre. The average recall for these models separated by the African dataset used are shown in Figure 3 (left plot) for different number of training samples. The same procedure was followed for the second approach (right plot in Figure 3), which is based on fine-tuning a model, pre-trained with data from Spain, using all patients in Malawi, Egypt, Uganda, Ghana and Algeria. When combining different datasets for training, the number of samples from Spain needed to reach the maximum performance on the African clinical centres is in most of the cases around 500. However, this number is not accurate and the good performance of the model is highly dependent on its
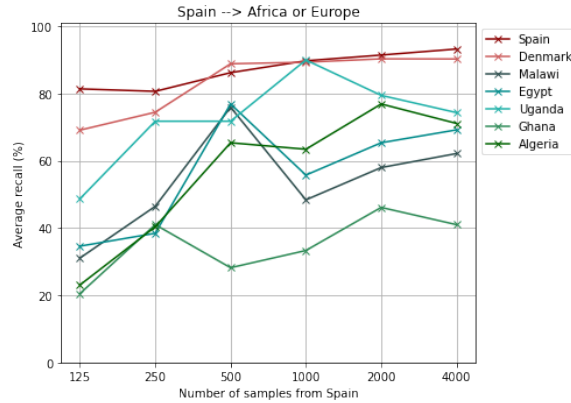
Figure 2: Generalisation performance of a model trained with different amount of samples from the Spanish dataset evaluated in the same clinical centre, in a new centre in Europe (Denmark) and in five datasets from Africa (Malawi, Egypt, Uganda, Ghana, Algeria).

choice, making it challenging to find a good balance between both datasets. On the other hand, in the transfer learning approach, the performance on the new clinical centres increases progressively with the number of cases from the Spanish dataset. In fact, the maximum recall is obtained when using transfer learning (in between 84.6% and 100%, depending on the target dataset). Only in the case of the Algerian dataset, acquired using a similar protocol to the Spanish dataset, the same maximal recall is achieved for both approaches.
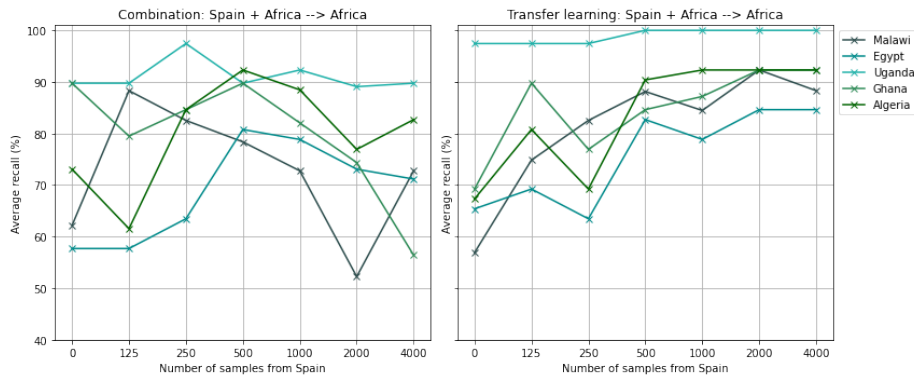

Figure 3: The average recall as a result of combining the dataset acquired in Spain with the US samples acquired in each African centre independently or after fine-tuning the Spanish model, trained with varying sample size, with US images from each African centre independently.

In terms of computational cost, Table 3 shows that training in the combination approach takes 8 times longer than when using transfer learning. Moreover, although the CPU and model storage is the same, the GPU resources needed are 3 times higher. In summary, the transfer learning approach is the most efficient considering the computational resources required and the final performance.

Given the potential of the transfer learning approach to result in a generalisable model we now study the optimal number of patients required to get a good performance. Using the model pre-trained with 4,000 images from the Spanish dataset, we obtain in Figure 4 the average test recall

8

Table 3: Summary of the computational resources needed for the combination ($n = 500$) and transfer learning approaches ($n = 4000$).

	Combination	Transfer learning
Inference time (s/epoch)	8	1
GPU memory (MB)	5,683	1,803
RAM memory (MB)	6,000	6,000
Model size (MB)	48.6	48.6
Batch size	24	24

for the different African datasets when using a different number of patients from the target domain. We compared this strategy with the case in which a big dataset is not available and so the model has to be trained using only the samples from the African centre. The results indicate that fine-tuning models trained on a large dataset with a small amount of target data is sufficient to reach a high classification performance. The generalisation performance increases progressively with the number of patients used from the target centre and reaches a maximum recall when using around 6 or 8 patients. On the contrary, when fine-tuning is not used a lower performance is obtained. For example, in the case of Malawi, Egypt and Algeria, the maximum average recall when training from scratch is 68.6%, 67.3% and 75.0%, respectively, while transfer learning increases the performance to 94.2%, 94.2% and 97.4%. In Uganda and Ghana the improvement given by transfer learning is smaller, since the maximum baseline performance is already high: 94.9% and 89.7% versus 100% and 92.3% with fine-tuning, respectively. The higher baseline performance in Ghana and Uganda could be explained by the fact that in these two centres the classes used for testing are only 3 (abdomen, brain and femur). As observed in Figure 5, the thorax plane is more frequently miss-classified when fine-tuning is not used. In particular, most of the thorax examples are wrongly classified as brain or 'other'. Figure 5 also shows that when a model pre-trained in a large dataset is used and fine-tuned, the model is able to adapt to the new configuration of the new centre.
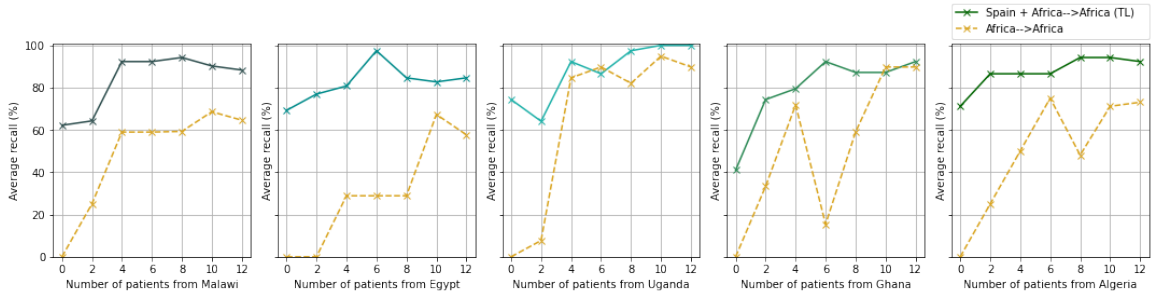

Figure 4: The evaluation of the model after being trained in Spain with all available samples and fine-tuned using a different number of patients from each African centre, as compared to the performance of the model directly trained with the African US images.

Finally, Figure 6 quantitatively summarizes the capacity of transfer learning to adjust existing models to new centres with minimal effort and reduced sample size as compared to models trained from scratch with African datasets. Transfer learning achieves an improved average recall, by reducing the amount of false negatives, while the average precision is stabilised and reaches a comparable value when compared to models trained with the Spanish dataset.
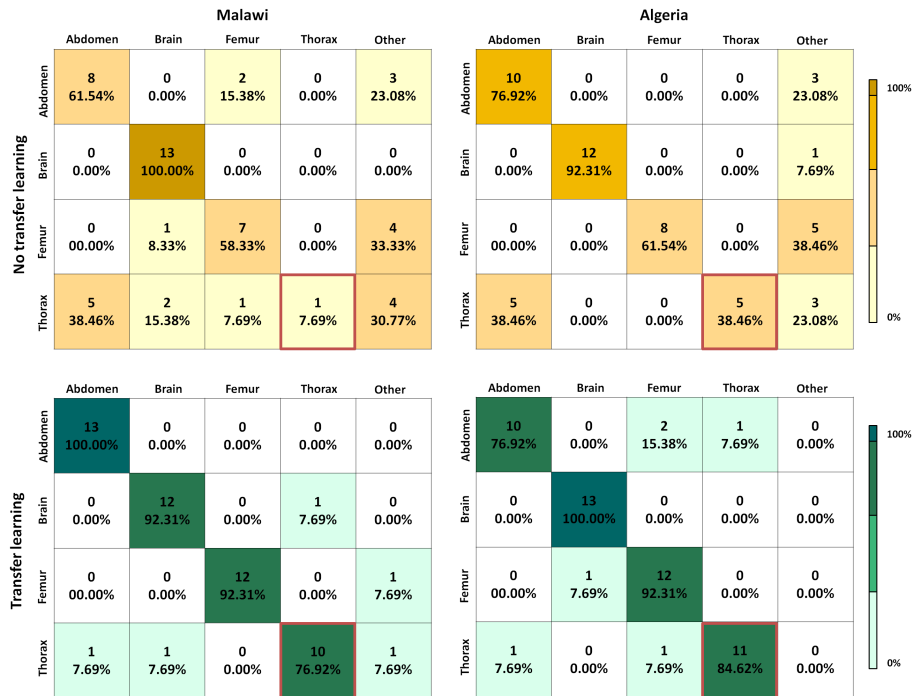
Figure 5: Results on common planes classification with and without transfer learning using 12 patients of each African centre in Malawi and Algeria.

4. Discussion

An important goal for the adoption of deep learning models in the medical domain is to develop inclusive solutions that consider contexts with limited resources. However, in order to build robust predictive models, neural networks usually require a large number of parameters and large datasets, which are especially difficult to obtain in low-resource settings. For this reason, our work investigated the transferability of deep learning models, developed in centres with a greater access to large clinical imaging datasets, to new centres with limited resources, where the generalisation ability is proved to be even more problematic.

Here, we used an existing large-scale data repository from Spain to build a model for US fetal plane classification that can be transferred to a new centre in the African continent. Our results showed that a model trained using a single-centre approach did not reach the desired performance when applied to a new centre in low-resource settings acquired in Malawi, Egypt, Algeria, Uganda or Ghana. In contrast, a multi-centre approach was able to bridge the gap between centres in Africa with less advanced equipment, and centres in Europe with high-quality images and similar acquisition. Specifically, we observed that by directly combining samples from both centres the performance was not as good and stable as with transfer learning, which was capable of effectively adapting the model to the new centre by only fine-tuning the last 4 layers of the model. Through shared representation learning, the neural network has favoured deep phenotypes that are common and consistent across populations/centres, while removing site-specific imaging patterns. With the optimisation of the final weights of the neural network the objective is to maximise the predictive performance for the African ultrasound datasets, in spite of their small sample size.
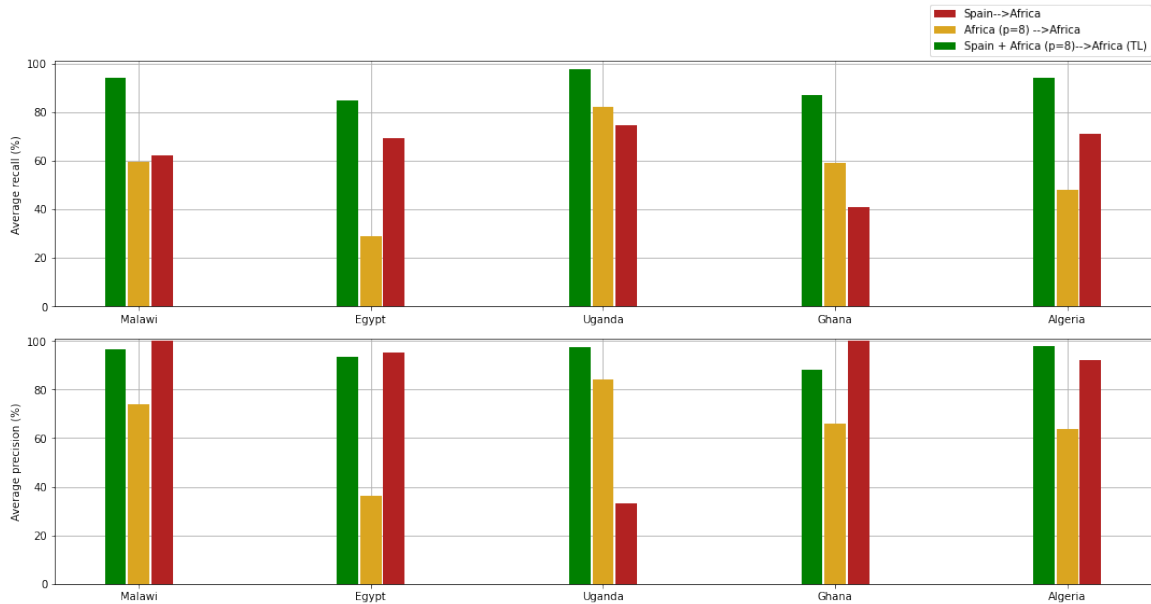
Figure 6: A summary of the performance achieved when only 8 patients from Egypt, Uganda, Ghana, Algeria and Malawi are used in terms of recall and accuracy. First, the model is directly trained with the African samples (yellow). Second, the model is trained using only the large Spanish dataset (red). Finally, the pre-trained Spanish model is fine-tuned with the new African samples (green).

Fine-tuning a model pre-trained with a large imaging dataset in Spain proved to be a potential solution to optimise the performance of a deep learning model in African settings by using small-size datasets and requiring minimal effort. This framework shows promise for generalisability across multi-centre African datasets with challenging and heterogeneous conditions. Moreover, it also encourages for further research to develop highly generalisable solutions for usability of AI in countries with less resources and consequently in higher need of clinical support.

One of the main limitations of the present study is that the datasets obtained in Africa are not complete, since they are less abundant and more difficult to compile. Future works should consider the evaluation of the model on the missing categories.

5. Conclusions

This work was motivated by the need for building new AI models generalisable across clinical centres with limited data acquired in challenging and heterogeneous conditions. This framework has shown to be an opportunity for its widespread application to real maternal-fetal clinical settings, especially as a supporting tool in US plane recognition in low-income countries. We believe our work is an important first step in this direction, and one that will encourage the development of more generalisable models based on transfer learning in centres with low-resource settings.

Funding

Abbreviations

Artificial Intelligence (AI); Convolutional Neural Network (CNN); General Electric (GE).

Availability of data and materials

The Spanish dataset is publicly available. The other datasers are available upon reasonable request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Design of the work: CSB, JTB, KL. Image collection: YAA, ME, BOB, PN, WS, MA, LNB, HNK, SGS, MT. Interpretation of results: CSB, JTB, KM, MT, KL. Manuscript draft: CSB, VMC, JTB, KL. Manuscript review: all authors.

References

[1] C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, K. Lekadir, Image-based cardiac diagnosis with machine learning: a review, Frontiers in cardiovascular medicine 7 (2020) 1.

[2] E. H. Houssein, M. M. Emam, A. A. Ali, P. N. Suganthan, Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review, Expert Systems with Applications 167 (2021) 114161.

[3] L. Zhang, M. Wang, M. Liu, D. Zhang, A survey on deep learning for neuroimaging-based brain disorder analysis, Frontiers in neuroscience (2020) 779.

[4] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, R. M. Summers, A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, Proceedings of the IEEE (2021).

[5] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis, NPJ digital medicine 4 (1) (2021) 1–23.

[6] WHO, Newborn mortality.
URL https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021

[7] L. De Bernis, M. V. Kinney, W. Stones, P. ten Hoope-Bender, D. Vivio, S. H. Leisher, Z. A. Bhutta, M. Gülmezoglu, M. Mathai, J. M. Belizán, et al., Stillbirths: ending preventable deaths by 2030, The lancet 387 (10019) (2016) 703–716.

[8] WHO, Number of licensed qualified obstetricians actively working.
URL https://www.who.int/data/gho/data/indicators/indicator-details/GHO/number-of-licensed-qualified-obstetricians-actively-working

[9] J. Torrents-Barrena, G. Piella, N. Masoller, E. Gratacós, E. Eixarch, M. Ceresa, M. Á. G. Ballester, Segmentation and classification in mri and us fetal imaging: recent trends and future prospects, Medical Image Analysis 51 (2019) 61–88.

[10] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, T. Wang, Deep learning in medical ultrasound analysis: a review, Engineering 5 (2) (2019) 261–275.

[11] M. Komatsu, A. Sakai, R. Komatsu, R. Matsuoka, S. Yasutomi, K. Shozu, A. Dozen, H. Machino, H. Hidaka, T. Arakaki, et al., Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning, Applied Sciences 11 (1) (2021) 371.

[12] J. Matthew, E. Skelton, T. G. Day, V. A. Zimmer, A. Gomez, G. Wheeler, N. Toussaint, T. Liu, S. Budd, K. Lloyd, et al., Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time, Prenatal diagnosis 42 (1) (2022) 49–59.

[13] S. Płotka, T. Włodarczyk, A. Klasa, M. Lipa, A. Sitek, T. Trzciński, Fetalnet: Multi-task deep learning framework for fetal ultrasound biometric measurements, in: International Conference on Neural Information Processing, Springer, 2021, pp. 257–265.

[14] J. Yin, J. Li, Q. Huang, Y. Cao, X. Duan, B. Lu, X. Deng, Q. Li, J. Chen, Ultrasonographic segmentation of fetal lung with deep learning, Journal of Biosciences and Medicines 9 (1) (2021) 146–153.

[15] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz, D. Rueckert, Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound, IEEE transactions on medical imaging 36 (11) (2017) 2204–2215.

[16] M. A. Maraci, C. P. Bridge, R. Napolitano, A. Papageorghiou, J. A. Noble, A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat, Medical image analysis 37 (2017) 22–36.

[17] Y. Li, B. Khanal, B. Hou, A. Alansary, J. J. Cerrolaza, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, et al., Standard plane detection in 3d fetal ultrasound using an iterative transformation network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 392–400.

[18] X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, E. Gratacós, Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes, Scientific Reports 10 (1) (2020) 1–12.

[19] H. Stephens, Ge healthcare unveils ai-enabled ultrasound system.
URL https://axisimagingnews.com/radiology-products/imaging-equipment/ultrasound/ge-healthcare-ai-enabled-ultrasound-system

[20] A. Montero, E. Bonet-Carne, X. P. Burgos-Artizzu, Generative adversarial networks to improve fetal brain fine-grained plane classification, Sensors 21 (23) (2021) 7975.

[21] X. Wang, Z. Liu, Y. Du, Y. Diao, P. Liu, G. Lv, H. Zhang, Recognition of fetal facial ultrasound standard plane based on texture feature fusion, Computational and Mathematical Methods in Medicine 2021 (2021).

[22] T. Sushma, N. Sriraam, P. Megha Arakeri, S. Suresh, Classification of fetal heart ultrasound images for the detection of chd, in: Innovative Data Communication Technologies and Application, Springer, 2021, pp. 489–505.

[23] M. Tolsgaard, M. Svendsen, J. K. Thybo, O. Petersen, K. Sundberg, A. N. Christensen, Does artificial intelligence for classifying ultrasound imaging generalize between different populations and contexts?, Ultrasound in Obstetrics and Gynecology (2021).

[24] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC medical imaging 22 (1) (2022) 1–13.

[25] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).