Development of Artificial Intelligence Image Classification Models for Determination of Umbilical Cord Vascular Anomalies

Byron C. Calhoun, MD, FACOG, FACS, FASAM, MBA, Heather Uselman, DO, Eric W. Olle, PhD

Received December 1, 2023, from the Department of Obstetrics and Gynecology, WVU School of Medicine, Charleston Division, Charleston, West Virginia, USA (B.C.C.); Maternal-Fetal Medicine, WVU School of Medicine, Charleston Division, Charleston, West Virginia, USA (B.C.C.); Resident Department of Obstetrics and Gynecology, Charleston Area Medical Center, Charleston, West Virginia, USA (H.U.); and Research and Development, SynXBio Inc., Charleston, West Virginia, USA (E.W.O.). Manuscript accepted for publication January 7, 2024.

The authors would like to acknowledge Dara Seybold (CAMC research institute) for aiding in preparation and submission of the IRB proposal. Additionally, we also would like to acknowledge Peter Skach and Shawn Romano for their engaging discussions and manuscript suggestions.

Address correspondence to Byron C. Calhoun (Clinical question to), Department of Obstetrics and Gynecology and Professor of Maternal-Fetal, Medicine, WVU School of Medicine, Charleston Division, Charleston, WV, USA.

E-mail: byron.calhoun@vandaliahealth.org

Eric W. Olle (AL/ML methodology to), Research and Development, SynXBio Inc., Charleston, WV, USA.

E-mail: eric@synx-bio.com

Abbreviations

2v, two-vessel umbilical cord; 3v, three-vessel umbilical cord; AI, artificial intelligence; BMI, body mass index; CFI, colorized flow image; CNN, convolution neural network; GB, gigabyte; IRB, Institutional Review Board; ML, machine learning; OCR, optical character recognition; PHI, protected health information; RGB, red, green and blue; SUA, single umbilical artery; TF, TensorFlow; v., version

doi:10.1002/jum.16418

Objective—The goal of this work was to develop robust techniques for the processing and identification of SUA using artificial intelligence (AI) image classification models.

Methods—Ultrasound images obtained retrospectively were analyzed for blinding, text removal, AI training, and image prediction. After developing and testing text removal methods, a small n-size study (40 images) using fastai/PyTorch to classify umbilical cord images. This data set was expanded to 286 lateral-CFI images that were used to compare: different neural network performance, diagnostic value, and model predictions.

Results—AI-Optical Character Recognition method was superior in its ability to remove text from images. The small n-size mixed single umbilical artery determination data set was tested with a pretrained ResNet34 neural network and obtained and error rate average of 0.083 (n = 3). The expanded data set was then tested with several AI models. The majority of the tested networks were able to obtain an average error rate of <0.15 with minimal modifications. The ResNet34-default performed the best with: an image-classification error rate of 0.0175, sensitivity of 1.00, specificity of 0.97, and ability to correctly infer classification.

Conclusion—This work provides a robust framework for ultrasound image AI classifications. AI could successfully classify umbilical cord types of ultrasound image study with excellent diagnostic value. Together this study provides a reproducible framework to develop AI-specific ultrasound classification of umbilical cord or other diagnoses to be used in conjunction with physicians for optimal patient care.

Key Words—artificial intelligence; image processing; prenatal screening; single umbilical artery; ultrasound; umbilical cord

The application of artificial intelligence/machine learning (AI/ML) to medicine continues to rapidly increase. It has proven useful in maternal-fetal medicine (MFM) in prediction modeling and diagnosis. New applications are continually emerging. Applications of machine learning (ML) include: predicting preterm birth, low birth weight, pre-eclampsia, infant mortality, hypertensive disorders, and postpartum depression.[1] An area of great potential is the application of AI to better diagnose birth defects both more precisely and sooner than conventional methods.[1,2] To further develop these methodologies, a collabora-

tion between artificial intelligence scientists and maternal-fetal medicine specialists is crucial. In this retrospective study, blinded prenatal ultrasound images were used to develop a versatile set of protocols to test advanced computer modeling for classifying and identifying umbilical cord anomalies.

AI/ML's application in the field of medicine including MFM is growing. AI/ML are a class of computer algorithms that are based upon different mathematical models where the computer can learn or detect patterns.[3–5] Some of these models are simple mathematical equations while others use automated feedback loop mechanisms originally theorized several decades ago. Current applications of AI/ML include: predicting fetal distress in pregnancy-induced hypertension, classification of amniotic fluid levels, fetal anatomy colorization, and the umbilical coiling index.[2,6–9] A practical application of AI image classification should involve increased maternal-fetal risk and be of added benefit to sonographers and obstetricians. Therefore, a clinical exemplar using standard ultrasound images that correlates to increased maternal-fetal risks should be pursued.

One potential fetal pathology for AI image classification models is the identification of umbilical cord anomalies. Several options for image classification research include: umbilical implantation, umbilical cord diameter, umbilical artery anatomy, cord knots, hematomas, or velamentous cord insertion. For a comprehensive review of different umbilical cord anatomical pathology and screening see Bohîlțea et al or Bethune et al.[10,11] Ultimately, a retrospective image classification model to identify two- or three-vessel umbilical cords was chosen.

Worldwide the formation of a 2-vessel umbilical cord (2v) ranges from under one-half to 6% of pregnancies and correlates to increased fetal mortality.[12–14] The initial identification of a single umbilical artery (SUA) is often diagnosed in the first trimester with follow-up studies done in the second trimester.[10] The rate of this diagnosis can vary due to the underlying population and different risk factors. SUA correlates with several risk factors such as: maternal age, body mass index (BMI), assisted reproductive technologies, diabetes, smoking, hypertension, and twin pregnancies. However, the exact correlations between different SUA cord anatomy risk factors remain elusive due to covariance with other factors, statistical test selection, and population differences.[13,15] While risk factors are actively studied, the outcome associated with this cord pathology is well established. The common clinical outcomes associated with single umbilical artery are: intrauterine growth retardation, early neonatal death, preterm birth, low birth weight, and a range of fetal anomalies.[12,13,16] The use of AI to classify 2v/SUA cords was chosen because the pathology is associated with high fetal risk, and requires additional testing. This tool may aid in identifying a low-probability but high-risk diagnosis.

Images were obtained retrospectively from maternal-fetal medicine patients stored in an ultrasound database. Once the images were obtained and blinded they were analyzed using standard image-classification models. Briefly, there are two general architectures for AI image classification. These models are Convolutional Neural Networks (CNN) and transformers. They are inspired by neuroanatomy and use multiple layers of inputs that effectively parse a numerical representation of an image into simpler patterns that develop filters leading to potential classification(s).[3,17] For further explanation of CNN's see the book "Deep Learning for Coders with Fastai and PyTorch" and references therein.[17] Next, transformer models were described by Vaswani et al created a "simpler" network based on self-attention.[18] The goal of this research is to develop a robust workflow and AI image classification models with transfer learning on fetal ultrasounds.

Materials and Method

The research proposal was sent to the Charleston Area Medical Center—Institutional Review Board for approval. The study used old ultrasound images that were part of a normal wellness check and would be blinded to contain no protected or identifying information. The study was considered exempt and approved (IRB No: 23-928). After approval, two different image data sets were collected. The inclusion criteria included all pregnant patient ultrasound images available with a 2v or a 3v (normal anatomy) umbilical cord. Exclusion criteria included images that lacked the required anatomy or were unable to be blinded. If data was transferred, it was encrypted,

password protected, and compressed using 7zip in a way that no information was visible without a password. Before the start of the study, correctly classifying 85% (error rate ≤0.15) or more images was considered successful.

Data Sets

Two data sets and two novel inference test images were used for this study. The first data set was considered a proof of concept study and contained both transverse and lateral colorized flow images. This was a small n-size study with 22 - 2 vessel and 21- 3 vessel, images. Images had potentially protected health information removed manually or using automated methods. After blinding, low-quality images were removed and the total number of images was 20 in each group. The second data set was composed of lateral aspect colorized flow images (lateral-CFI). It was blinded using a range of techniques. Final studies utilized the AI-optical character recognition method (AI-OCR see below). This data set started with 230 and 154 in the 2 and 3 vessel cohorts, respectively. Poor-quality images were removed resulting in 142 in 2v and 154 in 3v groups for a total of 286 images. In an attempt to minimize overfitting of these limited data sets, image augmentation was added to the training images and training epochs were limited.

Image Preprocessing: Removal of Header and Annotating Text

All images were preprocessed to remove potential PHI and annotating text. There was no standardized method to "clean" fetal ultrasound images from the header and annotating text so several were examined. The text removal methods tested were: manual physician blinding with "filled-in" rectangles, cropping, grayscale threshold, color (RGB) filtering, and Keras-OCR text box recognition with background in-painting (AI-OCR).[19,20] The manual method, an expert using Microsoft Paint™ program to place a block over the image. The cropping method used a standard Python Image library crop coordinates (in pixels): $X_{min} = 40$, $X_{max} = 550$, $Y_{min} = 80$, and $Y_{max} = 466$ to cut out potential PHI. The grayscale and RGB methods used a cropped image using the above coordinates and were used to attempt to remove annotating text. The OTSU and triangle automated methods to obtain a grayscale threshold value were used to create a mask image. The mask image was then combined with the original using *merge* or *bitwise_and* methods to remove annotating text. Similar methods were used to remove the annotating text using RGB filtering. Briefly, one or a range of RGB values associated with the annotating text were obtained. These values were input into an RGB filtering function to create a masking image. The masking image was then used as above to create a new image.

Finally, the AI-OCR image system used TensorFlow 2.12.0 and Keras-OCR 0.9.2 packages to identify blocks of text. Once the image block coordinates were obtained, they were used to in-paint the blocks of text with the average background using the openCV library (v. 4.7.0.72).[19–21] The different methods were compared on the ability to remove header text and annotate text while minimizing image information loss. Once the PHI and annotating text were removed and the images had novel names, they were used for image classification CNN training and testing.

Image Preprocessing: Adjusting Image Size for Classification

After text-removal, some images went through preprocessing to prepare for image classification using pretrained neural networks with transfer learning and validation.[22] The proof-of-concept images were tested at full-size (640 × 480 pixels), header removal cro, or center cropped (448 × 448 pixels). The center cropping used the python image library with the pixel coordinates $X_{min} = 96$, $X_{max} = 544$, $Y_{min} = 16$, and $Y_{max} = 464$. After image size preprocessing by cropping, the saved images may have been resized by the fastai data loader function to 224 × 224 pixels using the "squish" method. The full-size images were not resized or cropped then resized. All experiments used data augmentation in the fastai data loader with values ranging from 0.8 to 1.2.

Proof of Concept Image Classification Using Small n-Size Mixed Method Data Set

All image-classification experiments used fastai to interface with the PyTorch library.[22] The first experiment was to determine the applicability of an image-classification CNN to correctly categorize

2-vessel and 3-vessel umbilical cords using manually blinded images. This data set contained images of transverse or lateral-CFI cord images used by the obstetrician to identify cord anatomy. The manually blinded images used a pretrained ResNet18 with 2-epochs unfrozen and 3-epochs frozen. The data set was divided into training and validation sets (80:20 split) with the fastai data loader image augmentation.

In the next set of experiments images were blinded and text was removed using the AI-OCR method. These images were used to test different preprocessing methods and the ability to correctly classify the validation set image. Slightly different methods were used for the AI-OCR cleaned images. The AI-OCR cleaned images used pretrained ResNet34 default weights. Both full-size and cropped AI-OCR cleaned images were tested. In the 448 × 448 pixel cropped images, the fastai data loader resized to 224 × 224 pixels using the "squish" method. Training of the network used an 80:20 split for training and validation sets. The training used 1-epoch unfrozen and up to 10-epochs frozen with the standard learning rate.

Expanded Data Set Using Lateral-CFI for Image Classification

Lateral-CFI images with the umbilical cord located by the urinary bladder is the currently accepted technique, therefore these ultrasounds were chosen for the expanded study. The images were obtained and all text was removed using the AI-OCR method, center cropped, and saved. The fastai data loader was used with 80:20 training, validation random split, image augmentation, and resized to 224 × 224 pixels using the squish method. The neural network architectures with default pretrained weights were trained using: 1 epoch unfrozen and 10 epochs frozen. Training was set at 10 epochs to allow for comparison to different neural network models. Several pretrained image recognition networks such as: convnext_small.in12k_ft_in1k_384, LeViT_256.fb_dist_in1k, ResNet34.Imagenet1k_v1 (*aliases* "Default" or "ResNet34"), resnet34.a1_in1k, resnet50.a1_in1k, ResNetv2_50.a1h_in1k, and ResNet101.a1h_in1k were tested.[22–26] To compare training variability, each of the image-recognition neural networks were trained five times with a random assignment to training:validation sets using fixed seeds. The error rate and training loss by network and iteration were saved and later analyzed using R.[27] The plots show mean (point), error bars ± standard error of the mean (SEM), confidence interval in gray shading, locally optimized scatter plot smoothing (LOESS) trend line (dotted line) and red dashed line indicating a priori error rate maximum (0.15).[28] A subset of the trained models was used to calculate diagnostic value (sensitivity, specificity, and predictive values) as a probability using standard methods.[29]

Image Inference

Two novel images were used to test inference using different lateral-CFI trained models. One image was in the 2-vessel and the other was a 3-vessel. These images were not used as part of the training or validation. Images were obtained from the same source and underlying demographics. The AI-OCR cleaned images were tested at: full-size 640 × 480, center cropped 448 × 448 and center cropped resized to 224 × 224, pixels. The prediction, probability 2v and probability 3v were inferred using standard fastai methods.

Computer Resources

Two different computers were used for AI and statistical analysis. A Linux based desktop computer (Pop_OS 22.04) with an Nvidia 1070TI GPU for training, using an AMD ×2700 processor which also contains 48 GB of system RAM. Anaconda was used to install separate virtual environments loaded with: 1) Python v. 3.9.16 with TensorFlow (v. 12.2.0) or 2) Python (v. 3.8.16) with PyTorch (v. 1.12.1) and fastai (v. 2.7.12).[19,22,30] Figures were generated using fastai and saved using standard methods.[31] Statistical analysis and plotting of the different neural networks training or error rates was performed on a System 76 Darter pro with 16 GB of system RAM using Pop_OS 20.04 with R 4.3.0, EMACS/ESS. The packages tidyverse (specifically dplyr, ggplot, and readr) and ggpubr were used.[27,32–35]

Additional Data

Representative jupyter-notebooks will be available showing: AI-OCR image preprocessing, example training and inference. The two AI-OCR processed data sets will also be made available: https://github.com/Eric43/MFM_AI.

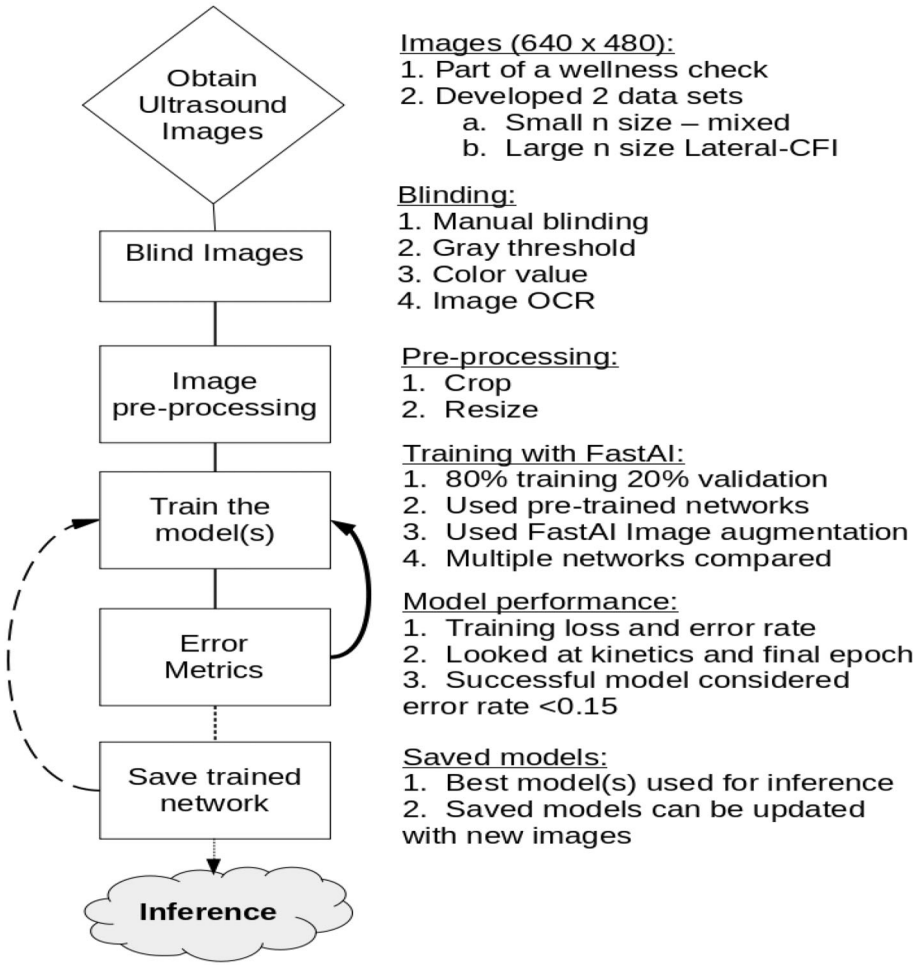
Results

The overall process followed a standardized work flow seen in Figure 1. Within each work block, there were several different conditions tested. The initial data set was small with a mixture of transverse and lateral-CFI umbilical cord images. These are common methods used in clinical practice to identify cord vessel anatomy. This small n-size data set was used for technique development and initial testing later to be applied to the expanded data set. This data set also tested the applicability of transfer learning to minimize the number of images or training epochs. The first step in an image recognition project with clinical samples is to remove protected health information to minimize identification risk to the patient.

Image Preprocessing to Remove Identifying and Annotating Text

To use the ultrasound images for exempt research and to train an AI image classification network, a standardized system to remove text was needed. There are several potential methods to remove text that include: 1) manual blinding, 2) cropping out the identifying text, 3) gray scale threshold masking, 4) RGB value-range masking, and 5) AI-OCR methods.

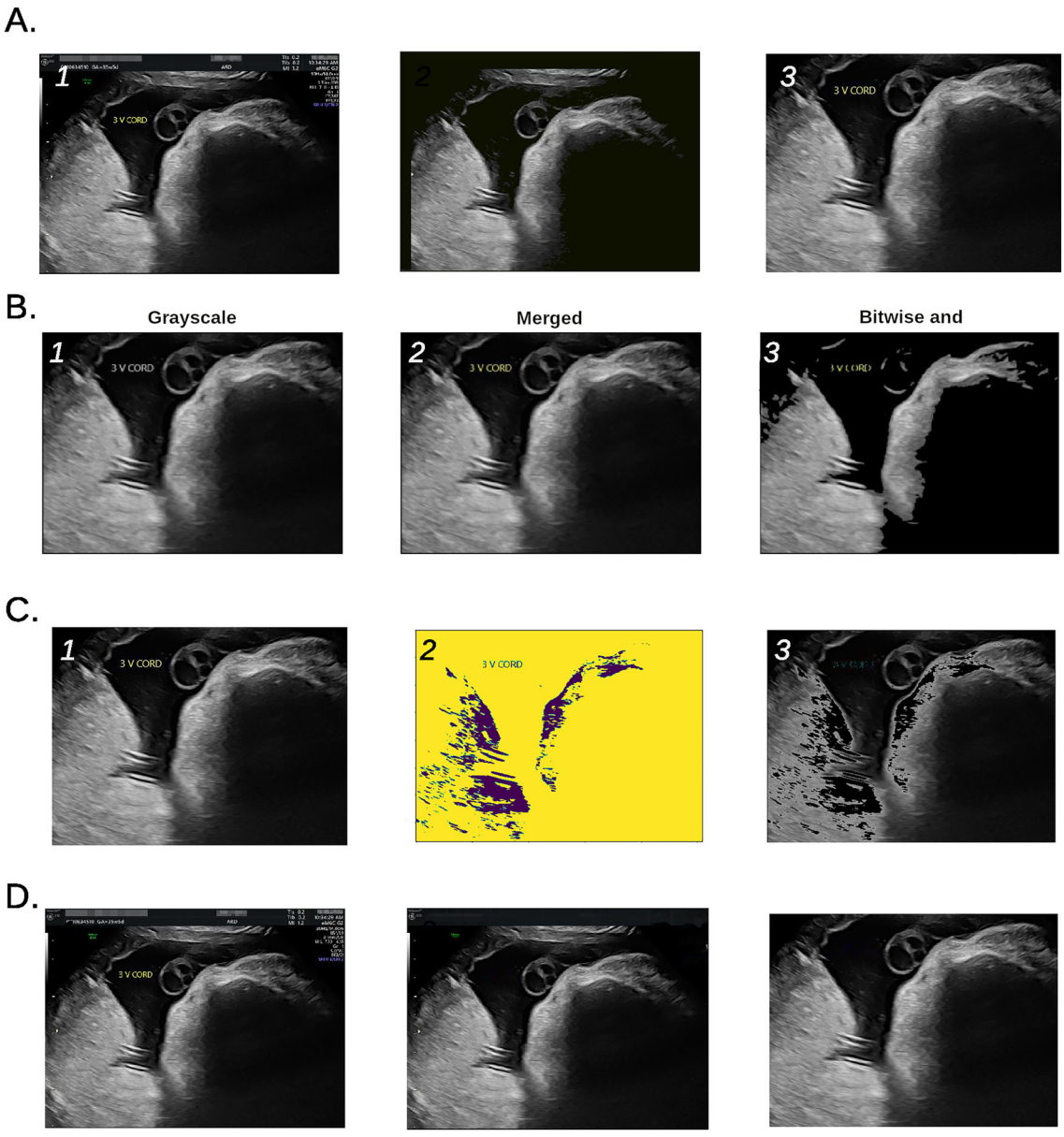
Figure 1. Workflow of ultrasound image analysis project. Graphical summary of the workflow used for this research. Once the images were obtained, text removal methods were tested. Then the images were kept full size or resized for training. Training randomly split the images into training and validation sets followed by augmentation and transfer learning with pretrained models. The models error rate was analyzed on the validation set and saved. Finally, different trained models were tested for predictive ability using two novel images at three different sizes.



The outcomes of these methods are visually summarized in Figure 2. Each one of these methods have advantages and disadvantages.

First, using the manual blinding method, an obstetrician or qualified individual was used to blind each image (Figure 2A, Image 2). This entailed loading the image, drawing rectangles over the text that were filled in black and then saving the image. It took several minutes per image to complete the task. This is the only method that performed well using

Figure 2. Methods to remove PHI and annotation text. Comparison of the different methods for blinding/PHI removal. Image **A1** shows the "raw image" with header blurred to obscure PHI. Image **A2** shows a manual blinding method. Image **A3** show the use of cropping to remove header information. Panel **B** shows cropping with gray scale threshold using image A3. Image **B1** shows gray scale image used for thresholding. The OTSU masking image was use to *merge* or *bitwise_and* methods (**B2** and **B3**). Panel **C** shows RGB filtering on image A3 (C1) with generated mask (C2) and merged image (C3). Panel **D** shows the use of AI-Optical Character Recognition with average background to fill in the identified text. Image C1 is the full-size image, C2 and C3 are the full size or cropped image (A3) after processing.



ResNet18 (error rate 0.11 or 1/9 misclassified) but when used for inference or repeatability it performed worse than other methods (data not shown). This was the first approach used and due to time-consuming nature it led to the development of automated methods that were used for the remainder of the experiments.

Next method tested was to crop the identifying text located in the ultrasound header image (Figure 2A, Image 3). This method could be automated to quickly remove the header information containing PHI but not annotating text. This method was very fast albeit leaving annotating text (ie, "3v CORD"). When used to train the neural network, the annotating text appeared to skew the results (data not shown). The cropping method was useful to remove PHI contained in the image header but required additional methods to attempt to remove annotating text.

As a way to augment the crop to remove PHI method, attempts were made to filter out the annotating text using either gray scale threshold or RGB filtering (Figure 2, B and C). These methods are similar but use different values to develop a mask. The gray scale method use automated thresholds obtained with the triangle (threshold = 13–17) or OTSU (threshold = 64–72). The threshold is used to create a mask image that is merged with the original image (Figure 2B, Images 1–3). The RGB method is similar but used RGB values for mask generation (Figure 2C, Images 1–3). Both of these methods appeared to modify the background and could not completely remove annotating text. They were not used for image classification. The removal of text using an image AI-OCR was tested next.

The use of AI-OCR method was tested on full size and cropped images that contained text (Figure 2D, Images 1–3). The AI-OCR method was able to recognize the vast majority of the text with a few minor exceptions of single characters. Additionally, this method was very selective to text, as it did not remove the ultrasound manufacturer logo in the upper left of the image (Figure 2D, Image 2). The proof of concept data set containing 40 images, took ∼3 minutes to process. This method was slower than the crop methods but removed nearly all text while retaining background information. The AI-OCR method proved to be acceptable for text removal and was chosen as the primary blinding method. The next workflow boxes are image size preprocessing and training the image.

Use of AI-OCR Blinded Ultrasound Images to Determine Optimal Neural Network Training

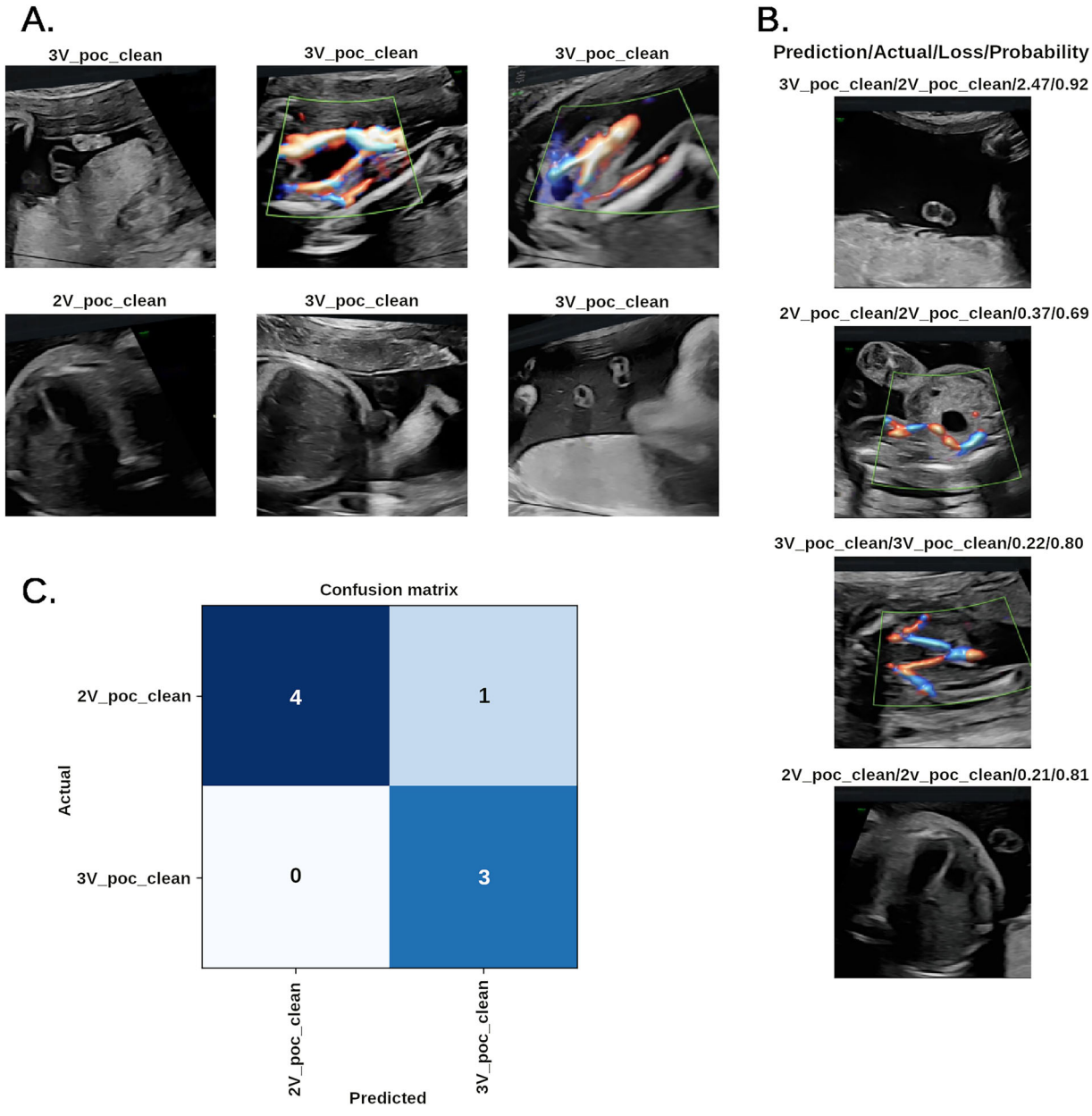
The first set of experiments concentrated on optimizing image size for model training. The images were kept full size (640 × 480), learning rate was not adjusted and trained using 1-epoch unfrozen, 5-epochs frozen. The validation set error rate was 0.125 or one out of eight images incorrectly identified. After testing the full-size images, the next step was to determine the effect of adjusting image size through a combination of cropping and resizing.

The final part for the proof-of-concept study was to adjust the image size to the optimal size for the pretrained neural networks. Using the squish method on the full-size images did not return acceptable error rates (data not shown). The images were then center cropped to 448 × 448 pixels and resized to 224 × 224 pixels using the squish method. An example of the training images, worst performing images and confusion matrix is seen in Figure 3, A–C. This method was repeated three times and an average error rate of 0.083 was obtained (error rates: 0.125, 0.000, and 0.125). The worst performing validation images is shown in Figure 3B. One of the three validation set confusion matrix is shown in Figure 3C. Our a priori threshold for maximum error rate was set a ≤0.15 and three experiment with errors ranging from 0.00 to 0.125 with an average of 0.083 is a successful proof of concept for the use of AI in umbilical cord vessel anatomy classification. This is a small n-size study, therefore, an expanded study comprised of only lateral-CFI was performed.

Image Classification Using Lateral-CFI Ultrasound Images

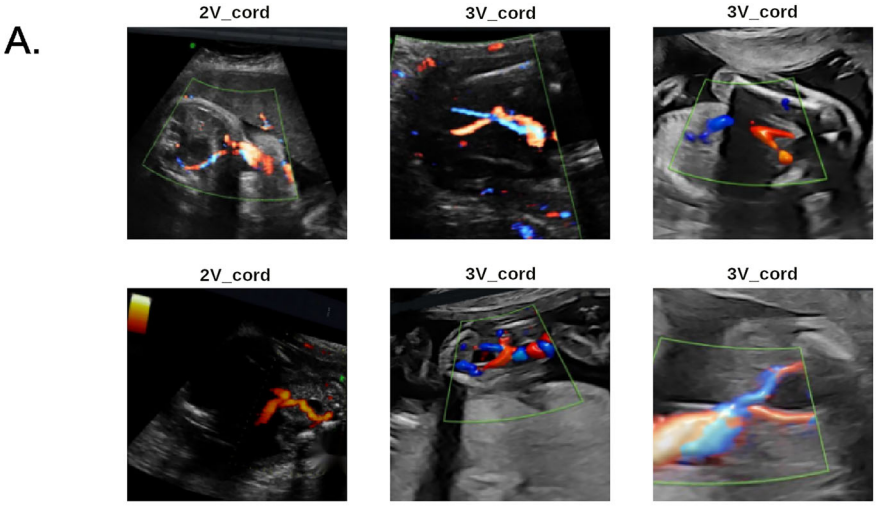
The use of lateral-CFI ultrasound images around the urinary bladder is the current accepted best practice for the identification of umbilical cord vessel anomalies when colorized flow information is available. The expanded data set included only AI-OCR blinded lateral-CFI center cropped ultrasound images. These were initially tested using the ResNet34-default with the squish method for resizing (Figure 4). This method used a random data set with six images

Figure 3. Small N-size study with traverse and colorized umbilical artery images. The mixed small n-size data set was subjected to standard image classification transfer learning with fastai and the pretrained ResNet34-default. A sample of the training split group is displayed (n = 6) with randomized image augmentation. After training, the lowest performing images were extracted and displayed (Panel **B**). Finally, the confusion matrix from the validation set is in Panel **C**. This experiment was repeated three-times with an average error rate of 0.083 (not shown).



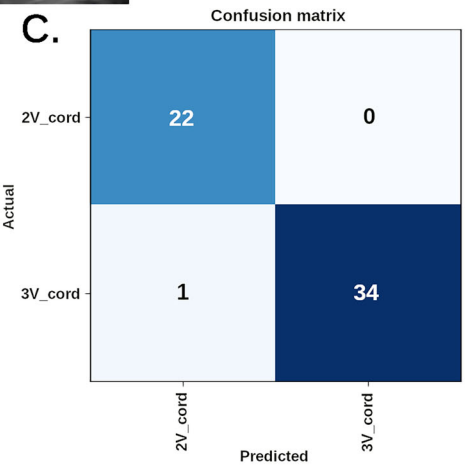
extracted from the training set shown in Figure 4A. Each epoch was trained in just over 20 seconds with the lowest error rate occurring in epoch five (Figure 4B). The image-classification model performed well with an error rate of 0.0175 (1/56 misclassified) (Figure 4C). The one error was a false positive result for SUA. Finally, the worst four performing validation images show that outside one image the other three

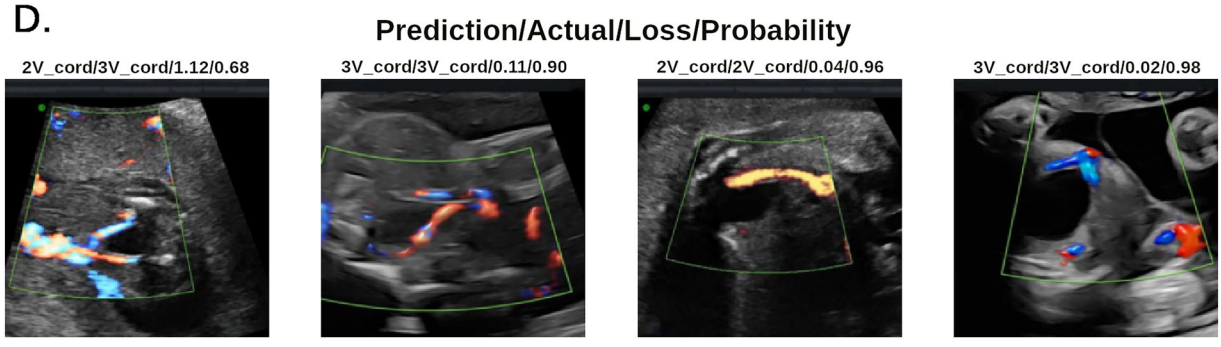
Figure 4. Image classification of lateral colorized-flow umbilical arteries with ResNet34-default. A collection of 286 images lateral aspect colorized flow images was used to train a ResNet34 network with default weights. Panel **A** shows a random subset of six images with augmentation using the fastai data loader. The image recognition network was trained for 10-epochs with error rate and training loss used as performance metrics (Panel **B**). The confusion matrix from the validation set is shown in Panel **C**. The worst performing images in the validation set are shown in Panel **D**. This method was used for comparison between different pretrained neural network models.



B.

Epoch	Training loss	Training error rate	Valid loss	Valid error rate	Time
0	0.296636	0.102679	0.188347	0.052632	21 secs
1	0.280996	0.071429	0.096254	0.052632	20 secs
2	0.246773	0.066964	0.062983	0.017544	21 secs
3	0.229309	0.075893	0.069786	0.035088	21 secs
4	0.212036	0.040179	0.532874	0.192982	21 secs
5	0.208810	0.049107	0.026906	0.000000	20 secs
6	0.195428	0.031250	0.024075	0.017544	21 secs
7	0.177681	0.040179	0.026062	0.017544	21 secs
8	0.158013	0.026786	0.025856	0.017544	20 secs
9	0.137254	0.017857	0.023174	0.017544	20 secs





were correctly called. The next experiment looked at how do different image-classification CNN's perform on the lateral-CFI data set.

A set of experiments were designed to test the ability of different pretrained CNN's to classify the lateral-CFI images. The architectures chosen to test

were: convnext small, LeViT256, ResNet34-default, ResNet34a1h-in1k, ResNet50a1h-in1k, ResNet v250a1h-in1k, and ResNet101a1h-in1k. To test the performance of the different pretrained neural networks, each was run following a standardized protocol five times. The average training and error rates by epoch are shown in Figure 5. The different networks performed well with the exception of convnext and ResNet50v2 that had error rate averages >0.15. These two networks had an average error rate of >0.15, although individual experiments were able to obtain less than a priori value. The majority of the models were analyzed for diagnostic performance and time per epoch (Table 1). The mean training time per epoch ranged from 67.52 (ResNet101) to 17.02 (LeViT256), seconds. In the ResNet family, ResNet34-Default had the highest diagnostic value with: sensitivity = 1.00, specificity = 0.97, PPV = 0.96, and NPV = 1.0 (Table 1). The difference between ResNet with default and the a1h-in1k can be seen by comparing the two diagnostic values. The LeViT256 pretrained network was the fastest per epoch but yielded only average diagnostic performance. ResNet50v2 was excluded from this table due to not meeting established error rate threshold.

To determine training differences, the four worst performing images were extracted. The default ResNet34 was already shown in Figure 4D and was not included. There was one image that performed poorly in nearly all the models and was the worst or

Figure 5. Comparison of different pretrained neural networks. The lateral-CFI data set was used to test performance of different pretrained neural networks. The image classification networks using transfer learning were subjected to same procedure as Figure 4 (n = 5 per model). The mean training loss (Panel **A**) and mean validation set error (Panel **B**) were graphed to show training over time. The error bars show SEM, gray shading is the confidence interval and the dotted line is the overall trend using the LOESS method. In panel B the red dashed line shows 0.15.

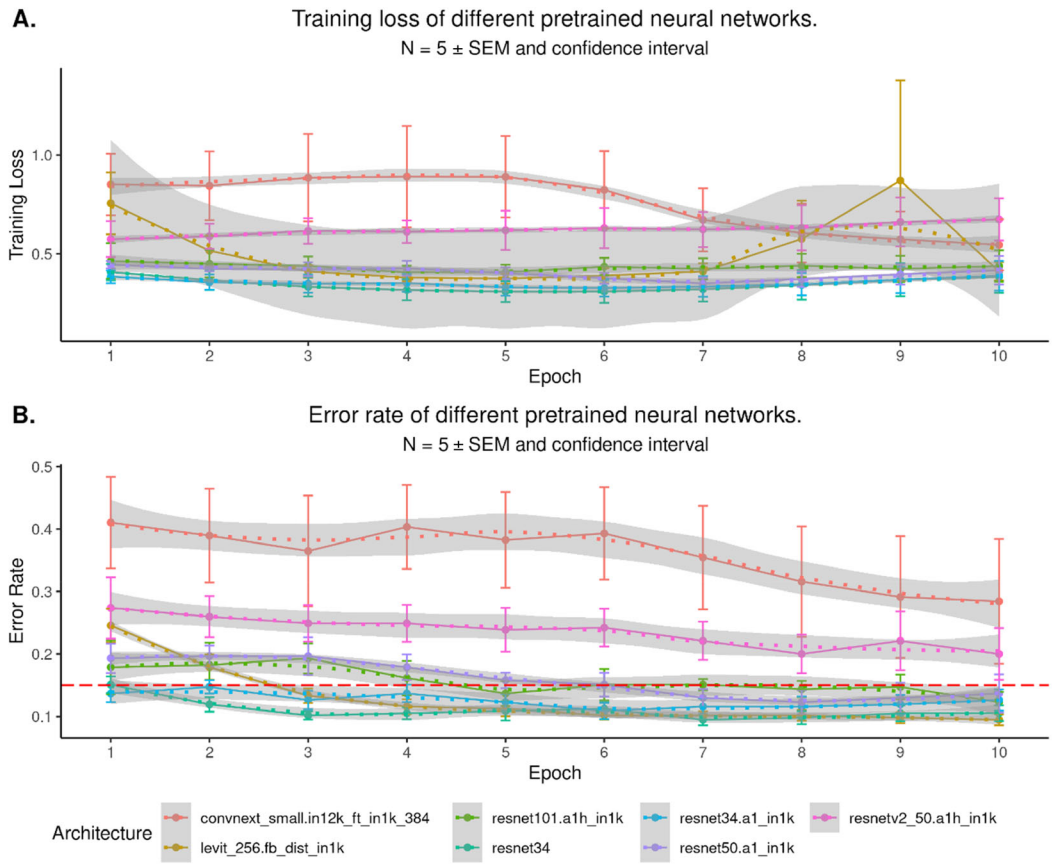


Table 1. Diagnostic Value and Mean Epoch Training Time

	Sensitivity	Specificity	PPV	NPV	Average Epoch (Seconds)	SD
Resnet34-Default	1.00	0.97	0.96	1.00	24.60	0.670
Resnet34-a1-in1k	0.95	0.94	0.91	0.97	24.20	0.404
Resnet50-a1_in1k	1.00	0.89	0.85	1.00	44.88	0.521
Resnet101-a1h_in1k	0.95	0.91	0.88	0.97	67.52	0.505
Levit_256.fb_dist_in1k	0.95	0.91	0.88	0.97	17.02	0.141
Convnext_small	1.00	0.86	0.81	1.00	63.08	0.752

Note: To quickly summarize the performance of different image-classification models on the lateral-CFI data set, the diagnostic value and average epoch training time were calculated. Sensitivity, specificity positive predictive value (PPV) and negative predictive value (NPV) were calculated on the best performing networks. Accuracy is not shown due to using a data set with nearly equal distribution that differs from normal clinical practice. The average epoch training time in seconds and standard deviation (n = 5) are included for comparison.

second worst in all models (Figure 6). The second observation was when loss is high the probability of classification is usually incorrect despite prediction probability. While there are commonalities between the worst images (ie, ResNet34 class), other neural networks had difficulties with other images. Overall, the models performed well using AI-OCR cleaned images and demonstrated the applicability of transfer learning to ultrasound images. Next, the ability to infer image classification was examined.

Inference from Trained Models

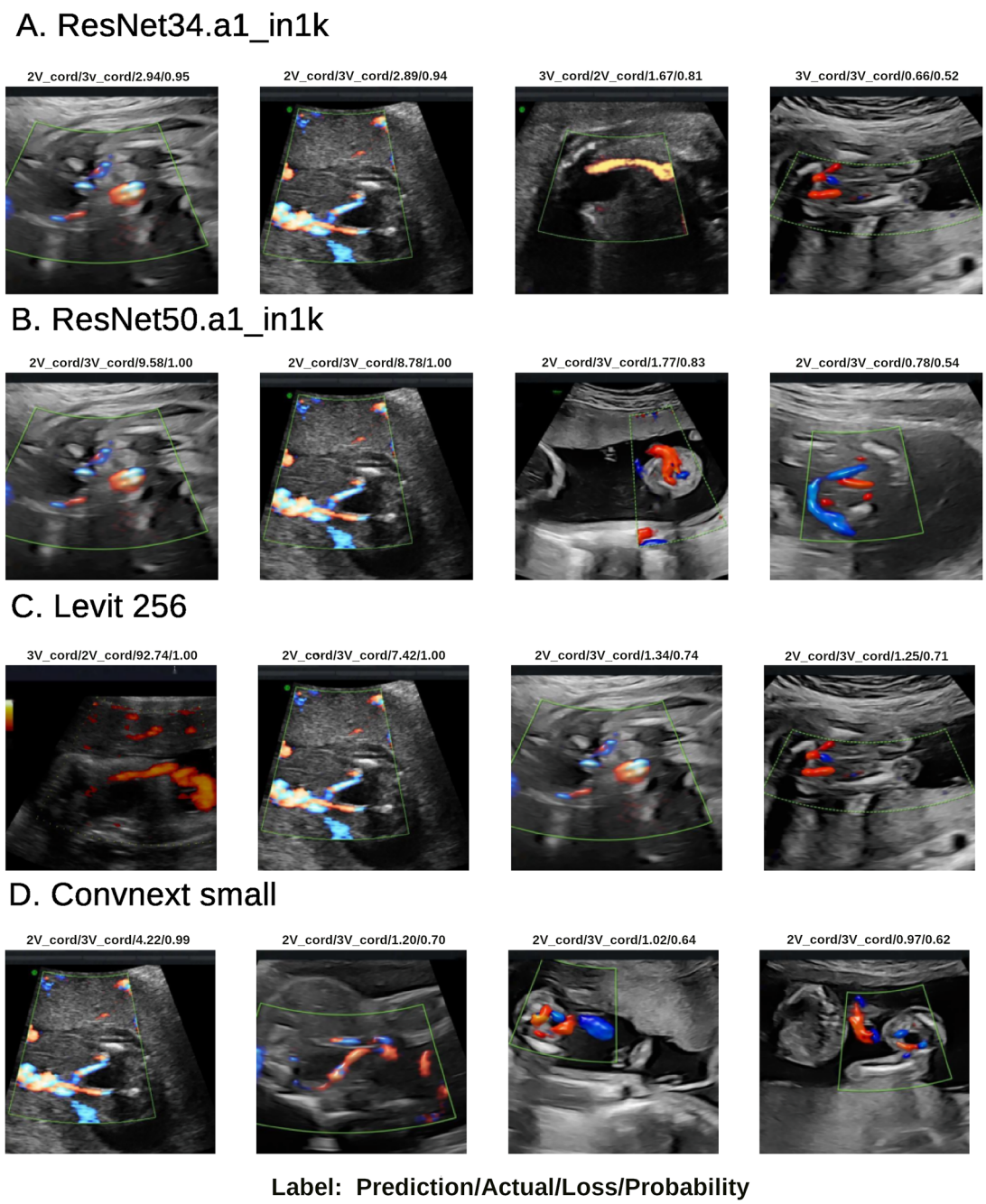
To test the ability to correctly classify images following training, the different lateral-CFI networks were challenged with one 2v and one 3v image. This is only a basic test but meant to demonstrate how the AI networks can be used to classify novel images. These images were then kept full-size, center cropped (448×448) or center cropped and resized (224×224). All of the networks were able to correctly classify the images, including convnext (Table 2). The ResNet family of networks performed similar for probability of classification except for ResNet50v2 on the full size 3v image. Due to similarity of performance to the ResNet34-default in inference, ResNet34_a1h-in1k and ResNet101_a1h-in1k were excluded. The ResNet50v2 was kept to show the ability to infer image classification despite poor training/error rate performance. The LeViT 256 model correctly classified all images with slightly lower probabilities than the ResNet family. Finally, the convnext model had the smallest difference between the different class probabilities thereby decreasing prediction confidence.

Discussion

Early detection of potential high-risk pregnancies provides patients and obstetricians time to potentially mitigate negative outcomes. Due to the morbidity and mortality associated with single umbilical artery cord pathology, this is a perfect application of AI on fetal ultrasound images. These protocols can be used to create additional tools for physicians. The goal of this research was three-fold with: 1) text removal from ultrasound images (blinding and cleaning), 2) determine optimal image size, and 3) application of image-classification AI models to classify umbilical cord vessel anatomy. This research provides the first beginning-to-end workflow for blinding, removing annotating text, image preprocessing, image classification training and inference for umbilical artery anatomical anomalies. This work was divided into several sets of experiments from a basic proof of concept study to an expanded data set containing only lateral-CFI images.

The initial experiment used manually blinded images. This work was the only experiment successful using ResNet18. A few issues occurred with the manually blinded images. First, the images are time consuming to produce and were a drain on obstetrician's time. Next, the blinding was variable in the amount of information removed from the original image. This variability can lead to blinder bias or unexpected errors. Finally, when the trained model was tested with AI-OCR cleaned images it performed poorly. Although not tested per se, it is possible that the blinding blocks were used in part to classify the image instead of the fetal ultrasound image. The manually blinding method was slow, required expert resources and did not perform well for inference. The next

Figure 6. Comparison of the lowest performing images by neural network. Comparison of the lowest performing four images corresponding to ~7% of the validation set. The ResNet34.a1_in1k (Panel **A–D**) was used to compare to the default weights in Figure 4D. The images show images that may affect the different network architectures. The labels above the images show prediction, actual, loss and probability. All images were derived from the same training:validation image set.



logical step was to develop ultrasound image cleaning methods that minimized expert's time and this removed image text effectively.

Ultrasound images obtained from clinical practice have protected health information and annotating text that needs to be effectively removed.

Table 2. Comparison of Inference Prediction and Probabilities at Different Input Resolutions

	2 Vessel Umbilical Cord (2V_cord) Image								
Image Size	**224 × 224**			**448 × 448**			**640 × 480**		
Inference	**Prediction**	**p 2V**	**p 3V**	**Prediction**	**p 2V**	**p 3V**	**Prediction**	**p 2V**	**p 3V**
Resnet34_default	2V_cord	0.9985	0.0015	2V_cord	0.9999	0.0001	2V_cord	0.9998	0.0002
Resnet50	2V_cord	0.9180	0.0820	2V_cord	1.0000	0.0000	2V_cord	0.9996	0.0004
Resnet50v2	2V_cord	0.9601	0.0399	2V_cord	0.9993	0.0007	2V_cord	0.9943	0.0057
Levit256	2V_cord	0.8931	0.1069	2V_cord	0.8382	0.1618	2V_cord	0.9636	0.0364
Convnext	2V_cord	0.5986	0.4014	2V_cord	0.5396	0.4604	2V_cord	0.8686	0.1314
	3 Vessel Umbilical Cord (3V_cord) Image								
Image Size	**224 × 224**			**448 × 448**			**640 × 480**		
Inference	**Prediction**	**p 2V**	**p 3V**	**Prediction**	**p 2V**	**p 3V**	**Prediction**	**p 2V**	**p 3V**
Resnet34_default	3V_cord	0.0001	0.9999	3V_cord	0.0000	1.0000	3V_cord	0.0002	0.9998
Resnet50	3V_cord	0.0000	1.0000	3V_cord	0.0011	0.9989	3V_cord	0.0000	1.0000
Resnet50v2	3V_cord	0.0211	0.9789	3V_cord	0.0006	0.9994	3V_cord	0.4790	0.5210
Levit256	3V_cord	0.0249	0.9751	3V_cord	0.0686	0.9314	3V_cord	0.0173	0.9827
Convnext	3V_cord	0.2416	0.7584	3V_cord	0.2094	0.7906	3V_cord	0.3867	0.6133

Note: The saved lateral-CFI models were tested with novel AI-OCR cleaned images. The classification was inferred on two different images (2v and 3v) using three different sizes of: full-size, 448 center cropped, or 224 cropped and resized. The classification and probability of prediction by class are shown.

Efficient text removal is necessary to prevent exposing patient PHI or skewing the AI model with annotating text. Several methods were tested to remove text. The first set of automated methods tested were cropping and gray or color filtering to remove PHI and annotating text. Cropping out the PHI using standard pixel coordinates was easy to automate and the fastest method but was unable to remove annotating text. One issue with leaving annotating text is that the AI may learn the text and not the anatomical anomalies. Due to the high-speed and limited strain on compute resources, cropping was combined with different methods in an attempt to remove annotating text.

The first method to remove annotating text tested was the gray scale threshold. Gray scale threshold develops a mask and merged image able to partially remove annotating text, but also removed background. Finally, the unique color used for annotating text was used to generate a filter based upon RGB values. The RGB method was able to remove slightly more text than gray scale thresholding but also had a greater effect on the image background. The gray scale and RGB masks were non-specific and removed background information making them unacceptable for AI use. Therefore, another method that was able to remove text but preserve background image integrity was needed.

The final text removal method tested was AI-OCR followed by filling this block with average local background.[19,20] This method was able to remove annotating text in cropped and all text full size images. The image had no obvious signs of blinding or "bare spots" or background removal. Although, the header region appeared gray due to the local background values. The AI-OCR function has since been modified to allow for a user selection of mean background or black fill to remove text. Since nearly all images had the header filled in with a gray it is expected to have little impact on AI training. Additionally when AI-OCR image cleaning was followed by center cropping, the gray bar was minimized. While this method is superior to manual and cropping it is not without drawbacks.

The main issues with the AI-OCR method are: individual characters remained, equipment logo remained, requiring a dedicated GPU and much slower than cropping or filtering. The inability to remove all text appeared to be random and had minimal impact on AI training. The ultrasound manufacturers logo in the upper left corner could be problematic if different

companies are included but shows the selectivity of this method and was usually cropped out. In the future, company logos and stray text identification may require additional training of Keras-OCR. The GPU requirement can be overcome by using the CPU version of TensorFlow but this process is significantly slower.[19,20] The downsides to using AI-OCR for ultrasound text removal is offset by its abilities.

There are several positive aspects to the AI-OCR method such as fully automated and can be deployed locally or on a server. Automation can be designed to blind images without direct access to images containing PHI. The AI-OCR function has been recently modified to allow for either average background or black. Once the images were cleaned of header and annotating text, they were used for training and image classification models.

The small n-size data set contained AI-OCR images of traverse and lateral-CFI umbilical cords. This study used full size images or size adjusted images. Overall, the image recognition models worked with an error rate <0.15. ResNet34 was used for this set of experiments due to poor performance of ResNet18 (data not shown). The full-size images were acceptable and had an error rate of between 0.11 and 0.125. On the other hand, the center-cropped and resized images performed well with an average error rate of 0.083. These experiments indicated that a range of image sizes could be used, but optimal training occurred using a 224 × 224 pixel 1:1 squished image. This is expected, since this size was used to obtain the pretrained weights. An interesting note is one experiment attempted to squish the full size to 224 × 224 pixels and did not have the same performance as the center-cropped and resized method (data not shown). While this effect was not tested per se, a potential reason is the 4:3 ratio of the full-size images skewed underlying data when resized using the squish method. These sets of experiments clearly demonstrate that image-classification models can be successfully applied on a mixed image type data set with a small data set. The problem with the small data set is the potential for model over fitting and a lack of error rate resolution. An expanded data set was needed to resolve the overall performance of neural network performance.

An expanded data set comprised of lateral-CFI images was used to test different pretrained neural networks and provide a better error resolution. The increase in images changed the minimum error rate from 1/8 to 1/56 leading to a minimum error resolution of 0.0175. The expanded data set error resolution allows for additional tuning, model comparison and testing of different neural networks. Several different pretrained neural networks were tested for SUA classification performance. These were chosen partially based upon the comparison performed by Jeremy Howard in a Kaggle notebook "Which image models are best?"[23,36] Each model was repeated five times using the same training:validation seed values. In these sets of experiments, only convnext and ResNet50v2 did not have an average error rate of <0.15. However, individual experiments were able to perform better than the a priori error rate threshold. The average training loss of ResNet family and LeViT 256 seemed to reach a minimum around epoch five to seven. The convnext training loss did not seem to lower until after epoch five and may benefit from additional training epochs. It was tested once out to 20-epochs and reached a minimum training loss and error rate around epoch 18 (data not shown). When comparing average training time per epoch, the LeViT 256 was the fastest with ResNet101 being the slowest. The LeViT uses a modified Vision Transformer (ViT) architecture and explains the fast training time.[25] When looking at the ResNet family, the more layers of did not equate to better performance, therefore, image classification networks need to be compared.

One potential limitation of this research is the limited data sets may cause model over fitting. Over fitting is where the model "learns" the images and not a generalized pattern for classification. To minimize over fitting standard methods were used. First, the training data used image augmentation. Image augmentation modification decreases the probability of over fitting. Next, the number of training epochs was limited to minimize over fitting. Finally, the inference images were not used in training or validation and obtained high class probability potentially indicating that the learned pattern was generalizable. These steps do not negate the need for greater number of images and patient diversity as this research continues.

The clinically accepted method to evaluate the performance of a diagnostic test in a clinical setting is

the use of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).[29] The diagnostic values were compared and showed nearly all of the ResNet family performed extremely well on the validation set. Overall the ResNet34-default performed the best. Generally, all the models had a good sensitivity and a positive predictive value of >0.8, however, this data was generated with the validation sets. As the work progresses, diagnostic value should be calculated with a separate data set that has around 5% SUA images. Using a separate but population representative data set allows for the ability to robustly calculate diagnostic value along with accuracy.

The ability for AI to correctly identify an image depends on the training, training set, and the quality of the image. The four worst performing images or around 7% of the validation set showed some interesting trends. One 3v image performed poorly in all tested models. This image was verified as a 3v cord. One potential reason for the incorrect classification maybe the non-specific information. This data was from the near field ultrasound window containing interference or nonclinical anatomic information/images/signals (ie, placental tissue/uterine wall). This ancillary information skewed the ability of the AI to separate the near field window interference from the actual umbilical cord image. Several other images appeared in more than one model. These errors maybe corrected by increased data set. A greater n-size with image diversity (ie, >1000 per group) may provide the AI the ability exclude non-specific noise while retaining the diagnostic value. Additionally, if an increase training size is obtained it should include greater geographic, racial, and demographic diversity to minimize errors due to patient population homogeneity.

To test the different models to infer from new images, a simple experiment was designed. The ability for the saved model to infer cord classification from AI-OCR cleaned images was tested using two unused images. All of the models were able to correctly predict the image class but had a range of different inference probabilities. The convnext model performed the worst but this may be improved by increased epochs or data set size. The predictions showed that lateral-CFI trained models can be applied to new images at a range of different sizes. The ability for a these models to classify cord pathology shows how obstetricians can add another diagnostic tool to their clinical practice.

There has been prior research applying AI to fetal ultrasound images. This work is the first to provide a workflow, AI-OCR image blinding, transfer learning and application to a umbilical cord anatomy. The prior research apply AI to fetal ultrasound studies has been recently reviewed.[1,2] First, there has been segmentation or "coloring" of the different anatomies to aid in identification and quantification.[7,9] Next, AI and ML have been applied to measurement and fetal risk assessments.[6,9,37] Finally, there was an AI model looking at umbilical cord coiling but when analyzing the figures included both 2v and 3v cords to classify coiling index.[8] The inclusion of 2v umbilical cords to calculate coiling index could skew results. The biophysical properties of a 2v cord will have different coiling characteristics due to potentially smaller centroid radius. Occam's razor indicates that AI-based coiling research correlation to outcomes maybe explained by inclusion of SUA cords. The ability to properly classify umbilical cord artery anatomy may improve the predictive nature of coiling index.

The early detection of umbilical cord anomalies could provide obstetricians and patients with essential information to allow for a successful pregnancy. The application of an AI classification model could be used to aid obstetricians to consult maternal fetal medicine specialists or needed confirmation to schedule follow up visits. SUA pathologies can occur in under one half to 6% of the population and is associated with increase fetal risk making this is an excellent model to develop.[10,12,38] The continued development of this AI-model would be required for clinical applications. To do this a multi-stage process should be developed. First, a comprehensive blinded fetal ultrasound image repository with demographic and geographic diversity would be needed. A CT-image repository has already been successfully used in AI applications.[39] Next, an expert working group assembled to create an independent image repository for calculating diagnostic value. With a large enough image repository and expert working group guidance this model could be refined to SUA detection followed by type I through IV classification. The addition of automatic classification may provide obstetricians additional clinical outcome information. The clinical outcome information

combined with clinical experiences can aid in developing a treatment strategy. The continued development of AI/ML to develop diagnostic tools to be applied by obstetricians could help decrease unforeseen maternal or fetal complications.

This research successfully developed standardized ultrasound image blinding and AI-image classification methods. This work showed that transfer learning using models pretrained on a standard image library can be applied to ultrasounds. The trained models were able to classify 2 or 3-vessel umbilical cords and proved to be of high diagnostic value. Taking into account the training, validation, diagnostic value and inference the ResNet34 appears to work the best with this data set. In the future, a fetal ultrasound repository should be developed along with a standardized diagnostic value calculation set. Due to potential for misclassifications, this AI should be a tool but not a replacement for expert medical care. The creation of obstetrician lead AI tools for the detection of fetal anomalies or high-risk patients could help decrease negative maternal and/or fetal outcomes.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/Eric43/MFM_AI.

References

1. Ramakrishnan R, Rao S, He JR. Perinatal health predictors using artificial intelligence: a review. *Womens Health (Lond)* 2021; 17. https://doi.org/10.1177/17455065211046132.
2. Ramirez Zegarra R, Ghi T. Use of artificial intelligence and deep learning in fetal ultrasound imaging. *Ultrasound Obstet Gynecol* 2023; 62:185–194. https://doi.org/10.1002/uog.26130.
3. Rumelhart DE, McClelland JL. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Cambridge: The MIT Press; 1986.
4. Geffner H, Dechter R, Halpern JY (eds).. *Probabilistic and Causal Inference: The Works of Judea Pearl*. Vol 36. 1st ed. New York: Association for Computing Machinery; 2022.
5. Scutari M, Denis JB. *Bayesian Networks: With Examples in R*. Boca Raton: CRC Press; 2021 https://books.google.com/books?id=8VQwEAAAQBAJ.
6. Liu S, Sun Y, Luo N. Doppler ultrasound imaging combined with fetal heart detection in predicting fetal distress in pregnancy-induced hypertension under the guidance of artificial intelligence algorithm. *J Healthc Eng* 2021; 2021:4405189. https://doi.org/10.1155/2021/4405189.
7. Khan IU, Aslam N, Anis FM, et al. Amniotic fluid classification and artificial intelligence: challenges and opportunities. *Sensors (Basel)* 2022; 22:4570. https://doi.org/10.3390/s22124570.
8. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH. Machine learning model for umbilical cord classification using combination coiling index and texture feature based on 2-D doppler ultrasound images. *Health Informatics J* 2022; 28. https://doi.org/10.1177/14604582221084211.
9. Gofer S, Haik O, Bardin R, Gilboa Y, Perlman S. Machine learning algorithms for classification of first-trimester fetal brain ultrasound images. *J Ultrasound Med* 2022; 41:1773–1779. https://doi.org/10.1002/jum.15860.
10. Bohîlţea RE, Dima V, Ducu I, et al. Clinically relevant prenatal ultrasound diagnosis of umbilical cord pathology. *Diagnostics* 2022; 12:236. https://doi.org/10.3390/diagnostics12020236.
11. Bethune M, Alibrahim E, Davies B, Yong E. A pictorial guide for the second trimester ultrasound. *Australas J Ultrasound Med* 2013; 16:98–113. https://doi.org/10.1002/j.2205-0140.2013.tb00106.x.
12. Murphy-Kaulbeck L, Dodds L, Joseph KS, van den Hof M. Single umbilical artery risk factors and pregnancy outcomes. *Obstet Gynecol* 2010; 116:843–850. https://doi.org/10.1097/AOG.0b013e3181f0bc08.
13. Vafaei H, Rafeei K, Dalili M, Asadi N, Seirfar N, Akbarzadeh-Jahromi M. Prevalence of single umbilical artery, clinical outcomes and its risk factors: a cross-sectional study. *Int J Reprod Biomed* 2021; 19:441–448. https://doi.org/10.18502/ijrm.v19i5.9253.
14. Terry M, Calhoun BC, Walker W, et al. Aneuploidy and isolated mild ventriculomegaly. Attributable risk for isolated fetal marker. *Fetal Diagn Ther* 2000; 15:331–334. https://doi.org/10.1159/000021031.
15. Siargkas A, Giouleka S, Tsakiridis I, et al. Prenatal diagnosis of isolated single umbilical artery: incidence, risk factors and impact on pregnancy outcomes. *Medicina (Kaunas)* 2023; 59:1080. https://doi.org/10.3390/medicina59061080.
16. Peacock JL, Bland JM, Anderson HR. Preterm delivery: effects of socioeconomic factors, psychological stress, smoking, alcohol, and caffeine. *BMJ* 1995; 311:531–535. https://doi.org/10.1136/bmj.311.7004.531.
17. Howard J, Gugger S, Chintala S. *Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD*. Sebastopol: O'Reilly Media Inc; 2020 https://books.google.com/books?id=xd6LxgEACAAJ.
18. Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need. 2023. https://doi.org/10.48550/arXiv.1706.03762.

19. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. https://www.tensorflow.org/.
20. keras-ocr. keras_ocr documentation. 2023. https://keras-ocr.readthedocs.io/en/latest/index.html.
21. Bradski G. The OpenCV library. *Dr Dobb's J Softw Tools* 2000. https://www.drdobbs.com/open-source/the-opencv-library/184404319
22. Howard J, Gugger S. Fastai: a layered API for deep learning. *Information* 2020; 11:108. https://doi.org/10.3390/info11020108.
23. Wightman R, Touvron H, Jégou H. ResNet strikes back: an improved training procedure in timm. *arXiv abs/2110.00476* 2021.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016; 770–778.
25. Graham B, El-Nouby A, Touvron H, et al. LeViT: a vision transformer in ConvNet's clothing for faster inference. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021; 12239–12249.
26. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. *arXiv* 2022. https://doi.org/10.48550/arXiv.2201.03545.
27. R Core Team. *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing; 2023 https://www.R-project.org/.
28. NIST/SEMATECH. NIST/SEMATECH Engineering Statistics Handbook. 2002. https://books.google.com/books?id=vXXjwEACAAJ.
29. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health* 2017; 5:307. https://doi.org/10.3389/fpubh.2017.00307.
30. Anaconda Software Distribution. Anaconda Documentation. 2020. https://docs.anaconda.com/.
31. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds). *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Virginia: IOS Press; 2016:87-90.
32. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw* 2019; 4:1686. https://doi.org/10.21105/joss.01686.
33. Kassambara A. Ggpubr: ggplot2 based publication ready plots. 2020. https://CRAN.R-project.org/package=ggpubr.
34. ESS – Emacs Speaks Statistics. https://ess.r-project.org/index.php?Section=home
35. GNU Emacs - GNU Project. https://www.gnu.org/software/emacs/
36. Which image models are best? | Kaggle. https://www.kaggle.com/code/jhoward/which-image-models-are-best/.
37. Davey MA, Watson L, Rayner JA, Rowlands S. Risk-scoring systems for predicting preterm birth with the aim of reducing associated adverse outcomes. *Cochrane Database Syst Rev* 2015; 2015: CD004902. https://doi.org/10.1002/14651858.CD004902.pub5.
38. Catanzarite VA, Hendricks SK, Maida C, Westbrook C, Cousins L, Schrimmer D. Prenatal diagnosis of the two-vessel cord: implications for patient counselling and obstetric management. *Ultrasound Obstet Gynecol* 1995; 5:98–105. https://doi.org/10.1046/j.1469-0705.1995.05020098.x.
39. Bressem KK, Adams LC, Erxleben C, Hamm B, Niehues SM, Vahldiek JL. Comparing different deep learning architectures for classification of chest radiographs. *Sci Rep* 2020; 10:13590. https://doi.org/10.1038/s41598-020-70479-z.