The American Journal of Pathology, Vol. 190, No. 10, October 2020

The American Journal of
PATHOLOGY
ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Decidual Vasculopathy Identification in Whole Slide Images Using Multiresolution Hierarchical Convolutional Neural Networks

Daniel Clymer,* Stefan Kostadinov,† Janet Catov,‡ Lauren Skvarca,† Liron Pantanowitz,§ Jonathan Cagan,* and Philip LeDuc*

From the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh; the Departments of Pathology† and Obstetrics, Gynecology, and Reproductive Sciences,‡ University of Pittsburgh Medical Center (UPMC) Magee-Womens Hospital, Pittsburgh; and the Department of Pathology,§ UPMC Shadyside Hospital, Pittsburgh, Pennsylvania*

Accepted for publication June 22, 2020.

Address correspondence to Philip LeDuc, Ph.D., or Jonathan Cagan, Ph.D., Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: prl@andrew.cmu.edu or cagan@cmu.edu.

After a child is born, the examination of the placenta by a pathologist for abnormalities, such as infection or maternal vascular malperfusion, can provide important information about the immediate and long-term health of the infant. Detection of the pathologic placental blood vessel lesion decidual vasculopathy (DV) has been shown to predict adverse pregnancy outcomes, such as preeclampsia, which can lead to mother and neonatal morbidity in subsequent pregnancies. However, because of the high volume of deliveries at large hospitals and limited resources, currently a large proportion of delivered placentas are discarded without inspection. Furthermore, the correct diagnosis of DV often requires the expertise of an experienced perinatal pathologist. We introduce a hierarchical machine learning approach for the automated detection and classification of DV lesions in digitized placenta slides, along with a method of coupling learned image features with patient metadata to predict the presence of DV. Ultimately, the approach will allow many more placentas to be screened in a more standardized manner, providing feedback about which cases would benefit most from more in-depth pathologic inspection. Such computer-assisted examination of human placentas will enable real-time adjustment to infant and maternal care and possible chemoprevention (eg, aspirin therapy) to prevent preeclampsia, a disease that affects 2% to 8% of pregnancies worldwide, in women identified to be at risk with future pregnancies. *(Am J Pathol 2020, 190: 2111–2122; https://doi.org/10.1016/j.ajpath.2020.06.014)*

Rendering a microscopic anatomic pathology diagnosis involves either inspecting stained tissue sections of specimens on a glass slide under the microscope or viewing a digitized version of the slide on a computer monitor. This inspection is a complicated process requiring a highly trained pathologist. It is also time consuming, not only because of the relative complexity of the histopathology but because of the need to screen all the tissue on the slides at a microscopic level. Today, commercial technology is available to digitize pathology glass slides using whole slide scanners. Acquisition of such whole slide images (WSIs) generates digital slides that are typically on the scale of gigapixels. WSIs offer numerous applications not possible with glass slides alone, such as the ability to employ artificial intelligence to triage, screen, and provide diagnostic assistance, as introduced in this work.

Recent advances have been made in the field of machine learning (ML) for histopathologic imaging. Deep learning has been used effectively in applications such as cancer tumor detection[1] and segmentation.[2] The most common type of task for ML in digital pathology is computer-assisted diagnosis, a supervised learning task that attempts to assist the pathologist in the diagnostic process of labeling a WSI within some category of disease. Because of the large size of

Supported in part by The Center for Machine Learning and Health at Carnegie Mellon University through the Pittsburgh Health Data Alliance; Office of Naval Research grant N00014-17-1-2566; Air Force Office of Scientific Research grant FA9550-18-1-0262; NIH grant 5-R01-AG061005; and Pennsylvania Department of Health grant SAP4100077084.

Disclosures: None declared.

https://doi.org/10.1016/j.ajpath.2020.06.014

WSIs, typical ML pipelines sample smaller selected regions from the WSI, performing analysis on each region, and then employ some method of aggregating the predictions generated. Often, each one of these regions, called image patches, ranges from 128 × 128 to 512 × 512 pixels in size to be computationally tractable. Because a typical WSI will have thousands of patches, even highly accurate classification algorithms will have many false positives per image. Some of the most successful implementations of ML to automate diagnosis using digital pathology images have come from applications where relevant features exist in many patches across the image, which can be aggregated to minimize the impact of false predictions on a small number of patches. For example, in problems such as cancer severity classification from the Camelyon 2017 competition,[3] most successful teams made a global disease probability estimation based on the list of patch-level probabilities from the entire image. In the field of placental histopathology, some work has made use of deep learning,[4] although overall the use of deep learning for noncancer histopathologic image analysis has remained limited, partially because of the limited availability of large data sets.

Microscopic analysis of the human placenta has been advocated in certain clinical settings to determine the anatomic basis of pregnancy-related complications.[5] When correlated with clinical findings, the results of a placental examination may provide actionable information to optimize treatment of both the mother and newborn. This is particularly important when an adverse pregnancy outcome occurs, and in these cases, a major role of the placental examination is to provide supporting histopathologic evidence of the disease process. For example, preeclampsia is a major pregnancy complication characterized by new-onset maternal hypertension and is associated with many serious acute and chronic adverse consequences for both the mother and the newborn. Preeclampsia affects 2% to 8% of pregnancies and is the leading cause of preterm birth and consequent neonatal morbidity in the developed world.[6,7] There are many known patterns of chorionic villous morphology, vasculature, and lesions that pathologists look for related to preeclampsia. In particular, the presence of a placental lesion called decidual vasculopathy (DV) is often found in cases of preeclampsia, and when detected in an uncomplicated pregnancy, it has been correlated to the occurrence of preeclampsia and other adverse outcomes in subsequent pregnancies.[8,9] Microscopic detection of DV, often characterized by hypertrophy of decidual arterioles, is vital for providing physicians with the information they need to move forward with treatment of the mother and newborn, especially with recent research showing that regular doses of aspirin during the first trimester can help to prevent preeclampsia in women who are determined to be at risk for the disease.[10,11]

In most hospitals, there are often so many deliveries being performed that there are not enough resources to examine every placenta microscopically. Furthermore, not all features may be reliably detected by general pathologists,[12] which justifies the need to employ perinatal pathologists to examine these placentas. Most placentas from uncomplicated pregnancies are typically discarded, with no microscopic inspection being performed.[13–15] There has been recent research working toward partial automation of some of these analyses, including the use of image processing, such as texture analysis[16] or morphometry,[17,18] to perform tasks such as vessel detection or villi counting.[19] The aim of this current work is to provide a microscopic placental analysis service to most mothers and infants who do not currently have access to this service, through both the automated detection and diagnosis of DV lesions, allowing many more placentas to be efficiently inspected in a more standardized manner and enabling diagnoses that could save lives during future pregnancies and lower health care costs.

From an image analysis perspective, the DV lesion is minute compared with the size of a WSI, and requires viewing the image at high resolution to accurately detect and diagnose. Moreover, the focal occurrence of even one DV lesion in a placental image containing numerous unaffected vascular segments could be indicative of possible future health problems, meaning that any practical implementation of an automated placental lesion detection algorithm requires stringent levels of both sensitivity and specificity. In this work, to minimize false positives from a WSI placental analysis while maintaining a high level of diagnostic accuracy, we propose a multiresolution deep learning framework in which high-resolution regions for classification are informed by a broader low-resolution examination for regions of interest.

Additionally, a method of aggregating local patch estimations from the latent space, or the learned hidden feature representation, of our classification framework is investigated. These aggregated features are combined with patient metadata for the purpose of learning a global classification of disease for each patient, which can inform the pathologist about which WSIs should be analyzed in more detail. This method can help achieve effective results when training with comparatively small data sets, as are often found in biomedical applications as well as the current work. The presented algorithm, shown at a high level (Figure 1), is designed to be used as a low-cost early microscopic detection method for predicting which mothers are most at risk of developing preeclampsia in future pregnancies, and can be treated to prevent this from occurring.

Materials and Methods

The overall deep learning pipeline has three stages: object detection, classification, and aggregation. The overall process (Figure 1) is built with two separately trained neural networks as well as a final aggregation step. In the object detection stage, a WSI is fed into the pipeline, split into a grid of patches, and analyzed for the detection of blood

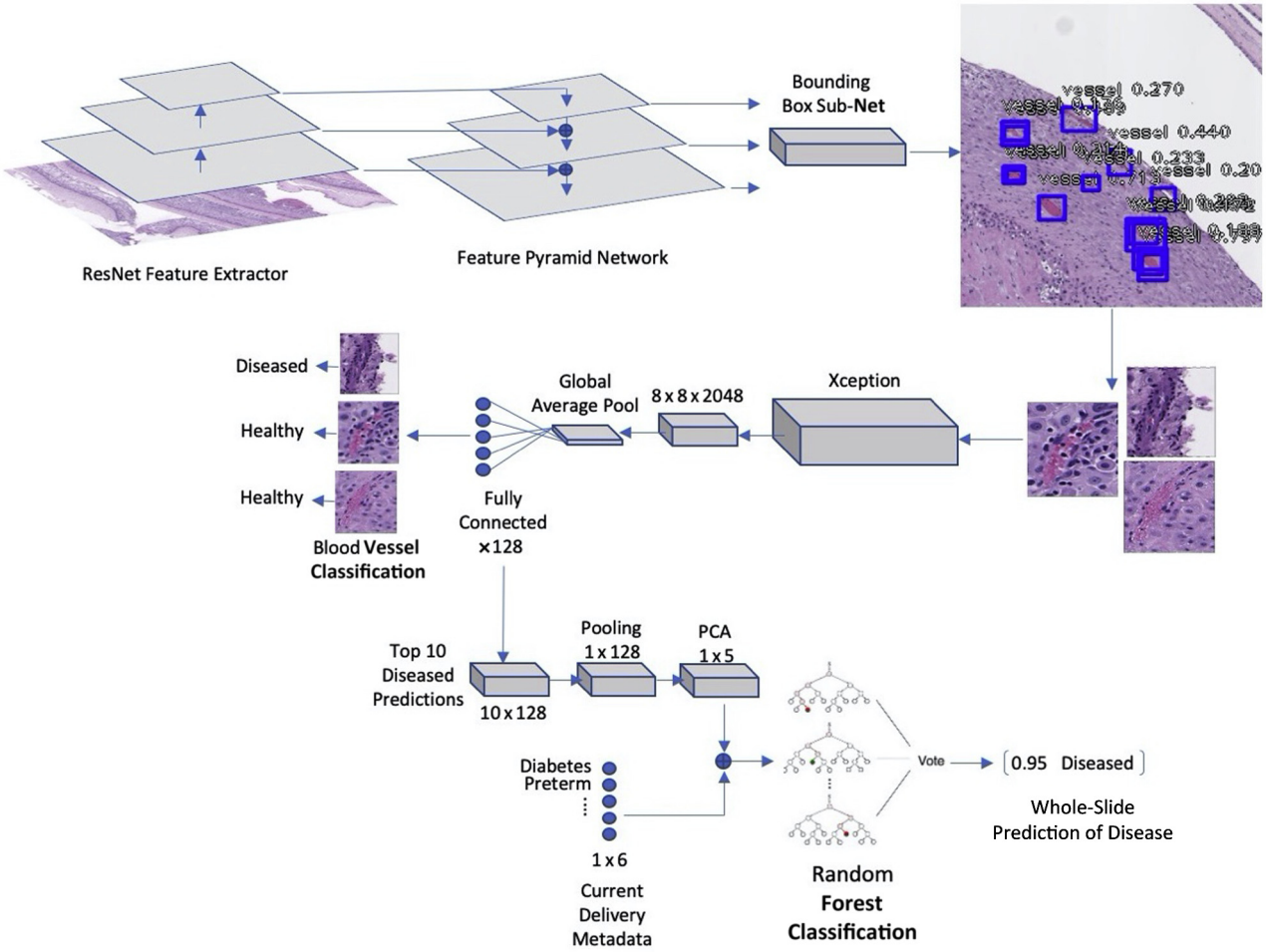

Figure 1 Overview of multiresolution machine learning framework used for whole slide placental image analysis of blood vessel lesions. The top level is an object detection framework trained to detect blood vessels at a low resolution within a whole slide image. The middle level is a classification framework that analyzes and detects disease inside a high-resolution patch of each blood vessel. The bottom level aggregates the latent space vectors of these patch-wise predictions, along with patient delivery metadata, for the prediction of disease at the patient level. PCA, principal component analysis.

vessels. The output of this provides a list of bounding boxes with expected probabilities for each patch. From each bounding box location, a higher-resolution patch is taken and used in the next stage of blood vessel classification, which outputs a binary classification for each blood vessel, as either diseased or healthy. The latent space vectors from these classifications are aggregated and concatenated with patient delivery metadata (Table 1). These patient-level data are classified with a random decision forest model, which outputs a diseased/healthy classification for the WSI.

Clinicopathologic and Image Data Sets

A total of 181 archival placenta cases from UPMC Magee-Womens Hospital (Pittsburgh, PA), obtained between the years 2008 and 2012, were enrolled in this study. The whole slide images used in this study are available from Univeristy of Pittsburgh Medical Center (*http://image.upmc.edu:8080/cmu%20placenta%20project/view.apml?*, last accessed May 17, 2020). Glass slides with hematoxylin and eosin stained tissue sections, cut at 4 to 5 μm thickness, were scanned on an Aperio AT whole slide scanner (Leica Biosystems, Wetzlar,

Table 1 Description of Patient Metadata Features Used in This Study

Feature	Description
Hypertensive disorder of pregnancy	Including preeclampsia, eclampsia, and HELLP (hemolysis, elevated liver enzymes, and a low platelet count) syndrome
Placental weight	Placental weight at delivery, in grams
Diabetes	Maternal diabetes status
Lupus	Maternal lupus
Infant growth	Infant growth using Alexander Growth Chart: 0 for average or large birth weight, 1 for small birth weight for gestational age
Preterm birth	Preterm delivery, defined as <37 weeks gestation

Germany) at ×20 magnification, acquiring digital slides with a resolution of 0.50 μm/pixel, using bright-field microscopy. Cases that were selected for review were analyzed by a blinded perinatal pathologist (S.K.). There were 46 cases (25%) with confirmed DV. DV lesions are characterized by abnormalities of decidual arterioles that may include a combination of fibrinoid necrosis of vessel walls, hypertrophy of the media, subendothelial lipid-laden macrophages, and possible thrombi within the lumen. Each slide contained many normal (nonlesional) microscopic blood vessels (approximately 30 per slide), whereas in the cases with DV, approximately five of these blood vessels per slide displayed signs of DV. All identifying information was removed from the images through an honest broker system, and the study was approved by the University of Pittsburgh Institutional Review Board (number STUDY19050188). The images were matched to clinical data through the Magee Obstetric Maternal and Infant database. The clinical data features used along with the images in this study were placental weight, diabetes status, lupus status, hypertensive disorder of pregnancy, infant growth, and preterm birth (Table 1).

Samples for histologic evaluation were taken from several regions of the placenta, including the umbilical cord, the placental disc, and the fetal membranes. For this study, digital images of the membranes from which a strip is taken rolled up tightly and cross-sectioned were analyzed. An example of the membrane roll is shown (Figure 2). The membrane roll provides opportunity to examine a large cross-sectional area of the decidual region of the placenta, which contains the distal portion of decidual spiral arterioles and is the region where DV lesions are most likely to be found. The images were reviewed by two pathologists (L.S. and S.K.) who labeled 710 instances of DV that were used as the primary labels for training. These labels were curated with the VGG Image Annotator.[20] In addition, healthy regions of the image were annotated by a graduate student (D.C.) trained to identify blood vessels in WSIs, who provided 6095 annotations for training. Because the presented algorithm was trained to identify diseased blood vessels, variance in the labels of the healthy class should not have affected the performance metrics of our algorithm, which was compared with the ground truth labels from physicians.

Object Detection

The purpose of this stage of the framework is to localize blood vessels in the WSI, analyzed at a low resolution, to feed these localizations into the next stage of the classification pipeline. To accomplish this, a localization framework called RetinaNet, published by Facebook,[21] is used. This framework introduced focal loss for training, which reduces the influence of well-classified background examples on the weight updates during training, and has been shown to be effective for object detection frameworks, particularly in cases such as ours where the number of background pixels vastly outweighs the pixels occupied by objects during training. The algorithm in this article uses a ResNet backbone as the feature extractor for this framework,[22] which uses residual learning to alleviate the vanishing gradient problem when training deep networks. The entire framework was initialized with network weights pretrained on the MS-COCO data set.[23] The MS-COCO data set used for this task is available from Common Objects in Context (COCO, *http://images.cocodataset.org/zips/train2017.zip*, last accessed May 17, 2020). All layers of the network were then fine-tuned through training on WSI data set.

For training, each WSI is split into a grid of 256 × 256-pixel patches at four times resolution, with a 10% overlap between patches to help account for blood vessels that would be split between patches. Because this overlap will sometimes cause the algorithm to detect the same blood vessel on two different patches, nonmaximum suppression is used to only keep the most confident among overlapping outputs before feeding into the next stage of the ML pipeline, to avoid having duplicate predictions on the same blood vessel. As a preprocessing step, slide patches that contained bubble artifacts were excluded from analysis because of the changes these bubbles cause in imaging features. Patches with folds that went through an annotated blood vessel were also removed because these could potentially affect the analysis. This patch generation typically results in around 50 to 100 patches per WSI. Similarly to many other state-of-the-art medical image analysis algorithms,[24,25] the data were augmented with random flips, rotations, and translations during training. In addition, a method of stain normalization introduced by Macenko

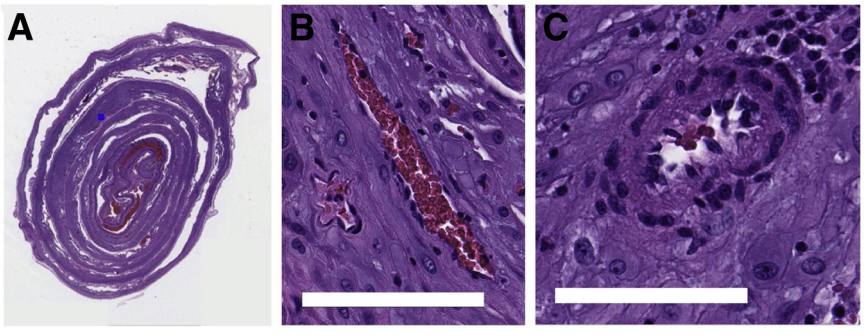

Figure 2 Example image and blood vessel patches from data set. **A:** A digitized whole slide image of a placental membrane roll [low magnification; hematoxylin and eosin (H&E) stain]. To illustrate relative scale, the **blue square** indicates a single blood vessel. **B:** Image patch showing an example of a healthy blood vessel (high magnification; H&E stain). **C:** Image patch showing a decidual arteriole affected by early-stage decidual vasculopathy, characterized by smooth hypertrophic muscle around the blood vessel lumen (high magnification; H&E stain). Scale bars: 125 μm (**B**); 100 μm (**C**). Original magnification, ×20 (**A**).

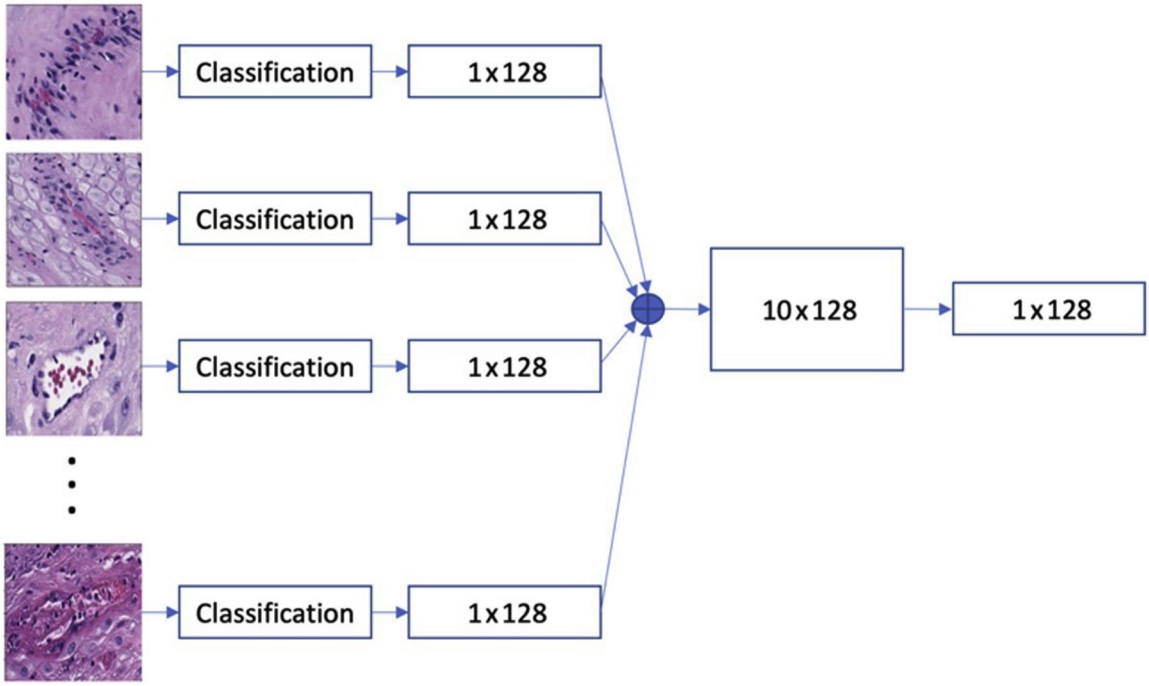

Figure 3 Method for aggregating latent feature representations from numerous blood vessels in one image. First, all detected blood vessels are evaluated with the classification algorithm. Then, the 128-node latent spaces from the top 10 predicted diseased blood vessels are concatenated into a single 10 × 128 matrix. Finally, a pooling operation is performed on each column of this matrix, where either the maximum or minimum value is taken, depending on if that node was maximized or minimized by the diseased image set during training.

et al[26] was used, which maps the individual stain contributions in an RGB (red, green, blue) image through an optical density transformation. The network was trained for 75 epochs, with a learning rate of 1×10^{-5} and a batch size of 1.

Classification

The blood vessel classification stage of the pipeline analyzes each blood vessel patch that has been identified from the previous object detection stage, for the purpose of classifying vessels with DV. To feed an image into this network, an image patch is taken around each localized blood vessel at 10× resolution (2.5 times higher than the previous stage), with 20 pixels of padding added around each side of the blood vessel. Because blood vessels can exhibit a range of shapes and sizes, each image is rescaled to be square, and then resampled to 299 × 299 pixels before feeding into the network.

Our model utilizes the Xception convolutional neural network backbone,[27] which uses depth-wise separable convolutions to reduce the computational complexity required to train a deep convolutional neural network. Initialized weights learned from training on the ImageNet data set were used.[28,29] The data set is available from ImageNet (*http://image-net.org/download-images*, last accessed May 17, 2020, a free account registration is required). The 1024 feature maps learned by network are fed through a global average pooling layer, which is a method to effectively reduce the trainable parameters during classification to avoid overfitting.[30] The output from this layer is fed through a 128 dense layer before the final classification layer. The network is trained with the Adam optimizer,[31] and dropout[32] of 50% is applied to the 128-node fully connected layer.

Standard data augmentation of flips, shear, rotation, and translation is used. Flips were performed with 50% likelihood, shear was applied between -15 and 15 degrees, rotation was applied between -45 and 45 degrees, and translation was applied between −15% and 15% of the size of the image in both the x and y directions. These augmentations are used because it is expected that because the orientation of a blood vessel is irrelevant to its classification, these types of affine translations would generate images that would still be considered valid blood vessels and diagnosable by a physician. The shear transformation is used because it is a method of simulating a blood vessel taken in an out-of-plane cross-section, providing more diversity in the training set. To account for the class imbalance in this stage of training, the data augmentation pipeline is weighted so that the underrepresented diseased class receives more augmentations during each batch.

Aggregation

After obtaining localized blood vessels and diseased classifications for each blood vessel in a WSI, the next step is to

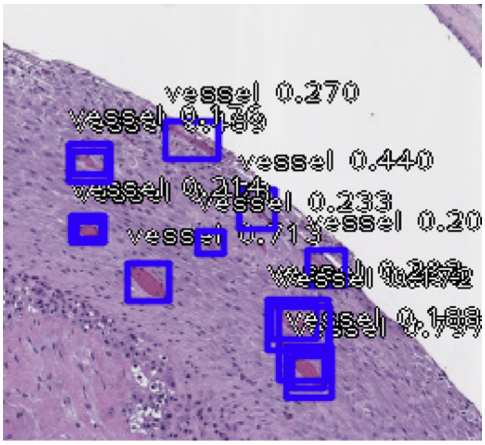

Figure 4 Example output from object detection framework for blood vessel detection, showing bounding boxes being drawn around each blood vessel in the image patch, along with prediction outputs from the algorithm for each detection.

aggregate these predictions to obtain a single vector of data for classification of disease at the image level. This classification can be used to alert the physician about which images would benefit from more in-depth pathologic inspection.

To maximize classification accuracy at this stage, it is desirable to incorporate information from many regions of the image rather than relying on image features detected from a single blood vessel. That is, it is expected that a more holistic representation of classified disease in the image will result in a better global-scale classification accuracy. However, each WSI has a different number of blood vessels, and any number of these vessels can have or not have disease.

From the previous stages of the analysis pipeline, each blood vessel has been identified and classified through a deep network. The method for aggregating these is described (Figure 3). First, the blood vessels are classified with the deep learning network. The latent feature representation of each classified blood vessel is 128 nodes in length. The blood vessels are sorted in a list based on the classification network's output confidence, from most to least diseased (note that even healthy blood vessels can be ranked in terms of diseased confidence). In this work, algorithm confidence is considered to be proportional to the deviation from 0.5 of the scalar classification output, with a diseased classification of 1.0 being maximally confident. Next, the latent vectors are stacked into one matrix of shape Nx128, where N is the number of analyzed blood vessels included in the analysis. In this work, the top 10 ranked blood vessels are used for the analysis. This number is chosen to minimize the potential negative impact of having a large number of healthy blood vessels in the downstream analysis, because most diseased images in the training set had $\leq$10 diseased blood vessels. The latent spaces are aggregated instead of the individual blood vessel classifications so that the full algorithm pipeline has the opportunity to make decisions from the more holistic feature sets learned during training, such as morphologic patterns across blood vessels, and not solely from the scalar classifications. After this aggregation, the data are pooled by calculating either the maximum or the minimum of the data for each node of the feature map; this is determined for each node based on if that node was being maximized or minimized by the diseased image class during training. As expected for a well-balanced classifier, about half of the nodes were being maximized and half were being minimized during training.

Once the aggregated latent vector, of size 1 $\times$ 128, has been generated for each image in the training set, principal component analysis is performed to reduce the dimensionality to 1 $\times$ 5. Principal component analysis is an unsupervised transformation method that linearly maps data to a lower dimensional space while maximizing the amount of variance explained in the original data.[33] This transformation is used to reduce the number of dimensions of the training data, to reduce the risk of overfitting on a small data set, while still keeping as much of the variance in the data as possible.

After dimensionality reduction, the pooled latent representation is concatenated with a vector of patient metadata describing the mother's health and outcome of the delivery. The metadata features used in this work are described (Table 1).

The combined vector of latent and metadata features is zero centered and scaled by the SD. The resulting data are used to train a random decision forest classifier to perform a binary classification between diseased and healthy slides. A random decision forest[34] is a method of ensemble learning in which a large number of shallow decision trees are constructed to provide an output that is the mode of the predictions from each tree. This type of model was selected for its robustness to overfitting, particularly on small data sets.[35] A diseased slide is defined as one with at least one example of a diseased blood vessel, which is considered a clinically relevant indication of potential hypoxia-related disease. At this stage of the pipeline, each whole slide image, as opposed to each blood vessel patch, is considered

Table 2 Results of Blood Vessel Object Detection, Trade-Off between High Disease Class Recall, and Overall Precision as the Discriminative Threshold Is Lowered

Total predictions made on 2524 test patches, *n*	Overall precision	Diseased class recall
28,834	0.08	0.95
16,663	0.13	0.94
12,400	0.22	0.92
6574	0.45	0.9
5028	0.59	0.86
4228	0.69	0.8
3660	0.77	0.77

Table 3 Performance Metrics for Blood Vessel Patch Classification

Variable	Validation, %	Test, %
Sensitivity	95	94
Specificity	96	96
Accuracy	96	96

$n = 989$ validation; $n = 1341$ test.

to be one data sample, which drastically lowers to available training data compared with the previous stages of the pipeline. To help avoid overfitting, a fivefold cross validation is used to select the best model parameters for the training set, which are then applied to the test set.

Results

Object Detection

The whole slide data set of 181 slides (46 with identified DV, 135 without) was split into a set of 11,610 low-resolution patches for the blood vessel detection step. For this set, 7281 patches were used for training, 1805 were used for validation, and 2524 were used for testing. Because the number of DV annotations is small compared with the total number of blood vessel annotations, and to prevent overfitting, this stage of the network treats blood vessels as a single class and does not make a classification, instead only outputting bounding boxes for any predicted blood vessels. An example of this is shown (Figure 4).

The purpose of this stage of the pipeline is to narrow down the number of regions that need to be analyzed in higher resolution as much as possible, without missing regions of DV in the WSI. To this end, one feature of the presented algorithm is that the discriminative threshold can be tuned to be more lenient and find a higher percentage of diseased blood vessels, at the expense of making more predictions on incorrect regions as well. The results of sweeping through many of these thresholds are shown (Table 2). Although this stage of the network is being trained to detect blood vessels without differentiating between healthy and diseased classes, the primary metrics that are considered important for this research are the network's ability to capture the disease cases, because the purpose of the framework is to identify DV within the digital slide. Using a lenient cutoff value (such as the 28,834 predictions) (Table 2), the recall of the diseased class (the number of diseased vessels captured in the predictions divided by the total number of diseased vessels) is high, whereas the total class precision (the number of total annotations captured in the predictions divided by the total number of predictions made) is low. However, using a stricter cutoff value (such as the 3660 predictions) (Table 2) results in fewer overall predictions, raising the total class prediction while lowering the diseased recall.

Classification

At this stage in the ML pipeline, patches are taken at high resolution around each annotated blood vessel. Of the 6095 total annotated blood vessels, 3765 (3313 healthy and 452 diseased) were used for training, 989 (879 healthy and 110 diseased) were used for validation, and 1341 (1193 healthy and 148 diseased) were used for testing. These images were augmented throughout training, which is described in detail above (*Materials and Methods*). The algorithm was trained for 40 epochs (cycles through full training set), and the trained model with the highest validation accuracy was selected for testing.

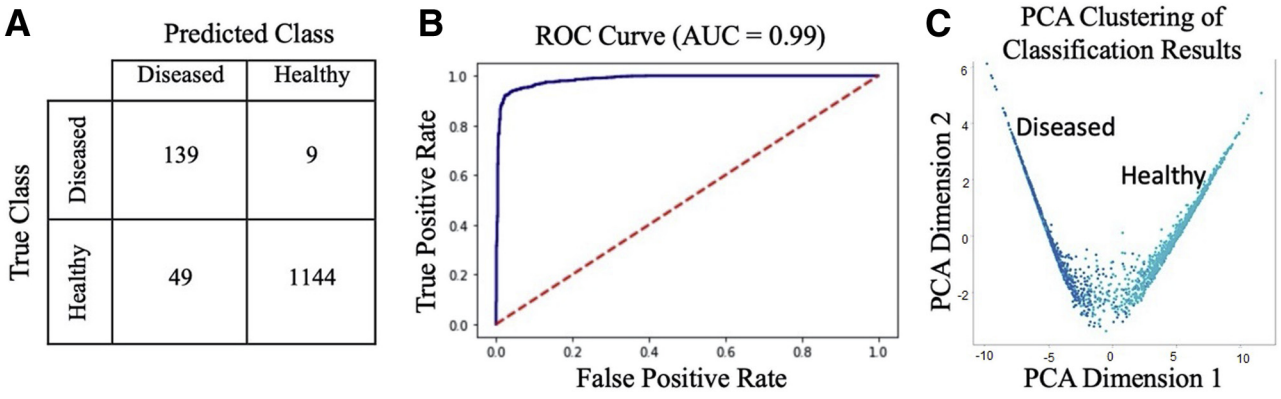

Figure 5 Blood vessel classification results. **A:** Confusion matrix of results on test set (positive predictive value = 0.74, and negative predictive value = 0.99). **B:** Receiver operating characteristic (ROC) curve of blood vessel classification results on the test data set, illustrating the achievable trade-off between true- and false-positive rates as the discriminative threshold of the classifier is varied. The **dashed line** shows the comparison to a random classifier. **C:** Results of two-dimension principal component analysis (PCA) on the 128-node latent space of classified blood vessel patches from the test set, to help visualize clusters being formed by the classification algorithm. Dark points indicate diseased blood vessels; and light points, healthy placenta vessels. AUC, area under the curve.

Table 4 Results of Test Slides from the Combined Object Detection and Classification Pipeline

Predictions from object detection, *n*	Fraction of diseased blood vessels located	Diseased recall on object detection outputs	Overall diseased recall	Healthy recall on object detection outputs	Overall diseased precision
28,834	0.95	0.92	0.87	0.91	0.19
16,663	0.94	0.92	0.86	0.91	0.29
12,400	0.92	0.92	0.85	0.91	0.35
6574	0.9	0.92	0.83	0.92	0.53
5028	0.86	0.93	0.80	0.92	0.58
4228	0.8	0.93	0.74	0.92	0.61
3660	0.77	0.93	0.72	0.92	0.64

The overall performance metrics for the blood vessel classification stage of the ML pipeline is shown (Table 3). For the test set, a sensitivity (true positives divided by total positives) of 94% and a specificity (true negatives divided by total negatives) of 96% was achieved. A slightly higher weighting toward the negative classification rate in both the validation and test set is observed, which may be due to the large class imbalance between the diseased and healthy sets. The confusion matrix for the test results is shown (Figure 5A), demonstrating the specific results for each class.

To demonstrate the expected trade-off between the true- and false-positive rates from the binary classification, a receiver operating characteristic curve was generated for the results on the test data. This curve is made by sweeping through every possible discriminative threshold value of a binary classifier and plotting the corresponding true-positive and false-positive rate for each point. The area under the curve is a commonly reported metric in binary classification, and is interpreted as the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample.[36] A receiver operating characteristic curve with an area under the curve equal to 1 is considered a perfect classifier. The area under the curve for the blood vessel classification task applied to the test set was 0.99 (Figure 5B).

A two-dimensional principal component analysis of the latent space of the validation set was also performed, mapping the data to the two orthogonal dimensions corresponding to the highest variance in the training set. This mapping can be used to visualize the algorithm's separation of the class populations and provide some measure of validation that a useful latent representation has been learned. Two visible distributions emerge between the healthy and diseased classes (Figure 5C).

Results from Combined Object Detection and Classification Framework

Both the classification and the object detection phases of the ML framework were trained separately from one another during the training phase. However, to get a more accurate estimate for the performance of the entire pipeline in a clinical setting, the held-out test slides were run through the entire ML pipeline, with the predictions made from one phase being used as the inputs for the next phase. To ensure

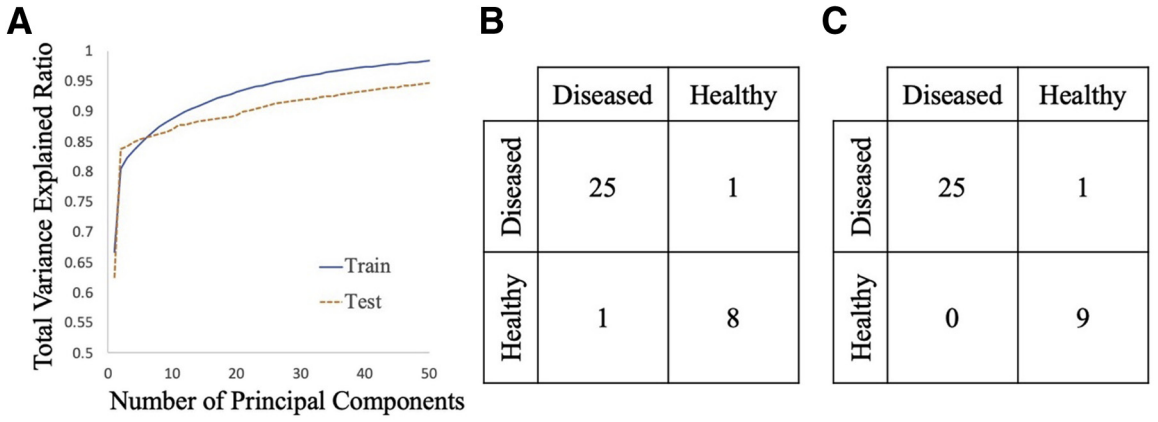

Figure 6 Results from combined object detection + classification analysis. **A:** Examination of variance explained by principal components used to trade-off between the number of principal components used and the ratio of explained variance (or the ratio of the cumulative sum of eigenvalues) in the blood vessel latent space analysis, with principal components fit to the train set and applied to both the train and test set. Five principal components were selected to be used for the final analysis, as higher numbers of principal components begin to display evidence of overfitting between the train and test set. **B:** Confusion matrix showing global disease classification results on test set, using latent features only. **C:** Confusion matrix showing global disease classification results using latent features and patient metadata.

Table 5 Global Disease Classification Results based on Aggregated Classification Latent Space Vectors, Both with and without Inclusion of Patient Metadata Features

	Latent features, %		Latent + metadata, %	
Variable	Validation	Test	Validation	Test
Sensitivity	98	89	99	100
Specificity	98	96	99	96
Accuracy	98	94	99	97

unbiased assessment, no model parameters at any stage of the pipeline were influenced by any data from these held-out test slides. For labeling purposes, any output from the object detection algorithm that had an intersection over union of >0.15 with a diseased blood vessel was given a diseased label, and all other predictions were given a healthy label. This intersection over union was chosen because, with a margin of padding being added around each object before being analyzed with the classification algorithm, it was expected to cover a sufficient area of diseased tissue for a classification to be made, a claim that has been validated qualitatively through shadowing sessions with a pathologist.

Similar to the results of the object detection stage of the pipeline, the discriminative threshold can be tuned to be either more strict, and make fewer incorrect predictions at the expense of finding fewer of the diseased blood vessels, or more lenient, and make more incorrect predictions while also finding a higher percentage of the total number of diseased blood vessels. The results of this trade-off for seven different discriminative thresholds are shown (Table 4). The main two performance metrics for comparison of these results are the overall diseased recall, which is the fraction of total diseased blood vessels that were located and correctly classified by the pipeline, and the overall diseased precision, which is the fraction of disease predictions that were actually diseased blood vessels.

Whole Slide Classification through Aggregated Latent Space Analysis

To obtain a single whole slide classification of disease, the latent space features from the blood vessel classifications are sorted, aggregated, and pooled for each patient, using a pooling technique described (*Materials and Methods*). This results in a vector of 1×128, on which dimensionality reduction via principal component analysis is performed. The obtainable explained variance ratio between using 1 and 50 principal components (eigenvectors) that were fit to the training data set, applied to both the train and test set, is shown (Figure 6A). The explained variance ratio for N principal components is equal to the sum of the N largest eigenvalues of the covariance matrix of the data, divided by the total sum of all eigenvalues. The first several principal components increase the explained variance significantly, and it starts leveling off at a higher number of principal components. Five eigenvectors were selected to be used in the analysis, to avoid overfitting between the training and test data that starts to become more apparent at higher numbers of principal components.

The resulting feature set is concatenated with patient metadata features and used as the input data to train a random decision forest[34] classification algorithm, which is a method of ensembling numerous shallow decision trees to prevent overfitting. A hyperparameter grid search was performed to scan through a set of potential algorithm parameters, in which the following parameters were explored: number of estimators, maximum depth per tree, and maximum features to consider at each split. Feature splits were selected using the maximum information gain criterion. A five-fold cross validation was performed for each parameter combination in the grid search, and the parameters with the best average performance across all five folds in the cross validation were selected to be used for testing the model performance. The final chosen parameters were as follows: number of estimators, 50; maximum depth, 2; and maximum features, 3. For testing, all five folds were combined and used to train a random decision forest algorithm with the selected parameters, which was applied to the held-out test set. The results from this are shown (Table 5) for both the case of only using latent image features, as well as the combined analysis with latent image features and patient metadata. The confusion matrix of these results is shown (Figure 6, B and C).

Discussion

In the whole slide classification through latent feature analysis, one case from the diseased and one case from the healthy class were misclassified. In the combined latent feature and patient metadata analysis, only one healthy case was misclassified. At the whole slide level, the algorithm's approach of aggregating blood vessel predictions appears to mitigate the risk of misclassification by incorporating the highest activated features from the set of classified blood vessels. This algorithm design was partially motivated by comments from a perinatal pathologist during shadowing sessions, that accurate DV classification could be made by a clear observation of DV in a small number of blood vessels, regardless of less certain classification of most blood vessels.

Although there has been recent work in artificial intelligence classification for other types of digital pathology,[3,37] there has been less work in algorithmic DV classification that could provide a baseline for comparison. Redline et al[38] performed a group survey of eight pathologists classifying placental lesions from 20 whole slide images, reporting that collective individual sensitivity and specificity for finding mural hypertrophy was 88% and 92%, respectively, which is somewhat lower

than our individual blood vessel classification rates of 94% sensitivity and 96% specificity. Furthermore, the area under the curve of 0.99 for individual test set blood vessel classifications in this study is within the range often considered to be excellent.[39,40]

The blood vessel classification results (Figure 5) show higher classification accuracy compared with the combined object detection and classification pipeline results (Table 4). This is anticipated because of the variability induced when the predicted blood vessel bounding boxes do not line up with the ground truth bounding box locations, which is caused by error in the object detection stage's bounding box predictions. This forces the classification algorithm to make predictions on partially cut off blood vessels. As such, the algorithm appears to be overfitting slightly toward classifying blood vessels that are fully visible and centered in an image. In addition, to achieve a high percentage of identified diseased blood vessels, a low discriminative threshold for the object detection is required, resulting in a high number of overall predictions being made. Despite both a high sensitivity and specificity, the large number of healthy patches compared with diseased patches causes a low diseased precision (true positives divided by total predicted condition positive). However, the number of false positives is still low, accounting for the size of an individual blood vessel compared with the size of a whole slide image. For example, for an average blood vessel size approximately 150 pixels in length, there would be around 40,000 patches of this size in an average test slide. A balanced classifier with even 99% accuracy, tested on all possible patches, would have hundreds of false positives that could result in a diseased precision of $\leq$0.01. The current study's pipeline is able to achieve diseased precisions of between 0.19 and 0.64, in part because of the low-resolution object detection stage that limits the number of patches that need to be analyzed in the classification stage. Depending on the importance of minimizing false negatives versus false positives, the discriminative threshold can be tuned to meet the user requirements.

One of the main sources for error in the combined ML pipeline seems to come from partially truncated blood vessels in the object detection stage. One potential method to account for this would be to add random crops in the augmentation pipeline during training, to simulate partially visible blood vessels. Another aspect to explore could be further analysis of image regions containing artifacts, such as bubbles or tissue folds, which were removed from training in the current study. With a larger data set, the ML algorithm could potentially be trained to perform a quality check to recognize these artifacts, as either a preprocessing module at the beginning of the ML pipeline or additional classifications through the algorithm's regular inference. There has been recent research in the literature dedicated to whole slide digital image quality checks,[41] which could potentially be incorporated with the current study's pipeline to improve robustness.

The data used in this study consisted of cases from a single institution. Because variations in scanning techniques have the possibility of causing biases between local institutions, future development would benefit from cross-institutional data sources to help validate the generalizability of the algorithm. However, some recent research[42,43] has found that regularized convolutional neural networks trained on single-institution data were robust to cohort variations during validation on data from other institutions.

Another matter for consideration when applying this algorithm clinically is the method in which the output is shown to the user who may or may not be familiar with artificial intelligence, to increase explainability and confidence, as well as efficiency in analysis. One possible method of output would be to provide a gallery of high-resolution image patches to the physician, that are sorted by the confidence level of the algorithm, so that the user is first shown the blood vessels that are most strongly considered to show disease, to allow the physician to efficiently sort through a large amount of imaging data. Furthermore, more data could allow for additional complexity in placental analysis, both in terms of expanding beyond a binary classification to account for multiple grades and morphologies of DV as well as expanding the placental regions, structures, and pathologies to be analyzed. In addition, in the future, it would be of interest to compare the current study methods with a set of new experiments using only the clinical metadata, to learn the types of cases that are best served by WSI analysis.

In summary, the proposed ML framework introduces a hierarchical method to analyze histologic digital images, for the purpose of automating placental DV lesion inspection. Results from this data set show the algorithm's ability to discriminate key features and candidate locations within a high-resolution WSI while keeping false positives minimized. This type of artificial intelligence approach can allow many more placentas to be screened with fewer pathologists, increasing DV detection for mothers who are at risk for preeclampsia in subsequent pregnancies.[9] When this risk of preeclampsia is identified earlier, it allows for preventative treatment, such as low-dose aspirin to delay the progression of the disease during pregnancy.[44,45] This can accordingly reduce health care expenses and reduce both mother and neonatal morbidity, particularly in the developing world where perinatal pathologist expertise may not be available.

Acknowledgments

We thank Danielle Sharbaugh, Cassandra Berkey, and the Magee Obstetric Maternal and Infant (MOMI) research group for selecting, curating, and managing the clinical data set and Matt O'Leary for managing the data transfer between the institutions throughout this research.

References

1. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non−small cell lung cancer histopathology images using deep learning. Nat Med 2018, 24:1559−1567
2. Xu Y, Jia Z, Wang L-B, Ai Y, Zhang F, Lai M, Chang EI-C: Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 2017, 18:281
3. Bandi P, Geessink O, Manson Q, Van Dijk M, Balkenhol M, Hermsen M, Bejnordi BE, Lee B, Paeng K, Zhong A: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE T Med Imaging 2018, 38:550−560
4. Ferlaino M, Glastonbury CA, Motta-Mejia C, Vatish M, Granne I, Kennedy S, Lindgren CM, Nellåker C: Towards deep cellular phenotyping in placental histology. Amsterdam, the Netherlands: The 1st Conference on Medical Imaging with Deep Learning, 2018
5. Salafia CM, Vintzileos AM: Why all placentas should be examined by a pathologist. Am J Obstet Gynecol 1990, 163:1282−1293
6. ACOG practice bulletin no. 202: gestational hypertension and preeclampsia. Obs Gynecol 2019, 133:e1−e25
7. Roberts JM, August PA, Bakris G, Barton JR, Bernstein IM, Druzin M, Gaiser RR, Granger JP, Jeyabalan A, Johnson DD: Hypertension in pregnancy: executive summary. Obstet Gynecol 2013, 122:1122−1131
8. Stevens DU, Al-Nasiry S, Bulten J, Spaanderman MEA: Decidual vasculopathy and adverse perinatal outcome in preeclamptic pregnancy. Placenta 2012, 33:630−633
9. Hauspurg A, Redman EK, Assibey-Mensah V, Tony Parks W, Jeyabalan A, Roberts JM, Catov JM: Placental findings in nonhypertensive term pregnancies and association with future adverse pregnancy outcomes: a cohort study. Placenta 2018, 74:14−19
10. Rolnik DL, Wright D, Poon LC, O'Gorman N, Syngelaki A, de Paco Matallana C, Akolekar R, Cicero S, Janga D, Singh M, Molina FS, Persico N, Jani JC, Plasencia W, Papaioannou G, Tenenbaum-Gavish K, Meiri H, Gizurarson S, Maclagan K, Nicolaides KH: Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. N Engl J Med 2017, 377:613−622
11. Hauspurg A, Sutton EF, Catov J, Caritis NS: Aspirin effect on adverse pregnancy outcomes associated with stage 1 hypertension in a high-risk cohort. Hypertension 2018, 72:202−207
12. Hecht JL, Zsengeller ZK, Spiel M, Karumanchi SA, Rosen S: Revisiting decidual vasculopathy. Placenta 2016, 42:37−43
13. Spencer M, Khong T: Conformity to guidelines for pathologic examination of the placenta. Arch Pathol Lab Med 2003, 127:205−207
14. Sills A, Steigman C, Ounpraseuth ST, Odibo I, Sandlin AT, Magann EF: Pathologic examination of the placenta: recommended versus observed practice in a university hospital. Int J Womens Health 2013, 5:309−312
15. Curtin W, Krauss S, Metlay L, Katzman P: Pathologic examination of the placenta and observed practice. Obs Gynecol 2007, 109:35−41
16. Swiderska-Chadaj Z, Markiewicz T, Koktysz R, Kozlowski W: Texture analysis to trophoblast and villi detection in placenta histological images. Conf Inf Technol Biomed 2016, 2:183−192
17. Mukherjee R: Morphometric evaluation of preeclamptic placenta using light microscopic images. Biomed Res Int 2014, 2014:1−9
18. Ptacek I, Smith A, Garrod A, Bullough S, Bradley N, Batra G, Sibley CP, Jones RL, Brownbill P, Heazell AEP: Quantitative assessment of placental morphology may identify specific causes of stillbirth. BMC Clin Pathol 2016, 16:1
19. Swiderska-Chadaj Ż, Markiewicz T, Koktysz R, Kozlowski W: Automatic method for vessel detection in virtual slide images of placental villi. Recent Glob Res Educ Technol Challenges 2017:519:175−181
20. Dutta A, Zisserman A: The VIA annotation software for images, audio and video. Nice, France: Proceedings of the 27th ACM International Conference on Multimedia, 2019
21. Lin T-Y, Goyal P, Girshick RB, He K, Dollár P: Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision; 2017. pp. 2980−2988
22. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. arXiv 2015, arXiv:1512.03385v1
23. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P: Microsoft COCO: common objects in context. Zurich, Switzerland: European Conference on Computer Vision, 2014. pp. 740−755
24. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. Munich, Germany: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015. pp. 234−241
25. Milletari F, Navab N, Ahmadi S-A: V-net: fully convolutional neural networks for volumetric medical image segmentation. Stanford, CA: Fourth International Conference on 3D Vision, 2016. pp. 565−571
26. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, Schmitt C, Thomas NE: A method for normalizing histology slides for quantitative analysis. Boston, MA: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. pp. 1107−1110
27. Chollet F: Xception: deep learning with depthwise separable convolutions. Honolulu, HI: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. pp. 1251−1258
28. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: ImageNet: a large-scale hierarchical image database. Miami, FL: IEEE Conference on Computer Vision and Pattern Recognition, 2009. pp. 248−255
29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M: Imagenet large scale visual recognition challenge. Int J Comput Vis 2015, 115:211−252
30. Lin M, Chen Q, Yan S: Network in network. arXiv 2013. arXiv: 1312.4400v3
31. Kingma DP, Ba J: Adam: a method for stochastic optimization. arXiv 2015, arXiv:1412.6980v9
32. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R: Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014, 15:1929−1958
33. Hotelling H: Analysis of a complex of statistical variables into principal components. J Educ Psychol Warwick York 1933, 24:417
34. Ho TK: Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE, 1995. pp. 278−282
35. Hastie T, Tibshirani R, Friedman J, Franklin J: The elements of statistical learning: data mining, inference and prediction, ed 2. Berlin, Germany: Springer Science & Business Media, 2009
36. Fawcett T: An introduction to ROC analysis. Pattern Recognit Lett 2006, 27:861−874
37. Bizzego A, Bussola N, Chierici M, Maggio V, Francescatto M, Cima L, Cristoforetti M, Jurman G, Furlanello C: Evaluating reproducibility of ai algorithms in digital pathology with dapper. PLoS Comput Biol 2019, 15:1−24
38. Redline RW, Boyd T, Campbell V, Hyde S, Kaplan C, Khong TY, Prashner HR, Waters BL: Maternal vascular underperfusion: nosology and reproducibility of placental reaction patterns. Pediatr Dev Pathol 2004, 7:237−249
39. Obuchowski NA: Receiver operating characteristic curves and their use in radiology. Radiology 2003, 229:3−8
40. Metz CE: Basic principles of ROC analysis. Semin Nucl Med 1978, 8:283−298
41. Kothari S, Phan J, Wang M: Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J Pathol Inform 2013, 4:22

42. Vizcarra JC, Gearing M, Keiser MJ, Glass JD, Dugger BN, Gutman DA: Validation of machine learning models to detect amyloid pathologies across institutions. Acta Neuropathol Commun 2020, 8:59
43. Wei JW, Suriawinata AA, Vaickus LJ, Ren B, Liu X, Lisovsky M, Tomita N, Abdollahi B, Kim AS, Snover DC, Baron JA, Barry EL, Hassanpour S: Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open 2020, 3:e203398
44. Wright D, Nicolaides KH: Aspirin delays the development of pre-eclampsia. Am J Obstet Gynecol 2019, 220:580.e1−580.e6
45. Wertaschnigg D, Reddy M, Mol BWJ, da Silva Costa F, Rolnik DL: Evidence-based prevention of preeclampsia: commonly asked questions in clinical practice. J Pregnancy 2019, 2019:2675101