Appendix 1. Glossary of Machine Learning

Supervised learning:	The machine is trained using labelled data (known outcomes)
Unsupervised learning:	The machine is trained using unlabeled data, to ultimately sort inputs based on their shared characteristics
Reinforcement learning:	The machine acts to maximize an outcome within an environment that provides feedback (trial and error approach)
Features:	Input variables of machine learning model
Algorithm:	The group of operations, used by the machine to determine interactions between the features and the output, thus creating the final “model”
Dimensionality reduction:	Machine learning preprocessing step that aims at reducing the size of features when they are so numerous that they may impair algorithm performance
Feature selection:	A dimensionality reduction process, made by reducing the number of tested features by eliminating features that are less important to the model
Feature extraction:	A dimensionality reduction process that replaces existing features with new ones that contain the most informative data
Train-test split:	Splitting a database into a train set, used to develop the model, and a test set, used to validate the model at a certain ratio. The two sets are not overlapping

K-fold cross validation:	Splitting data into a number of folds (e.g. 10), each serves as a test set once with subsequent repetition of the process and calculation of average results
Hyperparameters:	Properties of machine learning algorithm that are modifiable by the user
Cost function:	A calculation of algorithm error, which is the difference between predicted and true outputs of the model
Gradient descent:	The stepwise approach, made by the algorithm, to reach the lowest cost function (global optimum)
Convergence:	The algorithm successfully settles to the lowest cost function
Classification:	The act of predicting categorical output by machine learning
Clustering:	The act of sorting unlabeled data into groups based on their shared characteristics
Low bias:	Ability of machine learning model to predict outputs
High bias:	Poor ability of machine learning to predict outputs (indicating underfitting)
Low variance:	Reproducibility of predictions by the model
High variance:	Inconsistent predictions of the model on repetition (may indicate overfitting)
Underfitting:	Inability of the model to fit the training data, resulting in poor prediction accuracy

Overfitting:	The model strictly follows the train set, including the noise of data, so that it cannot reproduce its performance when applied to another dataset
Learning curve:	A graph that illustrates model performance over number of samples
Precision:	The possibility that a predicted positive output, made by the model, is truly positive
Recall:	The ability of the model to predict truly positive outputs
F1 score:	A score that represents a combination of precision and recall
Confusion matrix:	A 2 X 2 table of true positive, true negative, and false positive, and false negative values of the model
Jaccard index:	The size of intersection divided by the size of union of sample sets

Appendix 2. Traditional Machine Learning Algorithms

Linear regression:	A statistical model that models the relationship between the predictor variables and the outcome as a linear function.
Logistic regression:	A statistical model that assumes a linear relationship between the predictors and the log-odds of a binary event taken as Bernoulli random variable.
Decision Tree:	Data are dissected in a flow-chart-like format, so that decisions can be tested for possible outcomes. It is used for both categorical and continuous outputs e.g., triaging emergency room according to symptoms and age
Support vector machines:	Data are mapped onto a higher dimensional a space where the machine creates the best separating hyperplane that can classify the data.
Naive Bayes:	This classification algorithm assumes independency of features in manging classification problems. It is more commonly used when data are large, and the output consists of multiple classes
k- Nearest Neighbors:	It classifies new cases to the previous data that are the closest to them. The algorithm is used more in classification problems.
K-Means:	It is an unsupervised algorithm that is used for clustering problems e.g., classifying articles into groups based on the proximity of their contents
Random Forest:	A supervised algorithm that creates a forest of decision trees. These decision trees are merged to enhance accuracy and precision of the final model. The word 'random' comes from the random selection of

The authors provided this information as a supplement to their article.

	observation and features in creating these trees. It is suitable for both classification and regression problems
Gradient Boosting algorithms (e.g., XGboost):	Similar to random forest, the algorithms are designed to enforce model performance by combining multiple weak algorithms (weak learners), typically decision trees. The process is done sequentially and therefore, more weight is added to observations that yielded the worst predictions. It can be applied to both classification and regression problems

Appendix 3. Common Deep-Learning Algorithms

Artificial Neural Network:	In this classic neural network, each neuron is connected to all neurons from the previous layer. Information from inputs is proceeded in only one direction (towards output). Also known as Feedforward Neural Networks (FNN)
Convolutional Neural Networks:	Convolutional Neural Networks have been designed to improve image recognition using a function called "convolution", which helps to recognize the relation and arrangement of pixels
Recurrent Neural Networks (RNNs):	In addition to feedforward flow of data from input to output, RNNs allow flow of information from the output back to the input as a method of feedback. RNNs retain previous data and thus, they can be used in time series forecasting where previous values can be used to predict future values e.g., prediction of number of new patients based on current trend over time

Appendix 4. Examples of Artificial Intelligence Applications in Obstetrics and Gynecology

Author	Year	Sample size	Study	Type of AI	Model performance
Hoffman et al (1)	2021	20,032 deliveries	A prediction model was created to predict maternal readmission within 42 days postpartum due to complications of hypertensive disorders of pregnancy	Conventional machine learning (XG boost)	AUC was 0.85 (derivation cohort) and 0.81 (validation cohort)
Chill et al (2)	2021	98,463 deliveries	Using maternal and fetal characteristics at admission to labor, a model was designed to predict obstetric sphincter injury	Conventional machine learning (gradient boosting model)	AUC was 0.76 (95% CI 0.73–0.78)
Shazly et al (3)	2021	727 deliveries	A series of prediction models were used to predict peripartum massive blood loss, postpartum maternal admission to ICU, and prolonged hospital	Conventional machine learning (logistic regression model)	AUC of antepartum prediction models were 0.84 (massive blood loss), 0.81 (prolonged

			admission in women with placenta accreta spectrum		hospitalization), and 0.82 (admission to the ICU). Peripartum models performed at 0.86, 0.90, and 0.86, respectively
Guedalia et al (4)	2021	73,868 deliveries	Second stage variables were incorporated in a model that predicted severe adverse neonatal outcomes (umbilical cord pH levels ≤7.1 or 1-minute or 5-minute Apgar score ≤7)	Conventional machine learning (gradient boosting model)	Model AUC was 0.761 (95% CI 0.748–0.774)
Akazawa et al (5)	2021	75 patients	A model was designed to predict probability of recurrence in women with early-	Conventional machine learning (support vector	Model accuracy was 0.82. AUC was 0.53

			stage endometrial cancer	machine)	
Asali et al (6)	2021	336 patients	A model was used to predict women with intrahepatic cholestasis without bile acid measurement	Conventional machine learning (XG boost)	Model AUC was 0.9. maximum sensitivity and specificity were 86% and 75%, respectively
Raja et al (7)	2021	1,300 deliveries	A prediction model was established to predict preterm labor	Conventional machine learning (support vector machine)	Model accuracy was 90.9%
Ahn et al (8)	2021	2,949 mother–newborn pairs	A model used sonographic measures to predict actual newborn weight and weight/height	Conventional machine learning (linear regression, random forest) and deep learning (artificial neural networks)	Linear regression model was associated with the lowest mean squared error (0.077) for estimation of newborn weight

Maraci et al (9)	2020	3000 images	A model analyzed ultrasound images to estimate trans-cerebellar diameter and fetal gestational age	Deep learning (convolutional neural networks)	Trans-cerebellar diameter automated detection was 0.99 accurate. Mean manual gestational age assessment was 19.7 ± 0.9 weeks vs. 19.5 ± 2.1 weeks with automated estimation
Liu et al (10)	2020	66,706 entries	A machine learning model was structured to predict early pregnancy loss following in vitro fertilization-embryo transfer	Conventional machine learning (random forests model)	AUC of the model was 0.97
Signorini et al (11)	2020	120 fetuses	A model was based on heart rate features of antepartum fetal	Conventional machine learning	Mean classification accuracy of the

			monitoring to predict actual fetal growth restriction at birth	(random forests model)	model was 0.91 (95% CI, 0.86 - 0.96)
Venkatesh et al (12)	2020	152,279 deliveries (7,279 had postpartum hemorrhage)	A prediction model was created to predict postpartum hemorrhage (> 1000 ml) using Consortium for Safe Labor Study (2002–2008) database	Conventional machine learning (Extreme gradient boosting algorithm)	Model discriminative ability to predict postpartum hemorrhage was 0.93 (95% CI: 0.92 to 0.93)
Lipschuetz et al (13)	2020	9,888 deliveries	A machine learning model was created to predict success of trial of labor after cesarean delivery	Conventional machine learning (gradient boosting algorithm)	AUC for the model was 0.79 (95% CI, 0.78–0.81)
Guedalia et al (14)	2020	94,480 deliveries	A personalized machine learning model was created based on dynamic data acquired during the first stage of labor to predict successful	Conventional machine learning (gradient boosting algorithm)	Predictive performance increases from 0.82 (95% CI, 0.81–0.82) on admission to 0.92 (95% CI,

			vaginal deliveries		0.91–0.92) by end of the first stage
Marić et al (15)	2020	16,370 deliveries	Clinical and laboratory data from routine care in early pregnancy were used to establish a prediction model for development of preeclampsia	Conventional machine learning (elastic net algorithm)	AUC of the model was 0.79 (95% CI, 0.75–0.83), sensitivity was 45.2%, and false-positive rate was 8.1%. AUC for early-onset preeclampsia was 0.89 (95% CI, 0.84–0.95)
Naimi et al (16)	2019	18,517 pregnancies (31,948 ultrasound visits)	Using data available at delivery, population-specific fetal weight curves were generated, and a model was created to predict fetal weight over the course of pregnancy	Conventional machine learning (generalized boosted models)	Median absolute deviation from the actual weight was 88.3 (95% CI, 86.0 - 90.6)

Betts et al (17)	2019	422,509 deliveries	Maternal data throughout pregnancy and delivery and neonatal data were used to create models to predict postpartum complications, that require inpatient care	Conventional machine learning (gradient boosting algorithm)	The model predicted postpartum hypertensive disorders (AUC 0.88, 95% CI, 0.85–0.91), and surgical wound infection (AUC 0.86, 95% CI 0.84–0.87). The model performed poorly for prediction of postpartum sepsis and haemorrhage
David et al (18)	2019	559 patients	A model was created to predict likelihood of women with	Conventional machine learning	Model accuracy was 80.3% (95% CI 79.1–81.3),

			overactive bladder to respond to anticholinergic medications during the 3-month treatment period	(random forest model)	AUC was 0.77 (95% CI 0.74–0.79). AUC in women younger than 40 was 0.84 (95% CI 0.81–0.84)
Bahado-Singh et al (19)	2019	32 patients	A model used amniotic fluid metabolomics and proteomics clinical, sonographic, and demographic variables to predict perinatal outcomes in asymptomatic women with short cervices	Deep learning neural networks	AUC of the model was 0.89 (95% CI, 0.81–0.97), and 0.89 (0.79–0.99) for delivery before 34 weeks' gestation and < 28 weeks, respectively
Tsur et al (20)	2019	686 deliveries (derivation cohort), 2,584 deliveries	A machine learning model was built to predict shoulder dystocia using maternal characteristics,	Conventional machine learning (Lasso regression)	AUC of the model was 0.79 ± 0.04

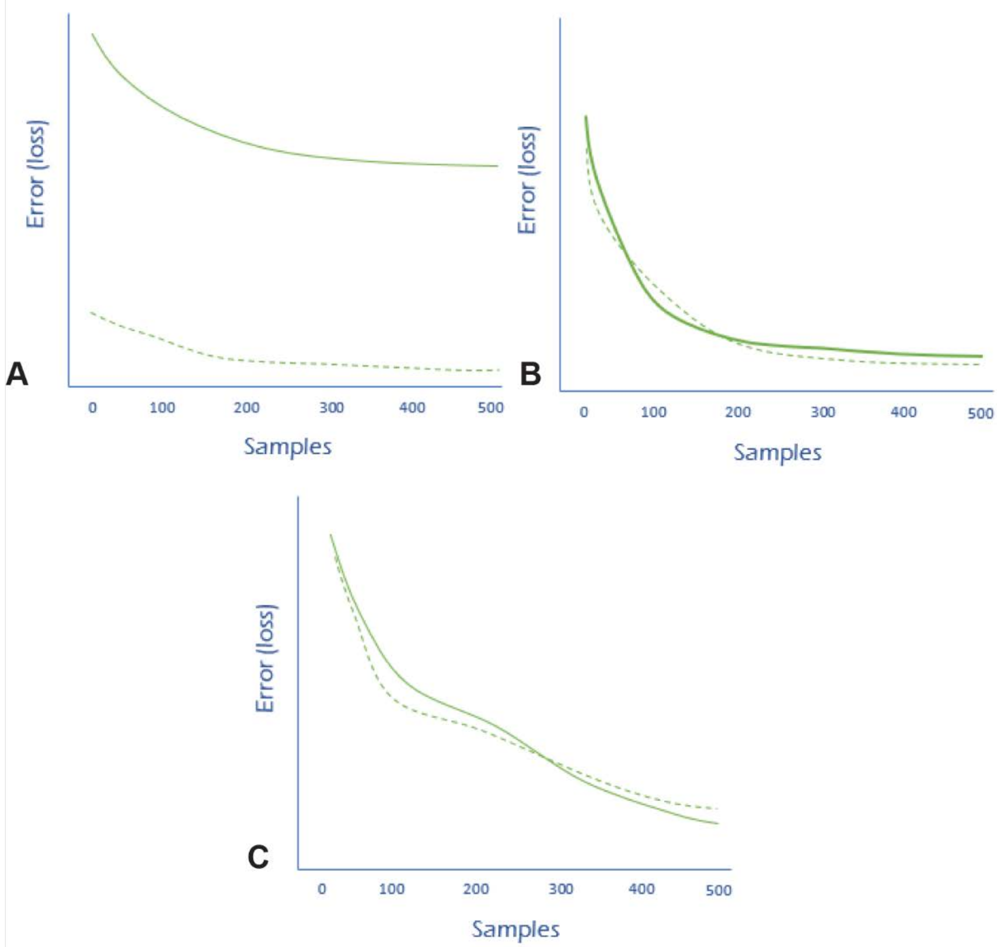
		(validation cohort)	obstetric data, and fetal biometry by ultrasound		
Idowu et al (21)	2015	262 records	A model predicted preterm contractions using electrohysterogram signals	Conventional machine learning (Random Forest)	Model AUC was 0.94. Model sensitivity and specificity were 85% and 97%, respectively
Tseng et al (22)	2014	168 patients	A model was made to determine women who were prone to cervical cancer recurrence	Conventional machine learning (C 5.0 classifier)	Correct classification rate was 96%
Bonet-Carne et al (23)	2014	900 fetal lung images, 144 neonates (validation cohort)	A model used analysis of fetal lung on ultrasound to predict probability of neonatal respiratory distress syndrome after birth	Regression models, classification trees and neural networks	Sensitivity, specificity, positive predictive and negative predictive values were 86%, 87%, 63% and 96%, respectively
Fergus et al (24)	2013	300 deliveries	Electrohysterogram features were used to	Conventional machine	The model performed at

	(38 preterm and 262 term)	predict preterm delivery	learning (polynomial classifier)	a sensitivity of 96%, a specificity of 90% (AUC 0.95) with 8% global error

1. Hoffman MK, Ma N, Roberts A, MFM AJoOaG. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. 2021;3(1):100250.
2. Chill HH, Guedalia J, Lipschuetz M, Shimonovitz T, Unger R, Shveiky D, et al. Prediction model for obstetric anal sphincter injury using machine learning. 2021:1-7.
3. Shazly SA, Hortu I, Shih J-C, Melekoglu R, Fan S, Ahmed FuA, et al. Prediction of clinical outcomes in women with placenta accreta spectrum using machine learning models: an international multicenter study. 2021:1-10.
4. Guedalia J, Sompolinsky Y, Novoselsky Persky M, Cohen SM, Kabiri D, Yagel S, et al. Prediction of severe adverse neonatal outcomes at the second stage of labour using machine learning: a retrospective cohort study. 2021.
5. Akazawa M, Hashimoto K, Noda K, Yoshida KJO, Science G. The application of machine learning for predicting recurrence in patients with early-stage endometrial cancer: a pilot study. 2020;64(3):266-73.
6. Asali A, Ravid D, Shalev H, David L, Yogev E, Yogev SS, et al. Intrahepatic cholestasis of pregnancy: machine-learning algorithm to predict elevated bile acid based on clinical and laboratory data. 2021:1-7.
7. Raja R, Mukherjee I, Sarkar BKJJoHE. A Machine Learning-Based Prediction Model for Preterm Birth in Rural India. 2021;2021.
8. Ahn KH, Lee K-S, Lee SJ, Kwon SO, Na S, Kim K, et al. Predictors of Newborn's Weight for Height: A Machine Learning Study Using Nationwide Multicenter Ultrasound Data. 2021;11(7):1280.
9. Maraci MA, Yaqub M, Craik R, Beriwal S, Self A, Von Dadelszen P, et al. Toward point-of-care ultrasound estimation of fetal gestational age from the trans-cerebellar diameter using CNN-based ultrasound image analysis. 2020;7(1):014501.
10. Liu L, Jiao Y, Li X, Ouyang Y, Shi DJCM, Biomedicine Pi. Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. 2020;196:105624.
11. Signorini MG, Pini N, Malovini A, Bellazzi R, Magenes GJCm, biomedicine pi. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. 2020;185:105015.
12. Venkatesh KK, Strauss RA, Grotegut C, Heine RP, Chescheir NC, Stringer JS, et al. Machine learning and statistical models to predict postpartum hemorrhage. Obstetrics gynecology. 2020;135(4):935.
13. Lipschuetz M, Guedalia J, Rottenstreich A, Persky MN, Cohen SM, Kabiri D, et al. Prediction of vaginal birth after cesarean deliveries using machine learning. 2020;222(6):613. e1-. e12.

14. Guedalia J, Lipschuetz M, Novoselsky-Persky M, Cohen SM, Rottenstreich A, Levin G, et al. Real-time data analysis using a machine learning model significantly improves prediction of successful vaginal deliveries. 2020;223(3):437. e1-. e15.
15. Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. 2020;2(2):100100.
16. Naimi AI, Platt RW, Larkin JCJE. Machine learning for fetal growth prediction. 2018;29(2):290.
17. Betts KS, Kisely S, Alati RJBAIJoO, Gynaecology. Predicting common maternal postpartum complications: Leveraging health administrative data and machine learning. 2019;126(6):702-9.
18. Sheyn D, Ju M, Zhang S, Anyaeche C, Hijaz A, Mangel J, et al. Development and validation of a machine learning algorithm for predicting response to anticholinergic medications for overactive bladder syndrome. 2019;134(5):946-57.
19. Bahado-Singh RO, Sonek J, McKenna D, Cool D, Aydas B, Turkoglu O, et al. Artificial intelligence and amniotic fluid multiomics: prediction of perinatal outcome in asymptomatic women with short cervix. 2019;54(1):110-8.
20. Tsur A, Batsry L, Toussia-Cohen S, Rosenstein MG, Barak O, Brezinov Y, et al. Development and validation of a machine-learning model for prediction of shoulder dystocia. 2020;56(4):588-96.
21. Idowu IO, Fergus P, Hussain A, Dobbins C, Khalaf M, Eslava RVC, et al., editors. Artificial intelligence for detecting preterm uterine activity in gynecology and obstetric care. 2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing; 2015: IEEE.
22. Tseng C-J, Lu C-J, Chang C-C, Chen G-DJNC, Applications. Application of machine learning to predict the recurrence-proneness for cervical cancer. 2014;24(6):1311-6.
23. Bonet-Carne E, Palacio M, Cobo T, Perez-Moreno A, Lopez M, Piraquive J, et al. Quantitative ultrasound texture analysis of fetal lungs to predict neonatal respiratory morbidity. 2015;45(4):427-33.
24. Fergus P, Cheung P, Hussain A, Al-Jumeily D, Dobbins C, Iram SJPo. Prediction of preterm deliveries from EHG signals using machine learning. 2013;8(10):e77154.

Appendix 5. Machine learning curve illustrating log loss over samples. *Dashed line* indicates training *train set*; *solid line* indicates validation *test set*. A. Log loss progressively decreases over training samples and it plateaus at a low loss among the train set. However, log loss is not as good among the test set. This is indicative of overfitting. B. Log loss is comparable among the train and test sets. In both curves, log loss decreases progressively to an appropriately low level and then plateau. This is indicative of good model fitting. C. Although log loss progressively decreases to a satisfactory low level in both curves, they do not plateau at the end of the curve. This indicate that training process is ongoing, and the data size may be insufficient to reach the best model.



The authors provided this information as a supplement to their article.