From Point Estimates to Distributions: GMM Pooling for MIL in Preterm Birth Prediction

Hussain Alasmawi[1], Numan Saeed[1], Soha Said[2], and Mohammad Yaqub[1]

[1] Division of Computing and Mathematical Sciences, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates
[2] Corniche Hospital, Abu Dhabi Health Services Company (SEHA), Abu Dhabi, UAE
hussain.alasmawi@mbzuai.ac.ae

Abstract. Preterm birth (PTB) prediction can enable targeted surveillance and timely intervention, yet most ultrasound-based models use a single selected transvaginal ultrasound (TVUS) frame per patient despite routine exams acquiring multiple cervical images. We formulate PTB prediction as a multiple instance learning (MIL) problem, representing each patient as a variable-sized bag of TVUS images with a single outcome label. To move beyond standard MIL aggregators that collapse a bag into a point estimate, we propose a Gaussian Mixture Model (GMM) pooling, which summarizes all images in a bag into a fixed-length representation by modeling their feature distribution. This design captures intra-patient variability. We evaluate the method on a private clinical cohort and on a public lymph node metastasis benchmark. For PTB prediction, GMM pooling improves over the instance-based model PR-AUC from 0.44 to 0.56. On the lymph node benchmark, it achieves state-of-the-art performance with 0.91 F1-score and 0.89 ROC-AUC for classification and 0.18 MAE for regression. The code is publicly available at https://github.com/HussainAlasmawi/GMM_Pooling.

Keywords: multiple instance learning · preterm birth · distribution pooling · ultrasound

1 Introduction

Preterm birth (PTB), defined as delivery before 37 weeks of gestation, remains a major contributor to neonatal morbidity and mortality worldwide and affects approximately 10% of pregnancies [23]. Importantly, three-quarters of preterm-related deaths could be prevented with timely and appropriate medical management [5]. Early risk stratification can inform preventive therapy, antenatal surveillance, referral, and neonatal readiness. In routine practice, PTB risk assessment relies on clinical factors and TVUS, most notably measuring cervical length as a biomarker of PTB. However, cervical length measurement is operator-dependent and exhibits inter- and intra-observer variability in cervical view acquisition [9]. Reflecting this uncertainty, clinical guidelines (e.g., ISUOG) recommend acquiring multiple cervical images and reporting the shortest valid measurement [7].

Recent machine learning work has explored PTB prediction from electronic health records [11,13,8,14] or ultrasound imaging [22,21,16]. More broadly, deep learning-based fetal health assessment from ultrasound imaging has been extensively studied for tasks such as congenital heart disease detection [3,19], synthetic fetal view generation [20,2], and ultrasound view classification [4,15] or clustering [1]. In contrast, comparatively less attention has been given to the prediction of preterm birth from ultrasound data. Moreover, existing PTB image-based approaches generally operate on a single selected image per patient, even though multiple images are often acquired during a routine exam. This selection can discard informative within-exam variability (e.g., view differences and acquisition conditions) and may miss subtle cues that appear only in a subset of images, limiting PTB prediction. While cervical length is the standard ultrasound biomarker for PTB risk assessment, it summarizes the examination using a single measurement, may not fully capture the information contained across multiple acquired images, and has shown limited predictive power for PTB prediction [6].

To leverage all available images under patient-level supervision, we formulate PTB prediction as a multiple instance learning (MIL) problem, where each patient corresponds to a bag of TVUS images with a single outcome label. A fundamental challenge in MIL is aggregating a variable number of instance-level features to produce a fixed-dimensional representation while ensuring permutation invariance to instance ordering. Standard pooling modules (max, mean, attention [12]) summarize a bag as a single point estimate in feature space. While effective, point-estimate pooling can create an information bottleneck by collapsing within-bag variability that may be informative for bag-level prediction [17]. This motivates distributional pooling approaches that aim to preserve richer information beyond a single summary vector.

Distribution-based MIL pooling, which models feature-wise marginal densities, has been proposed to retain richer information by representing a bag through estimated feature distributions [17]. In this work, we refer to such approaches as density-based pooling. In these formulations, the bag representation is constructed by modeling each feature dimension independently. While this strategy captures variability beyond point-estimate pooling (e.g., max or mean), it does not explicitly account for cross-dimensional dependencies within the feature space. To address this limitation, we propose a **Gaussian mixture model (GMM) pooling**, which models the within-bag feature distribution using a mixture model estimated end-to-end. The method learns (i) soft instance-to-component responsibilities and (ii) instance importance weights, and then produces a fixed-length bag embedding by evaluating the learned mixture density at a set of learnable probe vectors.

In summary, our contributions are: (i) a GMM pooling module for MIL that preserves informative within-bag feature distributions, addressing the limitations of point-estimate pooling; (ii) a differentiable distribution embedding constructed via log-density evaluations at learnable probe vectors; and (iii) a comprehensive empirical evaluation on a transvaginal cervical ultrasound cohort and a public lymph node metastasis MIL benchmark, demonstrating that MIL

improves over instance-level prediction and that GMM pooling is competitive across classification and regression tasks.

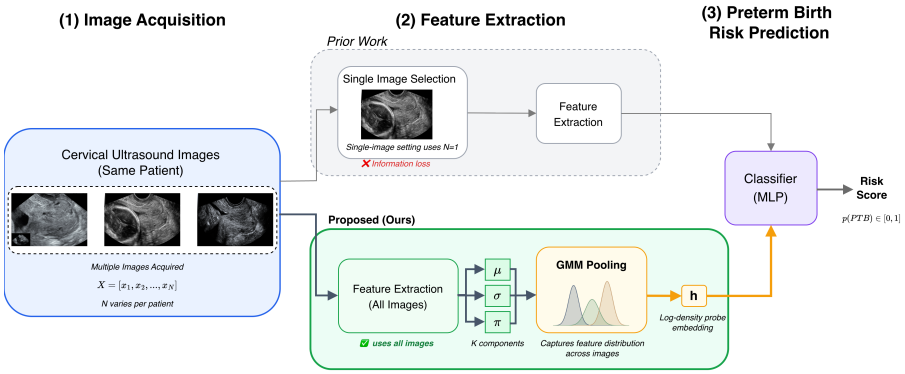


Fig. 1: Clinicians acquire multiple cervical ultrasound images per patient; prior work uses only one image, whereas we leverage all images using MIL with GMM pooling to capture feature distributions for PTB prediction. Here, μ, σ, and π denote the mean, (diagonal) standard deviation, and mixing weight of each Gaussian component, respectively; K is the number of mixture components; and $\mathbf{h}$ is the log-density probe embedding computed from the GMM and used for risk prediction.

2 Methodology

We model PTB prediction from TVUS as an MIL problem. Each patient corresponds to a bag of N cervical ultrasound images $X = \{x_1, \ldots, x_N\}$ with a single bag-level label Y indicating the delivery outcome (PTB vs. term). The overall pipeline is illustrated in Figure 1: we extract per-image features, aggregate them into a patient-level representation using GMM pooling, to predict the PTB risk.

2.1 Problem Formulation and MIL Framework

Given a bag $X \subseteq \mathcal{I}$ from the instance space I and a bag label $Y \in \mathcal{Y}$, MIL aims to learn a function f such that $\hat{Y} = f(X)$. We propose a standard three-stage architecture:

1. Feature extraction. A shared encoder $\theta_{feature} : I \to \mathbb{R}^J$ maps each image x_i to an instance feature vector $f_{x_i} \in \mathbb{R}^J$. Stacking features yields $F_X = [f_{x_1}, \ldots, f_{x_N}] \in \mathbb{R}^{J \times N}$. In the PTB experiments, the feature extractor is implemented using a U-Net [18] backbone, jointly trained for cervical canal segmentation and bag-level classification. The segmentation task serves as an

auxiliary objective that encourages the encoder to learn anatomically meaningful cervical representations, providing an inductive bias for the shared feature space used for bag-level classification. The segmentation branch predicts the cervical mask, while intermediate encoder features serve as instance representations for MIL aggregation.

2. Bag aggregation. A pooling module $\theta_{filter} : \mathbb{R}^{J\times N} \to H$ aggregates instance features into a fixed-dimensional bag representation $h \in \mathcal{H}$. The aggregation must be permutation-invariant and able to handle variable bag sizes. Figure 2 contrasts common MIL pooling modules (max/mean/attention [12] and density-based [17] pooling) with our GMM pooling.
3. Prediction. A head $\theta_{head} : \mathcal{H} \to \mathcal{Y}$ maps h to the predicted bag label $\hat{Y}$.

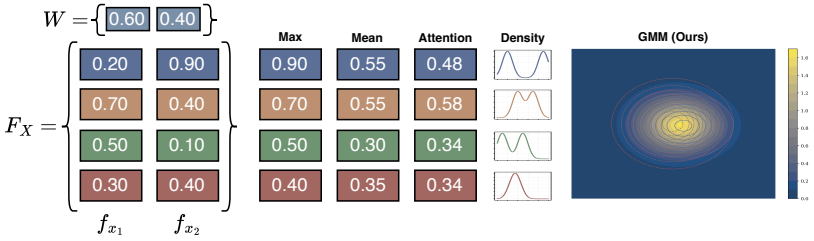


Fig. 2: Illustration of different MIL pooling strategies on a toy example. Given instance features F_X, max and mean pooling produce a single point-estimate representation. Attention pooling computes a weighted point estimate using learned instance importance weights W, but still collapses the bag into a single point-estimate representation. Density pooling models each feature dimension independently, whereas GMM pooling preserves the joint distributional structure across dimensions.

2.2 Gaussian Mixture Model Pooling

Standard pooling modules θ_{filter} reduce a bag to a point estimate in feature space, which can collapse informative within-bag variability. We instead estimate a GMM over instance features and embed the resulting distribution into a fixed-length vector. Algorithm 1 summarizes the full method.

Given $F_X \in \mathbb{R}^{J\times N}$, we construct a GMM with M components. Each component is parameterized by a mixing coefficient π_k, mean $\mu_k \in \mathbb{R}^J$, and diagonal variance $\sigma_k^2 \in \mathbb{R}^J$. We adopt a diagonal covariance parameterization for computational stability and scalability, as optimizing full covariance matrices within an end-to-end MIL framework would substantially increase the number of parameters and complicate training. We did not explore full or low-rank in this work.

Nevertheless, soft instance-to-component assignments and mixture modeling preserve heterogeneous intra-bag structure and relationships among instances. All parameters are estimated from the bag features in a differentiable manner.

Soft instance-to-component assignments. We obtain soft assignments of instances to mixture components using a responsibility network $\mathbf{f_r} : \mathbb{R}^J \to \mathbb{R}^{\mathcal{M}}$. For each instance feature f_{x_i}, the network produces a probability distribution over components:

$$r_{i,k} = \mathrm{softmax}_k\big(\mathbf{f_r}(f_{x_i})\big), \tag{1}$$

where $r_{i,k}$ denotes the responsibility of instance i for component k, satisfying $\sum_{k=1}^{\mathcal{M}} r_{i,k} = 1$. Thus, each instance contributes softly to multiple components. In practice, $\mathbf{f_r}$ is implemented as a lightweight two-hidden-layer multilayer perceptron MLP with ReLU activations, applied independently to each instance feature.

Instance importance weighting. In addition to instance responsibilities, we estimate the relative importance of instances within a bag using an instance importance network $\mathbf{f_w} : \mathbb{R}^J \to \mathbb{R}$. Instance weights are computed via a softmax over instances:

$$w_i = \mathrm{softmax}_i\big(\mathbf{f_w}(f_{x_i})\big), \tag{2}$$

ensuring $\sum_{i=1}^{N} w_i = 1$. Instance importance weighting adjusts how strongly each instance contributes to the estimated distribution, without enforcing hard instance selection. In practice, $\mathbf{f_w}$ is a lightweight single-hidden-layer MLP applied independently to each instance feature.

GMM parameter estimation. The final attention of instance i to component k is obtained by combining instance importance and component responsibility:

$$a_{i,k} = w_i \cdot r_{i,k}. \tag{3}$$

Using the attention weights $a_{i,k}$, the GMM parameters are estimated as

$$\pi_k = \sum_{i=1}^{N} a_{i,k}, \qquad \mu_k = \frac{\sum_{i=1}^{N} a_{i,k} f_{x_i}}{\sum_{i=1}^{N} a_{i,k}}, \qquad \sigma_k^2 = \frac{\sum_{i=1}^{N} a_{i,k}\,(f_{x_i} - \mu_k)^2}{\sum_{i=1}^{N} a_{i,k}}. \tag{4}$$

The mixing coefficients π_k are subsequently normalized across components to form a valid probability distribution. The square is applied element-wise, and all parameters are learned end-to-end.

Distribution embedding via learned probes. To produce a fixed-dimensional bag representation, we evaluate the log-density of the estimated GMM on a set of $\mathcal{P}$ learnable probe vectors $\{\mathbf{z}_p \in \mathbb{R}^J\}_{p=1}^{\mathcal{P}}$. For each probe, we compute

$$h_p = \log\left(\sum_{k=1}^{\mathcal{M}} \pi_k \mathcal{N}(z_p \mid \mu_k, \sigma_k^2)\right) / (J \times T). \tag{5}$$

where division by $J \times T$ controls the embedding scale. The resulting vector $\mathbf{h} = [h_1, \ldots, h_{\mathcal{P}}] \in \mathbb{R}^{\mathcal{P}}$ serves as the distribution-based bag embedding as the output of the pooling module θ_{filter}.

Algorithm 1 GMM Pooling (θ_{filter})

Input: Instance features $F_X = [f_{x_1}, \ldots, f_{x_N}] \in \mathbb{R}^{J \times N}$; number of components $\mathcal{M}$; number of probes $\mathcal{P}$; temperature T; features dimension J;
Output: Distribution embedding $\mathbf{h} \in \mathbb{R}^{\mathcal{P}}$, GMM parameters (π, μ, σ^2)

Soft responsibilities and instance importance:
for $i = 1$ **to** N **do**
  $w_i \leftarrow \text{softmax}_i\big(\mathbf{f_w}(f_{x_i})\big)\,, \quad r_{i,k} \leftarrow \text{softmax}_k\big(\mathbf{f_r}(f_{x_i})\big)\,, \quad k = 1, \ldots, \mathcal{M}$
end for
Combine importance and responsibilities:
for $i = 1$ **to** N **do**
  for $k = 1$ **to** $\mathcal{M}$ **do**
    $a_{i,k} \leftarrow w_i \cdot r_{i,k}$
  end for
end for
Mixing coefficients and attention-weighted moments:
for $k = 1$ **to** $\mathcal{M}$ **do**
  $\pi_k \leftarrow \sum_{i=1}^{N} a_{i,k} \quad \mu_k \leftarrow \frac{\sum_{i=1}^{N} a_{i,k} f_{x_i}}{\sum_{i=1}^{N} a_{i,k}} \quad \sigma_k^2 \leftarrow \frac{\sum_{i=1}^{N} a_{i,k} (f_{x_i} - \mu_k)^2}{\sum_{i=1}^{N} a_{i,k}}$
end for
$\pi_k \leftarrow \pi_k / \left(\sum_{k'=1}^{\mathcal{M}} \pi_{k'}\right)$
Distribution embedding via learned probes:
for $p = 1$ **to** $\mathcal{P}$ **do**
  $h_p \leftarrow \log\Big(\sum_{k=1}^{\mathcal{M}} \pi_k \mathcal{N}\big(z_p \mid \mu_k, \sigma_k^2\big)\Big) / (J \times T)$
end for
return $\mathbf{h}$, (π, μ, σ^2)

3 Experimental Setup

3.1 Dataset

We evaluate our method on a private transvaginal ultrasound (TVUS) dataset of cervical canal segmentations provided by clinicians. This study was approved by the Institutional Review Board of Corniche hospital (Protocol CH13032401). The cohort consists of 182 pregnant patients, including 44 preterm birth cases. Notably, all patients were prospectively classified as high-risk for PTB by the treating clinicians according to institutional clinical practice. This reflects the real clinical use of transvaginal ultrasound, where screening is not routinely performed in the general population but rather in women identified as being at high risk. This also makes the prediction task particularly challenging, as the elevated baseline risk reduces class separability between PTB and term cases. Each patient has between 1 and 43 ultrasound images (10±7 on average), acquired across one or multiple clinical visits. Most scans were acquired during the second trimester or near its boundaries (late first trimester and early third trimester).

We additionally evaluate on a public lymph node histopathology MIL dataset [17] for (i) binary classification (normal vs. metastasis) and (ii) regression of metastatic pixel percentage. The dataset consists of 933 training, 668 validation, and 736 test samples, of which 60% are positive.

3.2 Configurations, Implementation and Baselines

We use a unified training protocol across pooling methods to ensure fair comparison. We experiment with five different pooling strategies, i.e., max, mean, attention [12], density [17], and ours (GMM), and investigate the prediction performance. For PTB prediction, we use 5-fold cross-validation with patient-level splits and three random seeds per fold (15 runs). Models are trained for 80 epochs (batch size 3, learning rate 1×10^{-4}.). We use a U-Net [18] as in [21], training segmentation and classification jointly; the classification head aggregates images from the same patient to predict outcome and the segmentation mask predicts the cervical canal. For the lymph node dataset, we follow [17] using a ResNet18 [10] encoder with bag and batch size 32, repeating each experiment with three random initializations.

The number of mixture components $\mathcal{M}$, learnable probes $\mathcal{P}$, and temperature parameter T are treated as dataset-specific hyperparameters since they are empirically optimized. For the preterm birth dataset, we set $(\mathcal{M}, \mathcal{P}, T) = (4, 96, 30)$, and for the lymph node metastasis dataset, $(\mathcal{M}, \mathcal{P}, T) = (10, 5, 30)$.

4 Results

Table 1a compares different pooling strategies on the PTB dataset. The evaluation metrics are area under the precision–recall curve (PR-AUC) and area under the receiver operating characteristic curve (ROC-AUC) for classification, and the Dice score for segmentation. Given the class imbalance, PR-AUC is considered the primary metric.

Among MIL methods, max pooling, attention pooling, and GMM pooling achieve the highest PR-AUC scores, with differences within 0.01. To determine whether these differences are statistically meaningful, we conducted a paired t-test on PR-AUC across 15 paired experiments. The comparison between GMM and max pooling showed no significant difference, with a mean difference of -0.005 (95% CI: $[-0.070, 0.059]$, $p = 0.873$), indicating comparable performance. Notably, GMM pooling exhibits lower standard deviation across runs, suggesting improved stability.

Table 1b summarizes results on the lymph node benchmark. In contrast to the preterm dataset, max pooling shows reduced performance on this task, whereas GMM pooling achieves the best classification results (0.91 ± 0.01 F1, 0.89 ± 0.01 ROC-AUC) and competitive regression performance (0.18 ± 0.02 MAE), within 0.01 of the top regression method. These results suggest GMM pooling provides more consistent performance across tasks, achieving state-of-the-art results on the lymph node benchmark with lower variability.

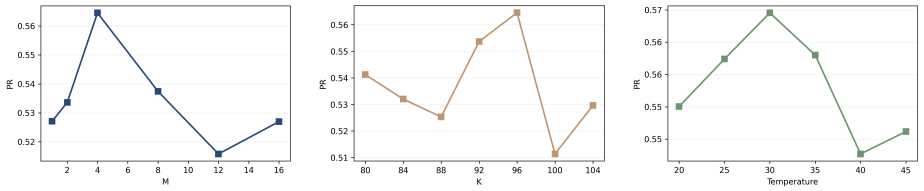
Table 1: Performance comparison on (left) preterm birth dataset and (right) lymph node dataset. Results are reported as mean ± std. Best results are in **bold**; second-best are underlined. Segmentation performance is reported for completeness.

(a) Preterm birth dataset.

Pooling Strategy	Classification PR-AUC ↑	Classification ROC-AUC ↑	Segmentation Dice ↑
instance-based	0.44±0.01	0.64±0.02	0.82
max	**0.57±0.05**	**0.71±0.03**	0.81
mean	0.54±0.01	0.68±0.04	0.82
attention	0.56±0.02	0.67±0.03	**0.83**
density	0.54±0.03	0.68±0.02	0.82
GMM (ours)	0.56±0.03	0.69±0.03	0.82

(b) Lymph node dataset.

Pooling Strategy	Classification F1 ↑	Classification ROC-AUC ↑	Regression MAE ↓
max	0.84±0.03	0.83±0.05	0.24±0.01
mean	0.86±0.04	0.86±0.03	0.23±0.02
attention	0.87±0.02	0.88±0.02	0.18±0.04
density	0.87±0.01	0.87±0.01	**0.17±0.01**
GMM (ours)	**0.91±0.01**	**0.89±0.01**	0.18±0.02



(a) Effect of M (b) Effect of P (c) Effect of T

Fig. 3: Ablations on GMM hyperparameter sensitivity on PTB (PR-AUC vs. $\mathcal{M}, \mathcal{P}, T$).

Figure 3 shows ablations on the hyperparameter sensitivity on the PTB dataset. Performance peaks at $\mathcal{M} = 4$ and decreases for larger values. The probe count $\mathcal{P}$ achieves its best result at $\mathcal{P} = 96$, while the temperature parameter T performs optimally at $T = 30$, with higher values reducing PR-AUC.

5 Discussion

Standard MIL pooling methods (max, mean, attention) summarize a variable-sized set of instance features into a single point estimate. While effective, this compression may discard informative intra-bag variability. In the context of transvaginal ultrasound, variability across images reflects differences in acquisition angle, anatomical visibility, and subtle cervical patterns. Modeling this variability explicitly can therefore be beneficial for patient-level risk prediction.

GMM pooling represents each bag as a learned mixture distribution rather than a single summary vector. On the PTB cohort, MIL substantially improves over instance-level training, confirming the benefit of leveraging all available images per patient. Although GMM pooling performs comparably to max and attention pooling in terms of PR-AUC, it exhibits lower variability across runs, suggesting improved stability. On the lymph node benchmark, GMM pooling

achieves the strongest classification performance and competitive regression accuracy, indicating that distributional modeling generalizes across tasks.

The main trade-off of the approach is the introduction of distribution-specific hyperparameters (e.g., number of mixture components and probes), which require dataset-dependent tuning. Nevertheless, results across two distinct medical imaging tasks suggest that modeling feature distributions provides a robust and flexible alternative to point-estimate pooling.

6 Conclusion

We formulate PTB prediction from TVUS as a multiple instance learning problem to leverage all cervical images acquired per patient. We introduce GMM pooling to model within-bag feature distributions. On a private high-risk PTB cohort, MIL improves substantially over instance-based training, and GMM pooling achieves competitive PR-AUC with lower variability across runs. On a public lymph node metastasis benchmark, GMM pooling attains state-of-the-art classification performance and strong regression results, suggesting good cross-task generality. Future work will explore adaptive strategies for selecting mixture complexity, conduct more comprehensive component-wise ablation studies to better understand the contribution of individual design choices, and evaluate the approach on additional clinical cohorts to further assess generalizability.

Disclosure of Interests. The authors have no competing interests in the paper as required by the publisher.

References

1. Alasmawi, H., Bricker, L., Yaqub, M.: Fusc: fetal ultrasound semantic clustering of second-trimester scans using deep self-supervised learning. Ultrasound in Medicine & Biology **50**(5), 703–711 (2024)
2. Arjemandi, M., Hassan, S., Wang, H., Valappil, S., Yaqub, M.: Difusal: Diffusion-based fetal ultrasound synthesis with active learning. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 130–139. Springer (2025)
3. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., Moon-Grady, A.J.: An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nature medicine **27**(5), 882–891 (2021)
4. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Real-time detection and localisation of fetal standard scan planes in 2d freehand ultrasound. arXiv preprint arXiv:1612.05601 (2016)
5. Behrman, R.E., Butler, A.S. (eds.): Preterm Birth: Causes, Consequences, and Prevention. National Academies Press, Washington, DC (2007)
6. Conde-Agudelo, A., Romero, R.: Predictive accuracy of changes in transvaginal sonographic cervical length over time for preterm birth: a systematic review and metaanalysis. American journal of obstetrics and gynecology **213**(6), 789–801 (2015)

7. Coutinho, C.M., Sotiriadis, A., Odibo, A., Khalil, A., D'Antonio, F., Feltovich, H., Salomon, L.J., Sheehan, P., Napolitano, R., Berghella, V., da Silva Costa, F.: ISUOG Practice Guidelines: Role of ultrasound in the prediction of spontaneous preterm birth. Ultrasound in Obstetrics & Gynecology **60**(3), 435–456 (2022). https://doi.org/10.1002/uog.26020, https://www.isuog.org/static/d88e5dff-ced3-43ee-aa2229c2679b9484/ISUOG-Practice-Guidelines-ultrasound-in-preterm-birth.pdf
8. Gao, C., Osmundson, S., Edwards, D.R.V., Jackson, G.P., Malin, B.A., Chen, Y.: Deep learning predicts extreme preterm birth from electronic health records. Journal of biomedical informatics **100**, 103334 (2019)
9. Gravett, M.G., Menon, R., Tribe, R.M., Hezelgrave, N.L., Kacerovsky, M., Soma-Pillay, P., Jacobsson, B., McElrath, T.F.: Assessment of current biomarkers and interventions to identify and treat women at risk of preterm birth. Frontiers in Medicine **11**, 1414428 (2024). https://doi.org/10.3389/fmed.2024.1414428, https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1414428/full
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Huang, C., Long, X., van der Ven, M., Kaptein, M., Oei, S.G., van den Heuvel, E.: Predicting preterm birth using electronic medical records from multiple prenatal visits. BMC Pregnancy and Childbirth **24**(1), 843 (2024)
12. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
13. Kloska, A., Harmoza, A., Kloska, S.M., Marciniak, T., Sadowska-Krawczenko, I.: Predicting preterm birth using machine learning methods. Scientific Reports **15**(1), 5683 (2025)
14. Koivu, A., Sairanen, M.: Predicting risk of stillbirth and preterm pregnancies with machine learning. Health information science and systems **8**(1), 14 (2020)
15. Maani, F., Saeed, N., Saleem, T., Farooq, Z., Alasmawi, H., Diehl, W., Mohammad, A., Waring, G., Valappi, S., Bricker, L., et al.: Fetalclip: A visual-language foundation model for fetal ultrasound image analysis. arXiv preprint arXiv:2502.14807 (2025)
16. Ohtaka, A., Akazawa, M., Hashimoto, K.: Deep learning algorithm for predicting preterm birth in the case of threatened preterm labor admissions using transvaginal ultrasound. Journal of Medical Ultrasonics **51**(2), 323–330 (2024)
17. Oner, M.U., Kye-Jet, J.M.S., Lee, H.K., Sung, W.K.: Distribution based mil pooling filters: Experiments on a lymph node metastases dataset. Medical Image Analysis **87**, 102813 (2023)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
19. Taratynova, D., Almsouti, A., Kalmakhanbet, B., Saeed, N., Yaqub, M.: Tpa: Temporal prompt alignment for fetal congenital heart defect classification. arXiv preprint arXiv:2508.15298 (2025)
20. Tian, Y., Ucurum, E., Han, X., Young, R., Chatwin, C., Birch, P.: Enhancing fetal plane classification accuracy with data augmentation using diffusion models. IET Image Processing **19**(1), e70151 (2025)
21. Włodarczyk, T., Płotka, S., Rokita, P., Sochacki-Wójcicka, N., Wójcicki, J., Lipa, M., Trzciński, T.: Spontaneous preterm birth prediction using convolutional neural

networks. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 274–283. Springer (2020)
22. Włodarczyk, T., Płotka, S., Trzciński, T., Rokita, P., Sochacki-Wójcicka, N., Lipa, M., Wójcicki, J.: Estimation of preterm birth markers with u-net segmentation network. In: International Workshop on Preterm, Perinatal and Paediatric Image Analysis. pp. 95–103. Springer (2019)
23. World Health Organization: Preterm birth. `https://www.who.int/news-room/fact-sheets/detail/preterm-birth/` (2023), accessed: 2026-02-11