FetalEDL: OOD-Aware Evidential Deep Learning for Reliable Fetal Ultrasound Plane Classification

Srivibha Parthasarathy[1], Lakshminarayanan M[2], Keerthi Ram[3], Shyam Ayyasamy[2], Suresh Seshadri[4], Manojkumar Lakshmanan[2], and Mohanasankar Sivaprakasam[1]

[1] Department of Electrical Engineering, Indian Institute of Technology Madras, India
[2] Healthcare Technology Innovation Centre (HTIC), IITM, India
[3] Sudha Gopalakrishnan Brain Centre, IITM, India
[4] MediScan Systems, Chennai, India
srivibha@htic.iitm.ac.in

Abstract. Standard plane classification in fetal ultrasound (US) is critical for prenatal diagnosis. However, for the model to be adopted in real-world clinical scenarios, a key challenge is to design a solution that is correctly uncertain about hard in-distribution (ID) samples and even acquisition-based Out-of-Distribution (OOD) data. Existing uncertainty quantification (UQ) methods such as Monte Carlo (MC) Dropout and Deep Ensemble require multiple passes and higher compute, limiting their suitability for real-world deployment. We propose **FetalEDL**, a ConvNeXt-based evidential deep learning (EDL) framework for classification of 20 standard second-trimester fetal planes, trained on B-mode fetal US images and exposed to clinically realistic acquisition-mode OOD samples, including Colour Doppler, Pulsed-wave Doppler, and non-fetal US images. EDL decomposes predictive uncertainty into vacuity and dissonance, which we analyse as clinically interpretable uncertainty signals. We incorporate feature-similarity guided OOD exposure sampling and vacuity supervision to address identified EDL failure modes. We also implement a vacuity-based selective classification strategy which allows us to defer hard ID images for sonographer review instead of spurious misprediction on those images. Abstaining on 2.2% of inputs catches 58.1% of gross inter-anatomical errors and achieves a 97.85% accuracy on retained predictions (AUROC 0.991, AUPRC 0.866). The abstained images were also clinically reviewed and found to be of poor quality, showing shadowing and acquisition artifacts. Trained on a privately acquired dataset of 1435 patients and evaluated on a separate cohort of 981 patients across 20 planes, FetalEDL demonstrates strong clinical reliability for both plane classification and OOD detection in realistic clinical conditions.

Keywords: Fetal Ultrasound · Standard Plane Classification · Uncertainty Quantification · Evidential Deep Learning · Uncertainty Decomposition · Vacuity

1 Introduction

The last decade has seen rapid growth of artificial intelligence (AI) in obstetric US, enabling automated detection of standard planes, biometric measurements, and anomaly screening [2, 3, 17]. During routine obstetric examinations, the sonographers acquire B-mode images to visualise fetal anatomy, Colour Doppler images to visualise blood flow and direction, and Pulsed-wave Doppler images to capture flow velocity. Additionally, there is also the potential risk of accidentally including non-fetal (maternal) images into the pipeline. We categorise fetal Colour and Pulsed-wave Doppler images (based on acquisition-mode), and non-fetal US images as OOD data. When exposed to these OOD data, traditional deep learning classifiers utilising Softmax layer exhibit overconfidence in incorrect predictions [6, 7, 12]. As Softmax forces outputs to normalize into a probability distribution, they cannot show ignorance towards OOD data. For example, a renal artery Doppler image could be mapped to the 'femur' class due to the visual similarity between the main axis of the artery and the femoral shaft. This highlights the need for interpretable UQ in fetal US AI systems [9].

Conventional UQ methods such as MC Dropout [5] and Deep Ensemble [10] require multiple stochastic forward passes or multiple models, limiting their application in the real-world. Therefore, we adopt evidential deep learning (EDL) [15] which estimates class probabilities and uncertainty simultaneously in a single forward pass [9, 13]. In EDL, the standard softmax layer is replaced with an activation function that outputs non-negative evidence for each class. These evidence values parameterize a Dirichlet distribution from which predictive class probabilities and uncertainty are computed. Predictive uncertainty is inversely proportional to the total accumulated evidence and can be decomposed into vacuity and dissonance for better interpretability. Vacuity reflects the absence of accumulated evidence and is associated with epistemic uncertainty arising from unfamiliar data like transitional views or OOD data, while dissonance measures conflicting evidence among competing classes as seen in noisy or ambiguous images. Images with severe shadowing that occludes key anatomical structures also yield high vacuity as lack of discriminative features prevent sufficient evidence accumulation. Therefore, the model behaves conservatively and abstains from making an incorrect prediction based on the partially visible structures alone. Consequently, EDL improves the reliability of fetal US AI systems by identifying highly uncertain predictions that can be deferred for sonographer review.

We present a reliable fetal US classification framework that is both anatomically aware and uncertainty-aware. Our main contributions are as follows:

1. We introduce **FetalEDL**, an OOD-aware EDL framework for 20-class fetal US plane classification, incorporating feature-similarity guided near-OOD exposure sampling and vacuity supervision to address identified EDL failure modes.
2. We analyse the decomposition of predictive uncertainty into vacuity and dissonance, and observe distinct uncertainty patterns across three clinically meaningful failure scenarios, namely, intra-class acquisition variance, visual class similarity, and image quality artifacts.

3. We demonstrate a vacuity-based selective classification mechanism that abstains on just 2.2% of inputs to catch 58.1% of gross inter-anatomical errors and achieving 97.85% accuracy on retained predictions.

2 Methodology

2.1 Dataset

FetalEDL was developed and evaluated using a privately acquired retrospective B-mode fetal US dataset collected from 2416 patients using 3 scanners (GE Voluson E6, Mindray, SonoScape), at a tertiary scan center. The dataset contains 20 International Society of Ultrasound in Obstetrics and Gynecology (ISUOG)-standard fetal US planes [4], excluding 3 Vessel Cord. Upper and lower limb classes were further divided into individual bones. In addition to this, the fetal Colour Doppler and Pulsed-wave Doppler images acquired from the same retrospective collection were used as OOD data (based on acquisition mode) for training and testing along with non-fetal US data obtained from open source datasets [1, 16]. To prevent leakage of patient-level information, we employed a split of patient-wise train-validation-test, resulting in 1078, 357, and 981 patients (17438, 5782, 11319 images, respectively) across the three subsets.

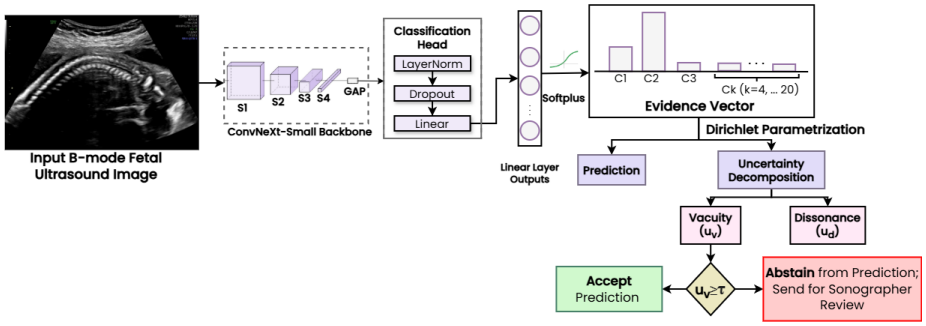

Fig. 1: Overview of the FetalEDL framework. A ConvNeXt-Small backbone with a Softplus activation produces a per-class evidence vector which parameterizes a Dirichlet distribution, enabling prediction and UQ. Predictive uncertainty is decomposed into vacuity (u_v) and dissonance (u_d), where u_v is associated with OOD data, while u_d is associated with anatomically ambiguous images. During selective classification, images with $u_v \geq \tau$ are deferred for sonographer review.

2.2 EDL Formulation

Given an input B-mode US image $\mathbf{x} \in \mathbb{R}^{H \times W \times 1}$, a CNN backbone f_θ generates logits $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^K$ (for K classes), from which a classification head

with a Softplus activation produces the evidence vector $\mathbf{e}$. Softplus activation ensures that the per-class evidence e_k always remains positive for all classes $k \in \{1, ..., K\}$. This is required for a valid Dirichlet parameterisation. The magnitude of e_k carries the measure of evidence accumulated for class k. Instead of placing a point estimate over class probabilities as is the case with Softmax activations, EDL treats the class probability vector π which lies in a (K-1)-dimensional probability simplex, Δ^{K-1}, as a random variable following a Dirichlet distribution.

Here, Dirichlet is parameterised by a concentration parameter α and Dirichlet strength is S. α is derived from the network's output e such that $\alpha_k = e_k + 1 \geq 1$. Therefore, even if the model cannot collect evidence for any class, every class gets a baseline slot. The Dirichlet strength S is measured as $S = \sum_{k=1}^{K} \alpha_k$. A large S results in a sharper Dirichlet distribution and, therefore, high confidence in predictions. Conversely, a smaller S indicates insufficient evidence and results in high epistemic uncertainty. This property allows EDL to naturally distinguish between confident and uncertain predictions. The probability of the expected class is $\hat{p}_k = \mathbb{E}[\pi_k] = \frac{\alpha_k}{S}$ and the final prediction is $\hat{y} = \arg\max_k \hat{p}_k$.

Under the Subjective Logic Formulation [8], the Dirichlet parameters can be interpreted epistemically as measures of evidence, belief, and uncertainty. The total uncertainty can be decomposed as shown in Eq. 1 where b_k and u_v denote the belief mass of a class k and vacuity, respectively. This uncertainty decomposition is explained further in 2.3.

$$\sum_{k=1}^{K} b_k + u_v = 1, \quad \text{where } b_k = \frac{e_k}{S} = \frac{\alpha_k - 1}{S} \tag{1}$$

The total training loss in FetalEDL combines an ID EDL loss, $\mathcal{L}_{EDL}$, an OOD regularization term, $\mathcal{L}_{OOD}$, and a vacuity regularization term, $\mathcal{L}_{vac}$, as shown in Eq. 2. Here, β and γ control the relative contribution of the OOD regularization term and vacuity regularization term respectively. In our experiments, β was set to 0.5 to provide a good balance between fitting the ID evidence objective and suppressing evidence on the OOD samples, while γ was set to a smaller value of 0.05 to penalise high vacuity correct predictions without over-regularizing it in the evidence distribution.

$$\mathcal{L} = \mathcal{L}_{EDL} + \beta \cdot \mathcal{L}_{OOD} + \gamma \cdot \mathcal{L}_{vac}, \quad \beta = 0.5,\ \gamma = 0.05 \tag{2}$$

The $\mathcal{L}_{EDL}$, as shown in Eq. 3, composed of an evidence fitting term and a KL regularization term, encourages the model to accumulate evidence for the correct class while penalizing the formation of unnecessary evidence.

$$\mathcal{L}_{EDL} = \sum_{k=1}^{K} y_k(\psi(S) - \psi(\alpha_k)) + \lambda(\text{epoch}) \cdot \text{KL}[\text{Dir}(\pi|\tilde{\alpha}) \parallel \text{Dir}(\pi|1)],$$
$$\text{where } \lambda(\text{epoch}) = \min\left(0.5, \frac{\text{epoch}}{\text{total\_epochs}}\right) \text{ and } \tilde{\alpha} = y_k + (1 - y_k)\alpha_k. \tag{3}$$

where, y_k is a one-hot label (1 for true class, 0 for the rest), α_k and S are Dirichlet parameters, $\mathrm{Dir}(\pi|\tilde{\alpha})$ is the Dirichlet distribution over π with α, modeling the uncertainty over class probabilities rather than a point estimate, $\psi(\cdot)$ is the digamma function, KL[·] is the KL Divergence, λ(epoch) is the annealing coefficient that prevents the KL term from dominating too early.

The $\mathcal{L}_{OOD}$, on the other hand, discourages the accumulation of any evidence for OOD images and therefore, it pushes the Dirichlet towards the maximum uncertainty distribution [18].

$$\mathcal{L}_{OOD} = \mathrm{KL}[\mathrm{Dir}(\pi|\alpha^{OOD}) \parallel \mathrm{Dir}(\pi|1)] \tag{4}$$

2.3 Uncertainty Decomposition

Vacuity (u_v) captures the overall lack of evidence as shown in Eq. 5. As the total evidence S approaches infinity (∞), the model becomes increasingly confident, causing u_v to approach 0. **Dissonance (u_d)** captures conflicting evidence across classes. It is calculated as the belief-mass weighted average conflict as shown in Eq. 6. The inner sum is normalized by $b_j (j \neq k)$ to avoid sparse evidence from artificially inflating u_d.

$$u_v = \frac{K}{S} = \frac{K}{\sum_{k=1}^{K} \alpha_k} \tag{5}$$

$$u_d = \sum_{k=1}^{K} b_k \cdot \frac{\sum_{j \neq k} b_j \cdot Bal(b_k, b_j)}{\sum_{j \neq k} b_j}, \quad Bal(b_k, b_j) = 1 - \frac{|b_k - b_j|}{b_k + b_j} \tag{6}$$

Here, $Bal(b_k, b_j)$ is the pairwise balance between classes k and j that measures how evenly the evidence collected for each class is split.

2.4 Distance-aware Near-OOD Sampling

As OOD samples exhibit varying degrees of distribution shift, we quantify each OOD sample's proximity to the ID manifold using a continuous feature space similarity score. For each OOD training sample, we extract the CNN backbone features from the classifier trained on the ID plane classification task and compute its mean cosine similarity to its nearest neighbours in the ID training feature set. This similarity is then converted to a normalized distance score in [0,1] where Near-OOD samples (closer to the ID manifold and represent challenging ambiguous cases) receive lower scores and Far-OOD samples (substantially different from the ID data) receive higher scores. During training, a WeightedRandomSampler is weighted using the inverse of this score so that Near-OOD samples are sampled more frequently, exposing the model to harder boundary cases and improving uncertainty estimation where ID/OOD distributions overlap.

2.5 Vacuity Regularization

Preliminary experiments showed that the standard OOD-aware EDL behaves very conservatively, occasionally giving high vacuity correct predictions. This results in unnecessary abstentions. To mitigate this effect, we implement a vacuity regularization term as shown in Eq. 7 that penalizes correct predictions with no evidence (high vacuity). We define a vacuity threshold of τ=0.95 for this. Here, $\mathcal{C}_\tau$ refers to correctly classified samples exceeding this threshold, and u_v denotes vacuity of the sample.

$$\mathcal{L}_{vac} = \frac{1}{|\mathcal{C}_\tau|} \sum_{i \in \mathcal{C}_\tau} u_{v_i}, \quad \text{where } \mathcal{C}_\tau = \{i : \hat{y}_i = y_i,\ u_{v_i} > \tau\},\ \tau = 0.95 \tag{7}$$

2.6 Selective Classification via Vacuity Thresholding

During inference, the model may encounter images that are diagnostically uninformative owing to OOD acquisition modes, image degradation or transitional positions of the probe between standard views. To address this issue, we implement a selective classification strategy in which the model abstains from predicting when the evidence accumulated is insufficient. An image is deferred for sonographer review if its vacuity exceeds a fixed threshold τ. We use vacuity rather than dissonance as the deferral criterion as the two effectively capture different failure modes. While vacuity directly reflects the total absence of evidence and rises for both OOD inputs and diagnostically uninformative acquisitions, dissonance reflects conflicting evidence among classes. Therefore, dissonance is more suitable for identifying ambiguity within the ID label space, while vacuity is well-suited for a general purpose deferral mechanism. The threshold was set based on the vacuity distribution of the validation set to maximize selective accuracy while maintaining high coverage. By abstaining only when insufficient evidence is available, FetalEDL reduces the chance of propagating unreliable predictions into other downstream tasks.

2.7 Training Details

FetalEDL was trained using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 32, with weighted sampling to address class imbalance. The baseline classifier was trained on ID data with a ConvNeXt-Small [11] backbone, initialized with ImageNet-pretrained weights and selected for its strong hierarchical feature extraction and efficiency in dense visual recognition tasks. Ultrasound-specific augmentations, including gain variation, speckle noise simulation and simulated shadowing, were incorporated during training to improve robustness of the classifier to imaging variability. For the ID/OOD setting, samples were categorised based on the acquisition mode and whether the image represented fetal anatomy. Specifically, fetal B-mode images were identified as ID and these represent the main imaging modality for the fetal standard plane classification task. Fetal Colour Doppler and Pulsed-wave Doppler images were

defined as OOD based on their acquisition modes, while non-fetal ultrasound images were also defined as OOD images. The test sets contain images from previously unseen patients, although the corresponding OOD categories were observed during training. ID and OOD samples were drawn from the same clinical sites and scanners, differing only in acquisition mode. Therefore, the current evaluation reflects the model's ability to generalize to unseen instances of known OOD categories rather than to entirely novel acquisition sources. The EDL head was trained using both ID and OOD data using the pretrained backbone, with the backbone frozen for 15 epochs followed by 15 epochs of full fine-tuning. Experiments were conducted on an NVIDIA RTX 4090 GPU, with a total training time of ∼4 hours.

3 Results and Discussion

3.1 Classification and Calibration Performance, and OOD Discrimination

We evaluate classification performance in terms of accuracy, and OOD discrimination in terms of AUROC/AUPRC. A Negative Log-Likelihood-based temperature scaling (TS) technique was also applied during post-hoc calibration to ensure reliable probabilities. Here, a single scalar T is optimized on the validation set to minimize NLL and subsequently applied to logits and evidence for the respective cases at test time. As shown in Table 1 and Fig. 2a, 2b and 2c, FetalEDL and its variants progressively improve OOD discrimination at the cost of a small amount of raw accuracy and Expected Calibration Error (ECE). This suggests that the model tends to suppress evidence for hard ID rather than incorrect prediction. Before TS, FetalEDL exhibits the highest ECE, likely reflecting the effect of $\mathcal{L}_{OOD}$ on calibration performance. However, TS restores calibration making it comparable with the other methods. Theoretically, TS preserves the predicted class, and therefore, maintains the same accuracy pre- and post- TS. However, for our proposed FetalEDL variants, we observe small differences in the accuracy. This arises exclusively from samples with near-zero evidence across all classes (model is confidently uncertain in prediction) and the Dirichlet distribution is close to uniform. Here, floating-point perturbations introduced in the scalar division by T can alter the predicted class despite no meaningful accumulation of evidence for any class. These samples also correspond to maximum vacuity and therefore, belong to the abstention region based on our selective classification mechanism irrespective of TS. These results show that while MC dropout + TS achieves highest raw accuracy, FetalEDL provides a more reliable framework in terms of UQ, especially for OOD detection. To account for the class imbalance across 20 classes in our dataset, we also report macro-F1 and balanced accuracy across all methods (post TS) in Table 2. While FetalEDL variants exhibit improved macro-F1 and comparable balanced accuracy compared to the baselines, per-class analysis shows the performance drop in low-support classes (refer to Supplementary Material Table 1).

Table 1: Classification and calibration performance, and OOD detection across methods: MSP baseline, MC Dropout, EDL baseline, proposed FetalEDL; N: Near-OOD sampling; V: Vacuity Regularization; Uncertainty Gap ΔU: $U_W - U_C$.

Method	TS	Acc	ECE	AUROC	AUPRC	U_C	U_W	ΔU
MSP	–	**0.972**	0.017	0.904	0.437	0.006	0.183	0.177
	✓	0.972	0.011	0.914	0.517	0.029	0.334	0.305
MC-D	–	0.972	0.016	0.907	0.503	0.020	0.486	0.466
	✓	0.972	0.017	0.917	0.501	0.035	0.381	0.346
EDL	–	0.966	0.091	0.894	0.524	0.111	0.443	0.332
	✓	0.966	**0.008**	0.933	0.500	0.013	0.229	0.216
FetalEDL (ours)	–	0.957	0.172	0.980	0.655	0.199	0.820	**0.622**
	✓	0.958	0.025	0.982	0.687	0.030	0.633	0.603
FetalEDL+N	–	0.965	0.252	0.987	0.763	0.283	0.770	0.497
	✓	0.964	0.045	**0.992**	0.804	0.063	0.596	0.533
FetalEDL+N+V	–	0.968	0.198	0.990	0.844	0.227	0.642	0.415
	✓	0.969	0.029	0.991	**0.866**	0.043	0.407	0.365

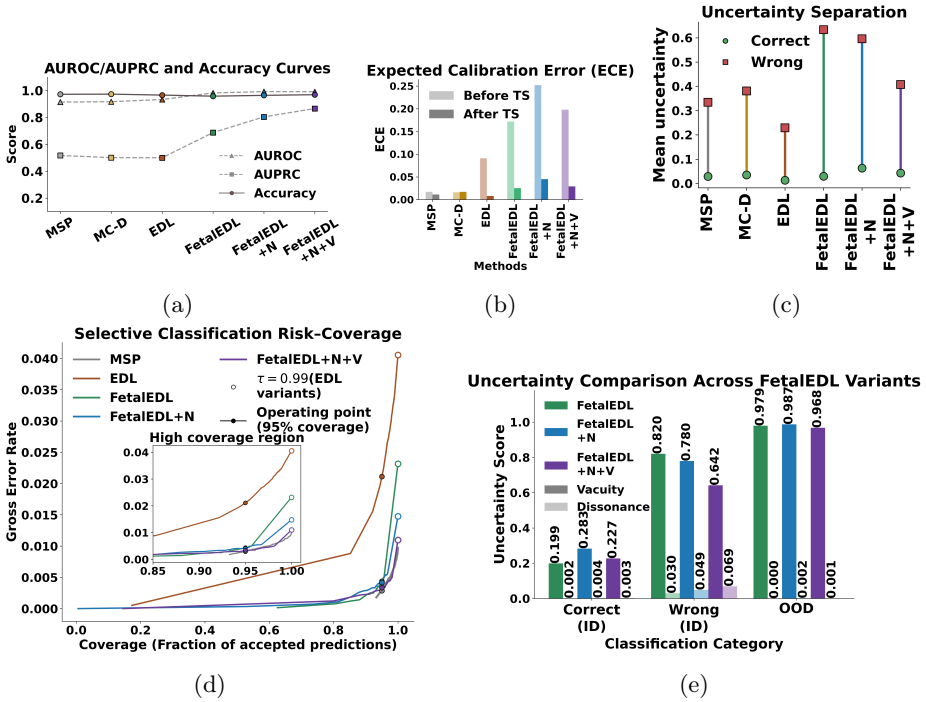

Fig. 2: Performance and Uncertainty Analysis across Methods. (a–c) show classification and calibration performance, (d) shows selective classification risk-coverage and (e) shows uncertainty decomposition across FetalEDL variants.

Table 2: Macro-F1 and balanced accuracy across methods (post-TS): MSP baseline, MC Dropout, EDL baseline, proposed FetalEDL; N: Near-OOD sampling; V: Vacuity Regularization

Method	Macro-F1	Balanced Accuracy
MSP	0.859	0.871
MC-D	0.859	0.871
EDL	0.861	**0.885**
FetalEDL (ours)	0.864	0.874
FetalEDL+N	0.863	0.882
FetalEDL+N+V	**0.865**	0.878

3.2 Analysis of Uncertainty Decomposition

FetalEDL uncertainty decomposition into vacuity and dissonance is consistent with clinical interpretation of the predictions, as shown in Fig. 2e. Correct ID predictions accumulate enough evidence with minimal class conflicts (low u_v; near-zero u_d) and wrong ID predictions occur due to insufficient evidence and class conflicts (higher u_v; relatively higher u_d). OOD inputs exhibit maximum u_v and no u_d, indicating evidence suppression. Since dissonance remains near-zero for OOD cases which is our primary deferral target, vacuity alone drives the abstention decision in our current pipeline substantiating evidence-based reliable selective classification and abstention in the absence of evidence. Dissonance values are comparatively higher for wrong ID predictions, reflecting class conflict rather than absence of evidence. This suggests it could serve as a complementary signal for flagging ID cases such as images containing multiple inter-anatomical structures which may require a secondary review. Fig. 3 further illustrates this qualitatively through GradCAM [14] visualizations of three representative clinical test cases.

3.3 Abstention Analysis

FetalEDL+N+V with a fixed vacuity threshold of $\tau = 0.99$ abstains on just 2.16% of the test set, retaining 97.8% coverage, as shown in Fig. 2d. Gross errors (GE) are defined as the inter-anatomy misclassifications. The GE rate drops from 1.1% to 0.47% after abstention, while overall accuracy improves from 96.86% to 97.85% on the retained images. Analysis of residual GE revealed that they occur primarily in anatomically similar planes (e.g., humerus/femur) and classes that exhibit acquisition variability due to fetal pose and motion. They reflect the ambiguity within single frame interpretations and highlight the need for contextual or temporal information. Qualitative observation, supported by clinical review, showed that many abstained images were those with visibly degraded image quality, despite the absence of explicit image quality supervision during training. Two reviewers with clinical expertise in fetal US acquisition

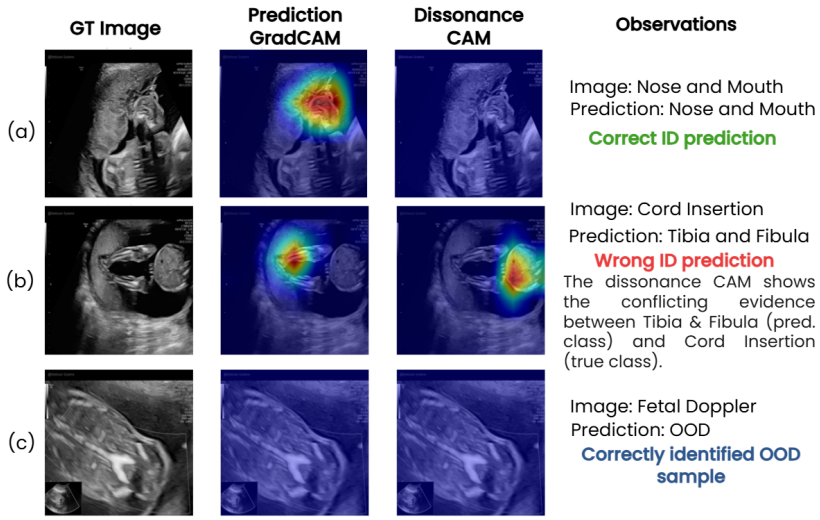

Fig. 3: GradCAM visualizations (Predicted class CAM and Dissonance CAM) for three representative test cases. (a) Correct ID prediction. (b) Wrong ID prediction. (c) Correctly identified OOD sample

assessed the subset of abstained images for clinical usability, evaluating shadowing and image degradation. The review was unblinded and the reviewers agreed upon their observations mostly. This behaviour suggests that $\mathcal{L}_{OOD}$ encourages to associate inconclusive and insufficient evidence with unreliable inputs. Consequently, FetalEDL acts as an efficient selective decision-support system while deferring diagnostically inconclusive cases to sonographer review, directly aligning with clinical workflows.

4 Conclusion and Future Works

We presented FetalEDL, an OOD-aware EDL framework for fetal US plane classification across the 20 standard planes, together with its variants for ablation and failure mode analysis. Identifying a failure mode between OOD exposure and poor-quality B-mode images, we introduced feature-similarity guided OOD sampling and vacuity supervision. Under vacuity-based selective classification, FetalEDL and its variants effectively identified and abstained on difficult images, capturing majority of the gross inter-anatomical errors while achieving high accuracy and demonstrating strong OOD detection. Future work will explore addressing failures driven by intra-class acquisition variability and visually similar anatomical planes by incorporating explicit image-quality assessment, and validating the framework through multi-expert annotation and prospective clinical studies.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**(104863), 104863 (Feb 2020)
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11), 2204–2215 (2017)
3. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
4. Chudleigh, T., Cohen-Overbeek, T.E., for the ISUOG Basic Training Sub-Committee: How to apply the 20 + 2-planes method for identification of 65 fetal abnormalities during routine second-trimester fetal ultrasound examination. Ultrasound in Obstetrics & Gynecology **66**(3), 383–396 (2025). https://doi.org/10.1002/uog.29299
5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059. PMLR (2016)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
7. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Hkg4TI9xl
8. Jøsang, A.: Subjective logic. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer International Publishing, Basel, Switzerland, 1 edn. (Nov 2016)
9. Laitinen-Fredriksson Lundström-Imanov, G.O.Y.: Uncertainty-calibrated explainable artificial intelligence for fetal ultrasound plane classification: A systematic review. arXiv e-prints (01 2026). https://doi.org/10.48550/arXiv.2601.00990
10. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems **30** (2017)
11. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022), https://arxiv.org/abs/2201.03545
12. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems **32** (2019)
13. Rahman, R., Alam, M.G.R., Reza, M.T., Huq, A., Jeon, G., Uddin, M.Z., Hassan, M.M.: Demystifying evidential dempster shafer-based cnn architecture for fetal plane detection from 2d ultrasound images leveraging fuzzy-contrast enhancement and explainable ai. Ultrasonics **132**, 107017 (2023). https://doi.org/https://doi.org/10.1016/j.ultras.2023.107017, https://www.sciencedirect.com/science/article/pii/S0041624X23000938
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7

15. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems **31** (2018)
16. Vitale, S., Orlando, J.I., Iarussi, E., Larrabide, I.: Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. International Journal of Computer Assisted Radiology and Surgery **15**(2), 183–192 (2019)
17. Xiao, S., Zhang, J., Zhu, Y., Zhang, Z., Cao, H., Xie, M., Zhang, L.: Application and progress of artificial intelligence in fetal ultrasound. Journal of Clinical Medicine **12**(9) (2023). https://doi.org/10.3390/jcm12093298, https://www.mdpi.com/2077-0383/12/9/3298
18. Zhao, X., Ou, Y., Kaplan, L., Chen, F., Cho, J.H.: Quantifying classification uncertainty using regularized evidential neural networks. arXiv preprint arXiv:1910.06864 (2019)