This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

SONIC: Sonographer-Inspired Conceptual Explanation for Neural Networks

Yingni Wang[1], Yihua Sun[1], Hee Guan Khor[1], Licong Dong[2], Qiongyu Ye[4], Huabin Zhang[3], Desheng Sun[2], Kehong Yuan[5], Jianwen Luo[1], Fang Chen[6], and Hongen Liao[1,6] ⋆

[1] School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China
[2] Department of Ultrasound, Peking University Shenzhen Hospital, Shenzhen
[3] Department of Ultrasound, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[4] Department of Ultrasound, Shenzhen Baoan Women's and Children's Hospital, Shenzhen, China
[5] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[6] School of Biomedical Engineering, Shanghai Jiaotong University, Shanghai, China

Abstract. Ultrasound is the primary imaging modality for evaluating fetal anatomy, and fetal standard ultrasound planes are essential for determining gestational age and assessing overall fetal health. Deep neural networks (DNNs) have achieved remarkable success in identifying fetal ultrasound standard planes. However, their black-box nature fundamentally lacks clinical interpretability and trustworthiness, limiting their real-world deployment. To bridge this gap, we propose SONIC (Sonographer-Inspired Conceptual Explanation for Neural Networks), an interpretable framework that aligns decisions with clinically defined anatomical concepts. Moving beyond pixel-level patterns, SONIC explicitly extracts anatomical concepts and models their spatial dependencies via a graph convolutional network. We further employ a graph-attribution method to quantify the importance of anatomical structures, ensuring transparent and trackable classification. We also analyze prediction–explanation consistency, showing that anatomically implausible evidence can serve as a practical safety signal for clinical verification. Extensive experiments on a public dataset and two private multi-center cohorts, including perturbation-based fidelity analysis and a blinded expert user study, demonstrate that SONIC achieves competitive or improved performance compared with pixel-based DNN baselines while providing concept-level interpretability.

Keywords: Interpretable Framework · Medical Anatomical Knowledge · Fetal Ultrasound Imaging · Deep Neural Networks (DNNs)

⋆ Corresponding author.

1 Introduction

Two-dimensional ultrasound remains the primary modality for fetal anatomical assessment. Standard planes are critical for estimating gestational age and evaluating fetal health. However, their identification is challenging, requiring interpretation of ISUOG-defined anatomical landmarks [16] and substantial clinical expertise. Therefore, automated analysis has become highly desirable to reduce clinical workload, and Convolutional Neural Networks (CNNs) have demonstrated remarkable success in fetal standard plane analysis [1, 14].

Deep learning models evaluate pixel-level patterns rather than the structural and anatomical reasoning employed by expert sonographers. This fundamental distinction creates a substantial gap between deep learning models and clinical experts, limiting their real-world applicability in medical settings. We refer to this discrepancy as a critical **reasoning gap**. Although interpretability is essential for establishing clinical trust in prenatal diagnosis [12], existing approaches remain insufficient to bridge this reasoning gap in clinical contexts and largely retain an inherent "black-box" nature [11].

To bridge the reasoning gap, we propose **Son**ographer-**I**nspired **C**onceptual Explanation for Neural Networks (SONIC), an intrinsically interpretable framework grounded in clinically defined anatomical concepts. Instead of directly predicting ultrasound plane labels, SONIC mimics expert sonographers by locating clinically relevant landmarks and grounding its decisions in anatomically meaningful evidence. This framework explicitly models fetal anatomy using graph convolutional networks (GCNs), mapping extracted anatomical concepts to nodes and spatial dependencies to edges. This design enhances diagnostic accuracy and facilitates transparent decision-making across three clinically crucial fetal standard planes: the fetal abdominal standard plane (FASP), fetal thalamic standard plane (FTSP), and fetal femur standard plane (FFSP). Our main contributions are as follows. (1) **Concept Discovery:** We introduce a robust prototype-based pipeline that extracts expert-validated anatomical landmarks. (2) **Relational Reasoning:** We develop a novel graph framework that explicitly models structural dependencies to enhance plane classification performance. (3) **Clinical Trustworthiness:** We conduct extensive evaluations demonstrating high model fidelity and strong alignment with expert sonographers. SONIC further incorporates an intrinsic failure-aware design that visually exposes undetectable anatomy, empowering clinicians to intercept potential silent diagnostic errors.

2 Method

To achieve interpretable classification of ultrasound standard planes, we first perform coarse classification to train a region-level feature extractor (Fig. 1-a). Then, image sub-regions are mapped to a set of recognizable medical concepts via feature matching (Fig. 1-b). To achieve this, we define and construct cluster centers for 12 key medical concepts across three planes (Fig. 1-I) and dynamically match them with region-level features during model forward pass (Fig. 1-II). A concept-based GCN is subsequently used to achieve high-performance, interpretable classification based on the medical concepts (Fig. 1-c). Grounding predictions in explicit medical concepts makes the decision process transparent and helps flag anatomically implausible evidence for verification.

2.1 Offline Anatomy-Driven Concept Construction

Our goal is to first construct four representative concept prototypes for each standard plane (12 prototypes in total), following the ISUOG guidelines [16] (Fig. 1-I). These prototype features can be used for subsequent automated matching from image sub-regions to medical concepts, including: (1) **FASP**: stomach (ST), spine (SP), umbilical vein (UV), and amniotic fluid wall (AFW); (2) **FTSP**: left/right thalamus (LTM/RTM), cavum septum pellucidum (CSP), and skull; (3) **FFSP**: femur (FM), two metaphyses (MPs), and surrounding tissues.

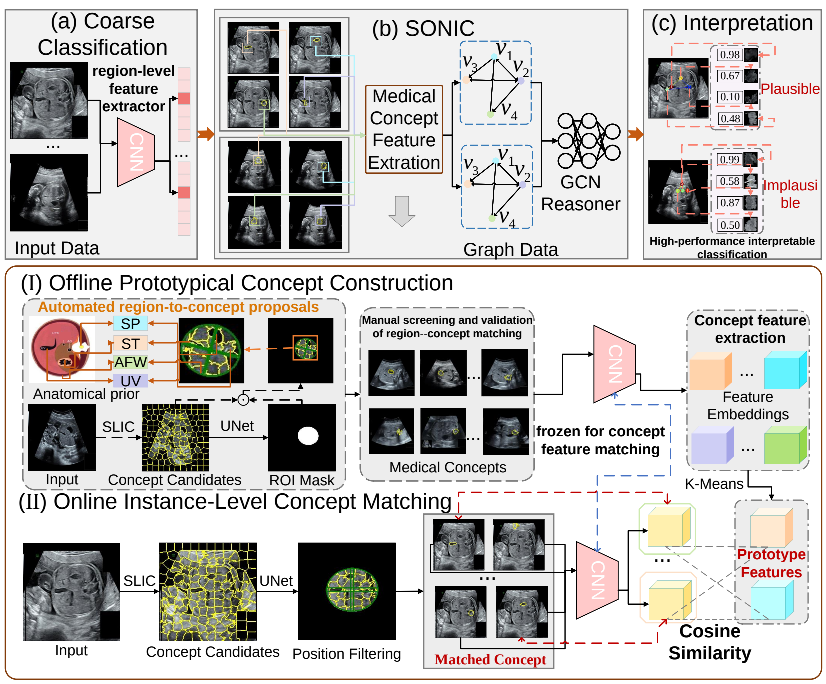

Fig. 1: SONIC pipeline. A coarse CNN provides region-level features for concept extraction (a). SONIC matches sub-regions to plane-specific prototypes built offline by K-means using cosine similarity (b, I–II), then a concept-graph GCN predicts the standard plane and yields concept-based explanations (c).

As shown in Fig. 1-I, based on ISUOG practice, we first employ an automated approach to group superpixels and automatically propose sub-region-to-concept pairings based on the echogenicity and texture of the sub-regions. Second, we manually verify these pairings to remove implausible matches, resulting in a gold-standard dataset of sub-region-to-medical concept pairs. Third, we extract features from each sub-region and perform clustering to obtain the feature centroids for each medical concept.

To improve automated annotation performance and streamline the workflow, we first localize regions of interest (ROIs) using segmentation or class activation maps. For FASP and FTSP, key anatomical structures are consistently distributed in the abdominal and cranial regions. We train a U-Net [15] on 1,000 annotated planes to obtain coarse abdominal and head segmentations and reference axes. Within the ROI, we automatically identify concept candidates by combining appearance and geometry cues [16] by ranking average gray value and Gray-Level Run-Length Matrix (GLRLM) texture statistics[4]: 1) In FTSP, structures aligned with the reference axes are prioritized as candidates for the LTM, RTM, and CSP; 2) In FASP, hypoechoic elliptical regions indicate the ST and UV, coarse-textured regions along the long axis suggest the SP, and high-intensity boundaries correspond to the skull or AFW. On the other hand, for FFSP, the FM exhibits large pose variability and tubular morphology, making rigid ROI segmentation unreliable. We thus derive a coarse FM ROI from GradCAM heatmaps and select the FM and MPs candidates using intensity cues and GLRLM texture statistics[4].

Finally, the selected sub-region-to-concept pairings are manually verified to remove inaccurate proposals. Each sub-region is then embedded using a pre-trained region-level feature encoder and clustered with K-Means to form plane-specific concept prototypes for subsequent online matching (Fig. 1-I).

2.2 Online Instance-Level Concept Matching

Once the prototype features for each medical concept are established, any image can be processed via the pre-trained region-level feature extractor. Each sub-region is then matched online to its corresponding medical concept (Fig. 1-II). Specifically, each sub-region is embedded using the pretrained CNN for prototype matching. Cosine similarity is computed between the candidate embedding and all concept prototypes, and the best match is selected as the final instance:

$$\hat{p} = \arg\max_{p \in \mathcal{P}_c} \cos\big(f(\text{cand}),\, p\big), \tag{1}$$

where $f(\cdot)$ denotes the CNN encoder, $\mathcal{P}_c$ is the prototype set for concept category c, and $\hat{p}$ indicates the selected concept instance.

2.3 Graph-based Anatomical Reasoning

Given the matched concept instances, we formulate standard-plane classification as graph-based anatomical reasoning (Fig. 1-b). Each ultrasound image is represented as a fully-connected directed concept graph $G = (V, E)$ with a fixed concept set ($|V| = 4$ and $|E| = 12$). Nodes $v_i \in V$ correspond to concept

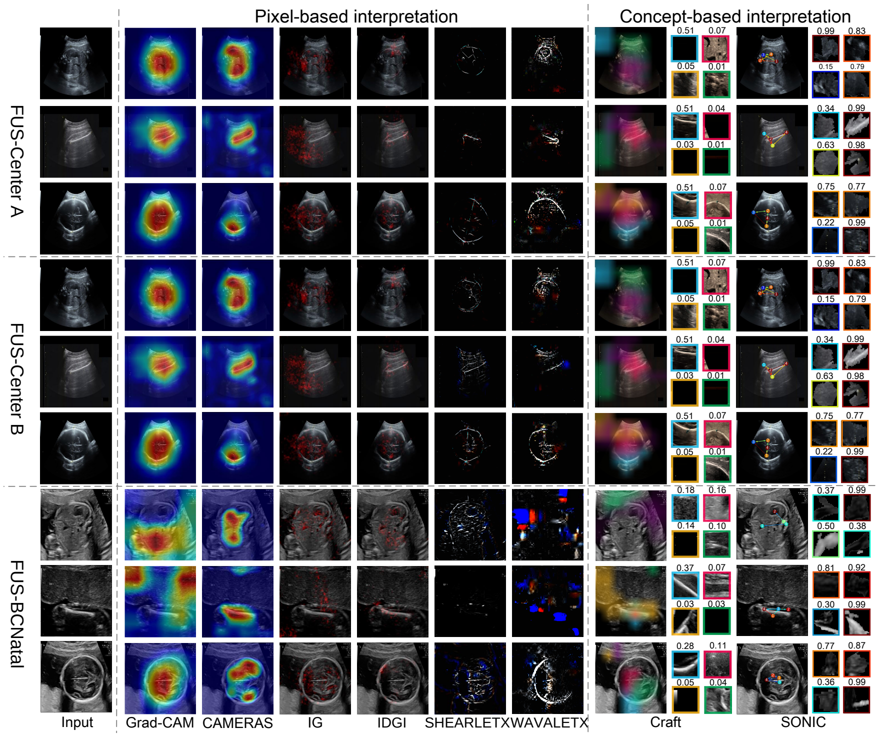

Fig. 2: Qualitative comparison with SOTA methods. For CRAFT and SONIC, each image shows the discriminative medical concepts (right). The number in the top of the concept denotes the importance score for the predicted plane. Higher scores indicate higher importance.

instances and are initialized with CNN embeddings $x_i^{(0)} \in \mathbb{R}^d$, while directed edges $e_{ji} \in E$ encode spatial relations using $r_{ji}^{(0)} = [\Delta x_{ji}, \Delta y_{ji}]$, the normalized relative displacement between concept centers. An edge-gated GraphConv [13] is adopted, where a learnable scalar γ_{ji} modulates messages along edge e_{ji}. After K layers, we concatenate all node and edge features into a graph vector and use an MLP for plane prediction.

We evaluate the anatomical relevance of the inference using Graph Integrated Gradients. This method attributes the prediction output to input concepts by accumulating gradients along an interpolation path. Formally, let x_i denote the feature vector of node v_i and x_i' be a baseline reference (e.g., a zero vector denoting concept absence). The attribution score $s(x_i)$ represents the accumulated sensitivity of the prediction function Φ and is computed as:

$$s(x_i) = (x_i - x_i^{'}) \odot \int_{\alpha=0}^{1} \nabla_{x_i} \Phi \left(x_i^{'} + \alpha (x_i - x_i^{'}) \right) \, d\alpha \tag{2}$$

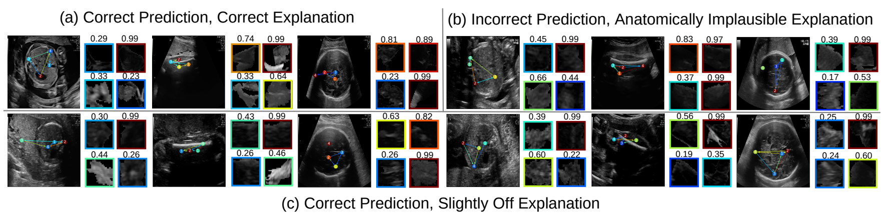

Fig. 3: Prediction-explanation consistency. The number at the top of the concept denotes the importance score. Higher scores indicate higher importance.

3 Experiments

3.1 Datasets and Implementation Details

Datasets. Experiments were conducted on one public dataset (FUS-BCNatal [2]) and two private datasets: FUS-Center A (7,547 images from 1,130 subjects) and FUS-Center B (380 images) for external validation. We consider a four-class setting (FASP, FFSP, FTSP, Other). For the private cohorts, gestational ages range from 16 to 39 weeks, and multiple pregnancies are excluded. For FUS-BCNatal, we follow its official evaluation protocol/split. To avoid patient-level leakage in the private data, we adopt a **subject-level** split on FUS-Center A with a 7/1/2 ratio for train/validation/test, assigning all images from the same subject to a single split. Models are trained and evaluated separately on FUS-BCNatal and on FUS-Center A using their respective protocols. FUS-Center B is used only for **external** validation of models trained on Center A, without any training or hyperparameter selection on Center B. The class distributions are Center A (938/1070/1337/4202) and Center B (66/51/63/200).

Implementation. Implemented in PyTorch on an NVIDIA RTX 8000, images were resized to 256×256 and normalized. We evaluated ResNet18 [6], DenseNet121 [7], VGG16 [21], and MobileNetV2 [17]. All models used the Adam optimizer [9] (initial $lr = 10^{-2}$). CNN baselines were trained for 100 epochs, whereas GCNs were trained for 200 epochs. Statistical significance ($p < 0.05$) of performance improvements was assessed via paired, two-sided non-parametric bootstrap resampling ($N = 1000$).

3.2 Visualizing Medical Concept Reasoning

Comparison with State-of-the-Art (SOTA). We qualitatively compare SONIC against leading pixel-level, frequency-based, and concept-based explanation methods (Fig. 2). Results demonstrate that SONIC yields highly interpretable predictions, offering explanations consistently grounded in anatomical reality. Standard pixel-based methods struggle with domain semantics: in FTSP cases, GradCAM [18] and CAMERAS [8] diffusely highlight irrelevant structures (e.g., the entire brain), whereas IG [19] and IDGI [20] degenerate into scattered noise. Furthermore, frequency-based approaches like SHEARLETX [10] over-index on textures rather than structural meaning. Even Craft [3], a concept-based method, exhibits alignment bias by misattributing concepts to incorrect spatial features. In contrast, SONIC leverages GCNs to explicitly model relational dependencies, achieving precise spatial localization and ensuring that diagnostic interpretations more align with clinical anatomical definitions.

Consistency under imperfect unsupervised concepts. Given the lack of dense annotations, occasional concept omissions are inevitable. We characterize the framework's practical reliability through three representative regimes (Fig. 3): (a) *Correct Prediction, Correct Explanation:* Perfect alignment between accurate localization and correct predictions supports transparent decision-making; (b) *Incorrect Prediction, Anatomically Implausible Explanation:* The model maintains predictive accuracy despite partial concept occlusion, demonstrating structural robustness; (c) *Correct Prediction, Slightly Off Explanation:* In extreme failure cases, incorrect predictions consistently co-occur with anatomically implausible explanations, providing a practical safety signal for clinical verification. Overall, these patterns suggest that SONIC remains useful under imperfect concept extraction and exposes its own failure modes via concept-level evidence, improving trustworthiness in practice.

3.3 Clinical Alignment and Fidelity Analysis

User Study: Clinical Relevance of Explanations. We conducted a blinded user study with four fetal ultrasound experts on 100 randomly sampled images from two centers to assess the clinical relevance of our explanations. Each image was rated by all four experts. Experts scored the anatomical relevance of highlighted structures on a 5-point Likert scale (1: Irrelevant, 5: Highly Relevant). As summarized in Table 1, most explanations received high ratings (4–5), with a mean score of 4.00/5 (computed by treating Likert ratings as numeric values). These results suggest that the proposed explanations are generally consistent with expert anatomical expectations.

Fidelity Analysis: Perturbation-based Faithfulness. Beyond subjective ratings, we evaluate the faithfulness of concept-based explanations using perturbation tests [5, 22]. Here we compare fidelity only with CRAFT [3]. To fairly compare concept-based methods, we use a unified image-space perturbation protocol: concepts are ranked by each method's attribution scores, and we perturb the image regions of the top-k concepts (defined as the associated superpixel unions) and re-evaluate the CNN baseline, reporting the accuracy drop relative

Table 1: User study results. The predominance of high ratings (4-5) confirms the high anatomical relevance of the proposed concept-based explanations.

	1	2	3	4	5
User 1	1 %	18 %	14 %	57 %	10 %
User 2	2 %	1 %	2 %	8 %	87 %
User 3	3 %	4 %	10 %	49 %	34 %
User 4	0 %	2 %	43 %	49 %	6 %
Average	1.5 %	6.25 %	17.25 %	40.75 %	34.25 %

Table 2: Fidelity analysis. "Per. most k" perturbs the k-th highest-ranked concept. The accuracy drop from ACC_{ori} indicates concept importance.

Method	Per Concept	FUS-Center A			FUS-Center B			FUS-BCNatal		
		FASP	FFSP	FTSP	FASP	FFSP	FTSP	FASP	FFSP	FTSP
Craft	Per. most 1	63.99	4.66	9.06	12.50	5.88	11.11	47.10	20.22	66.01
	Per. most 2	10.29	1.04	0	9.37	0	0	2.24	13.74	58.69
	Per. most 3	5.79	0.52	0	7.81	0	0	1.68	13.36	54.25
	Per. most 4	0	0	0	2.00	0	0	1.12	10.31	50.33
SONIC	Per. most 1	69.60	49.92	87.13	57.89	49.90	51.61	56.28	62.55	84.75
	Per. most 2	45.81	18.87	69.50	45.61	25.00	35.38	50.66	58.29	77.76
	Per. most 3	28.43	5.11	56.26	26.31	4.42	27.42	44.05	50.61	74.49
	Per. most 4	17.53	1.00	42.24	21.05	2.27	19.35	37.00	43.52	69.20
ACC_{ori}		99.84	99.54	99.73	92.19	98.04	95.24	92.46	94.27	96.73

Table 3: Performance comparison between pixel-based CNN baselines and the proposed concept-graph reasoning model. The GCN consistently achieves superior performance across three datasets and four backbones.

Dataset	Metric	ResNet18		VGG16		DenseNet121		MobileNetV2	
		CNN	GCN	CNN	GCN	CNN	GCN	CNN	GCN
FUS-Center A	ACC (%)	96.73	**97.20***	89.15	**97.08***	95.89	**98.67***	95.59	**99.11***
	PRE (%)	**95.33**	94.59	93.91	**96.61***	96.28	**98.29***	92.46	**98.71**
	REC (%)	93.42	**94.53***	78.86	**96.66***	93.44	**97.63***	95.17	**98.93***
	F1 (%)	94.22	**94.56***	84.86	**96.63***	94.55	**97.96***	93.35	**98.82***
FUS-Center B	ACC (%)	79.89	**95.03***	82.80	**96.31***	88.62	**94.75**	81.75	**97.24***
	PRE (%)	77.13	**91.95***	89.89	**95.71***	91.55	**92.88**	81.58	**97.54***
	REC (%)	70.83	**91.83***	72.51	**96.17***	83.87	**94.76***	75.61	**95.98***
	F1 (%)	71.77	**91.89***	78.57	**95.91***	86.81	**93.48***	76.66	**95.96***
FUS-BCNatal	ACC (%)	91.64	**94.43***	91.04	**94.84***	91.67	**97.89***	91.88	**98.95***
	PRE (%)	90.02	**92.71***	**88.45**	77.81	89.36	**96.82***	89.60	**98.84***
	REC (%)	**92.66**	91.43	**93.24**	82.00	93.49	**95.52***	93.16	**98.80***
	F1 (%)	91.22	**92.03**	**90.54**	79.51	91.20	**96.13***	91.14	**98.83***

* *denotes SONIC significantly outperforms the corresponding CNN baseline ($p < 0.05$).*

to ACC_{ori} (Table 2). We inject Gaussian noise into the selected regions instead of masking them to simulate subtle evidence corruption. CRAFT often shows abrupt accuracy drops when perturbing the top-ranked concept, suggesting region overlap among multiple discriminative cues. In contrast, SONIC degrades more progressively as k increases, consistent with more separable plane-specific concepts and a clearer importance ranking.

Effectiveness of the Concept Representation. Table 3 shows that the concept-graph model built on anatomical concepts achieves strong performance, and in many settings matches or exceeds the corresponding CNN baseline. This indicates that the selected concepts and their relational representation preserve key discriminative evidence for standard-plane recognition, rather than serving as purely qualitative visual cues. Consequently, the resulting concept-level attributions are grounded in task-relevant evidence and provide informative interpretations of the model's decisions.

4 Conclusion

In this work, we propose SONIC, an intrinsically interpretable framework for fetal standard-plane classification that combines prototype-based concept discovery with concept-graph reasoning. By explicitly extracting clinically meaningful concepts and modeling their spatial relations with a GCN, SONIC achieves competitive or improved performance over CNN baselines across datasets. Moreover, prediction–explanation coupled deviations provide a practical failure-awareness signal by revealing anatomically implausible evidence for clinical verification. A blinded expert study further supports the anatomical relevance of the proposed explanations. Collectively, these results indicate that concept-driven structural reasoning can narrow the gap between black-box recognition and clinically grounded evidence. Future work will focus on clinical integration and extension to other medical imaging tasks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (U22A2051, 62271246), the National Key Research and Development Program of China (2022YFC2405200), the Science and Technology Commission of Shanghai Municipality (24511104100).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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)
2. Burgos-Artizzu, Xavier P and Coronado-Gutiérrez, David and Valenzuela-Alcaraz, Brenda and Bonet-Carne, Elisenda and Eixarch, Elisenda and Crispi, Fatima and Gratacós, Eduard: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
3. Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., Serre, T.: Craft: Concept recursive activation factorization for explainability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2711–2721 (2023)
4. Galloway, M.M.: Texture analysis using grey level run lengths. Nasa Sti/recon Technical Report N **75**, 18555 (1974)
5. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in Neural Information Processing Systems **32** (2019)
6. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
7. Huang, Gao and Liu, Zhuang and Van Der Maaten, Laurens and Weinberger, Kilian Q: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)

8. Jalwana, M.A., Akhtar, N., Bennamoun, M., Mian, A.: Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16327–16336 (2021)
9. Kingma, D.P.: Adam: A method for stochastic optimization (2014)
10. Kolek, S., Windesheim, R., Andrade-Loarca, H., Kutyniok, G., Levie, R.: Explaining image classifiers with multiscale directional image representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18600–18609 (2023)
11. Li, Jintang and Gao, Zhan and Wang, Chunlian and Pu, Bin and Li, Kenli: A rule-guided interpretable lightweight framework for fetal standard ultrasound plane capture and biometric measurement. Neurocomputing **621**, 129290 (2025)
12. Mahapatra, D., Poellinger, A., Reyes, M.: Interpretability-guided inductive bias for deep learning based medical image. Medical Image Analysis **81** (OCT 2022). https://doi.org/10.1016/j.media.2022.102551
13. Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M.: Weisfeiler and leman go neural: Higher-order graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4602–4609 (2019)
14. Pu, Bin and Li, Kenli and Li, Shengli and Zhu, Ningbo: Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Transactions on Industrial Informatics **17**(11), 7771–7780 (2021)
15. Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
16. Salomon, LJ and Alfirevic, Z and Berghella, V and Bilardo, CM and Chalouhi, GE and Costa, F Da Silva and Hernandez-Andrade, E and Malinger, G and Munoz, H and Paladini, D and others: ISUOG Practice Guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics and Gynecology **59**(6), 840–856 (2022)
17. Sandler, Mark and Howard, Andrew and Zhu, Menglong and Zhmoginov, Andrey and Chen, Liang-Chieh: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
19. Sundararajan, M., Taly, A., Yan, Q.: Gradients of counterfactuals. arXiv preprint arXiv:1611.02639 (2016)
20. Yang, R., Wang, B., Bilgic, M.: Idgi: A framework to eliminate explanation noise from integrated gradients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23725–23734 (2023)
21. Zeiler, Matthew D and Fergus, Rob: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)

22. Zhang, R., Madumal, P., Miller, T., Ehinger, K.A., Rubinstein, B.I.: Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11682–11690 (2021)