# Effect of Knowledge Distillation on Generalization of Fetal Ultrasound Classifiers

**Carl Reilly*** [1] CARRE@DTU.DK
**Chun Kit Wong*** [1] CKWO@DTU.DK
**Pablo Delgado-Rodriguez** [1] PDERO@DTU.DK
**Zahra Bashir** [2] ZAB@REGIONSJAELLAND.DK
**Anders Nymark Christensen** [1] ANYM@DTU.DK
**Martin Tolsgaard** [2] MARTINTOLSGAARD@GMAIL.COM
**Aasa Feragen** [1] AFHAR@DTU.DK
[1] *Technical University of Denmark, Kongens Lyngby, Denmark*
[2] *CAMES, Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark*

**Editors:** Under Review for MIDL 2026

## Abstract

Deep learning has shown strong performance in fetal ultrasound standard-plane classification and quality assessment, but deployment in clinical settings remains limited by slow inference and limited interpretability. Progressive Concept Bottleneck Models (PCBM) address interpretability by predicting anatomical and property concepts prior to classification, yet their computational complexity hinders real-time use.

We investigate knowledge distillation as a means to obtain lightweight PCBM-based student models that preserve diagnostic behaviour while improving deployability. Two strategies are evaluated: **MiniPCBM**, a reduced-capacity concept-based student that maintains hierarchical concept reasoning, and **MicroPCBM**, a pure classification student trained using softened teacher logits. We further analyse robustness in terms of cross-domain generalisation and stability of model explanations.

Across three datasets — a private clinical cohort and two public sets from Barcelona and Africa — MiniPCBM achieves the strongest in-distribution performance on the Denmark dataset (7-class accuracy: 0.93 vs. PCBM 0.90), while MicroPCBM provides competitive accuracy (0.90) with a $\geq$30$\times$ reduction in parameter count. Both distilled models deliver substantial efficiency gains, reducing inference latency from 42.91,ms (PCBM) to 10.58,ms and 6.61,ms, respectively, supporting real-time guidance. These results show that structured domain knowledge from PCBM can be effectively compressed into lightweight architectures, enabling interpretable and deployable fetal ultrasound AI for point-of-care use.

**Keywords:** Knowledge Distillation, Robustness, Explainability, Fetal Ultrasound Classification

## 1. Introduction

Deep learning is a central tool for analysing fetal ultrasound images, with models achieving strong performance for tasks such as standard scan plane detection, scan quality assessment

* Contributed equally

and detection of anomalies and adverse outcomes. Nevertheless, many high-performing architectures remain opaque and computationally expensive, which limits their use as real-time decision support during routine screening. Obstetric ultrasound is a dynamic procedure, where navigating to high quality images is challenging, and only select images are saved. Acquisition is challenging, as sonographers work with noisy, artefact-prone data and substantial inter-patient variability (Fiorentino et al., 2023; Patra et al., 2020). To support the sonographers, AI assistants aim both to support their navigation to higher quality images, and to support automatic detection of sufficiently good images. Clinically useful AI-guided navigation and quality assessment systems therefore need to be not only accurate, but also interpretable and fast enough to operate at video frame rates.

**Explainability in fetal ultrasound AI.** Previous work on real-time standard scan plane detection (Baumgartner et al., 2016, 2017) has shown that convolutional networks can process freehand ultrasound at over 100 frames per second, enabling online annotation and plane retrieval. While these models provide post-hoc saliency maps that can be useful for model development, these explanations are limited in their utility for clinicians: They offer limited control over the model's internal reasoning and no explicit representation of the anatomical and measurement concepts that clinicians use. As a result, their interpretability is not useful for clinicians during acquisition time, and also does not give the clinicians feedback that is useful to support the development of their image acquisition skills.

Concept Bottleneck Models (CBMs) (Koh et al., 2020) address the opacity of deep neural networks by introducing an intermediate layer of human-interpretable concepts, enabling reasoning inspection and user intervention during prediction. Although effective across medical imaging tasks, traditional CBMs typically rely on a single bottleneck and may struggle to represent spatially structured concepts such as anatomical visibility. Progressive Concept Bottleneck Models (PCBM) extend this framework through hierarchical concept learning (Lin et al., 2022; Bashir et al., 2025), improving clinical alignment and decision interpretability. However, their multi-branch architecture imposes high computational cost, motivating the development of more efficient distilled models for real-time ultrasound deployment.

**Knowledge distillation for real-time inference.** In parallel, Knowledge Distillation (KD) has emerged as a standard technique for compressing large neural networks into smaller, deployment-ready student models. In its basic form, KD trains a lightweight student model to match the soft output distribution (consisting of either logits or temperature-scaled probabilities) of a high-capacity teacher, typically using a Kullback–Leibler divergence term alongside cross-entropy to match the ground-truth labels (Hinton et al., 2015). This strategy has been widely adopted to reduce latency and memory footprint while preserving much of the teacher's performance. In fetal ultrasound, KD has been used to build efficient models for anatomy classification from freehand sequences and for standard scan plane detection, demonstrating that substantial reductions in model size and inference time are achievable without catastrophic loss of accuracy. For example, Dapueto et al. (Dapueto et al., 2024) distil a heavy standard-plane detector into a compact network tailored for efficient fetal ultrasound scan plane detection. However, the effect of distillation on model robustness remain unexplored.

**In this work**, we build on these two strands: interpretable concept-based modelling and efficiency-oriented distillation for fetal ultrasound. Instead of distilling a conventional

black-box detector, we use a PCBM trained on fetal ultrasound standard planes as an intrinsically explainable teacher and transfer its class-level knowledge to a compact student classifier via logit-based KD. We employ KD in two different ways: firstly, we create a distilled PCBM which retains the two interpretable concept bottleneck layers but uses distillation to lighten their predictions, **MiniPCBM**. Secondly, we create a very lightweight predictor, **MicroPCBM**, which is no longer a PCBM model but which we hope will retain the diagnostic behaviour and cross-dataset robustness of the PCBM while achieving inference speeds compatible with real-time use on clinical ultrasound machines. **Our contribution** is a comprehensive evaluation of knowledge-distilled PCBM students — assessing the tradeoff between classification performance, cross-domain robustness, inference speed, and explainability — demonstrating that high-quality, concept-driven reasoning can be preserved while enabling real-time deployment on point-of-care ultrasound systems, supporting faster and more reliable decision-making in clinical practice

## 2. Methods

### 2.1. Progressive Concept Bottleneck Models (PCBMs) for Fetal Ultrasound

The PCBM (Lin et al., 2022; Bashir et al., 2025) serves as the teacher model in this study. PCBM follows a hierarchical architecture that progressively maps ultrasound images to clinically meaningful intermediate representations prior to classification.

The model, shown in the top row of Fig. 1 first extracts dense visual concepts through a segmentation block, predicting multiple anatomical structures present in the image. These segmentation-based concepts are then abstracted into higher-level property concepts representing clinically relevant criteria such as visibility, alignment, and caliper placement. Finally, the predicted property concepts are used as input to the classification head, which determines the anatomical plane label.

The segmentation block is a DTU-net (Lin et al., 2023), consisting of two mini-U-nets trained specifically for high fidelity in segmenting curvilinear structures such as bone boundaries or skin, whose visibility is important for measuring fetal ultrasound biometries. This block is computationally heavy, and provides spatially resolved segmentation logits for 14 fetal anatomical structures. The next block consists of three individually trained ResNet blocks with additional logical that combine original images and segmentation features to predict 27 continuous property concept scores representing image quality and anatomical correctness. The ResNet predictions are additionally subjected to rule-based adaptation based on the existence of certain wanted or unwanted organs in the segmentation model's prediction, which contributes to the complexity of the concept prediction model. The final classifier predicts 7 anatomical plane categories based solely on property concepts, allowing concept-level intervention and interpretation.

This progressive structure enables interpretable decision-making at multiple levels: segmentation attention provides spatial localization, and property concepts provide semantic reasoning aligned with clinical criteria. PCBM has demonstrated high classification performance and generalization capability also compared to black-box models of similar capacity (Lin et al., 2022), leading to strong user trust in ultrasound quality assessment tasks. However, its multi-branch design and segmentation inference introduce high computational cost, motivating the development of compressed student models for real-time deployment.

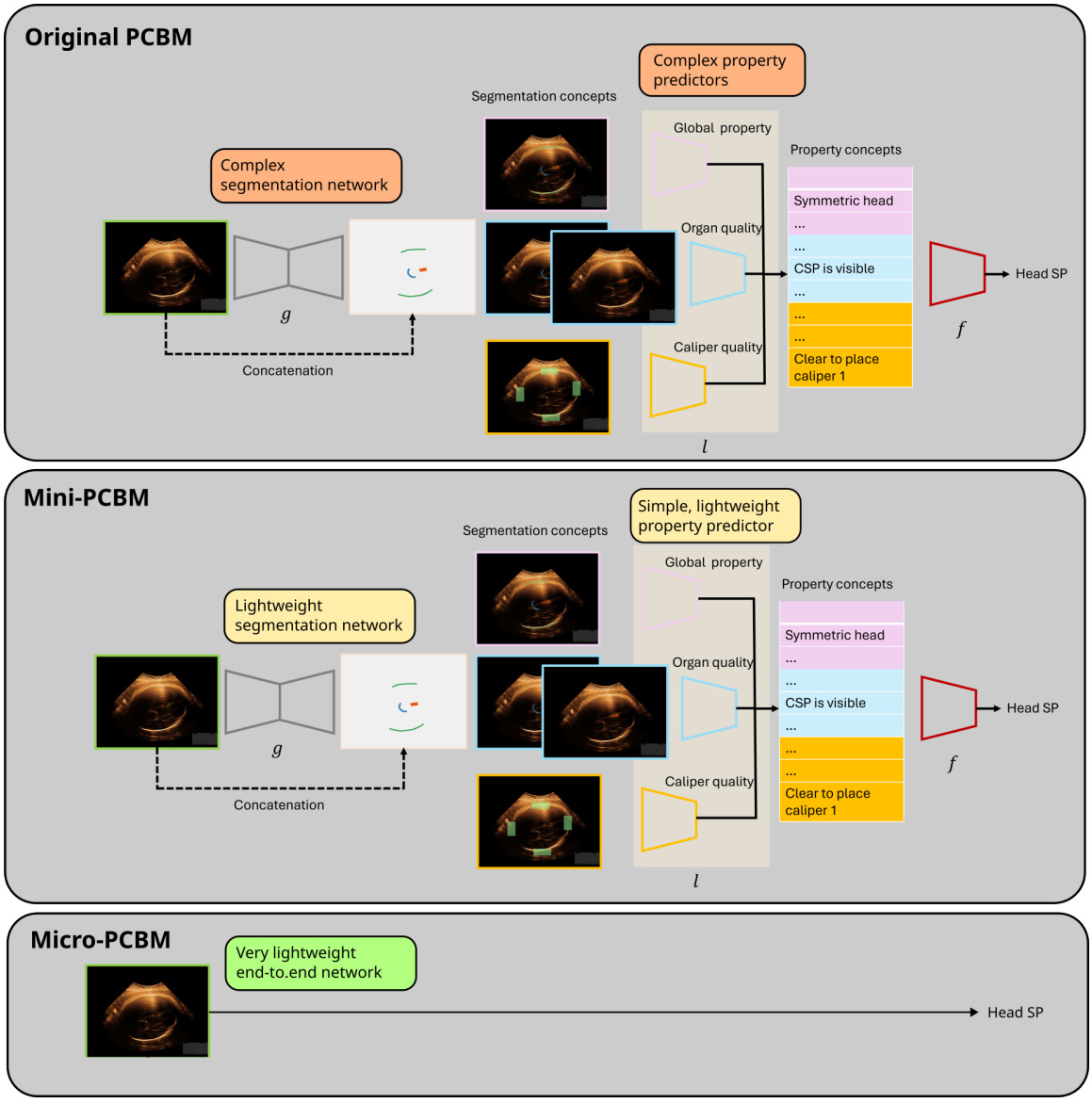


Figure 1: The PCBM architecture consists of three blocks: A segmentation block creating visual concepts, a property concept prediction block, and a classification block. We present two options for distillation: Distilling the three individual blocks for retained explainability (MiniPCBM), or distilling into an end-to-end predictive model (MicroPCBM).

### 2.2. MiniPCBM

MiniPCBM is a lightweight distilled variant of the PCBM that retains the two concept layers but replaces their predictors with substantially smaller networks. It thus allows a similar level of interpretability as the original PCBM, but at reduced computational cost.

The MiniPCBM segmentation block is a mini-U-net consisting of 4 encoder-decoder pairs with (32, 64, 128, 256) features each, and a bottleneck with 512 features. The concept prediction network is replaced with a RegNet (Xu et al., 2022), and the final classification network is a fully connected network with 3 layers, each with batch normalization and dropout. The three blocks are trained independently using the Dice, Huber and Cross Entropy loss, respectively. The concept and final classification predictors are trained with predicted labels from the segmentation and concept networks as input, respectively.

### 2.3. MicroPCBM

To obtain an even more lightweight student model for real-time deployment, we trained a black-box MobileNetV2-based classifier (referred to as MicroPCBM) using knowledge distillation from the pretrained PCBM teacher.

MobileNetV2 (Sandler et al., 2018)is a highly efficient convolutional architecture designed for mobile and edge deployment, using inverted residual blocks and depthwise separable convolutions to drastically reduce computational cost and parameter count while maintaining strong feature representation. This makes it an ideal student model for knowledge distillation in real-time fetal ultrasound applications, as it can inherit the teacher's diagnostic behaviour while achieving fast inference on resource-constrained ultrasound hardware.

The student was supervised both by the one-hot original labels but also by the teacher's softened logit predictions supplied by the teacher. This should enable the student to retain the hierarchical concept-based decision patterns from the PCBM, enabling the student to inherit clinically meaningful behaviour without needing to actually carry out segmentation or concept prediction.

## 3. Experimental Design

### 3.1. Datasets

Three fetal ultrasound datasets with differing image characteristics and class distributions were used to evaluate model generalisation. The primary training dataset was a private clinical set consisting of 3rd trimester growth scans collected from Danish hospitals (referred to as "Denmark" dataset). It includes 7 anatomical categories, separating standard-plane (SP) and non-standard-plane (NSP) views for Femur, Abdomen and Head, alongside an "Other" class representing non-diagnostic or poorly aligned views. As shown in Table 1, the dataset is notably imbalanced, with "Other" making up roughly 65% of the total samples. This reflects realistic scanning conditions but poses a challenge for minority class learning and fairness in evaluation. Denmark dataset was split 70:15:15; Training, Validation, Test.

Generalisation robustness was assessed using two public datasets with 4 clinically relevant anatomical classes are represented: Femur, Abdomen, Head, and Other. These datasets are not annotated with the additional SP/NSP plane quality label. The Fetal Plane

Database dataset (Burgos-Artizzu et al., 2020) , henceforce referred to as the "Barcelona" dataset originates from ultrasound examinations carried out in clinics in Barcelona, while the "Africa" dataset (Bano et al., 2023) consists of clinical ultrasound recordings from Egypt, Algeria, Uganda, Ghana and Malawi. These two datasets thus represent increasing geographical, sonographer and scanner diversity . These datasets contain different prevalence rates per class (see Table 2), enabling evaluation of performance under domain shift and reduced class granularity. The Africa dataset label Fetal Thorax was renamed "Other" for consistency with the other datasets.

Table 1: Class distribution for the Denmark dataset used to train MiniPCBM and MicroPCBM.

| **Class** | **Train** | **Val** | **Test** | **Total** |
|---|---|---|---|---|
| Femur SP | 208 | 69 | 69 | 356 |
| Abdomen SP | 133 | 44 | 44 | 221 |
| Head SP | 168 | 56 | 56 | 280 |
| Femur NSP | 253 | 84 | 85 | 422 |
| Abdomen NSP | 546 | 183 | 182 | 911 |
| Head NSP | 883 | 295 | 294 | 1472 |
| Other | 4245 | 1416 | 1415 | 7076 |
| **Total** | **6436** | **2146** | **2146** | **10728** |

Table 2: Class distribution for the Barcelona and Africa datasets used for external-domain evaluation.

| **Class** | **Barcelona** | **Africa** |
|---|---|---|
| Femur | 1040 | 125 |
| Abdomen | 711 | 125 |
| Head | 3092 | 125 |
| Other | 4213 | 75 |
| **Total** | **9056** | **450** |

### 3.2. Model selection and training

#### 3.2.1. MiniPCBM

MiniPCBM was trained end-to-end using the AdamW optimiser with a learning rate of $5\times10^{-5}$ and a weight decay of $5\times10^{-5}$. Training employed a ReduceLROnPlateau scheduler with a patience of five epochs, a batch size of 64, and a maximum of 75 epochs with early stopping based on validation accuracy. The best-performing checkpoint, as measured on the validation set, was retained for all subsequent evaluation experiments.

#### 3.2.2. **MicroPCBM**

Training combined the standard cross-entropy classification loss with respect to ground truth labels with a Kullback–Leibler (KL) divergence term matching the student's predictions to the PCBM's softened output distribution. KL divergence leverages class similarity structure that is not present in hard labels, promoting smoother decision boundaries and higher robustness to anatomical ambiguity. We performed a structured hyperparameter search to optimise MicroPCBM knowledge distillation, resulting in learning rate 0.002, momentum 0.85, weight decay 0.001, alpha 0.4, temperature 3.0, and label smoothing 0.1.

### 3.3. Performance Evaluations

All models were evaluated under a consistent protocol to assess diagnostic performance and robustness to domain shift. First, the PCBM teacher, MiniPCBM, and MicroPCBM were evaluated on the Denmark dataset using the full 7-class anatomical scheme, reflecting clinical differentiation of standard and non-standard planes. To test generalisation, the same models were then assessed on Denmark, Barcelona, and Africa in a 4-class setting (Femur, Abdomen, Head, Other), a predictive task aligned with the common clinical task of automatically verifying growth examination completeness. Finally, a binary Head vs. Non-Head evaluation, which is clinically important in determining fetal orientation (Wiśniewski et al., 2025) as real-time support for inexperienced ultrasound operators, isolated the most critical screening task.

Performance was reported using test accuracy, weighted-sensitivity, and weighted-specificity to ensure fair assessment under class imbalance, while confusion matrices were used to interpret error patterns relative to anatomical similarity.

## 4. Results

### 4.1. 7 Class Anatomy Classification on in-distribution dataset

Per-class ROC–AUC scores on the Denmark dataset (Table 3) demonstrate that all models provide highly reliable discrimination for the clinically required standard planes. MiniPCBM shows consistently strong performance, achieving AUC values above 0.97 for both Femur SP and Head SP classes. Notably, the teacher PCBM performs extremely poorly on the ambiguous Femur NSP category (AUC = 0.34), whereas MiniPCBM improves to 0.99 and MicroPCBM reaches 0.99. These results suggest that the segmentation-driven reasoning in the PCBM bottleneck may overfit to the structured anatomical criteria of standard planes, leading to unstable confidence when anatomy deviates from those expectations. In contrast, knowledge distillation encourages a smoother and more generalisable decision boundary in such challenging views.

These trends are reflected in the overall 7-class metrics (Table 3). MiniPCBM achieves the highest accuracy (0.93) and weighted sensitivity (0.93), while also maintaining the strongest weighted specificity (0.98). MicroPCBM provides competitive performance (accuracy = 0.90), despite being over 30× smaller in parameter count (Table 6) . Importantly, the preserved specificity across all three models ($\geq$0.93) indicates that false-positive misclassification of non-target anatomy remains rare. Together, these results confirm that the distilled student networks retain the reliability of the concept-based teacher while offering

Table 3: Performance on the Denmark dataset: per-class AUC-ROC and overall classification metrics.

| Metric / Class | PCBM | MiniPCBM | MicroPCBM |
|---|---|---|---|
| **Overall Performance** | | | |
| Accuracy | 0.90 | 0.93 | 0.90 |
| Sensitivity (Weighted Avg) | 0.90 | 0.93 | 0.90 |
| Specificity (Weighted Avg) | 0.97 | 0.98 | 0.93 |
| **Per-class AUC-ROC** | | | |
| Femur SP | 0.9859 | 0.9938 | 0.9906 |
| Abdomen SP | 0.9831 | 0.9768 | 0.9633 |
| Head SP | 0.9028 | 0.9562 | 0.9451 |
| Femur NSP | 0.3401 | 0.9943 | 0.9878 |
| Abdomen NSP | 0.9932 | 0.9892 | 0.9627 |
| Head NSP | 0.9689 | 0.9913 | 0.9872 |
| Other | 0.9967 | 0.9988 | 0.9867 |

improved robustness in non-standard anatomical presentations — a critical requirement for real-time scanning workflows where suboptimal views occur frequently.

### 4.2. Four-Class Anatomy Classification: Generalisation to external datasets

To further assess clinical utility across a broader diagnostic workflow, we evaluate the models on a four-class task distinguishing Femur, Abdomen, Head, and Other (Table 4).PCBM achieved 0.97 accuracy on the Denmark dataset, while MicroPCBM achieved 0.95 accuracy. Performance decays in Barcelona and Africa are consistent with the scarcity of standard-plane classes during student training. Notably, MiniPCBM performs comparatively well on these datasets despite lacking distilled teacher knowledge, suggesting that balancing capacity with reduced exposure to biased training priors may better support multi-class generalisation when distributions mismatch.

Table 4: Anatomy Classification Performance (Femur, Abdomen, Head, Other) across datasets.

| | **Denmark** | | | **Barcelona** | | | **Africa** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Metric** | PCBM | Mini | Micro | PCBM | Mini | Micro | PCBM | Mini | Micro |
| Accuracy | 0.97 | 0.93 | 0.95 | 0.86 | 0.77 | 0.76 | 0.72 | 0.68 | 0.54 |
| Sensitivity (Wt Avg) | 0.98 | 0.99 | 0.96 | 0.88 | 0.78 | 0.76 | 0.73 | 0.68 | 0.50 |
| Specificity (Wt Avg) | 0.97 | 0.99 | 0.97 | 0.97 | 0.88 | 0.87 | 0.78 | 0.83 | 0.75 |

### 4.3. Head vs. Non-Head Classification: Generalisation to external datasets

We first evaluate binary head versus non-head classification, which reflects the most clinically relevant task for ensuring correct acquisition of cranial standard planes during screen-

ing. Table 5 summarises results across the three imaging domains. MicroPCBM matched PCBM on the Denmark cohort with 0.99 accuracy and 1.00 sensitivity. This strong performance is expected given that the Denmark dataset used for training MicroPCBM contains a very large proportion of the "Other" class ( 65% of all images), allowing the model to reliably distinguish head from non-head views despite class imbalance. Under cross-domain shift, MicroPCBM achieves 0.94 accuracy on the Africa dataset, but a noticeably higher rate of false positives on the "Other" category. Since "Other" images form a minority class in the African cohort, this result indicates bias induced by the training distribution: the student model learns a prior that non-head views are common and therefore tends to overpredict this outcome in low-resource domains. On the Barcelona dataset, MicroPCBM achieved 0.91 accuracy, maintaining performance within 6% of the teacher despite increased variability in imaging protocols. Notably, MicroPCBM outperformed MiniPCBM on this challenging shift, demonstrating the contribution of inherited teacher knowledge to out-of-domain stability. On the most difficult setting, the Africa dataset, the student achieved 0.94 accuracy, representing a stronger balance between generalisation and efficiency than both alternatives. These results collectively indicate that the distilled model captures the essential discriminative cues used by the high-capacity teacher and retains them under cross-domain variation.

Table 5: Head vs Non-Head classification performance across datasets.

| | **Denmark** | | | **Barcelona** | | | **Africa** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Metric** | PCBM | Mini | Micro | PCBM | Mini | Micro | PCBM | Mini | Micro |
| Accuracy | 0.99 | 1.00 | 0.99 | 0.97 | 0.92 | 0.91 | 0.98 | 0.94 | 0.90 |
| Sensitivity | 1.00 | 1.00 | 1.00 | 0.97 | 1.00 | 0.91 | 0.98 | 0.94 | 0.91 |
| Specificity | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.97 | 1.00 | 0.99 | 1.00 |

### 4.4. Inference Efficiency and Deployment Suitability

The computational complexity and inference latency of all models are presented in Table 6. PCBM contains 66M parameters and requires 42.91 ms per prediction, yielding only 23 frames per second (FPS), which is below real-time sonographic frame-rate requirements. By contrast, MicroPCBM reduces the parameter count to just 2M, representing a 33× compression. This directly translates to practical speed: MicroPCBM achieves 6.61 ms latency and 151 FPS on standard GPU hardware, exceeding real-time requirements and enabling continuous video-based decision support. MiniPCBM hit 94 FPS at higher parameter cost, highlighting MicroPCBM as the most efficient architecture-to-performance trade-off. These gains support deployment not only on workstation-grade hardware but on embedded ultrasound compute units with far more constrained resources.

## 5. Discussion

We demonstrate that knowledge distillation from an intrinsically explainable and high-quality PCBM can successfully produce lightweight classifiers that retain the PCBM's

Table 6: Model complexity and inference efficiency for PCBM and students.

| Model | # Params (M) | FLOPs (GMac) | Latency (ms) | FPS |
|---|---|---|---|---|
| PCBM | 66 | 94.0 | 42.91 | 23.31 |
| MiniPCBM | 13 | 14.4 | 10.58 | 94.48 |
| MicroPCBM | 2 | 0.33 | 6.61 | 151.28 |

diagnostic behaviour while improving computational efficiency. Both student variants—MiniPCBM and MicroPCBM—achieved strong classification accuracy on the Danish test set, with MicroPCBM matching PCBM performance (0.90 accuracy) despite a 33× reduction in parameters and a 7× reduction in inference latency. These findings indicate that the hierarchical knowledge encoded within PCBM logits can be effectively transferred without requiring intermediate concept prediction or segmentation branches at inference time.

Evaluation across three datasets further highlighted how the two distilled models offer complementary advantages. MiniPCBM, trained with both concept and classification supervision, showed stronger robustness in classes where spatial structure is distinctive, providing a compelling compromise between interpretability and efficiency. MicroPCBM, a black-box model distilled emd-toend solely from softened class logits, achieved the fastest inference speed and remained competitive across domain shifts—although a tendency to mislabel ambiguous views as "Other" in the Africa dataset reflected training-set class imbalance. However, for the clinical task of supporting navigation by detecting fetal orientation, even the MicroPCBM obtained both accuracy, sensitivity and specificity $\geq 0.9$ even on the Africa dataset, indicating its potential utility for navigational support across locations.

Our observations suggest that the choice of distillation strategy enables tuning the deployment trade-off: The fastest inference speed is obtained by **MicroPCBM**: optimised for *real-time clinical scanning*, where speed is critical, but with reduced performance on the more fine grained task when moving to unseen populations. **MiniPCBM**, on the other hand, retained the full concept-based explainability from PCBM, with a smaller drop in performance on unseen data, and with a smaller reduction in computational speed. For situations such as training, or educational tools, where interpretability and error analysis are essential, the Mini-PCBM offers a strong alternative to the full PCBM. On in-distribution data or for the simpler task of head detection, the Micro-PCBM offers the most significant speed-up, at a limited cost in performance.

We see an interesting effect on generalizability: While both distilled models perform very well on in-distribution data, the cost of distillation is clearer when moving to unseen datasets, and the performance drop correlates with the reduction in complexity. It remains an open question whether the main contributing factor to the performance difference between Mini-PCBM and Micro-PCBM comes solely from the reduction in model complexity, or also from the retained concept guidance in training the Mini-PCBM.

Overall, however, these results show that progressive concept-based reasoning can be effectively compressed into lightweight, deployable models while retaining strong classification performance and explainable behaviour. The proposed two-track distillation strategy therefore provides practical flexibility for integrating reliable real-time AI support into both point-of-care ultrasound devices and medical simulation workflows.

## Acknowledgments

This work is supported by the Independent Research Fund Denmark (EASE project nr 4264-00151A), the Danish Pioneer Centre for AI (DNRF grant number P1) and SONAI - a Danish Regions' AI Signature Project.

## References

Sophia Bano, Lubomir Hadjiiski, Efthymia Karampati, Henri Ardon, Aris Papageorghiou, and J. Alison Noble. Fetal ultrasound standard plane images dataset (africa), 2023. URL https://doi.org/10.5281/zenodo.7540447. Dataset.

Zahra Bashir et al. Clinical validation of explainable ai for fetal growth scans through multi-level, cross-institutional prospective end-user evaluation. *Scientific Reports*, 2025. in press.

Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017. doi: 10.1109/TMI.2017.2712367.

Christian F Baumgartner et al. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. *MICCAI Workshop on Perinatal, Preterm and Paediatric Image Analysis*, 2016.

Xavier Burgos-Artizzu, David Coronado-Gutierrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fàtima Crispi, and Eduard Gratacós. Common maternal-fetal ultrasound images, 2020. URL https://doi.org/10.5281/zenodo.3904279. Dataset.

Jacopo Dapueto, Luca Zini, and Francesca Odone. Knowledge distillation for efficient standard scanplane detection of fetal ultrasound. *Medical & Biological Engineering & Computing*, 62(8):1801–1814, 2024. doi: 10.1007/s11517-023-02881-4.

Maria Chiara Fiorentino et al. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, 83:102640, 2023. doi: 10.1016/j.media.2022.102640.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

Manxi Lin, Aasa Feragen, Zahra Bashir, Martin Grønnæk Tolsgaard, and Anders Nymark Christensen. I saw, i conceived, i concluded: Progressive concepts as bottlenecks. *arXiv preprint arXiv:2211.10630*, 2022.

Manxi Lin, Kilian Zepf, Anders Nymark Christensen, Zahra Bashir, Morten Bo Søndergaard Svendsen, Martin Tolsgaard, and Aasa Feragen. Dtu-net: Learning topological similarity for curvilinear structure segmentation. In *International Conference on Information Processing in Medical Imaging*, pages 654–666. Springer, 2023.

Sritam Patra et al. Efficient ultrasound image analysis models with sonographer gaze assisted distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 952–953, 2020.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

Jakub Maciej Wiśniewski, Anders Nymark Christensen, Mary Le Ngo, Martin Grønnebæk Tolsgaard, and Chun Kit Wong. Determining fetal orientations from blind sweep ultrasound video. In *Scandinavian Conference on Image Analysis*, pages 254–263. Springer, 2025.

Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9562–9567, 2022.