

Evaluating the impact of measurement calipers on deep learning-based ovarian cancer diagnosis: a prospective, multicentre analysis

Axel Geysels, Giulia Garofalo, Stefan Timmerman, Francesca Buonomo, Daniela Fischerová, Antonia Carla Testa, Francesca Moro, Lil Valentin, Povilas Sladkevicius, Ligita Jokubkiene, Marek Kudla, Caroline Van Holsbeke, Artur Czekierdowski, Elisabeth Epstein, Yvette Groszmann, Bart De Moor, Ben Van Calster, Wouter Froyman, Dirk Timmerman

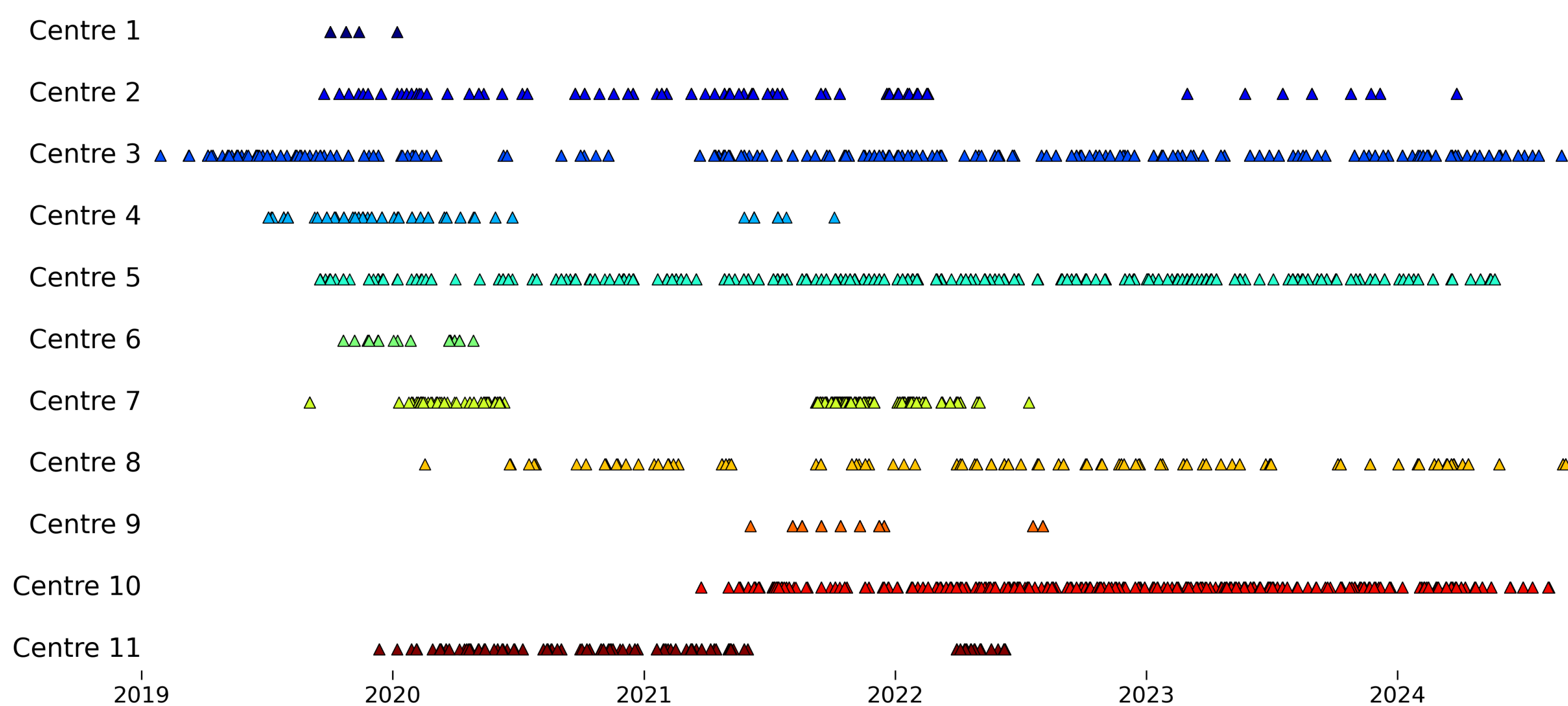
Objective

To determine whether measurement calipers in transvaginal ultrasound images can bias the performance of AI models in ovarian-cancer diagnosis

Design

- IOTA-AI: Prospective international observational study (2019–present)
- 1457 patients (37% malignant) with 8824 transvaginal 2D ultrasound images, all acquired on GE scanners
- 2790 images contained visible calipers (1034 patients, 42% malignant).

Centre contribution over time

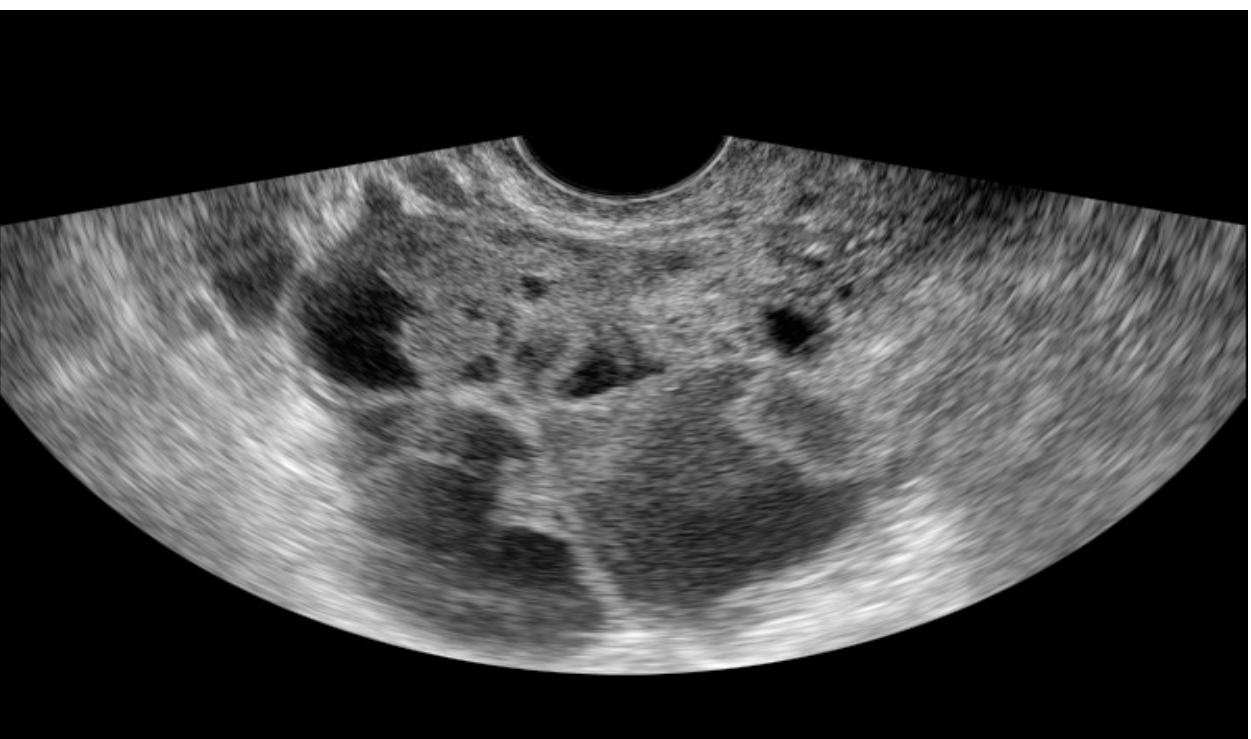


Contact:
axel.geysels@esat.kuleuven.be

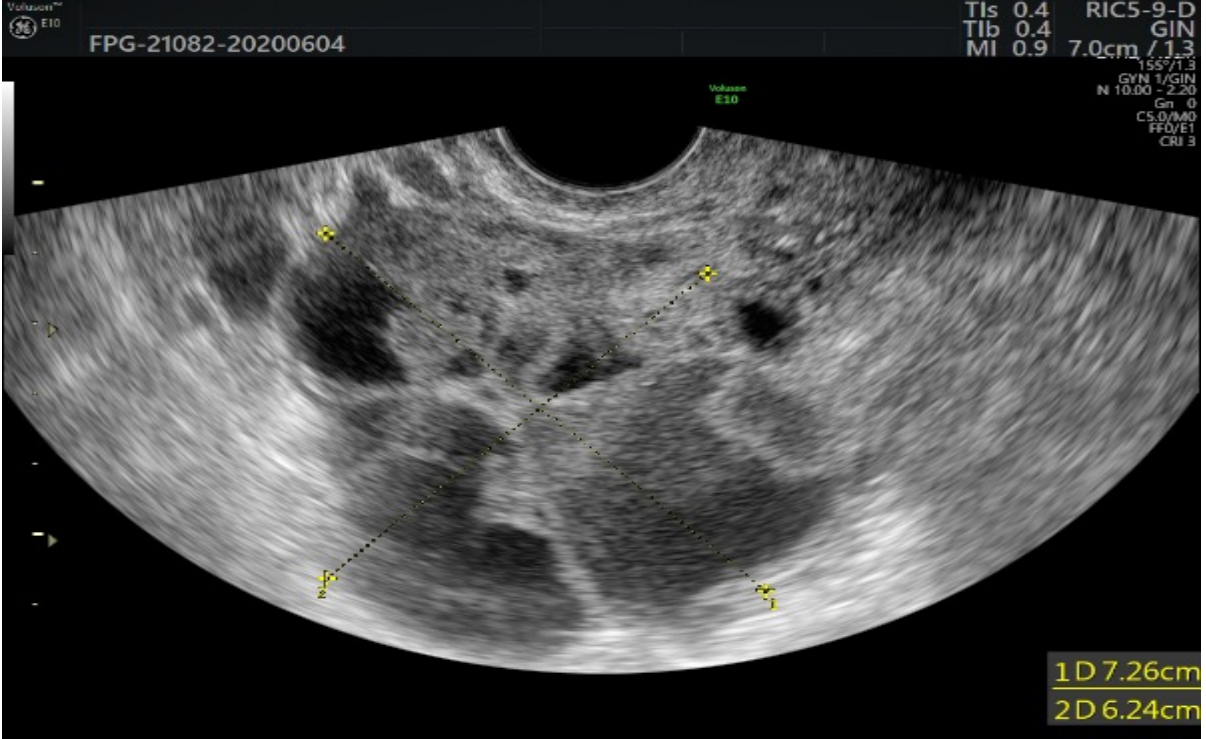
Method

Creation of two datasets:

Raw (without calipers)



Overlay (with calipers)



Assess four train → test scenarios (Raw→Raw, Overlay→Overlay, Raw→Overlay, Overlay→Raw):

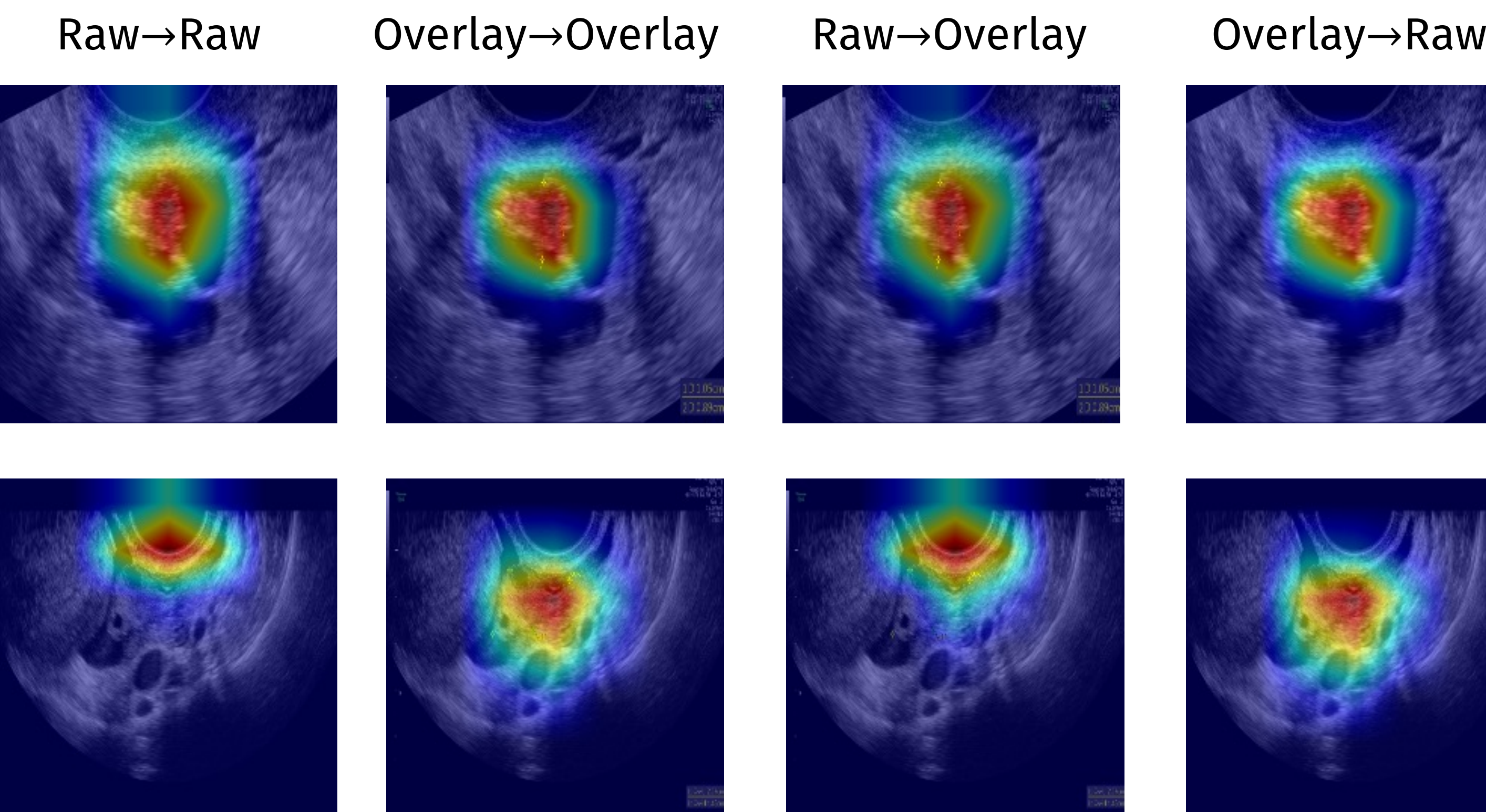
- Quantitative: upper bound AUC for diagnostic discrimination
- Qualitative: representative saliency maps showing model focus

Results (quantitative)

Upper bound AUC (95% CI)	Evaluated on	
	Raw	Overlay
Trained on Raw	0.909 (0.893-0.925)*	0.907 (0.891-0.923)
Trained on Overlay	0.913 (0.899-0.928)	0.914 (0.900-0.929)*

*p=0.26 (paired DeLong)

Results (qualitative)



- Overlay on/off leaves saliency maps mostly unchanged: Raw→Raw ≈ Raw→Overlay and Overlay→Overlay ≈ Overlay→Raw

Conclusions

- **AUC:** no significant change in training with or without calipers.
- **Saliency:** heatmap shifts reflect training noise more than caliper cues.
- **Future research:** validate using images from machines of different ultrasound vendors.

KU LEUVEN

