ISUOG 2025 Abstract Submission

Title: Evaluating the impact of measurement calipers in AI-based ovarian cancer diagnosis: a prospective international multicenter analysis

Authors: 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, Dirk Timmerman, Wouter Froyman

Objectives: To assess whether overlay annotations – particularly measurement calipers on ultrasound images – influence the performance of a deep learning classification model for diagnosing ovarian cancer.

Methods: We used data from IOTA7-AI, a prospective multicenter study including patients with ovarian masses from 2019 to 2024. Ultrasound images were collected at inclusion, and histological outcomes were recorded post-surgery. Two datasets were created: an "overlay dataset" (with measurement calipers, machine settings, and textual annotations) and a "raw dataset" (ultrasound only). We trained and validated separate deep learning models on each dataset – both within and across datasets – using a uniform five-fold cross-validation procedure. Each model's maximum validation performance was recorded, and all predictions were pooled to gauge potential performance gains. Diagnostic performance was evaluated via the area under the receiver operating characteristic curve (AUC), DeLong's test was used for comparisons.

Results: We analyzed 8824 ultrasound images from 1457 patients, of which 543 (37%) had malignant outcomes. Among these, 2790 images (32%) contained visible calipers (1034 patients, 42% malignant). On the overlay dataset, the model's patient-level AUC was 0.914 (95% CI, 0.900-0.929), compared with 0.909 (0.893-0.925) on the raw dataset (p=0.26). Across datasets, the AUC remained stable at 0.913 (0.889-0.927) when training on "overlay" and testing on "raw" and 0.907 (0.889-0.923) in the reverse scenario. Restricting the analysis to images containing calipers showed a similar trend but slightly lower AUC values.

Conclusions: Including calipers offered a small, non-significant AUC improvement over training on raw images, yet the model appeared to use these overlays more as supplemental guidance than a single confounding cue – demonstrated by stable performance on raw images. Further research is warranted to confirm these findings across different ultrasound devices and patient populations.

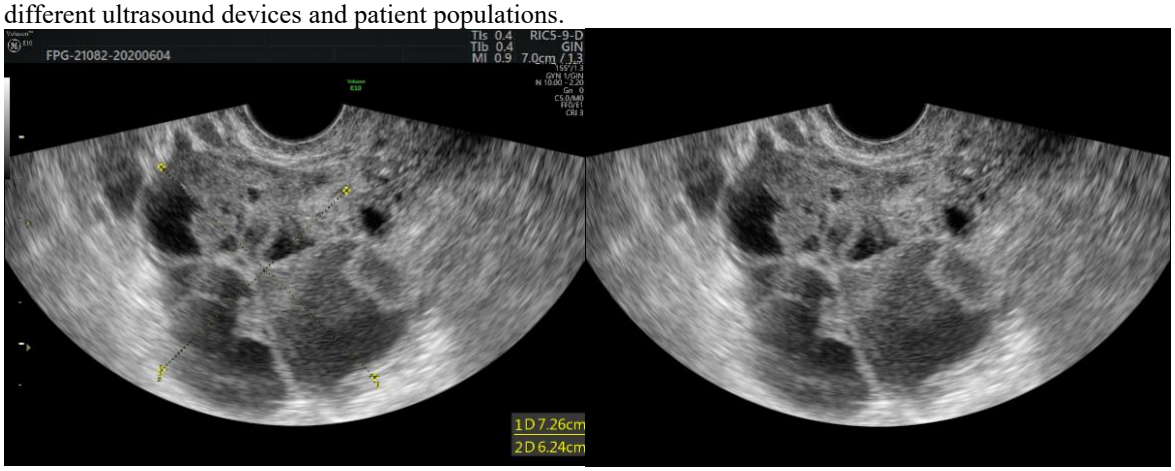

Figure 1*. (Left) An ultrasound image from the "overlay dataset," displaying measurement calipers. (Right) The corresponding image in the "raw dataset," revealing only the underlying ultrasound content.

*The figure will not be included in the abstract submission.