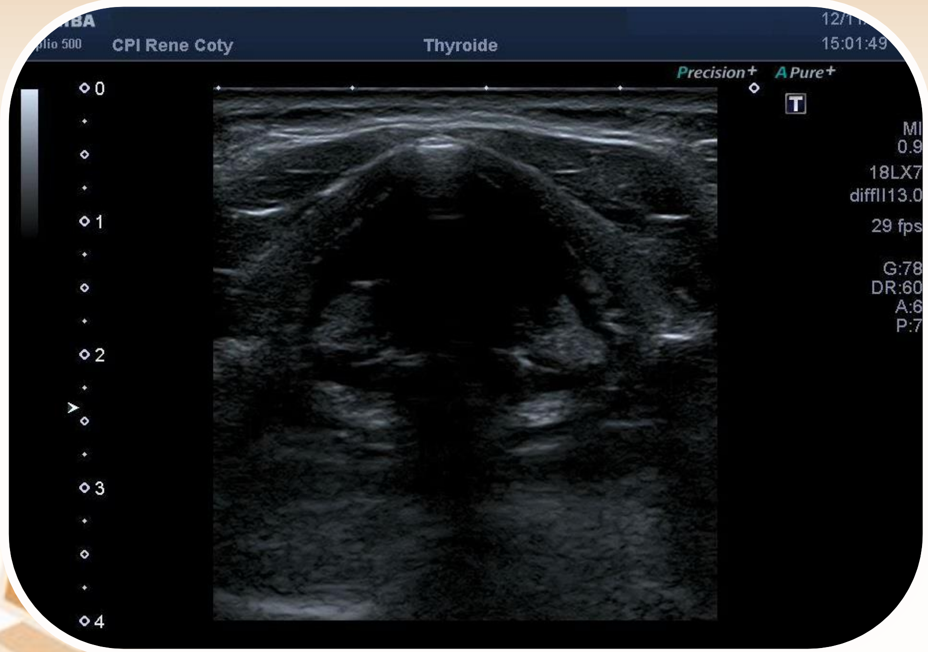
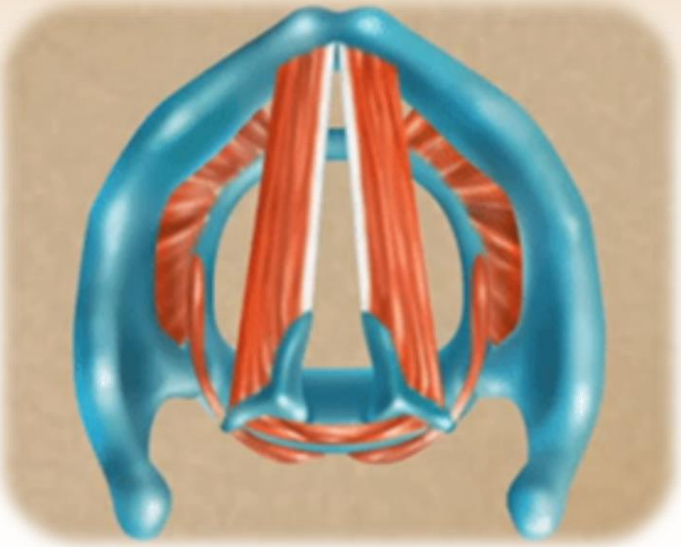
Laboratoire d'Imagerie Biomédicale

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : électrons, photons, sciences du vivant (EOBE)

VOCALISE : Non-invasive Measurement of Vocal Fold Motion based on Dynamic Translaryngeal Ultrasound Imaging



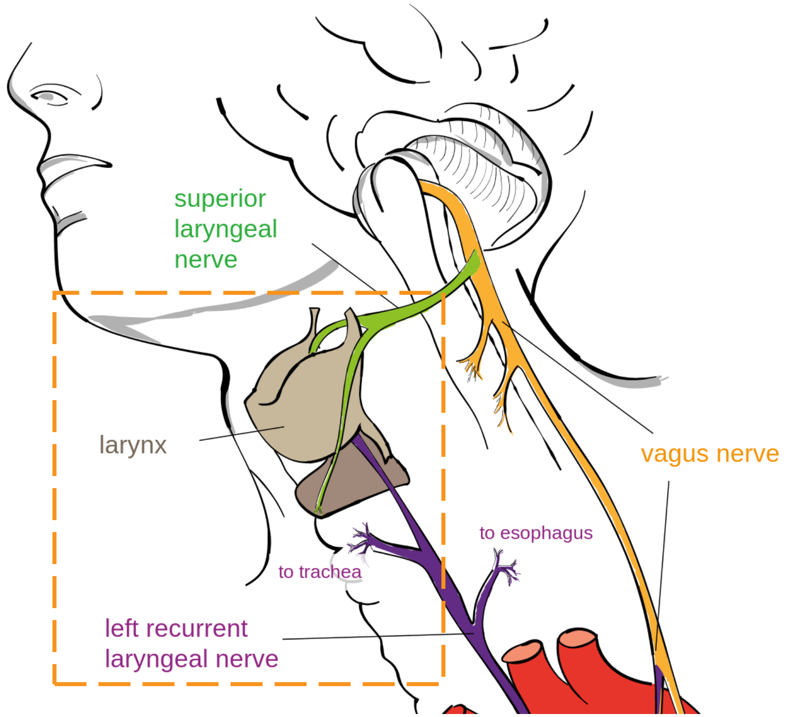


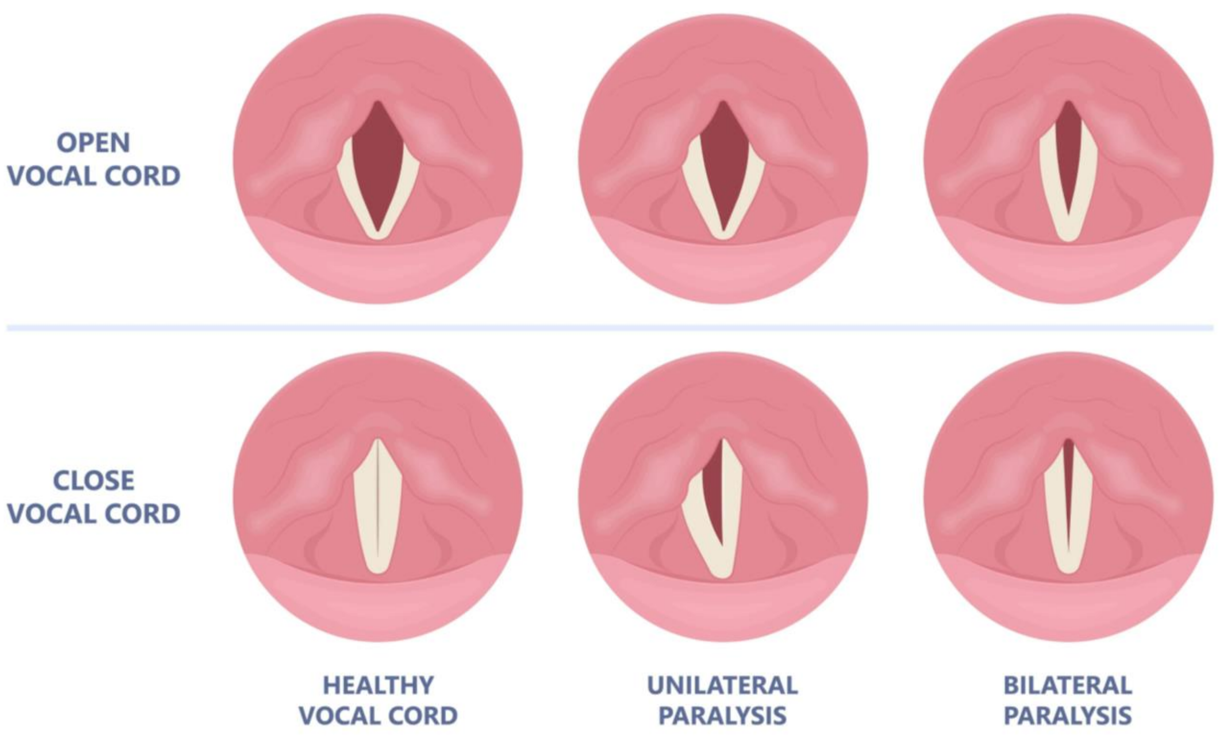
Presented by BUI Trung Kien

Research background

- In France, about 50,000 patients annually undergo thyroid/parathyroid surgery.
- Recurrent laryngeal nerve injury is one of the major complications, affecting the vocal cords

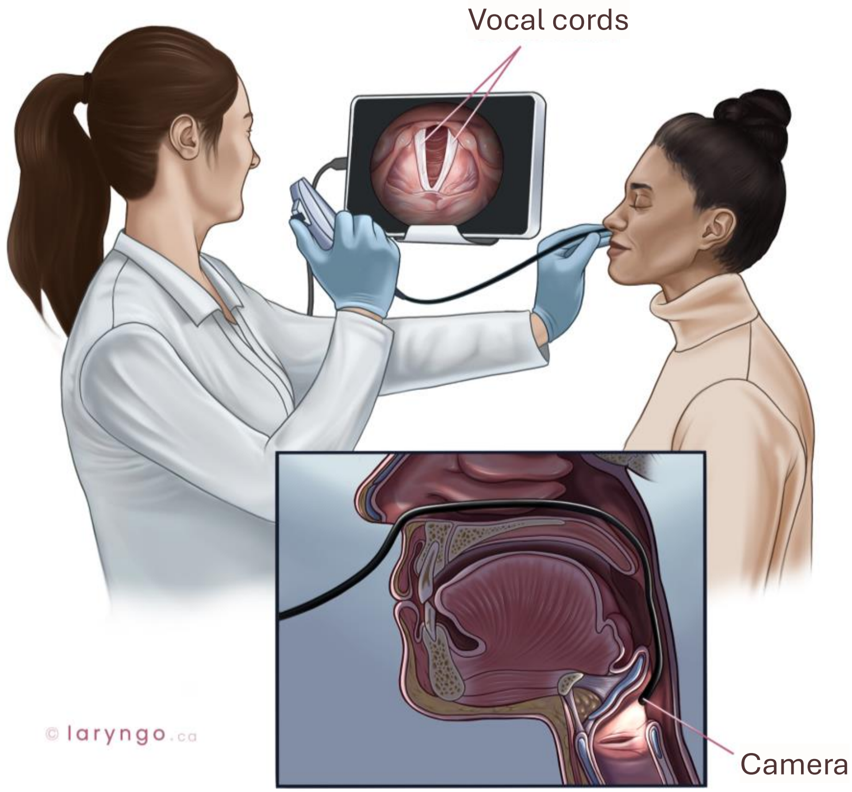
  ➔ Troubles in swallowing, breathing, phonation, etc.



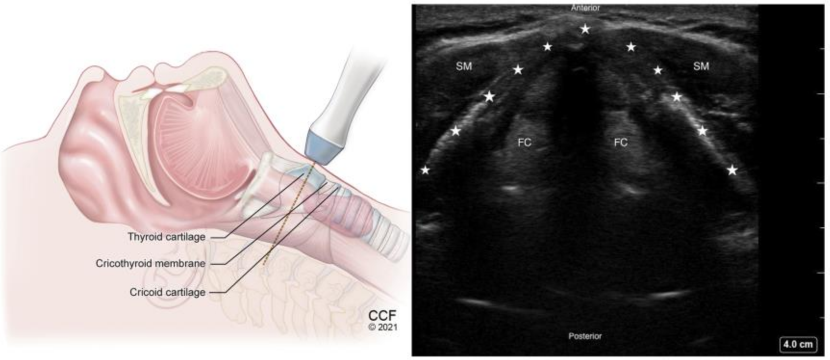


Research background

- Standard for the evaluation of the mobility of vocal cords: **Laryngoscopy**



- Alternative non-invasive method: **Translaryngeal ultrasound**

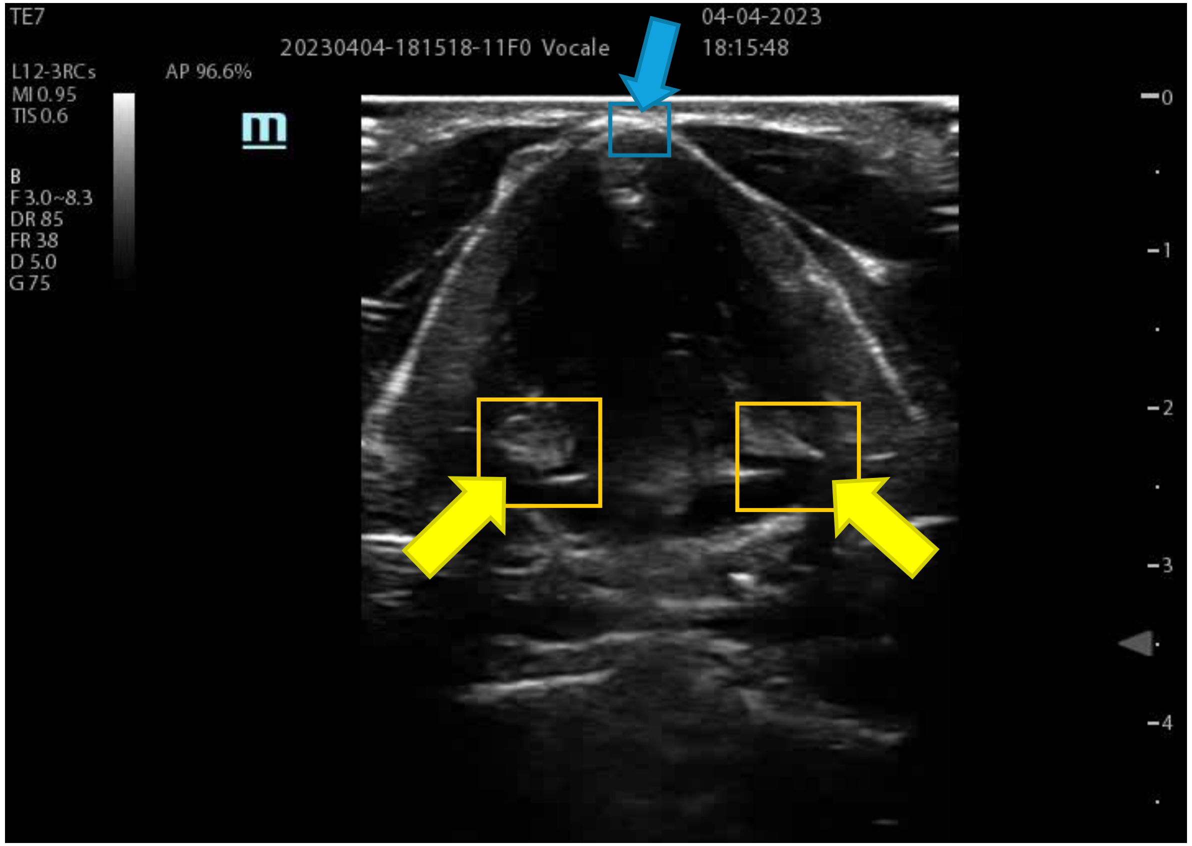


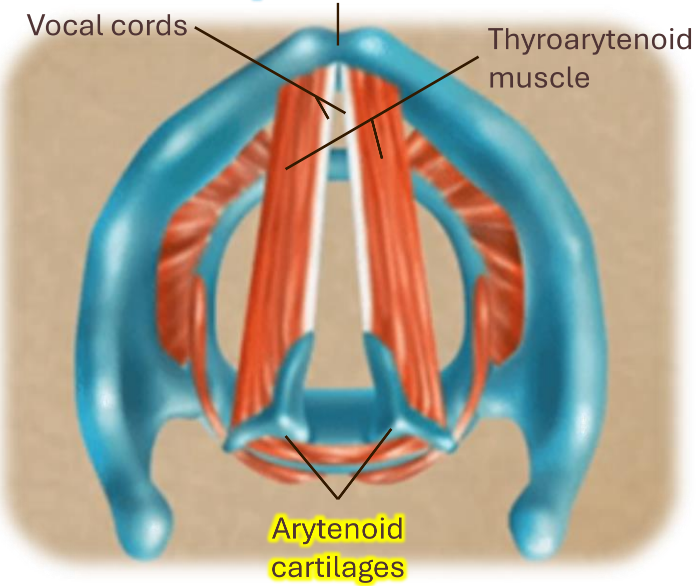
Thyroid cartilage inverted V-shaped, marked by stars ; SM - Strap muscle ; FC - False Cords

Research background

- Vocal fold motion tracking thanks to visible structures linked to vocal cords :

  The prominence of thyroid cartilage and two arytenoid cartilages





How vocal fold works

Objective of this study

Using machine learning approaches in the prediction of vocal fold paralysis
to support the therapist in the patient's voice rehabilitation after a neck surgery

- Detection and tracking of the vocal cord landmarks in ultrasound video
- Quantification of their motion to allow a diagnosis of vocal cords paralysis

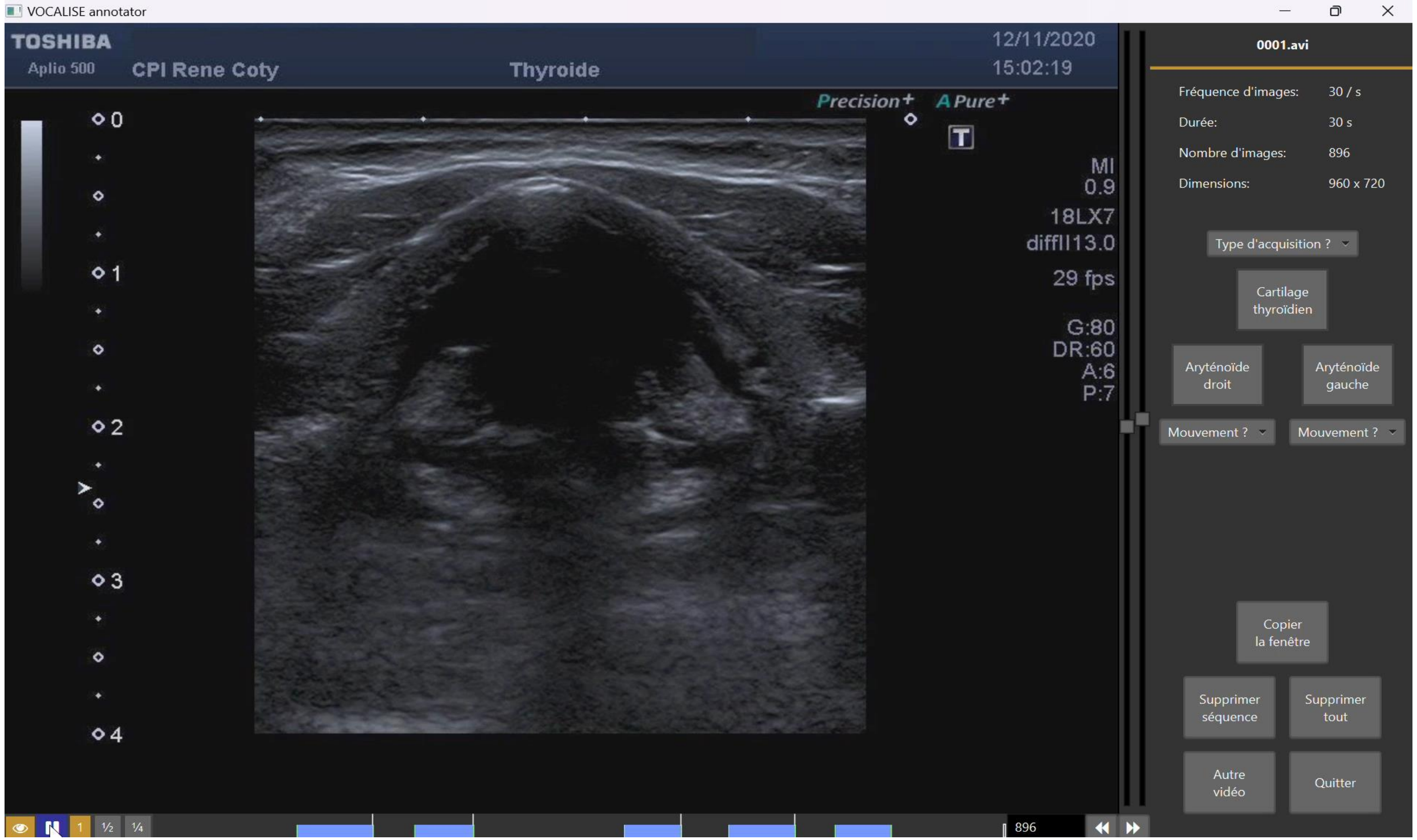
Data

- **3 retrospective datasets of ultrasound videos** of patients and healthy individuals
  - Underwent scans during free breathing and postoperatively for patients
  - Video duration of about 10-30 seconds with 30 frames per second
  - Acquired in 5 different French medical centers
  - Performed with 4 different ultrasound device constructors

Dataset	Number of individuals	Number of videos	Acquisition site					Ultrasound device		
			Chir. Endoc (PS)	IE3M (PS)	Med Nuc (PS)	CPI René Coty	Avicenne	UltraSonix	SSI (*different version)	Toshiba
BDD1	149	149	✓					✓		
BDD2	41	78		✓	✓				✓	
BDD3	67	67		✓	✓	✓	✓		✓*	✓

Detection and tracking of landmarks in ultrasound video

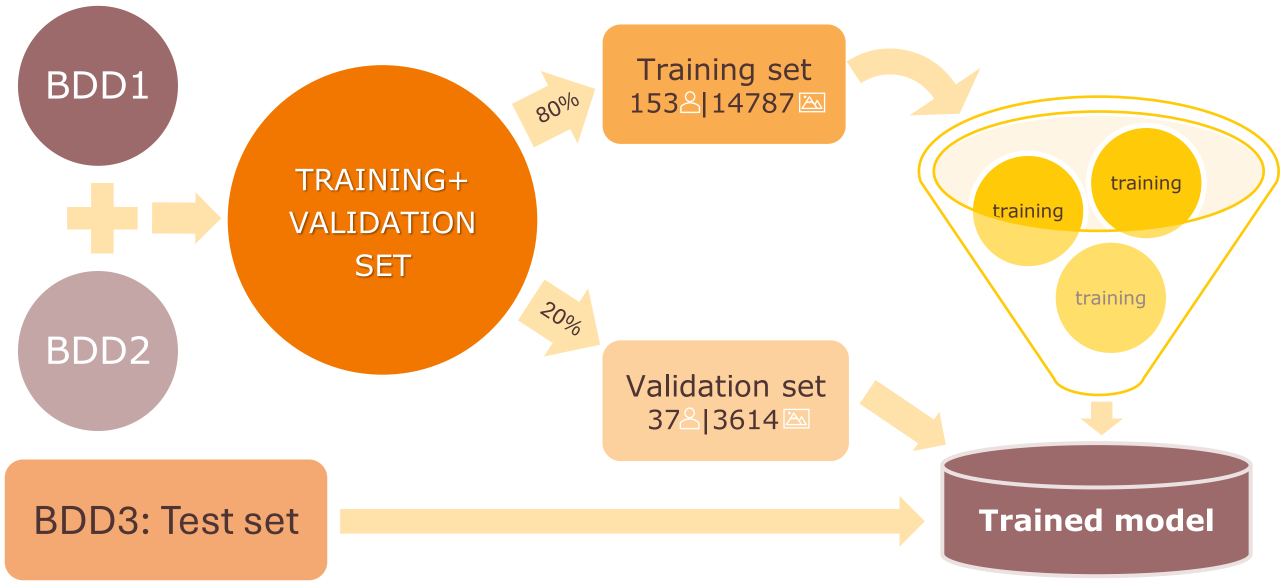
- Annotation of structures was semi-automatically done using the in-house software « VOCALISE annotator », with multiple 'closing–opening' motion subsequence.



Detection and tracking of landmarks in ultrasound video

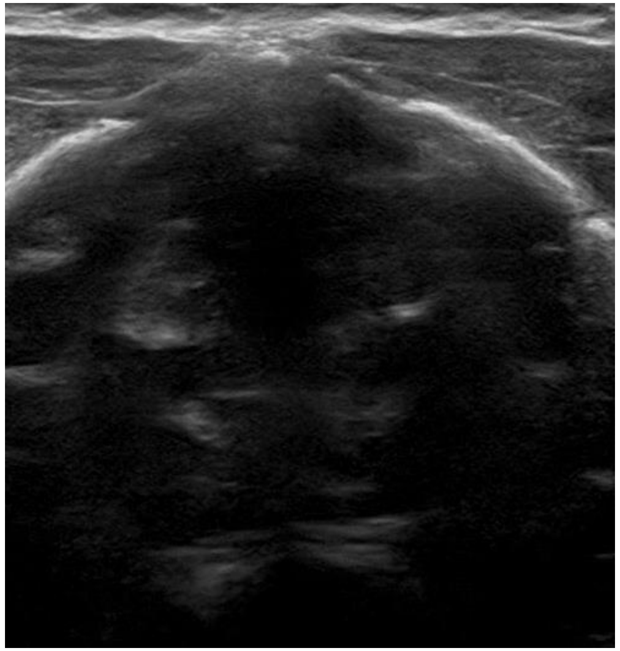
Dataset	Number of individuals	Number of videos	Number of annotations		Average image count per subsequence
			Sub-sequences	Images	
BDD1	149	149	194	8,067	42
BDD2	41	78	259	10,334	40
BDD3	67	67	161	8,115	50

Data splitting is based on the number of individuals to ensure that **all** images from an individual appear either in the training or validation sets.

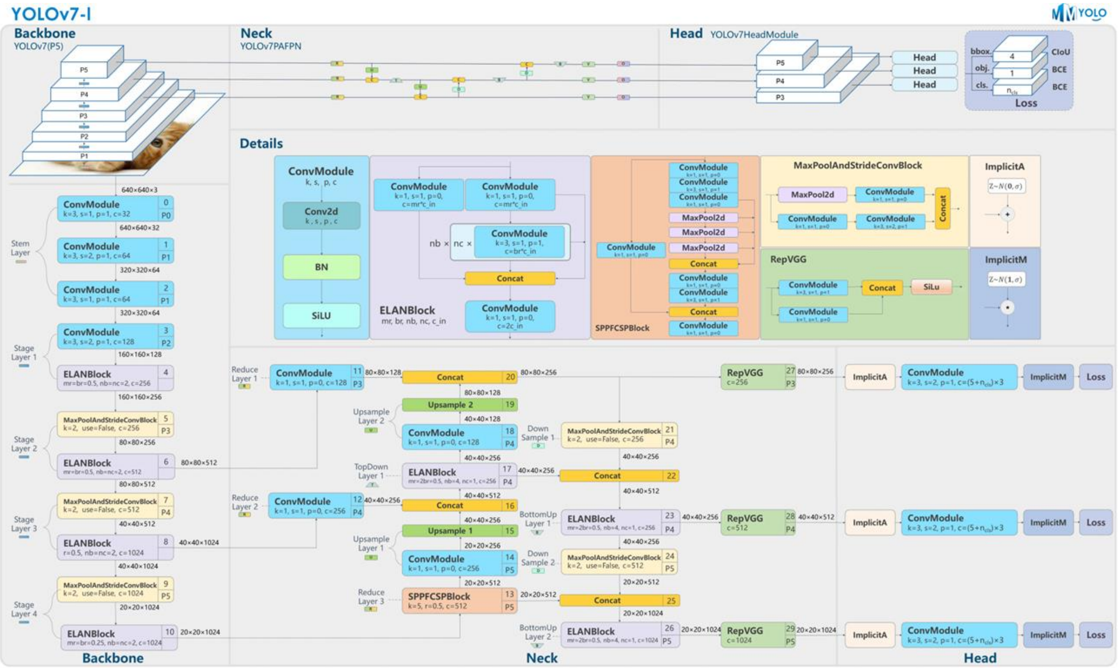


Detection and tracking of landmarks in ultrasound video

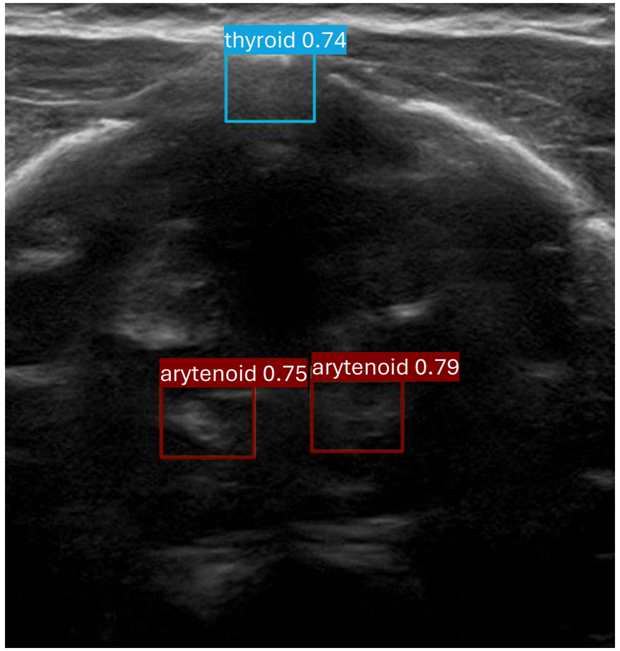
- Object detection model: **"You Only Look Once" (YOLO)**
  - State-of-the-art for real-time object detection tasks in multiple fields
  - To predict the location and the class of each object with a bounding box



Input



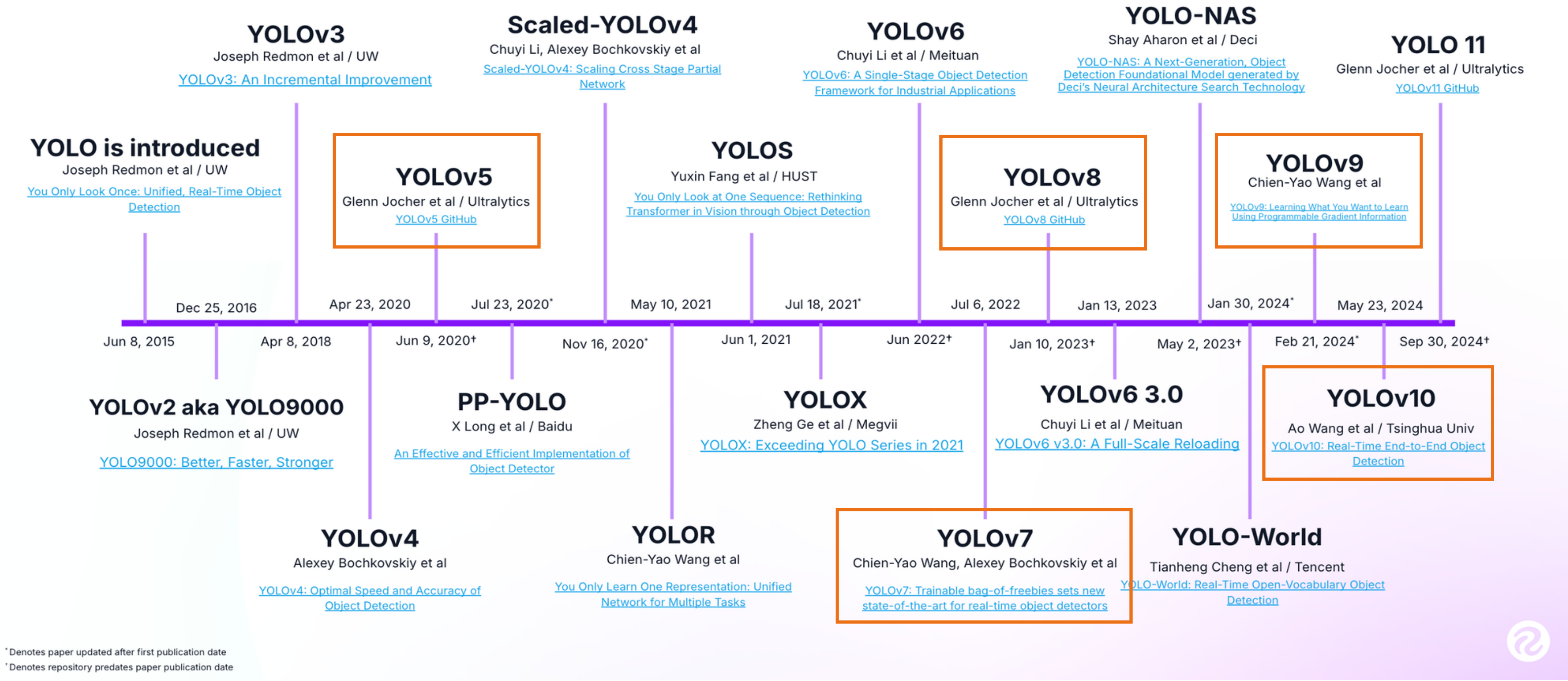
YOLOv7 architecture



Output

Detection and tracking of landmarks in ultrasound video

- Object detection model: **"You Only Look Once" (YOLO)**
  - For this study, we used different versions of YOLO.



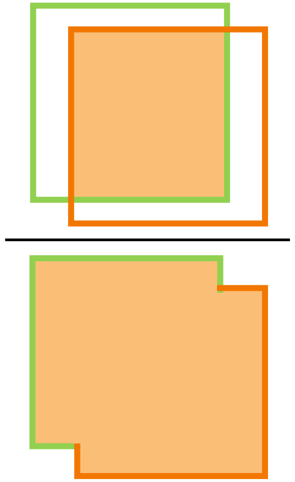
Detection and tracking of landmarks in ultrasound video

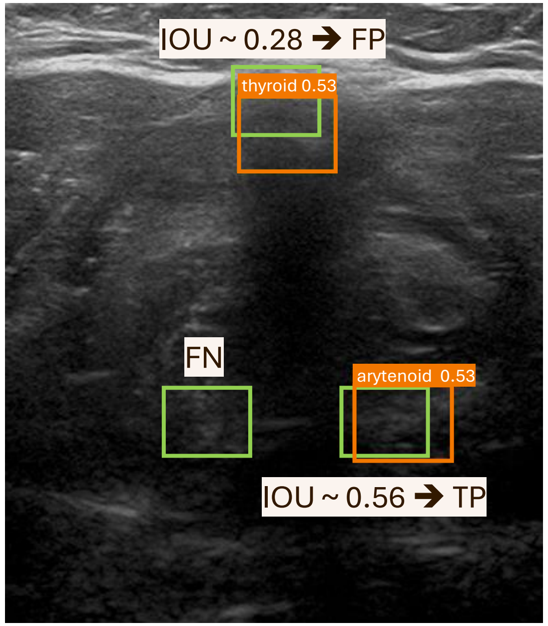
- **Performance evaluation metrics:**
  - Confidence score is calculated by C = Box confidence * Class confidence
    = (Objectness score * IoU) * Class confidence
  - Intersection over Union (IoU) evaluates object detection accuracy.

➔ These two metrics determine whether an output is correct or incorrect by the thresholds

C threshold = 0.25 ; IoU threshold = 0.50

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} =$$





ground truth box ; prediction box

Detection and tracking of landmarks in ultrasound video

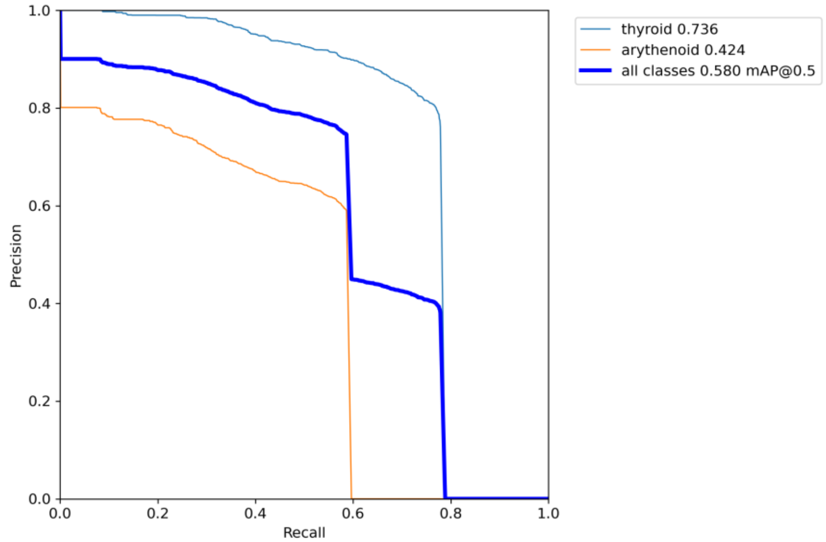
- **Performance evaluation metrics:**
  - Confidence score is calculated by C = Box confidence * Class confidence

    = (Objectness score * IoU) * Class confidence
  - Intersection over Union (IoU) evaluates object detection accuracy.

➔ These two metrics determine whether an output is correct or incorrect by the thresholds

  - True positive, false positive, false negative ➔ Precision, Recall, F1-score
  - Average precision at 50% IoU threshold (AP50) by varying the confidence score

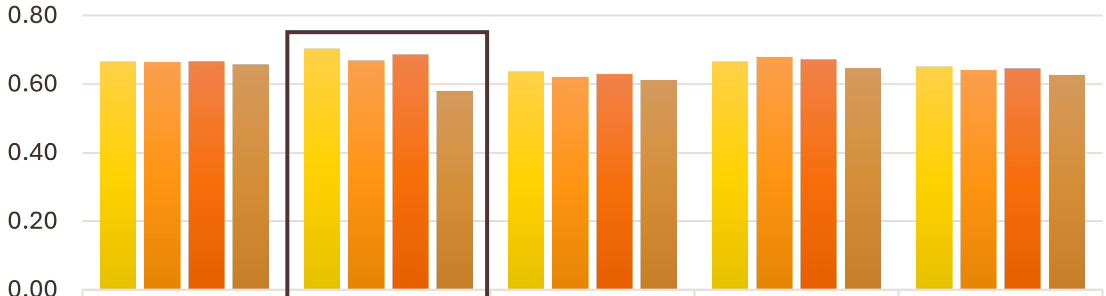
    ➔ AUC of the precision-recall curve.



Detection and tracking of landmarks in ultrasound video

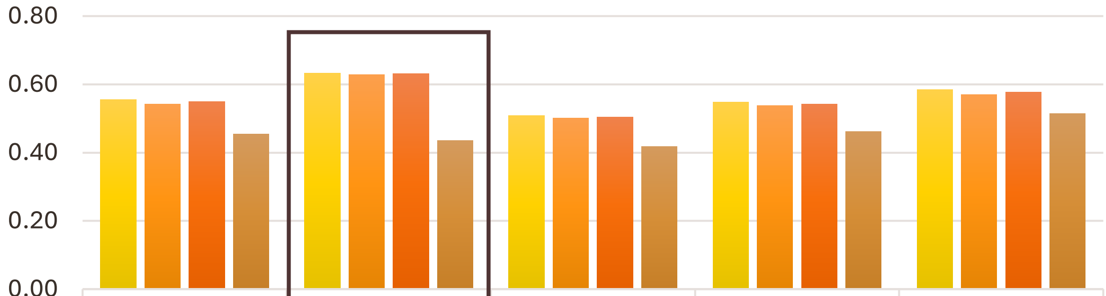
- **Comparison of performance by multiple YOLO models**
  - Performance scores were averaged across all classes.
  - Lower test set performance due to the variability of the validation and test data.
  - No single version demonstrates a clear performance advantage over the others.

Performance scores on VALIDATION set (n=3614)



	yolov5small	yolov7tiny	yolov8small	yolov9small	yolov10small
Precision	0.666	0.704	0.636	0.665	0.651
Recall	0.664	0.669	0.621	0.679	0.640
F1-score	0.665	0.686	0.628	0.672	0.645
AP50	0.657	0.580	0.611	0.646	0.626

Performance scores on TEST set (n=8115)



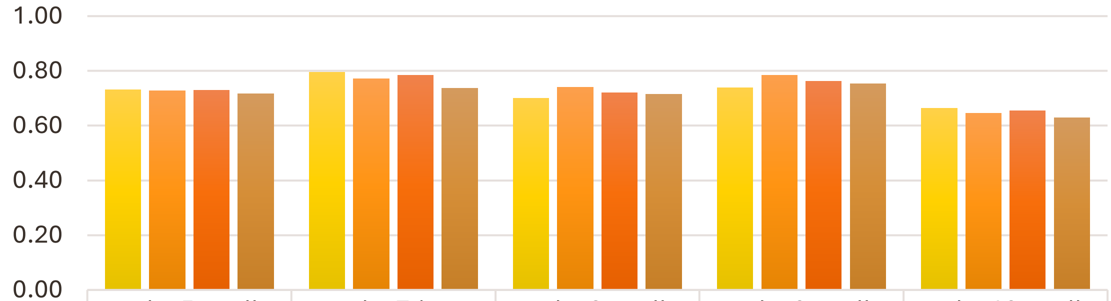
	yolov5small	yolov7tiny	yolov8small	yolov9small	yolov10small
Precision	0.555	0.633	0.509	0.548	0.585
Recall	0.543	0.629	0.501	0.538	0.570
F1-score	0.549	0.631	0.505	0.543	0.577
AP50	0.455	0.435	0.418	0.462	0.514

The YOLOv7-tiny model results have been maintained for this study thanks to its efficiency and model small size

Detection and tracking of landmarks in ultrasound video

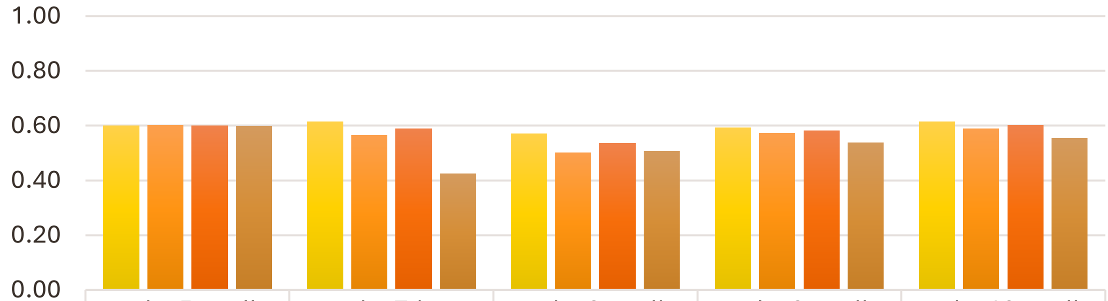
- **Comparison of performance by multiple YOLO models**
  - Higher performance scores in thyroid detection compared to arytenoid detection for all models.

Thyroid detection performance on VALIDATION



	yolov5small	yolov7tiny	yolov8small	yolov9small	yolov10small
Precision	0.731	0.795	0.701	0.738	0.664
Recall	0.728	0.772	0.740	0.785	0.646
F1-score	0.730	0.784	0.721	0.762	0.655
AP50	0.717	0.736	0.715	0.754	0.629

Arytenoid detection performance on VALIDATION

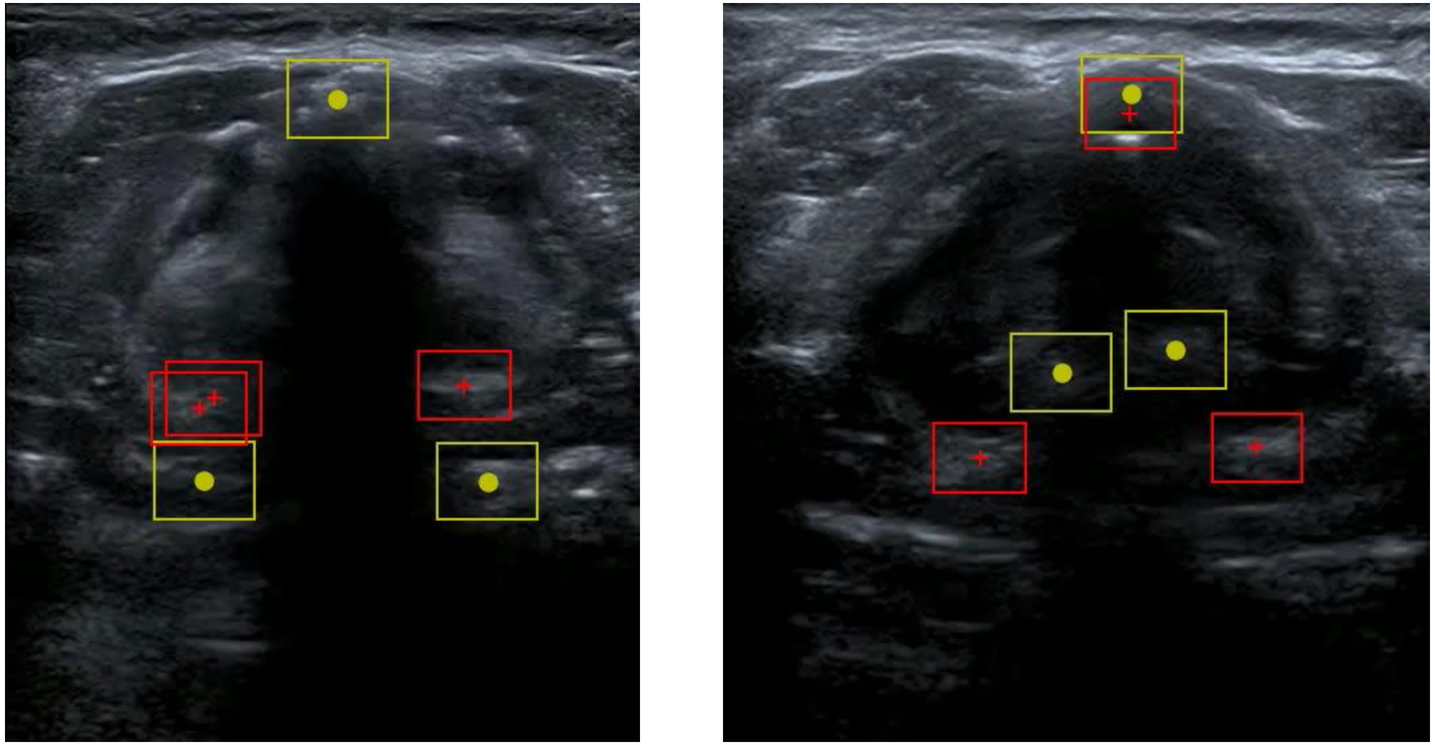


	yolov5small	yolov7tiny	yolov8small	yolov9small	yolov10small
Precision	0.600	0.614	0.570	0.592	0.615
Recall	0.601	0.565	0.502	0.572	0.588
F1-score	0.601	0.590	0.536	0.582	0.602
AP50	0.598	0.424	0.507	0.538	0.554

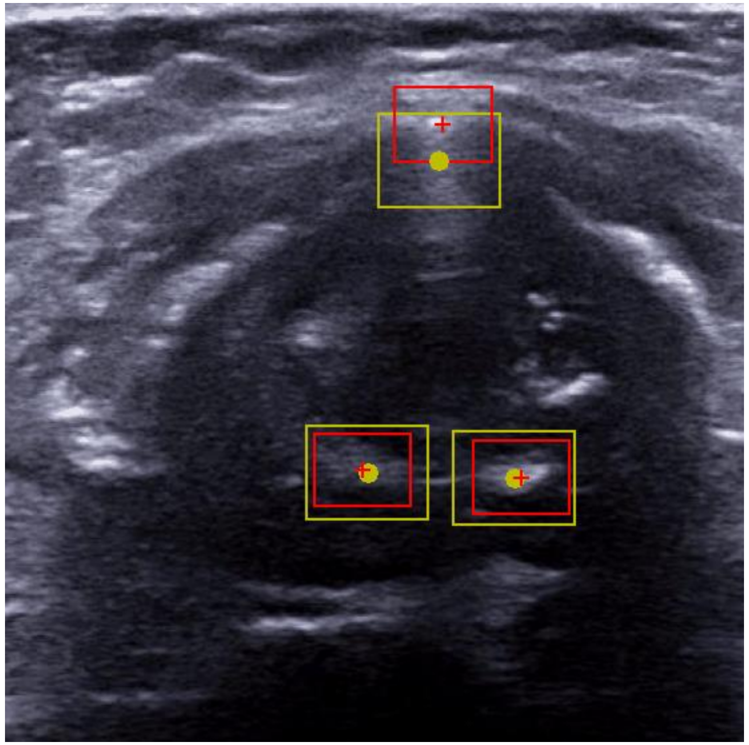
Detection and tracking of landmarks in ultrasound video

- Focusing on model output, arytenoid predictions are severely missing or misinterpreted as false positives due to low IoU with small box size.

Multiple predictions in a hyperechoic region



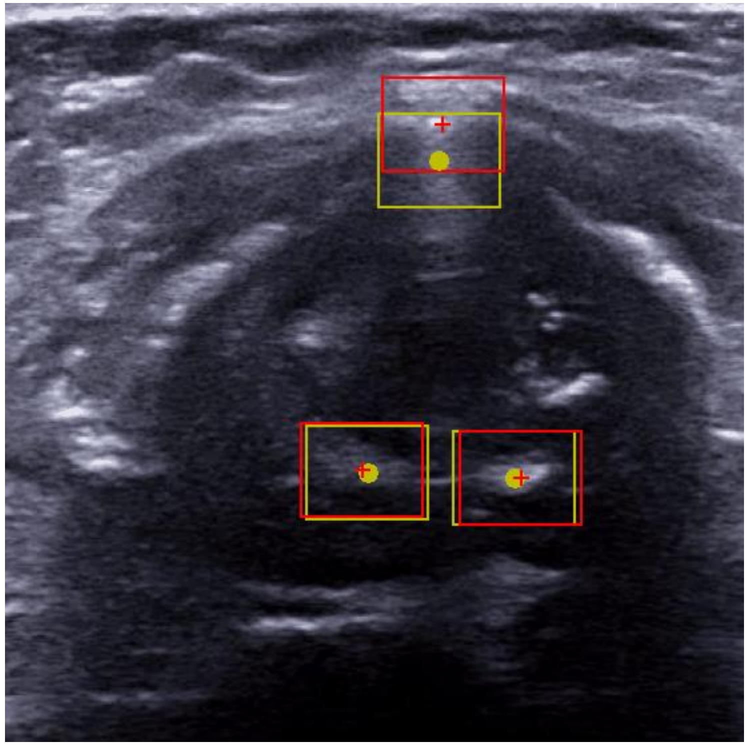
Non-uniformity in bounding box size



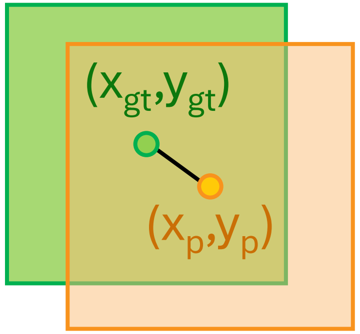
Yellow: reference Red: YOLO

Detection and tracking of landmarks in ultrasound video

- Focusing on model output, arytenoid predictions are severely missing or misinterpreted as false positives due to low IoU with small box size.
- To correct this, we propose two methods:



- **Bounding box normalisation:** Bounding box size was adjusted to match the ground truth box size before the IoU calculation, to focus on the accuracy of object location.
- **Center-point distance:** This metric is introduced with a predefined threshold to estimate the accuracy of object location, complementing traditional IoU-based evaluation.



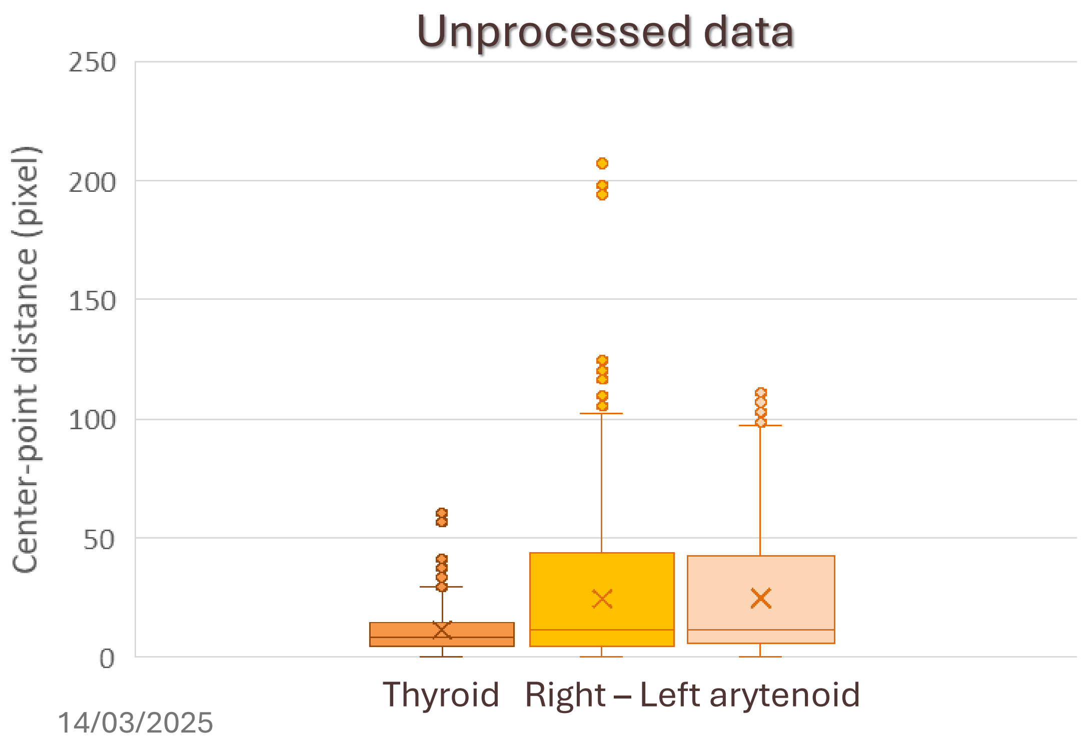
Detection and tracking of landmarks in ultrasound video

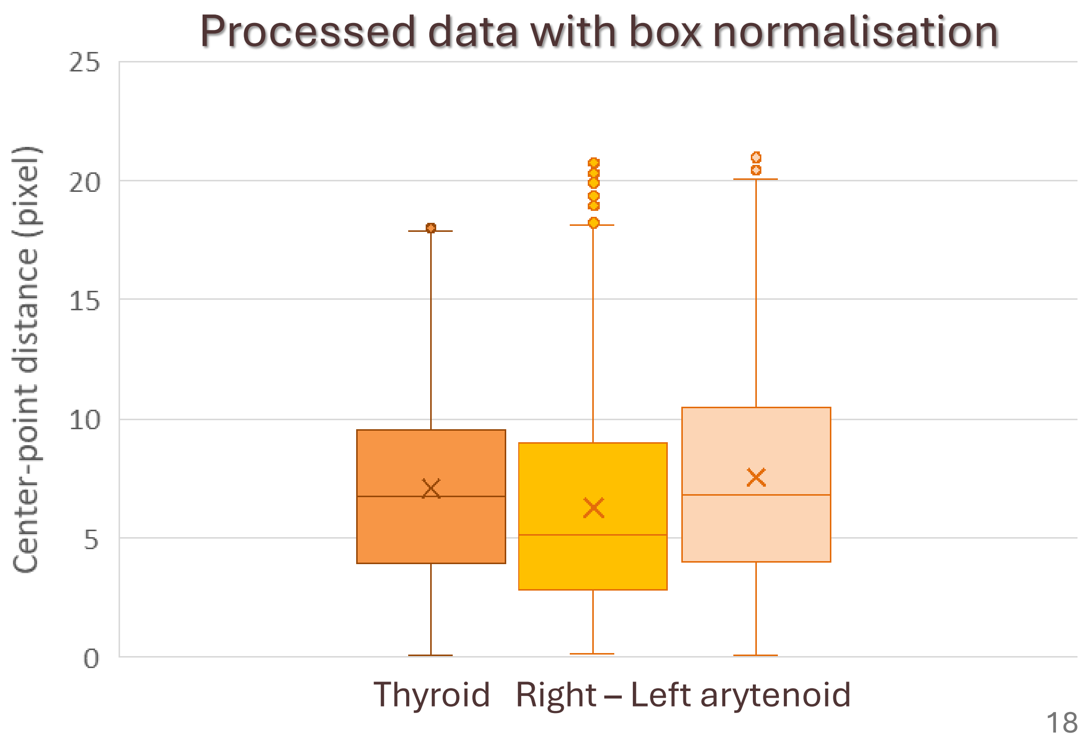
- **Bounding box normalisation:**
  - Validation set results show no significant variation in performance.
  - Test set results show the improved precision and recall.
  - Suggesting that post-processing enhances model performance on unseen and heterogeneous datasets.

Class	Images	Instances	VALIDATION SET				TEST SET			
			Precision		Recall		Precision		Recall	
			Original	Post-processing	Original	Post-processing	Original	Post-processing	Orginal	Post-processing
All	3614	10842	0.683	0.689 (+)	0.685	0.691 (+)	0.630	0.724 (+++)	0.640	0.728 (+++)
Thyroid	3614	3614	0.773	0.781 (+)	0.779	0.787 (+)	0.628	0.730 (+++)	0.642	0.746 (+++)
Arytenoid	3614	7228	0.592	0.597 (+)	0.590	0.594 (+)	0.631	0.718 (+++)	0.638	0.710 (+++)

Detection and tracking of landmarks in ultrasound video

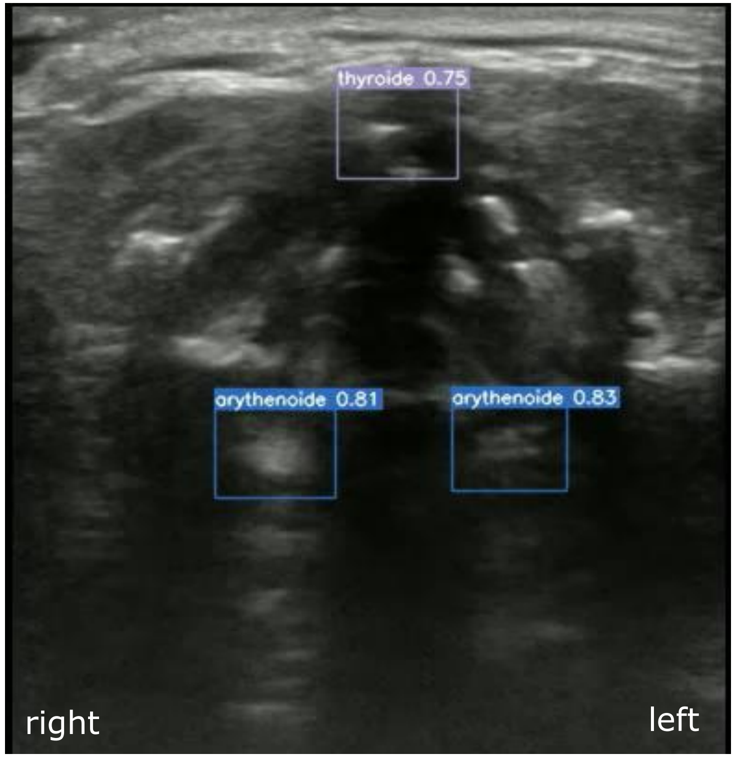
- **Center-point distance:** no standard value has been determined yet.
  - When calculated with unprocessed data, the distance values vary considerably.
  - With box normalisation and filtering (IoU < 0.5), the remaining detections exhibit a maximum distance of about 20 pixels, suggesting a potential localisation threshold.

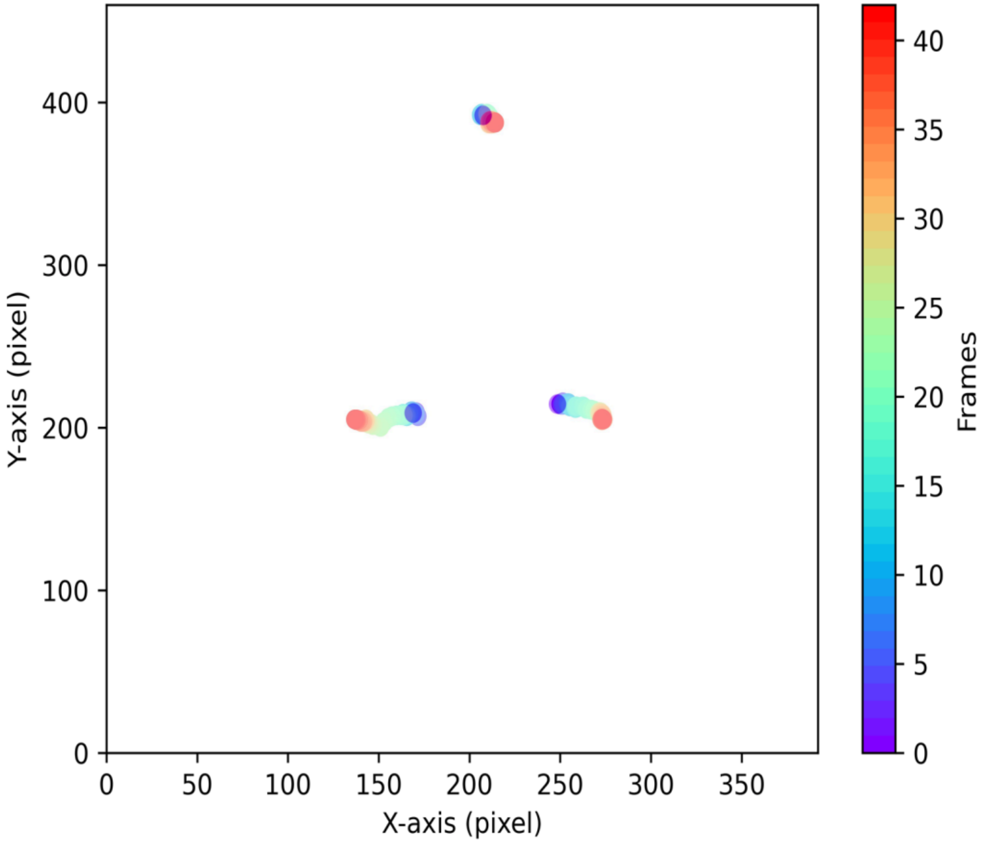




Detection and tracking of landmarks in ultrasound video

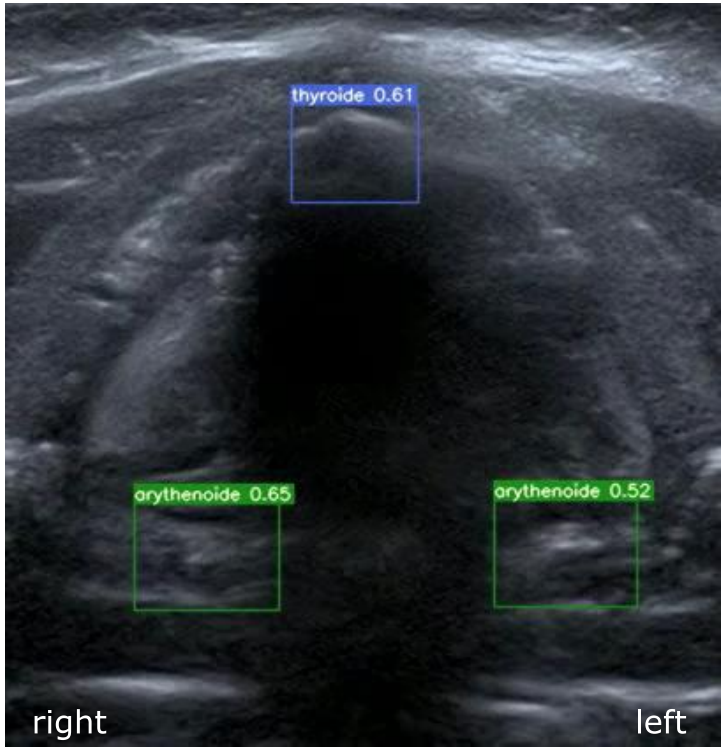
- **Color mapping of detections during vocal fold "closing – opening" time**
  - An example of a healthy vocal fold

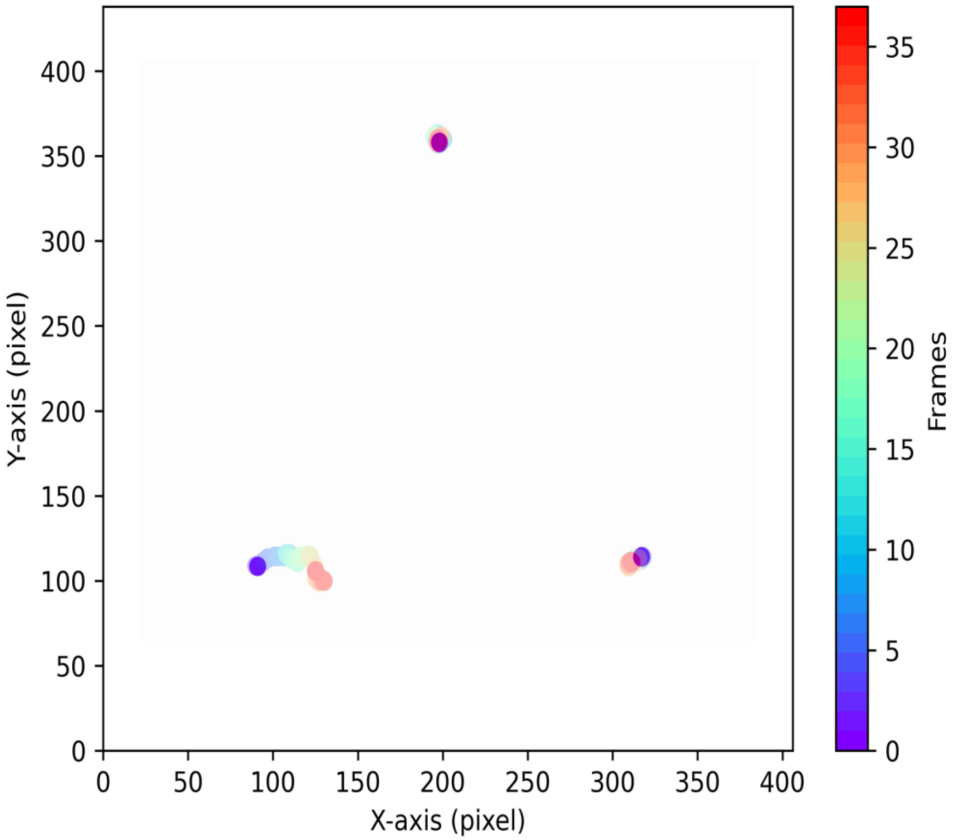




Detection and tracking of landmarks in ultrasound video

- **Color mapping of detections during vocal fold "closing – opening" time**
  - An example of left vocal cord paralysis in accordance with laryngoscopy assessment

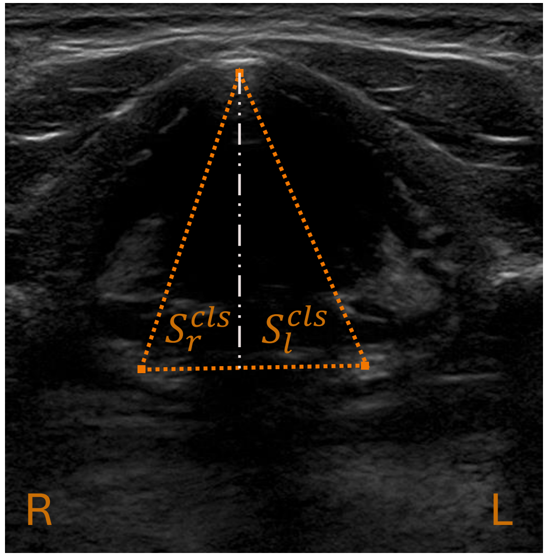




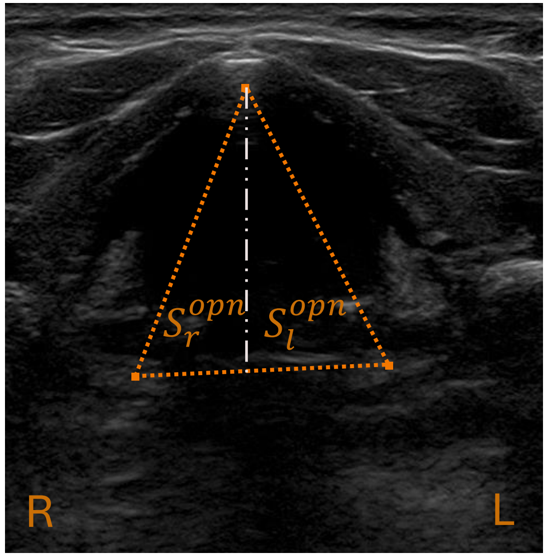
Characterization of vocal cords landmarks motion

- **Symmetry Index (SI):** the symmetry between the left and right sides of the vocal folds during both the closing and opening phases. It is calculated by dividing the smaller area by the larger area.
- **Mobility Fraction index (MF):** the relative change in area of the vocal folds between the opening and closing phases. It is calculated by dividing the difference between the opening and closing areas by the opening area.

$$\mathrm{SI}_{\mathrm{closing}} = \frac{\min(S_l^{closing}, S_r^{closing})}{\max(S_l^{closing}, S_r^{closing})} \; ; \; \mathrm{SI}_{\mathrm{opening}} = \frac{\min(S_l^{opening}, S_r^{opening})}{\max(S_l^{opening}, S_r^{opening})}$$



Closing position



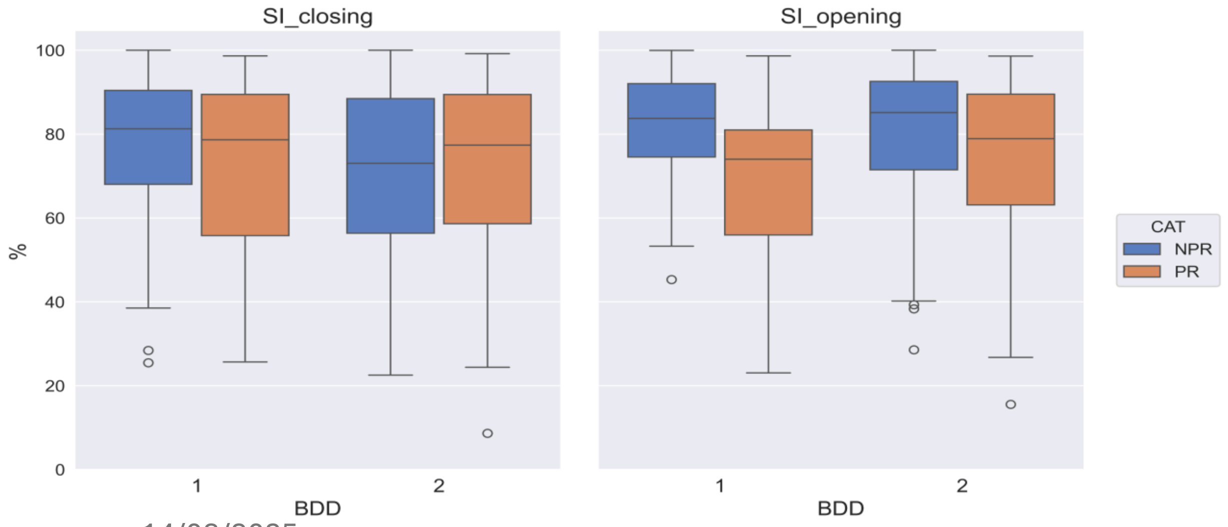
Opening position

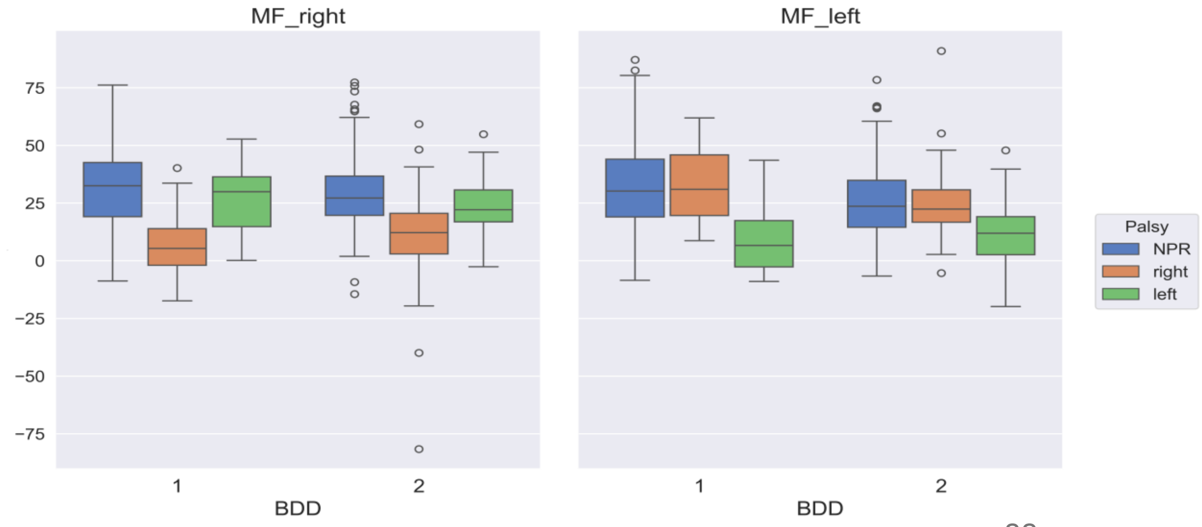
$$\mathrm{MF\_right} = \frac{S_r^{opening} - S_r^{closing}}{S_r^{opening}} \; ; \; \mathrm{MF\_left} = \frac{S_l^{opening} - S_l^{closing}}{S_l^{opening}}$$

Characterization of vocal cords landmarks motion

Dataset	Number of individuals	Number of sub-sequences	Vocal fold paralysis status evaluated by laryngoscopy
BDD1	149	194	Yes, all of dataset. 50/149 subjects with VC paralysis
BDD2	41	259	Under reevaluation
BDD3	67	161	Under reevaluation

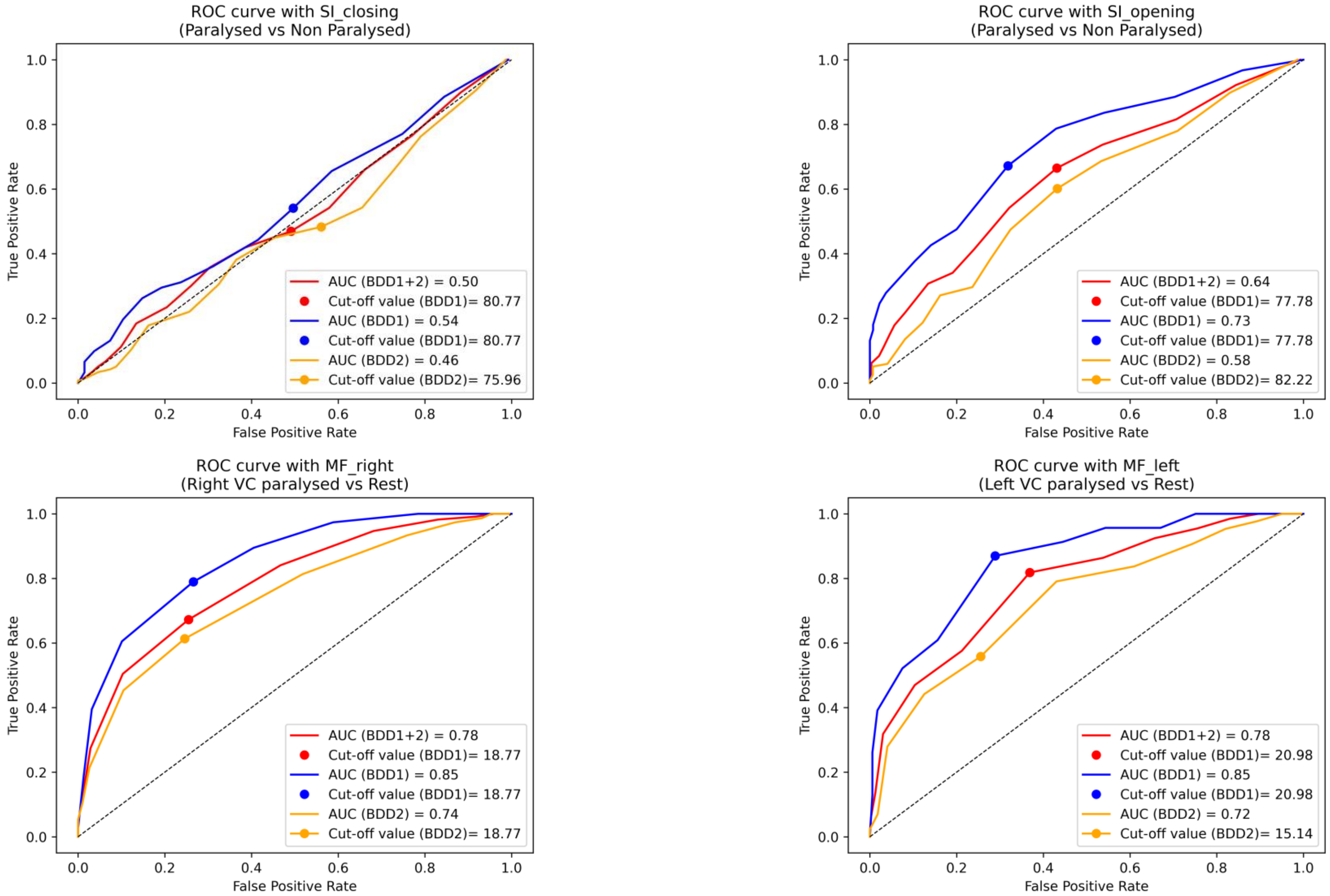
- Statistical analysis on annotated data





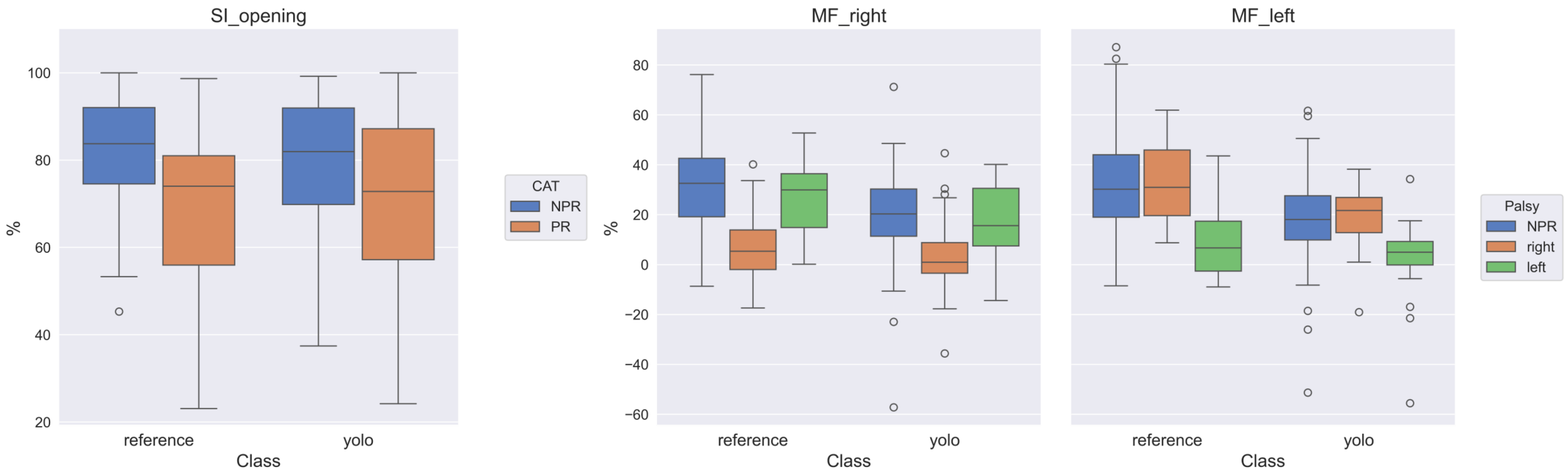
Characterization of vocal cords landmarks motion

- ROC curve analysis for evaluating the effectiveness of each variables as binary classifiers.



Characterization of vocal cords landmarks motion

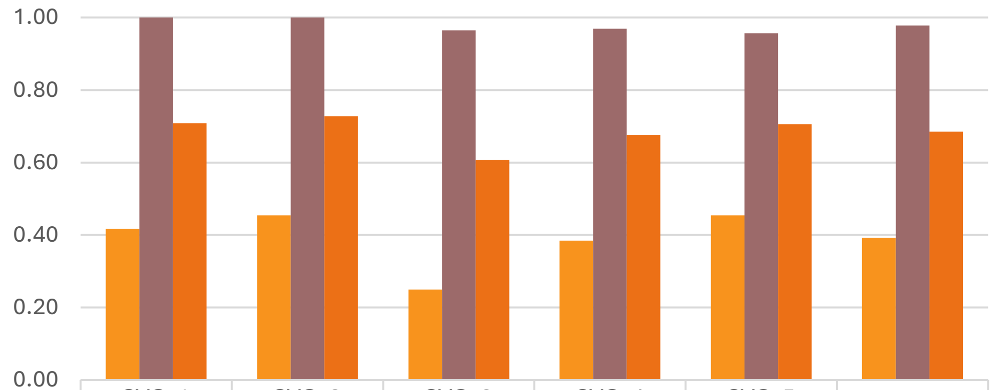
- Results comparison between BDD1 reference data and predictions by YOLO post-processing



Characterization of vocal cords landmarks motion

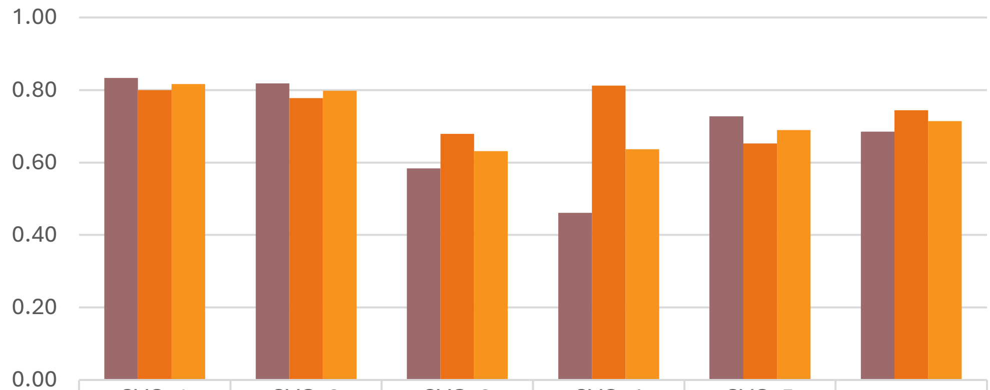
- **A support vector machine classifier (SVC)** was trained and validated using BDD1 data with 3 features "SI_opening + MF_right + MF_left" and a stratified 5-fold cross-validation strategy.
- **Synthetic Minority Over-sampling Technique (SMOTE)** was used to improve 'paralysed' case detection in an imbalanced dataset.

Validation performance



	SVC_1	SVC_2	SVC_3	SVC_4	SVC_5	average
recall/sensitivity	0.417	0.455	0.250	0.385	0.455	0.392
specificity	1.000	1.000	0.964	0.969	0.957	0.978
balanced Acc	0.708	0.727	0.607	0.677	0.706	0.685

Validation performance using SMOTE



	SVC_1	SVC_2	SVC_3	SVC_4	SVC_5	average
recall/sensitivity	0.833	0.818	0.583	0.462	0.727	0.685
specificity	0.800	0.778	0.679	0.813	0.652	0.744
balanced Acc	0.817	0.798	0.631	0.637	0.690	0.714

Conclusion

- **Detection & tracking task:**
  - The YOLO model, though effective in our training dataset, has limitations in generalization to the unseen dataset, requiring post-processing to enhance its performance.
  - Visual representation of the tracked positions was highly consistent with the structures' actual motion.
- **VC paralysis classification task:**
  - Results suggest that 'SI_opening,' 'MF_right,' and 'MF_left’ have promising discriminatory power for vocal cord paralysis classification.
  - An imbalanced dataset poses a challenge to classifier performance, requiring the augmentation of the minority class with adding new data or using SMOTE to generate synthetic samples.

Future works

- **Detection & tracking task:**
  - Evaluating results of the YOLO latest version
  - Incorporating center-point distance as a feature in bounding box regression to improve localisation accuracy in target detection.
  - Investigating the integration of attention modules into YOLO architecture by focusing on relevant features in the input image.

- **VC paralysis classification task:**
  - Embedding the training data with BDD2 and BDD3 data, which features double-assessed labels by experts for increased accuracy.
  - Ensembling individual models generated from each fold of the cross-validation process.
  - Expanding the set of movement-related features for paralysis quantification.

Thank you ☺