Peter Yao, Dan Witte, Alexander German, Preethi Periyakoil, Yeo Eun Kim, Hortense Gimonet, Lucian Sulica, Hayley Born, Olivier Elemento, Josue Barnes, et al. A deep learning pipeline for automated classification of vocal fold polyps in flexible laryngoscopy. *European Archives of Oto-Rhino-Laryngology*, 281(4):2055–2062, 2024.

Qian Zhao, Yuqing He, Yanda Wu, Dongyan Huang, Yang Wang, Cai Sun, Jun Ju, Jiasen Wang, and Jeremy Jianshuo-li Mahr. Vocal cord lesions classification based on deep convolutional neural network and transfer learning. *Medical physics*, 49(1): 432–442, 2022.

Appendix A. Additional Vocal Fold Images

Figure 1 in the main text shows examples of vocal folds in the adducted (closed) position. In Figure 6, we provide examples from our dataset showing vocal folds in the abducted (open) position.

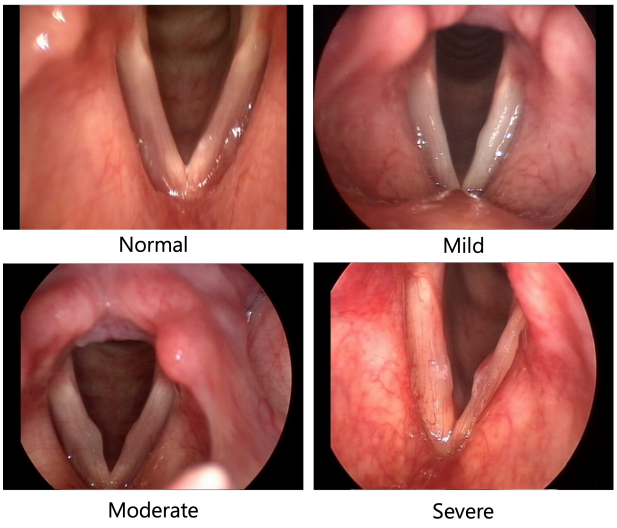


Figure 6: Images of vocal folds in the abducted (open) position, showing varying levels of phonotrauma severity. Normal indicates healthy control.

Appendix B. Experimental Details

B.1. Data Pre-Processing

The images in our dataset vary in size, orientation, and color. We standardize them by applying the following pre-processing steps: (1) *center cropping* along the width dimension by a factor 0.9 to remove parts of the image not relevant to severity prediction (e.g., anatomy surrounding the vocal folds), (2) *resizing with padding* to a fixed target size (554 by 544) while preserving the original aspect ratio, and (3) *color normalization* by min-max scaling.

B.2. Data Augmentations

During model training, we apply data augmentations to encourage robustness to aspects of the image that are not relevant to the severity prediction task. The augmentations we use include: cropping, rotation, horizontal flipping, adjusting brightness and contrast, Gaussian noise, Gaussian blurring, and gamma correction. In addition, we create a custom augmentation to simulate the black circular borders that appear in some but not all images (e.g., as in Figure 1).

B.3. Hyperparameters

We train all models for 1000 epochs. We use a batch size of 16 and a learning rate of 0.00001. We use the Adam Optimizer.

B.4. Model Selection

For each experiment, we split the training set into training examples (80%) and validation examples (20%). We select the best model from the 1000 training epoch based on the validation performance. We use MAE with uncertainty-weighting (cf. 5.3) as our model selection metric.

B.5. Evaluation

We apply five-fold cross validation. For each fold, we run experiments with three seeds. When evaluating performance on the held-out test data for a fold, we use an ensemble of the three models trained with different seeds. Specifically, we take the average prediction across the three models as our final prediction.

Appendix C. Statistical Significance Tests

We conduct statistical significance tests to assess whether the observed differences between methods are statistically meaningful and not attributable to sampling variability. Since we used 5-fold cross-validation, we have five observations per metric for

each method. This limited sample size constrains our statistical power, particularly when applying multiple hypothesis test corrections across all pairwise comparisons. We therefore focus our analysis on key comparisons that identify the best-performing methods.

We use the following statistical significance test: for a given metric (e.g., QWK), we compare two methods using a paired one-sided t-test (paired across test folds to ensure consistent comparisons). We use $\alpha = 0.05$ as the significance threshold. We focus on two hypotheses that help identify the overall best method:

1. **Does CORN outperform OR-Soft in terms of predictive performance?** We find that the difference between the two methods is not statistically significant in terms of QWK ($p = 0.256$) or MAE ($p = 0.135$).

2. **Does OR-Soft outperform CORN in terms of calibration (ECE)?** We find that OR-Soft achieves significantly lower ECE ($p = 0.025$).

These tests support our main finding that OR-Soft provides the best balance of predictive performance and calibration among the methods we evaluated.

Appendix D. Additional Results

In Table 3, we present results for cross entropy and Brier score, two metrics that measure the difference between true and predicted label distributions. In Table 4, we present additional metrics that capture predictive performance. We present the same confusion matrices shown in the main text but with the standard deviation across folds included in Figure 7. We present risk coverage curves for OR-Soft, OR-CNN, and CORN for each individual test fold in Figure 8.

Appendix E. Predictive Performance Versus Uncertainty Estimation Tradeoff

Among the methods examined in this study, we observed a tradeoff between predictive performance and uncertainty estimation quality. To better understand this tradeoff, in Figure 9, we present a scatter plot with MAE (UW) on x-axis and ECE on the y-axis showing where each method falls along this tradeoff.

Appendix F. Results from Full Multi-Rater Dataset

After the submission deadline, we collected annotations from three additional laryngeal surgeons for all subjects in the dataset. We conducted a preliminary analysis on the data with this new label set; results are shown in Table 5.

In this analysis, we took the annotations provided by the three new raters and combined them with the labels used in our initial analysis. For images that had annotations from three raters in the initial set (i.e, the Multi-Rater subset), we directly combined these with the three new annotations, yielding six ratings per image. For images in the initial set that did not have three independent ratings – specifically, the normal cases identified through comprehensive clinical screening and the severe cases labeled by a three-person consensus – we replicated their initial labels three times to maintain balanced weighting across the two annotation sets. This approach ensured that all images had exactly six ratings.

We derived soft labels from the empirical distribution over these six ratings. We created hard labels by choosing the mode rating. For some images, there was a 50-50 tie between two classes. We excluded these images from evaluation but retained them for training. To train with these images, for soft-label approaches, we used the full six-rater distribution directly. For hard label approaches, we randomly sampled from the majority classes at each training epoch.

The results in Table 5 largely align with those presented in the main text. OR-Soft achieves the best balance between predictive performance and uncertainty estimation, and the soft label methods have superior uncertainty calibration compared to their hard label variants. One difference is that OR-Soft performs slightly better than CORN in terms MAE, QWK, and Accuracy in this analysis. However, the results are not statistically significant ($p \geq 0.21$ for each of these metrics), which is consistent with our main findings.

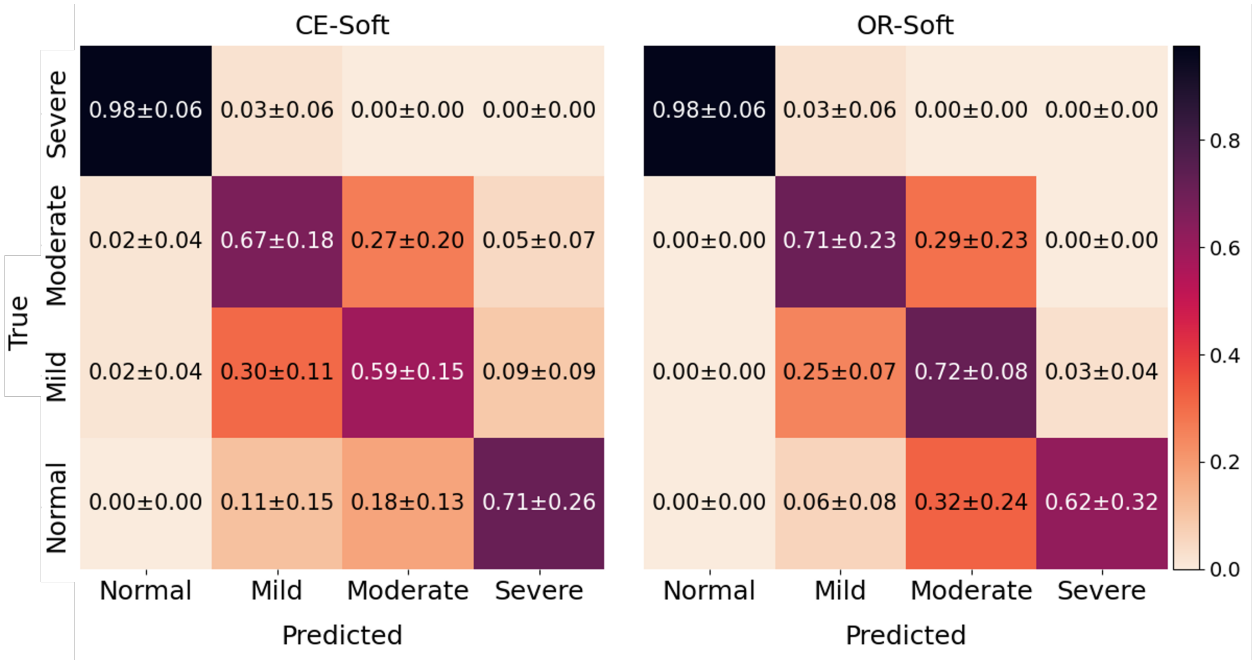
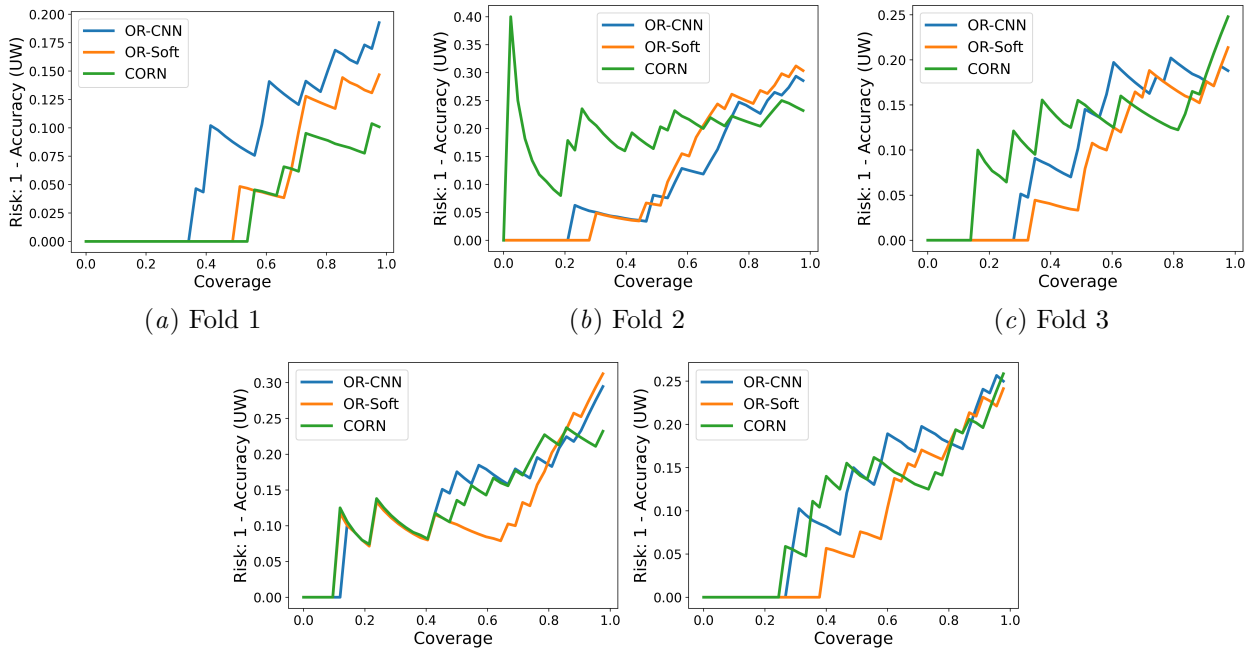


Figure 7: Confusion matrices for CE-Soft and OR-Soft. Confusion matrices are row-normalized. We show mean ± standard deviation across the five folds. Both methods have high accuracy in discriminating between normal and non-normal cases. OR-Soft makes fewer off-by-two errors than CE-Soft.



(*a*) Fold 1 (*b*) Fold 2 (*c*) Fold 3

(*d*) Fold 4 (*e*) Fold 5

Figure 8: Risk–coverage curves for the three methods with the lowest AURC (OR-Soft, OR-CNN, and CORN), shown per test fold.

Method	Cross Entropy	Brier Score
CE	0.93 ± 0.17	0.30 ± 0.05
CE-Soft	0.69 ± 0.11	0.24 ± 0.05
CORN	0.81 ± 0.16	0.25 ± 0.05
SORD-AE	0.95 ± 0.02	0.37 ± 0.03
SORD-SE	0.86 ± 0.03	0.33 ± 0.02
CORAL	1.65 ± 0.04	0.66 ± 0.03
CORAL-Soft	1.62 ± 0.05	0.65 ± 0.02
OR-CNN	0.80 ± 0.16	0.26 ± 0.04
OR-Soft	**0.64 ± 0.11**	**0.21 ± 0.04**

Table 3: Performance (mean ± standard deviation) across five folds. Lower is better for both metrics. The best average performance is in **bold**, second-best is underlined. OR-Soft performs best for both metrics, followed by CE-Soft.

Method	Coverage Error	AUC	Spearman ρ
CE	1.33 ± 0.08	0.90 ± 0.02	0.81 ± 0.08
CE-Soft	1.35 ± 0.09	0.91 ± 0.04	0.77 ± 0.06
CORN	**1.28 ± 0.10**	0.92 ± 0.04	**0.84 ± 0.07**
SORD-AE	1.34 ± 0.06	0.91 ± 0.02	0.80 ± 0.06
SORD-SE	1.35 ± 0.09	0.90 ± 0.02	0.80 ± 0.06
CORAL	2.54 ± 0.08	0.83 ± 0.03	0.73 ± 0.09
CORAL-Soft	2.51 ± 0.06	0.86 ± 0.02	0.75 ± 0.06
OR-CNN	1.33 ± 0.10	0.92 ± 0.02	0.82 ± 0.05
OR-Soft	1.31 ± 0.12	**0.92 ± 0.04**	0.83 ± 0.05

Table 4: Performance (mean ± standard deviation) across five folds. Lower is better for Coverage Error, higher is better for AUC and Spearman correlation. The best average performance is in **bold**, second-best is underlined. CORN and OR-Soft are the top-performing methods.

Method	MAE (UW)	QWK (UW)	Accuracy (UW)	Accuracy (AR)	ECE	AURC
CE	0.33 ± 0.13	0.76 ± 0.12	0.69 ± 0.11	0.88 ± 0.06	0.21 ± 0.04	0.18 ± 0.05
CE-Soft	0.34 ± 0.10	0.77 ± 0.07	0.67 ± 0.10	0.86 ± 0.05	0.15 ± 0.02	0.22 ± 0.07
CORN	0.30 ± 0.08	0.80 ± 0.09	0.72 ± 0.07	0.89 ± 0.05	0.17 ± 0.04	0.13 ± 0.05
OR-CNN	0.30 ± 0.07	0.81 ± 0.08	**0.73 ± 0.04**	**0.91 ± 0.02**	0.18 ± 0.04	0.13 ± 0.03
OR-Soft	**0.28 ± 0.06**	**0.83 ± 0.03**	0.73 ± 0.06	0.88 ± 0.05	**0.10 ± 0.01**	**0.13 ± 0.05**

Table 5: Results on the dataset with annotations from three additional clinical experts (obtained after the submission deadline). Performance (mean ± standard deviation) across five folds is shown. The method with the best average performance is in **bold**, second-best is underlined. UW = Uncertainty-Weighted, AR = Any-Rater Accuracy (see Section 5.3). OR-Soft performs best in terms of both ordinal metrics (MAE, QWK) and uncertainty estimation metrics (ECE, AURC).

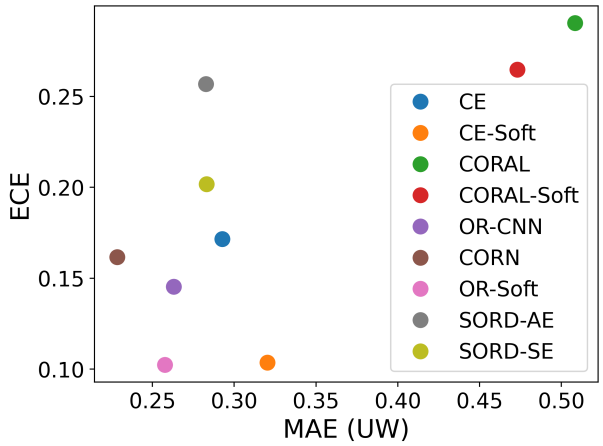


Figure 9: For each method, we plot its predictive performance, measured by MAE (UW), against its uncertainty calibration, measured by ECE. OR-Soft is the in the lower left, with the lowest ECE and a slightly larger MAE (UW) than CORN.