Evaluating AI for Voice Disorder Screening

From Research to Clinical Practice

6th of June 2025, Prague

31st Union of European Phoniatricians Congress

Mette Pedersen and Sneha Das

Pedersen, M.

MD, PhD, Ear-Nose-Throat Specialist, Head & Neck surgeon

Fellow of the Royal Society of Medicine UK

Danish Representative Union of European Phoniatricians

The Medical Center, Østergade 18, 1100 Copenhagen, Denmark

e-mail address: m.f.pedersen@dadlnet.dk

Phone: +45 31126184

Member UEP Biomarkers Committee

Sneha Das, Dsc. (Tech),

Assistant Professor,

Department of Applied Mathematics and Computer Science,

Technical University of Denmark, Denmark

sneha.das1991@gmail.com

+4591440864

Member UEP Biomarkers Committee

- We have no conflicts of interest

Why Voice-related Artificial Intelligence (Voice-AI) Matters in Healthcare

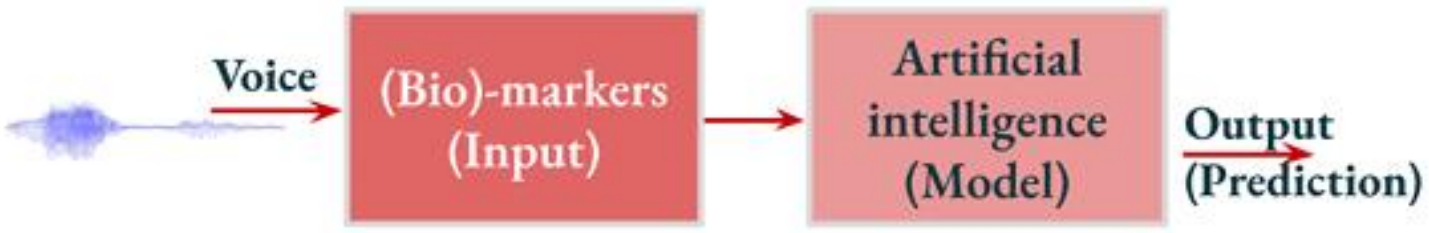
- Voice-AI represents a rapidly advancing field of research that enables non-invasive, scalable methods for
- voice disorders,
- with Parkinson's disease being a notable part in our library search on Voice-AI, by the Royal Society of Medicine Library UK.

- **However, adoption in clinical settings has seen limited clinical integration due to insufficient clinical evaluation.**

Components of a Voice-AI System

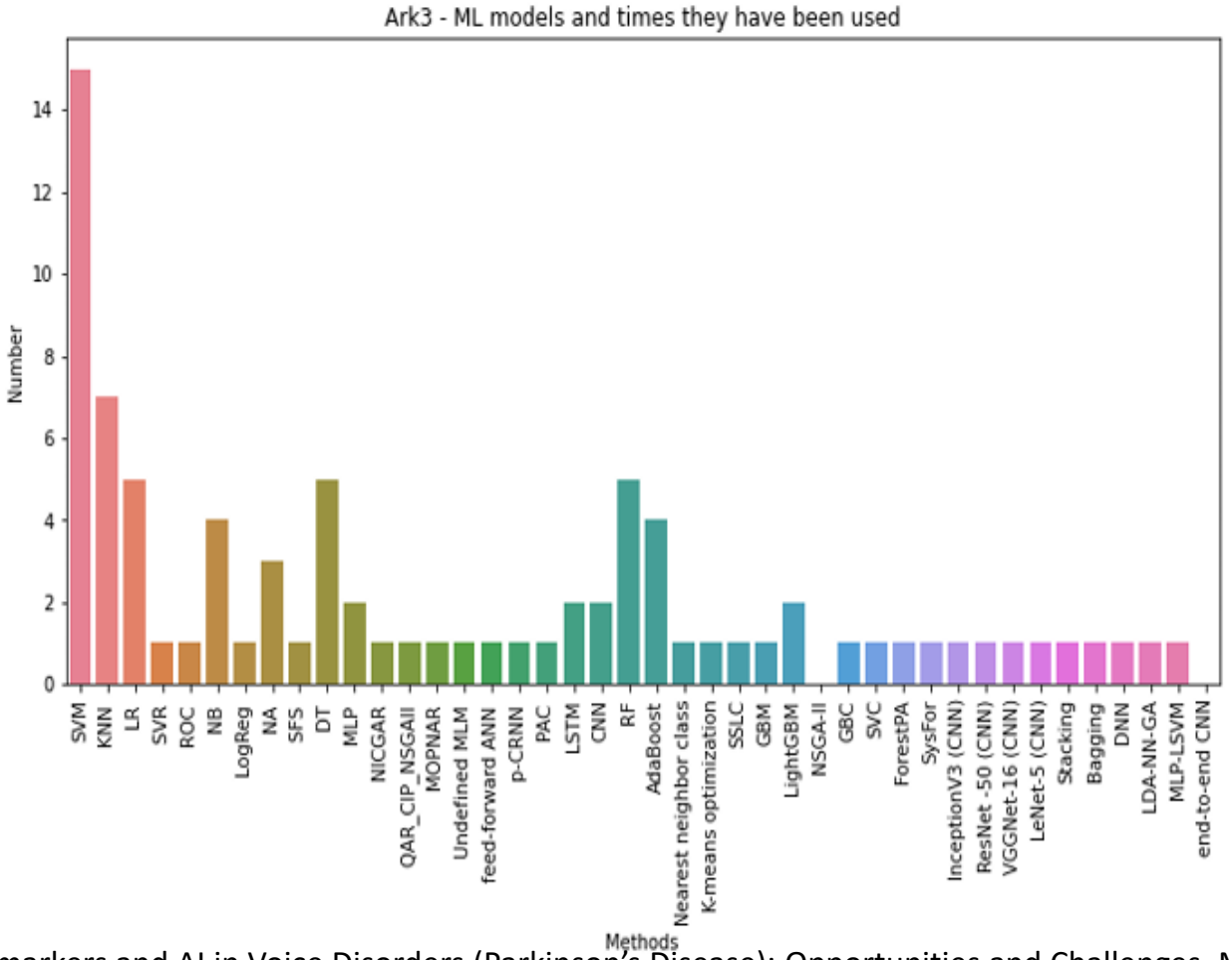
- The schematic illustrates the fundamental components of a minimal Voice-AI system.
- It maps the input audio signal through pre-processing and modeling to generate outputs such as:

  diagnostic classifications or quantitative scores.



Das S. Presentation: (Bio)-markers and AI in Voice Disorders (Parkinson's Disease): Opportunities and Challenges. May 20th 2025. Committee of Voice Related Biomarkers, Union of European Phoniatricians.

- Voice-AI models encompass both classical statistical techniques and advanced deep learning architectures.
- **This figure provides an overview of models applied across the 19 Parkinson ’s-related studies** (search of Voice-AI by the Royal Society of Medicine, UK, 2013-2023).



Das S. Presentation: (Bio)-markers and AI in Voice Disorders (Parkinson’s Disease): Opportunities and Challenges. May 20th 2025. Committee of Voice Related Biomarkers, Union of European Phoniatricians.

Model Performance –

- This table compares critical model performance across the 19 studies. Support Vector Machines (SVMs), CNNs, and Random Forests demonstrated robust accuracy and sensitivity,
  although performance varied regarding precision and F1.
- (not all articles give detailed data)

Model	Number of Studies	Accuracy (%) (No. of Studies)	F1-Score (%) (No. of Studies)	Recall/Sensitivity (%) (No. of Studies)	Precision (%) (No. of Studies)	Specificity (%) (No. of Studies)
SVM (Support Vector Machines)	11 out of 19	84-96 (10 out of 19)	80-95 (6 out of 19)	89-95 (8 out of 19)	85-94 (5 out of 19)	87-93 (8 out of 19)
CNN (Convolutional Neural Networks)	3 out of 19	85-97 (3 out of 19)	82-96 (2 out of 19)	89-96 (3 out of 19)	85-92 (2 out of 19)	88-92 (3 out of 19)
Random Forest (RF)	5 out of 19	83-88 (4 out of 19)	81-90 (3 out of 19)	86-91 (4 out of 19)	80-89 (3 out of 19)	85-90 (4 out of 19)

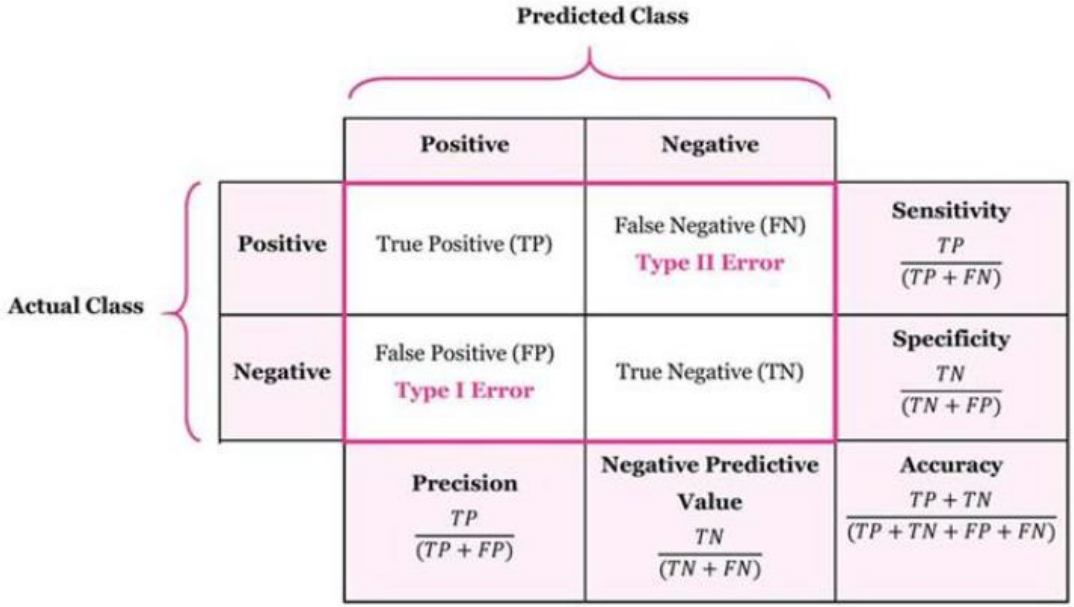
Models information across studies.

Pedersen M, Meiner VG. AI-Based Quality of Voice Analysis Models for Clinical Use, Insights of Quality of Models from 19 Parkinson's Disease Studies (2013-2023). Journal of Clinical Medical Research - Year 2024, Vol 6, Issue 1.
https://doi.org/10.46889/JCMR.2025.6107

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Confusion Matrix and Its Importance

- The confusion matrix facilitates contextual interpretation of accuracy metrics. High accuracy may yield misleading conclusions in imbalanced datasets unless paired with sensitivity, specificity, and precision.



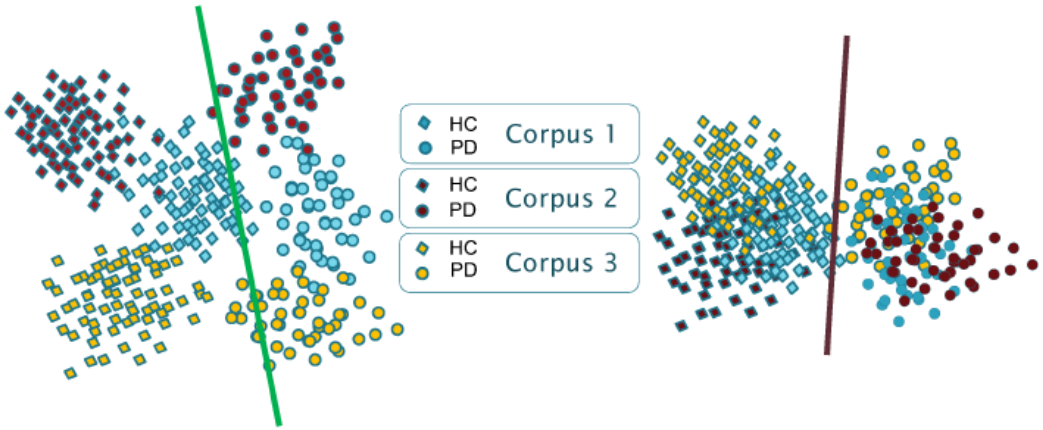
Confusion matrix.

Pedersen M, Meiner VG. AI-Based Quality of Voice Analysis Models for Clinical Use, Insights of Quality of Models from 19 Parkinson's Disease Studies (2013-2023). Journal of Clinical Medical Research - Year 2024, Vol 6, Issue 1. https://doi.org/10.46889/JCMR.2025.6107

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Evaluation Metrics for Classification

- For classification tasks, such as distinguishing between Healthy Controls (HC) and patients with Parkinson's Disease (PD), we assess model performance using standard metrics including accuracy, recall, and the F1-score

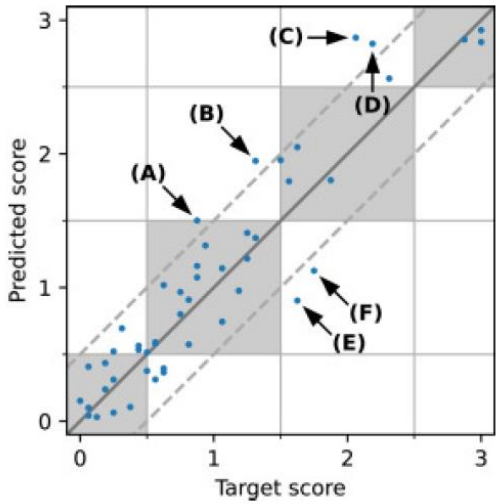


Metrics for classification

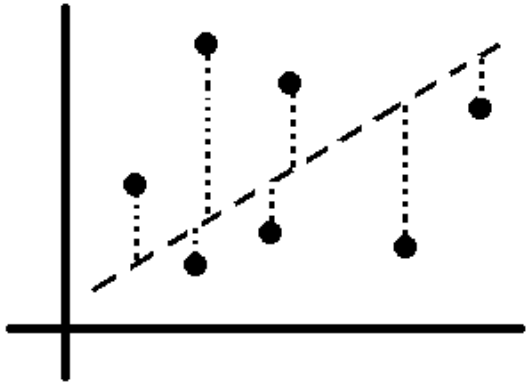
Ibarra EJ, Arias-Londoño JD, Zañartu M, Godino-Llorente JI. Towards a corpus (and language)-independent screening of parkinson's disease from voice and speech through domain adaptation. Bioengineering. 2023 Nov 15;10(11):1316.

Evaluation Metrics for Regression

- When predicting continuous scores like GRBAS ratings, we use regression metrics like Mean Squarred Error (MSE): $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and Coefficient of Determination R^2.
- MSE is the average squared distance between predicted and actual values, penalizing large errors more heavily.
- Points within the shaded regions show predictions within ±1 of the true score. Labeled examples (A–F, see below) highlight typical error types.
- R^2 indicates how well the model explains the variability in the target scores — with 1 being a perfect fit and 0 meaning no predictive power.



Regression metrics.



Example of MSE

Hidaka S, Lee Y, Nakanishi M, Wakamiya K, Nakagawa T, Kaburagi T. Automatic GRBAS scoring of pathological voices using deep learning and a small set of labelled voice data. Journal of Voice. 2022 Nov 25.

The 19 articles on Voice - AI in Parkinsons Disease

- **Methodological limitations in the Voice-AI papers** include
- lack of demographic diversity,
- insufficient dataset size, and
- incomplete metric reporting.
- **These limit real-world relevance**.
- (not all papers give sufficient data)

Category	Problems Identified	Articles That Provide Data	Articles That Do Not Provide Data	Reason for Missing Data
Usability	Insufficient dataset size	6 articles	12-13 articles	Articles focus on model performance
Precision	Variability in measurement protocols	5 articles	13-14 articles	Many articles assume standardized datasets
Content	Limited diversity in vocal tasks	7 articles	11-12 articles	Studies focus on specific phonemes
Population	Underrepresentation of demographic groups	4-5 articles	13-15 articles	Articles do not address demographic diversity
Disorder Characteristics	Limited characterization of specific vocal imparimetns	6 articles	12-13 articles	Some studies focus purely on classification

Acoustical Dataset

Pedersen M. (2025). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

The 19 articles on Voice - AI in Parkinson's Disease

Category	Challenges Identified	Articles That Provide Data	Articles That Do Not Provide Data	Reason for Missing Data
Microphone Placement	Distance of the microphone was identified as a challenge	5 articles	14 articles	Many articles assume ideal recording conditions
Noise Factors	Environmental noise variability	6 articles	12-13 articles	Articles often assume noise-free environments
Measurement Parameters	Frequency area measurement inconsistencies	4 articles	14-15 articles	Many studies do not report detailed frequency analysis
Feature Extraction	Challenges in signal processing	7 articles	11-12 articles	Some articles focus on algorithm testing or dataset creation without detailing feature extraction.

Challenges in Acoustic Data Processing
(note that not all give sufficient data)

Pedersen M. (2025). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

The 19 articles on Voice - AI in Parkinson's disease

Evaluation Criteria	Articles That Provide Data	Articles That Do Not Provide Data	Measurement Description
Sensitivity (recall)	7 articles	11-12 articles	Sensitivity ranged from 73% to 95%
Specificity	7 articles	11-12 articles	Specificity ranged from 60% to 96%.
Accuracy	6 articles	12-13 articles	Accuracy ranged between 84% and 96%
Cross-Validation	8 articles	10-11 articles	Common validation techniques included 10-fold cross-validation
Training Setup	7 articles	11-12 articles	An 80:20 split for training and testing was most common
Testing Setup	6 articles	12-13 articles	Evaluation metrics included F1-scores: 0.75–0.79
AI Model Choice & Description	7 articles	11-12 articles	Popular models included SVM, CNN, and AdaBoost

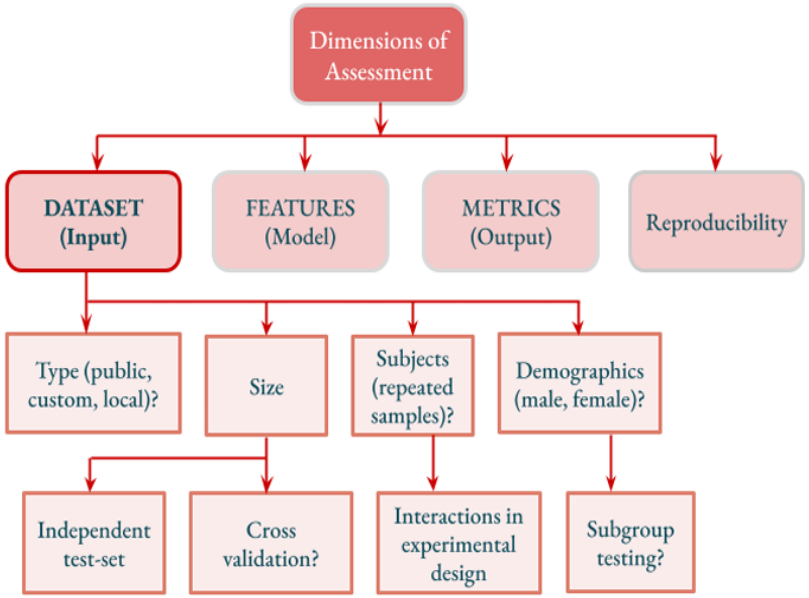
Evaluation Metrics and Frameworks for Parkinson's Disease Models.
(Note that not all give sufficient data)

Pedersen M. (2025). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Dataset and Quality Challenges

- Data quality and distribution heavily impact AI performance. Inadequate microphone positioning, background acoustic interference, and non-representative data distributions can all introduce bias and **reduce generalizability.**



Flow diagram depicting the sub-factors under the datasets that influences model choices.

Das S. Presentation: (Bio)-markers and AI in Voice Disorders (Parkinson's Disease): Opportunities and Challenges. May 20th 2025. Committee of Voice Related Biomarkers, Union of European Phoniatricians.

Foundation Models – The Next Step?

- Emerging foundation models such as GPT or BERT hold transformative potential voice-AI by enabling multi-task learning.

- Their adaptation to voice-related biomarkers is currently underexplored or in early investigational phases.

4

A foundation model trained on voice-related biomarkers

Voice Handicap Index SCORES

GRBAS (or CAPE – V)listening **TEST**

Acoustic parameters (F0, jitter, shimmer, Harmonics Noise Ratio), and

5

Maximun Phonation Time

could learn powerful representations of vocal health and impairment.

Here's what such a model could *conclude* or *infer* depending on its design and the task it's fine-tuned for:

Example Tasks Enabled by the Multiple Layer Perception Foundation Model

Task	Input	Output
Voice disorder screening	Audio + MPT	Binary or multiclass label
Automatic perceptual scoring	Audio	GRBAS or CAPE-V scores
Predict VHI	Audio + demographic info	VHI-10 score
Track therapy effect	Audio over time	Time-series of severity
Voice health embedding generation	Raw or processed audio	Fixed-size vector

Clinical Readiness and EU AI Act

- Quantitative performance metrics alone are insufficient for determining clinical applicability.
- Regulatory frameworks like the EU AI Act stress explainability, robustness, and real-world validation.

- The integration of AI applications in clinical practice remains subject to ongoing regulatory and practical evaluation.

Gstrein, O. J., Haleem, N., & Zwitter, A. (2024). General-purpose AI regulation and the European Union AI Act. Internet Policy Review, 13(3). https://doi.org/10.14763/2024.3.1790

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Limitations and Open Questions

- Despite its promise, Voice-AI remains hindered by a lack of standardized protocols and robust clinical trials.

  Trust, transparency, and diverse data remain key barriers.

Conclusion & Path Forward

- To make AI for voice disorders viable in clinics, it is imperative to design updated, reliable systems
- with input from both
- data scientists and clinical stakeholders,
- but with direct input from clinicians and patients.

Acknowledgements
Thank you to Vitus Girelli Meiner, MSc. in Computer Science,
IT-University of Copenhagen, Denmark
Vigga Girelli Meiner, High School Graduate,
Gefion Gymnasium, Denmark

No conflict of interest

Thank you for listening!

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Biomarker-Based Evaluation

- This table lists key voice biomarkers used in genetic disorder studies. F0, jitter, and MPT are central, but many studies are still in early stages of clinical relevance.

Assessment Method/ Feature	Number of Articles Reporting
VHI (Voice Handicap Index)	6
GRBAS (Listeners Test)	10
F0 (Fundamental Frequencies)	20
Jitter, shimmer	8
HNR/NHR (Harmonics to Nois Ratio/Noise to Harmonics Ratio)	6
MPT (Maximum Phonation Time)	6
ML (Machine Learning)	5

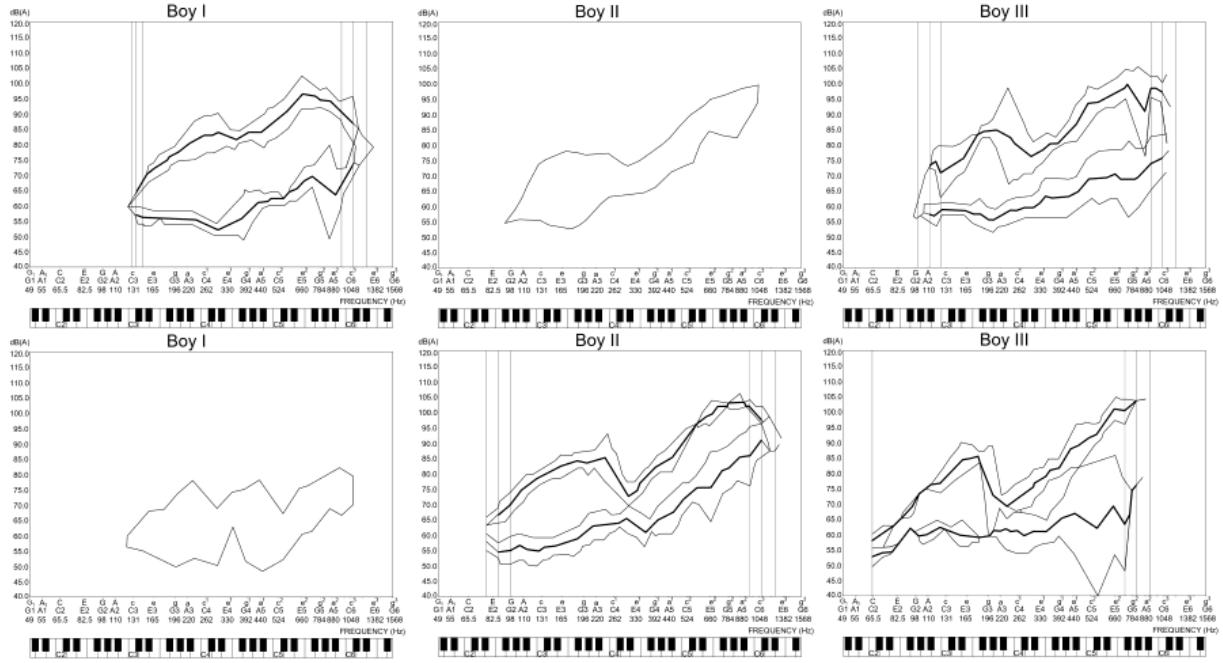
Frequency of Voice-Related Parameters in Papers on Genetics in the Last 5 Years

Pedersen M. (2025). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

The Role of Optical Coherence Tomography (OCT)

- Optical Coherence Tomography (OCT) provides a non-invasive modality to visualize vocal fold tissue, useful for tracking puberty, hormones, and voice recovery after injuries or radiation exposure.



Three boys average Voice Range Profiles and standard deviations for the highest and lowest tones

Pedersen M. Tissue Analysis of the Vocal Folds Cellular and Biochemical Aspects. J Biomed Res Environ Sci. 2025 Feb 18; 6(1): 153-156. doi: 10.37871/jbres2067

The Medical Center Østergade 18, 1100 Copenhaen, Denmark

Overview of a foundation model

🔧 **Objective**

❓ **Steps to Build the MLP Model**

1.Collect Data

- Each sample = [MPT, F0, jitter, shimmer, HNR] + severity label
- Labels: “Mild”, “Moderate”, “Severe”

2.Extract Features

- Use tools like **Praat**, **openSMILE**, or **Librosa**
- Normalize features for model input

3.Define the Model (MLP)

- Input layer: 5 features
- Hidden layers: 1–2 (e.g., 32 → 16 neurons)
- Output layer: 3 neurons (one per class)
- **Activation**: ReLU
- **Output**: Softmax → probabilities per class

4.Train the Model

- Loss function: CrossEntropyLoss
- Optimizer: Adam
- Input: feature vectors
- Target: class labels

5.Deploy

- Use model in a clinical app or desktop tool
- **Input features during laryngoscopy**
- Output severity level in real time

🧠 **Why Use MLP?**

- Simple, fast, and effective for tabular clinical data
- Can integrate with real-time systems
- Works well even without raw audio if features are accurate