Image Anal Stereol 2025;44:183-196 doi: 105566/ias.3741
Original Research Paper

COMPARISON OF DEEP LEARNING MODELS FOR VOICE DISORDER CLASSIFICATION USING PHONOVIBROGRAPHIC IMAGES

B PANCHAMI, S PRAVIN KUMAR✉
Department of Biomedical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
e-mail: panchamib@ssn.edu.in, pravinkumars@ssn.edu.in
(Received August 15, 2025; revised November 6, 2025; accepted November 10, 2025)

ABSTRACT

Accurate diagnosis of vocal fold disorders is difficult because of subtle variations between pathological conditions. Phonovibrography (PVG), generated from high-speed videoendoscopy (HSV), documents glottal vibration patterns as static images, allowing systemic analysis. In our study, we propose PVGNet, a hybrid deep learning model combining multiscale feature extraction and channel attention, designed specifically for PVG-based classification. We benchmark PVGNet against InceptionResNetV2, VGG19, DenseNet169, and X-ViT across binary, tertiary, and multi-class tasks. PVGNet continuously outperforms baselines in accuracy, F1-score, and AUC, by minimizing false negatives, which is important for reliable diagnosis. These results show PVG's potential as a diagnostic imaging modality and PVGNet's effectiveness in automated voice disorder classification.

Keywords: Classification; Deep Learning Models; Functional Voice Disorders; High-speed video endoscopy; Phonovibrogram; Voice Disorder.

INTRODUCTION

Voice production is the most important tool of human communication (Kamiloğlu and Sauter, 2021), and any abnormalities in the process can lead to voice disorders (Spina *et al.,* 2009).

Within the larynx, the vocal folds are located which are the central organs of voice production. They are responsible for necessary voice functions such as phonation, voice quality, pitch control, volume, loudness, and verbal expression (Jiang *et al.,* 2000).

Disruptions in the normal movement of the vocal folds can cause many voice disorders. These are grouped into organic, functional, structural, neurological and acquired types (Stemple *et al.,* 2020). Accurately identifying these disorders with traditional clinical methods can be difficult and mistakes in interpretation often delay diagnosis.

To make assessment more reliable several imaging techniques such as laryngoscopy, stroboscopy and high-speed videoendoscopy (HSV) have been introduced to let clinicians directly observe how the vocal folds move (Deliyski and Hillman, 2010). HSV provides very high temporal resolution and gives a detailed view of vocal-fold vibrations (Deliyski *et al.,* 2008;

Malinowski *et al.,* 2024). Many image-processing methods have tried to analyze HSV recordings and describe how the vocal folds vibrate. Still, because HSV produces a huge number of frames each second, reviewing and diagnosing from these sequences can be demanding.

This challenge led to the creation of visualization methods that summarize vocal-fold vibration into single images for easier interpretation. These visualization tools have improved the accuracy of clinical evaluation. Broadly, they fall into two categories: local and global.

Local approaches such as digital kymography (DKG), vocal-fold trajectories (VFT), mucosal-wave kymography (MKG), and optical-flow kymography (OFKG), track motion along a line that cuts across the glottis. Global approaches such as the glottal optical-flow waveform (GOFW), glottal area waveform (GAW), glottovibrogram (GVG), and phonovibrogram (PVG), show how the entire glottis behaves through the vibration cycle (Andrade-Miranda *et al.,* 2020).

PVG provides a 3-D view of vocal-fold motion and converts dynamic vibrations into static maps that can be studied both visually and quantitatively.

This method was Introduced by Lohscheller *et al.,* in which we can extract and visualize the vocal fold vibrations along the entire edge of the glottis. Their work shows both visual and quantitative analysis in various conditions like normal phonation, laryngeal nerve paralysis, and functional voice disorders such as vocal

nodules (Bohr *et al.,* 2013, Doellinger *et al.,* 2007, Kunduk *et al.,* 2012, Lohscheller and Eysholdt, 2008, Patidar *et al.,* 2016).

Recent advances in machine learning have improved more in the clinical of PVG by automated feature extraction and classification of vocal fold pathologies (Döllinger *et al.,* 2011, Lohscheller, 2009, Schlegel *et al.,* 2020, Voigt *et al.,* 2010a, Voigt *et al.,* 2010b). These computational methods improve diagnostic objectivity, minimize inter-rater variability, and support early intervention, which are the important factors in improving clinical outcomes for individuals with voice disorders. Although many studies have investigated the visual and quantitative features of PVG, the application of PVG images for classifying pathologic conditions is still unexplored.

A machine learning-based study that involved transforming PVG contour lines into numerical feature vectors, which were then analyzed using a Support Vector Machine (SVM) classifier.

They tested their SVM method on both functional voice disorders and vocal fold paralysis. They achieved 78.5% accuracy for functional disorders and 93% for paralysis (Lohscheller, 2009).

Another study used a similar method focusing specially on vocal fold paralysis. They have shown 93% accuracy for classifying healthy vs. paralysis and 73% for a 3-class task (healthy, left paresis, right paresis) (Voigt *et al.,* 2010b).

In another study, they found that features extracted from PVGs performed better than traditional glottal parameters, producing an overall classification accuracy of 81% for predicting functional dysphonia (Voigt *et al.,* 2010a).

These studies shows the potential of PVG-based features to support the automated diagnosis of voice pathologies, including non-organic disorders such as paresis and muscle tension dysphonia (MTD).

In addition to that, in a study, authors found that SVM based machine learning classification of PVG-derived vibratory features outperforms acoustic feature analysis, particularly in identifying subtle phonation-dependent variations (Döllinger *et al.,* 2011). Based on current literature, only one study has utilized deep learning for the vocal fold disorder classification using PVG. In that study, authors showed classification accuracies of 82% using CNN-based LeNet architecture for a binary classification (physiologic vs. pathologic) and 85% for a multi-class classification (Healthy, MTD, Paresis, and Polyp) (Fehling *et al.,* 2020).

In our study we explore novel applications of various deep learning architectures to analyze PVG images, focusing on both binary classification (Healthy and unhealthy), tertiary classification (Healthy, functional, and organic), and multi-class classification (Healthy, MTD, atrophy, nodule, and edema) tasks. An overview of the proposed workflow is presented in Fig. 1.

MATERIALS AND METHODS

Dataset

This study utilizes the BAGLS (Benchmark for Automatic Glottis Segmentation) dataset which consists of 640 HSV recordings collected in seven different hospitals. It is a multi-centered dataset, which has a demographically diversified patient population. Also, with variations in age, gender, and vocal fold pathology. All recordings collected by clinical experts, using high-quality data acquisition and professional validation (Gomez *et al.,* 2020). The dataset contains both healthy subjects and patients with various vocal fold disorders makes it useful for developing and testing classification models.

PVG Generation

A detailed procedure for the PVG computation is found in (Doellinger *et al.,* 2007). In this current study, we used the Glottal analysis tool (GAT) to generate PVGs.

The glottis was fixed as the region of interest within the vocal fold (Kist *et al.,* 2021). First the glottal contours were extracted from each HSV frame using an edge segmentation algorithm.

The centerline of the glottis, referred to as the glottal axis, was then identified.

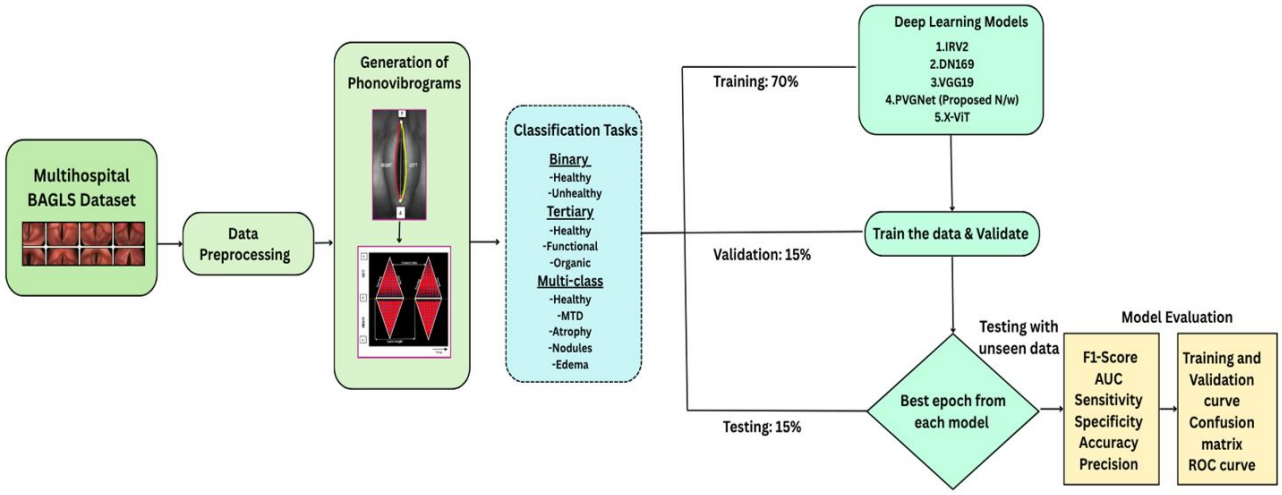

Fig. 1: *Workflow for phonovibrogram-based vocal fold disorders classification using the BAGLS dataset. The pipeline includes data preprocessing, PVG generation, classification (binary, tertiary, multiclass), model training, validation and testing, and evaluation using five different deep learning models and standard performance metrics.*

For each frame, the distances from points along the glottal axis to the corresponding points on the left and right fold contours were calculated and stored in a column vector.

The glottal axis was bisected, and the left contour was rotated 180 degrees around the posterior commissure to align it with the right contour.

Their respective distance vectors were then color-coded based on their magnitude: red indicates greater distances, black represents zero distance, and intermediate values were shown in gradations between red and black.

If a vocal fold contour crosses the glottal midline, it is marked in blue. This process was repeated across all frames, and the resulting vectors were concatenated into a 2-D matrix and visualized as the PVG, as illustrated in Fig. 2(a).

Data Preprocessing

Generated PVGs consist of multiple vibratory cycles, with each image having dimensions of 3000x720 pixels (length x width), corresponding to the HSV sequence.

Next, it was divided into a number of images with dimensions of 210x720 pixels, each capturing 3-4 cycles. This approach expanded the training dataset size while maintaining detailed analysis of temporal patterns. Fig. 2(b) shows the representative HSV frames for different vocal fold conditions and their corresponding kymograms. Table 1 presents the count of PVG images for each vocal fold condition.

Table 1. *Count of PVG images*

Conditions	PVG image counts
Healthy	5095
Functional/MTD	1626
Unhealthy	3016
Organic	954
Nodules	169
Edema	139
Atrophy	243

Resizing and Normalization

PVGs were resized to 128×128 for all models. After resizing the pixel values were converted to the range [0, 1] by normalization. Ensuring consistency across all images. To achieve a more balanced sample distribution, oversampling methods were applied, augmenting the underrepresented classes to achieve a more balanced distribution of samples. Only Deep Learning (DL)-based vibration features were used, and no manual descriptors were included.

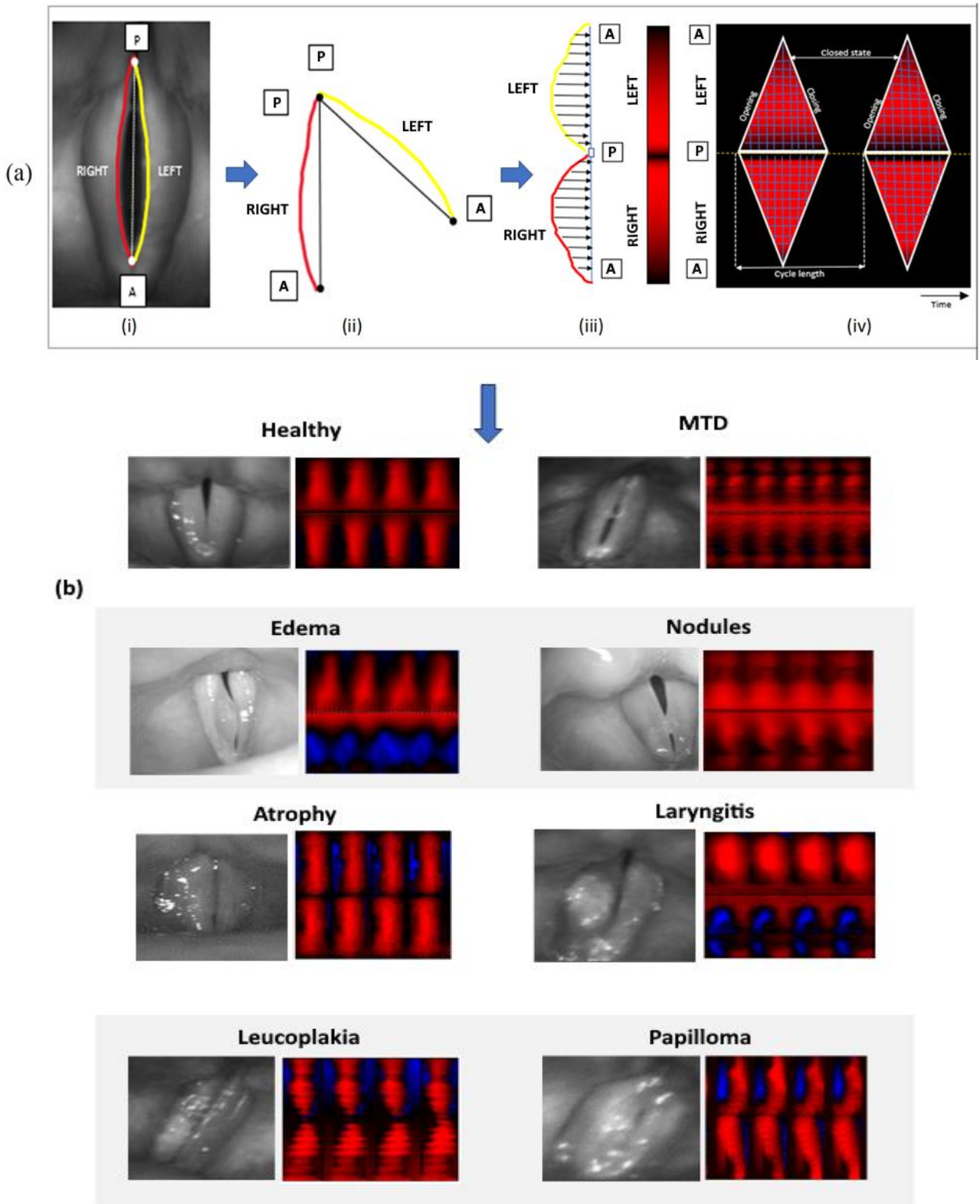

Fig. 2: *(a) Generation of PVGs from HSV recordings: (i) Glottal area segmentation in HSV frames; (ii) Contour extraction and splitting of the left and right vocal fold edges; (iii) Color coding of vibratory motion over time; (iv) Construction of the PVG 2(b) Representative samples of HSV endoscopic frames (left in each pair) and their corresponding PVGs (right in each pair) for different vocal fold conditions: Healthy, MTD, Edema, Nodules, Atrophy, Laryngitis, Leukoplakia, and Papilloma. These examples show the distinct morphological and vibratory patterns associated with each condition.*

Deep learning models

Our study focused on evaluating five different deep learning models for the classification of PVG images derived from HSV data.

The models include VGG19 (Simonyan and Zisserman, 2014), DenseNet169 (Huang et al., 2017), InceptionResNetV2 (Szegedy et al., 2017), a custom transformer ensemble model (Xception (Chollet, 2017) + Vision Transformer (Dosovitskiy et al., 2020)) named X-ViT, and a custom Hybrid CNN with an attention mechanism specifically designed for PVG, named PVGNet.

These models were chosen to investigate their effectiveness in analyzing vocal fold vibrations and classifying vocal fold pathologies based on PVG representations. Each model brings distinct advantages to the PVG classification task. VGG19 is a deep convolutional architecture with its simple yet improved hierarchical feature extraction capabilities. DenseNet169 uses dense connectivity between layers, helping efficient feature reuse and gradient flow. InceptionResNetV2 merge both the advantage of Inception modules and residual connections, improving both model performance and classification accuracy. Also the X-ViT model merges convolutional and transformer-based model using both local and global features for thorough PVG analysis (Ganaie et al., 2022).

Finally, PVGNet is a fully custom-built architecture developed specially for PVG classification. It combines the attention mechanisms to focus on important vibratory regions of interest (Wang et al., 2018). Unlike the other models, PVGNet does not rely on pretrained weights, allowing it to learn from scratch and making it highly specialized for this task.

Hybrid CNN-Attention Network (PVGNet)

The diagnostic accuracy of PVG depends on actively detecting subtle, multi-scale pathological patterns even in background noise (Doellinger and Berry, 2006).

Therefore, we developed PVGNet as a hybrid Convolutional Neural Network (CNN) improved with adaptive attention mechanisms, noise resilience, and multi-scale feature representation as shown in Fig. 3 (Hu *et al.,* 2020). PVGNet merges hierarchical feature extraction with attention modules to show diagnostically meaningful regions while repressing irrelevant information. The model processes input PVG images of size 128 × 128 × 3 via three convolutional blocks with increasing filter sizes of 64, 128, and 256.

Each block applies ReLU activation, then max pooling and batch normalization to stabilize convergence and improve training performance (Ioffe and Szegedy, 2015). This continuous convolutional block actively spots the layered complexity of PVG irregularities. Early layers documents intricate textures such as edema and vocal nodules using smaller receptive fields, deeper layers extract broader structural abnormalities like vocal fold atrophy and MTD via large filters (Chen *et al.,* 2018).

The attention mechanism is applied after the convolutional stages, warranting the model to selectively boost the most informative spatial regions pertinent to pathology.

Each block has max pooling and batch normalization to stabilize convergence, which is very important for handling PVG images. The convolutional operation at each layer is defined as:

$$F^l = \delta\left(W^l * F^{l-1} + b^l\right)$$

Where F^l is the feature maps at layer l, W^l and $b^l$are the learnable weights and biases, $*$ means the convolution operation, and δ-delta is the ReLU activation function. Increasing filter sizes continuously learns multi-scale features. The expanding filter hierarchy make sure that thorough pattern extraction across spatial scales innate. After the convolutional layers, PVGNet uses a squeeze-and-excitation (SE) attention block that dynamically improves diagnostically useful channels when repressing noise (Hu *et al.,* 2020).

This attention mechanism starts with compressing spatial information into compact channel descriptors via global average pooling

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{c(i,j)}$$

z_c is the c-th element of the channel descriptor. H, W are the feature map's height and width which compresses unwanted spatial noise, maintaining channel-wise discriminative information.

The excitation operation then models channel-wise dependencies:

$$s = \sigma\left(W_2\delta(W_1 z)\right)$$

Where $W_1 \in R^{\left(\frac{C}{r}\right)\times C}$ and $W_2 \in R^{C\times\left(\frac{C}{r}\right)}$ form a bottleneck with a reduction ratio r = 8. δ delta refers to ReLU, σ sigma is the sigmoid activation function, the recalibration stage weights each channel by multiplying the feature maps elementwise:

$$\tilde{F}_c = s_c \cdot F_c$$

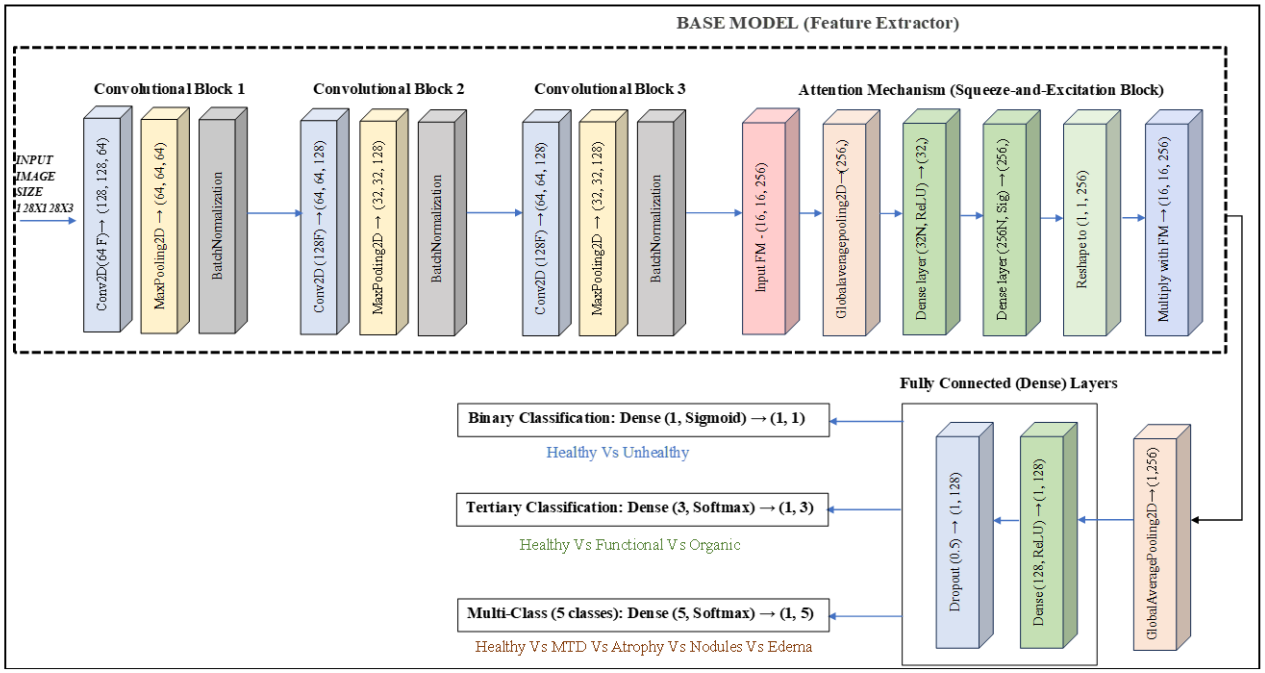

Fig. 3: *Proposed Hybrid CNN-Attention Network for PVG Classification (PVGNet), where F - Filters, N - Neurons, Sig - Sigmoid, and FM - Feature Maps*

The network applies global average pooling before sending features via two fully connected layers (256 and 128 neurons) with ReLU activation after the attention block. L2 regularization (λ=0.001) is applied to prevent overfitting (Krogh and Hertz, 1991):

$$\Omega(w) = \frac{\lambda}{2}\|w\|_2^2$$

Dropout (rate = 0.4) is set to improve generalization and prevent overfitting, helping the model become more robust to variations and noise in the PVG images (Srivastava *et al.,* 2014). The output layer uses a single neuron with sigmoid activation for binary classification of healthy vs. unhealthy PVG conditions (Bishop, 1995):

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The corresponding binary cross-entropy loss function(Murphy, 2012) is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

For tertiary and multi-class tasks, categorical cross-entropy loss (Heaton, 2017) is utilized:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(\hat{y}_{ij})$$

PVG images show different types of vibratory patterns, ranging from intricate vocal fold textures (e.g., edema, nodules) to broader structural alterations (e.g., atrophy, MTD) on their intensity variations.

We designed PVGNet specially to address this variability via a combination of multiscale feature extraction, an attention mechanism, and a noise-resilient design. Its hybrid architecture is intended to predict different vibratory features.

Experimental setup

Experiments were performed on an HP workstation, configured with an NVIDIA GeForce RTX 2080 Ti and 42.9 GB of GPU memory. Python was used with the TensorFlow 2.10 framework for executing the classification tasks.

The dataset consisted of HSV recordings of vocal fold vibrations with 376 recordings of Healthy and 163 Unhealthy.

A total of 101 recordings were excluded from our analysis: 50 due to missing health status information, and 51 due to the presence of multiple co-occurring disorders. 102 recordings were categorised as functional voice disorders, primarily MTD, and 56 were categorized as organic disorders.

The organic category included cases of scar tissue, papilloma, nodules, edema, carcinoma, laryngitis, polyps, cysts, and atrophy.

Data stratification for Binary Classification (Healthy vs Unhealthy)

For the binary classification task we worked with 8111 PVG images in total. To begin with, 15% of the data (1218 images) was set aside as the test set, containing 765 Healthy and 453 Unhealthy samples. This portion was never touched again and was used only for the final evaluation.

The remaining 6893 images were then divided into training (70%) and validation (30%) subsets. Before doing any balancing we separated a validation set of 1034 images (around 15% of the total) to keep its class distribution natural. The training set (5859 images) showed a clear class imbalance which could easily bias the model toward the majority category.

To fix this we used random oversampling but only on the training data. This increased the underrepresented class until both were evenly matched resulting in a balanced training set of 7360 images. The validation and test sets were left untouched so they will reflect the true distribution allow a fair and real world evaluation of performance.

Data stratification for Tertiary Classification (Healthy vs. Functional vs. Organic)

For the tertiary class setup we used 6522 images. From this, about 15% (1153 images) was kept aside as the test set. The remaining data was then split into a training set (5543 images) and a validation set (979 images). As before the validation set was separated before handling any imbalance to maintain its natural mix of classes.

The training data had uneven representation across the three classes so we again applied random oversampling. This brought all classes to roughly equal levels, a balanced training of 11040 samples. The validation and test sets were kept in their original form to provide a realistic check on how well the model would generalize.

Data stratification for Multi-class Classification (Healthy vs. MTD vs Atrophy vs Nodule vs Edema)

In the five-class we analyzed 7272 images in total, distributed as Healthy (5095), MTD (1626), Atrophy (232), Nodule (180), and Edema (139). This imbalance was quite pronounced, so special care was needed during training.

We first allocated 15% from each class to the test set to make sure every category was represented in evaluation. The rest formed the training and validation sets. As with the earlier tasks we used random oversampling to balance the training data. The smaller classes were increased till they matched the size of the Healthy group giving us a balanced training set of 18400 images with equal class.

The validation and test sets were not modified so that performance results would reflect how the model behaves on naturally imbalanced and real world data.

Training and Validation Phase

Random seed value of 42 was used for all classification tasks to maintain reproducibility and consistency. This warranted proper data partitioning into training and validation sets. The same applied in the random oversample to maintain consistency in class balancing through oversampling. The model trained using adam optimizer for 100 epochs using binary cross-entropy for binary class and categorical cross-entropy for tertiary and multi class tasks.

Hyperparameter tuning strategies such as learning rate scheduling and early stopping, reduce overfitting and improve performance. During training, model checkpoints were saved at the epochs producing the best validation performance. The final testing was conducted on a separate, previously unseen test set using the best-performing saved model.

Testing Phase

We used both qualitative and quantitative metrics to thoroughly evaluate PVG models. Quantitative measures such as accuracy, F1-score, sensitivity, specificity, and precision were used to objectively compare performance. Qualitative inspection with learning curves, confusion matrices, and ROC curves.

RESULTS

Evaluation of Binary Classification Performance

As shown in Fig. 4(a), PVGNet exhibits strong generalization capabilities, with a small aperture between training and validation accuracy shows a low risk of overfitting and reliable performance for real-world applications.

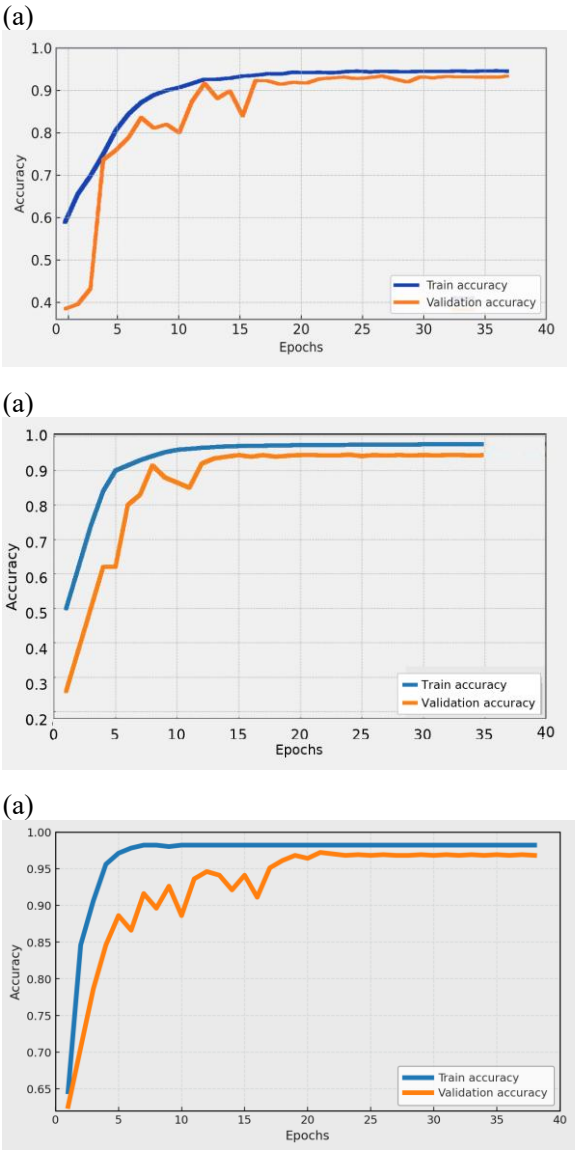

Fig.4. *Training and validation performance curves for proposed network (PVGNet) across three classification tasks:(a) Binary classification(top) ,(b) Tertiary classification (middle), and(c) Multi-class(bottom).*

In contrast, we observed in some models showing noticeably larger gaps, suggesting a tendency to overfit and reduced ability to generalize to unseen data. Others remain relatively stable, but display early performance plateaus, which can limit further learning and improvement.

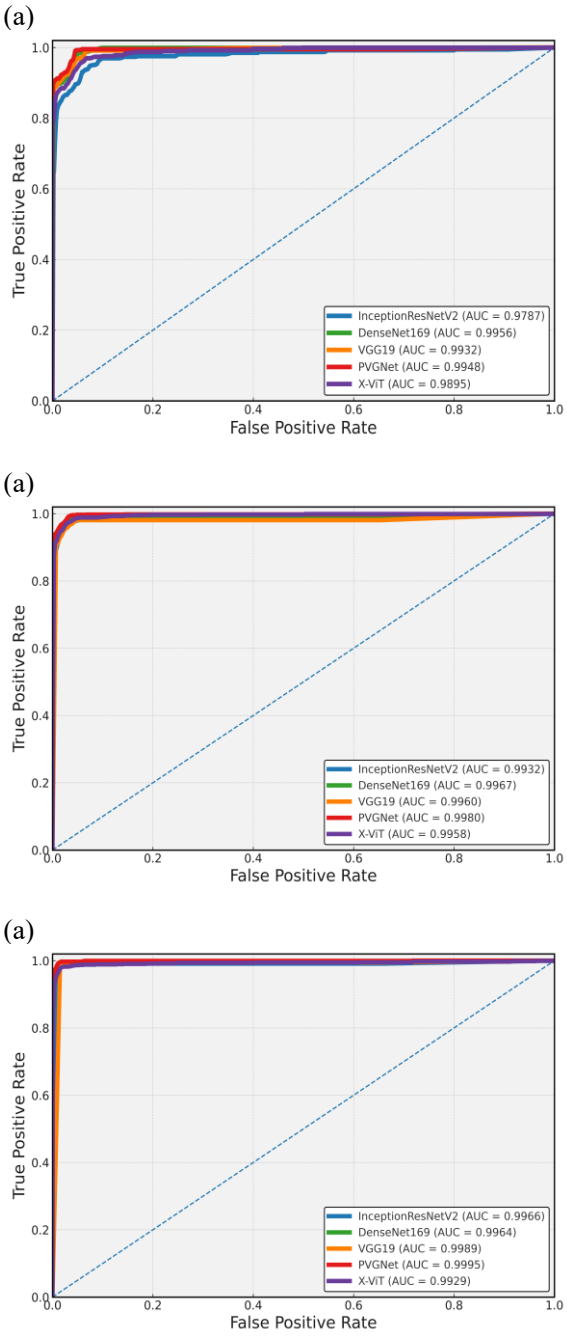

Fig.5. ROC curves for binary (top), tertiary (middle), and multi-class (bottom) classifications across IRV2, VGG19, DN169, PVGNet, and X-ViT.

Fig.6. *Confusion matrices across three classification tasks:(a) Binary classification,(b) Tertiary classification, and(c) Multi-class classification. Each subFig. presents results from five models (left to right): InceptionResNetV2 (IRV2), VGG19, DenseNet169 (DN169), PVGNet, and X-ViT.*

Table 2**.** *Binary classification performance of five models (IRV2 – InceptionResNetV2, DN169 – DenseNet169, X-ViT – Xception + Vision Transformer)*

Metrics	Condition	IRV2	VGG19	DN169	PVGNet	X-ViT
AUC (%)		97.87	99.32	99.56	99.48	98.95
F1 score (%)	Healthy	94.69	96.37	96.01	97.01	95.82
	Unhealthy	91.11	94.02	93.29	95.16	92.92
Precision (%)	Healthy	95.00	97.21	96.19	98.52	95.70
	Unhealthy	90.61	92.70	92.98	92.86	93.13
Sensitivity (%)	Healthy	94.69	95.56	95.82	95.56	96.96
	Unhealthy	91.11	95.36	93.60	97.57	95.02
Specificity (%)		91.61	95.36	93.60	95.59	92.71
Accuracy (%)		93.35	95.48	94.99	96.31	94.75

Table 3. *Tertiary classification performance of five models (IRV2 – InceptionResNetV2, DN169 –DenseNet169, X-ViT – Xception + Vision Transformer)*

Metrics	Condition	IRv2	VGG19	DN169	PVGNet	X-ViT
AUC (%)	Healthy	98.71	99.44	99.49	99.61	99.21
	Functional	98.43	99.37	99.40	99.74	99.25
	Organic	99.77	99.88	99.75	99.71	99.77
F1 score (%)	Healthy	97.03	96.73	97.12	97.82	96.85
	Functional	91.24	93.44	91.84	94.47	92.09
	Organic	94.85	92.81	95.17	94.81	94.12
Precision (%)	Healthy	98.13	98.91	97.37	99.06	97.23
	Functional	88.76	90.73	91.46	91.22	91.16
	Organic	93.88	87.65	94.52	94.48	93.79
Sensitivity (%)	Healthy	95.95	94.64	96.86	96.60	96.47
	Functional	93.85	96.31	92.21	97.95	93.03
	Organic	95.83	98.61	95.83	95.14	94.44
Specificity (%)	Healthy	98.05	97.94	94.85	98.19	94.59
	Functional	96.91	97.36	97.69	97.47	97.58
	Organic	99.20	98.02	99.21	99.21	99.11
Accuracy (%)		95.49	95.49	95.75	96.70	95.49

Table 4. *Multi-class classification performance of five models (IRV2 – InceptionResNetV2, DN169 – DenseNet169, X-ViT – Xception + Vision Transformer)*

Metrics	**Conditions**	IRv2	VGG19	DN169	PVGNet	X-ViT
AUC (%)	Healthy	98.95	99.57	98.64	99.79	98.33
	MTD	99.42	99.58	98.89	99.83	97.75
	Atrophy	97.80	99.95	98.01	99.96	99.91
	Nodules	100.0	100.0	100.0	100.0	100.0
	Edema	97.36	99.66	99.79	99.79	99.27
F1 score (%)	Healthy	96.52	97.63	97.13	98.22	97.30
	MTD	91.20	93.93	90.69	95.83	93.39
	Atrophy	91.67	94.74	84.51	92.11	89.47
	Nodules	100.0	100.0	100.0	100.0	95.83
	Edema	91.89	94.74	97.14	94.74	88.89
Precision (%)	Healthy	97.09	98.15	98.12	99.33	97.88
	MTD	89.06	92.80	89.62	93.05	91.37
	Atrophy	91.67	90.00	75.00	87.50	85.00
	Nodules	100.0	100.0	100.0	100.0	100.0
	Edema	100.0	100.0	100.0	100.0	100.0
Sensitivity (%)	Healthy	95.95	97.12	96.15	97.12	96.73
	MTD	93.44	95.08	91.79	98.77	95.49
	Atrophy	91.67	100.0	96.77	97.22	94.44
	Nodules	100.0	100.0	100.0	100.0	92.00
	Edema	85.00	90.00	94.44	90.00	80.00
Specificity (%)	Healthy	93.23	95.69	94.46	98.46	95.08
	MTD	96.69	97.87	97.99	97.87	97.40
	Atrophy	99.72	99.62	99.72	99.53	99.43
	Nodules	100.0	100.0	100.0	100.0	100.0
	Edema	100.0	100.0	100.0	100.0	100.0
Accuracy (%)		95.14	96.70	96.51	97.43	95.96

Table 2 shows that PVGNet performs best with an accuracy of 96.31%. VGG19 follows at 95.48%, and DenseNet169 comes close at 94.99%.

Although DenseNet169 gives the highest AUC of 99.56%, meaning it separates the two classes very well, PVGNet's AUC of 99.48% is almost the same. What makes PVGNet stand out is how well it balances accuracy, sensitivity, and specificity instead of being strong in just one area.PVGNet also gives the highest F1-scores for both classes 97.01% for Healthy and 95.16% for Unhealthy showing that it can handle both categories reliably without bias.The confusion matrices in Fig. 6(a) make this clearer.

IRV2 shown first from the left struggles to detect Unhealthy cases misclassifying 38 of them as Healthy which means a high false-negative rate. VGG19 shown next performs better with 432 correct Unhealthy detections and only 21 missed. DenseNet169 is close but misses slightly more (29). PVGNet shown fourth clearly performs the best. It misses only 11 Unhealthy cases and correctly finds 442 which shows improved learning.

Though DenseNet169 has a slightly higher AUC, PVGNet stays nearly close while keeping false negatives much lower. This balance detecting more Unhealthy cases without losing precision is what matters most in voice disorder detection. Overall, these results show that PVGNet's hybrid CNN–attention setup can capture fine visual details in PVG images more effectively than the other models. It doesn't just classify it learns the subtle cues that distinguish healthy from disordered patterns

Evaluation of Tertiary Classification Performance

In Fig. 4(b) PVGNet generalizes well since the training and validation accuracy curves almost overlap, which points to a low risk of overfitting. Some baselines do reasonably well but don't generalize as strongly, a few clearly overfit with wide gaps between training and validation. Others stay stable yet plateau early by limiting further gains. Table 3 shows the same. PVGNet shows the highest test accuracy at 96.70%, ahead of DenseNet169 (95.75%), VGG19 (95.49%), and X-ViT (95.49%). Also gives the high F1-scores for Healthy (97.82%) and Functional (94.47%). DenseNet169 leads on Organic with an F1 of 95.17%, and PVGNet is close at 94.81%. Among the baselines IRV2 and VGG19 show clear overfitting with large train–val gaps but DenseNet169 is steadier but levels off early. PVGNet shows slight signs of overfitting in place but keeps validation accuracy high, unlike models that struggle more generalization.

The confusion matrices in Fig. 6(b) shows DenseNet169 is marginally best (741 correct) for healthy cases. Then PVGNet and X-ViT (739 and 738) but IRV2 and VGG19 misclassify 24 and 22 Healthy. For Functional cases, PVGNet performs well with only 5 errors (4 Healthy, 1 Organic). X-ViT makes more mispredictions (15 Healthy, 2 Organic), DenseNet169 misclassifies 18, and IRV2 (13 Healthy, 5 Organic). For Organic cases, DenseNet169 again leads (2 Healthy, 4 Functional). PVGNet stays close (3 Healthy, 4 Functional).X-ViT shows slightly higher errors (6 as Healthy, 2 as Functional), and IRV2 trails (5 Functional, 1 Healthy).

Fig. 5(b) provides the ROC view. PVGNet's micro-average AUC is 99.80%. It also has the highest class AUCs for Healthy of 99.61% and Functional of 99.74%. Tertiary performance mirrors the binary case: high accuracy, balanced errors, and good generalization.

This behavior reflects the attention blocks focusing on the most informative PVG informations. That focus helps PVGNet keep false decisions low without giving up separability, which explains its strong accuracy and reliable class wise detection.

Evaluation of Multi-class Classification Performance

PVGNet shows smooth convergence for both training and validation with minimal fluctuation indicating stable learning and strong generalization in Fig. 4(c). Several baselines require more epochs to stabilize and improve more gradually others show marked variation in validation accuracy suggesting data sensitivity and potential instability.

PVGNet shows mild overfitting in places yet validation accuracy remains high unlike models with weaker generalization. Table 4 is consistent with these patterns. PVGNet reports the highest test accuracy (97.43%) and strong overall classification performance, and it attains the top AUC for most conditions. Class-wise results favor PVGNet: F1, precision, specificity, and sensitivity remain high, with F1 of 98.22% (Healthy), 95.83% (MTD), and 92.11% (Atrophy).

VGG19 is close overall but its sensitivity and F1 lags in some settings. DenseNet169 is competitive yet drops on MTD (90.69% F1) and Atrophy (84.51% F1), lowering its overall score. IRV2 is weaker for MTD (91.20% F1). X-ViT generalizes least well, particularly for Atrophy (89.47% F1) and Edema (88.89% F1).

The confusion matrices in Fig. 6(c) clarify these outcomes. PVGNet (fourth matrix) gives few errors, with correct counts of 763 Healthy, 241 MTD, 34 Atrophy, 25 Nodules, and 18 Edema. IRV2 (first) and DenseNet169 (third) misclassify a noticeable number of Healthy and MTD samples.

VGG19 (second) performs well but produces more false positives than PVGNet. X-ViT (fifth) shows several errors in MTD, Atrophy, and Edema. Considering the learning curves (Fig. 4(c)), test metrics (Table 4), and confusion matrices (Fig. 6(c)) together, PVGNet provides the most favorable balance of accuracy, generalization, and class-wise reliability for this multi-class task.

VGG19 is a strong comparator, but PVGNet's higher sensitivity, stronger F1 scores, and lower misclassification rates make it the more suitable choice for vocal-fold multi-class classification. Finally, Fig. 5(c) presents the ROC analysis. PVGNet's micro-average AUC is 99.95%, and it records the highest per-class AUCs 99.79% (Healthy), 99.83% (MTD), 99.96% (Atrophy), 100% (Nodules), and 99.79% (Edema) indicating excellent separability across all categories and a clear margin over IRV2, VGG19, DenseNet169, and X-ViT.

DISCUSSION

Binary classification result is an evident that PVGNet's hybrid CNN-attention mechanism extracts intricate and informative PVG patterns and improves feature recognition, performing better than other models.

This balance between precision and sensitivity is specially important in clinical voice diagnostics where false negatives may delay correct treatment. PVGNet also shows consistent reliability across tertiary and multiclass tasks.

PVGNet shows potential for real-world clinical application by maintaining solid performance across diverse voice disorder categories,

Though DenseNet169 produces high AUCs in some cases PVGNet maintains better balance in sensitivity and F1-scores across conditions mainly where misclassification can have major diagnostic consequences.

This further supports its suitability for use in automated screening or adjunct diagnostic tools for otolaryngologists.

CONCLUSION

This study presents PVGNet, a custom DL framework that produces high accuracy, good generalisation, and balanced performance across binary, tertiary, and multi-class classification tasks. Its hybrid architecture that combines multiscale feature extraction with attention mechanisms learns diagnostically relevant vibration features and reduces overfitting. These outcomes show the potential of attention-enhanced models like PVGNet in advancing automated voice-disorder assessment and their integration into clinical decision-support systems.

In future studies, visualization maps such as Grad-CAM will be explored to understand vibration regions, improving interpretability and clinician trust. Although this work is limited to a single HSV dataset, model behavior may vary under real-world clinical conditions with greater variability and noise. Future research will include domain-adaptation techniques, and multi-fold cross-validation to further strengthen generalizability. The model will eventually be expanded and refined into clinically deployable software tool.

ACKNOWLEDGEMENTS

This work was carried out with support from the SSN Trust grant (SSN/IFFP/January2019/1-12/08).

B. Panchami gratefully acknowledges the UGC Fellowship (UGCES-22-TAM-F-SJSGC-1873) for providing financial assistance during her doctoral research.

REFERENCES

Andrade-Miranda G, Stylianou Y, Deliyski DD, Godino-Llorente JI, Henrich Bernardoni N (2020). Laryngeal image processing of vocal folds motion. Appl Sci 10:1556.

Bishop CM (1995). Neural networks for pattern recognition. Oxford Univ Press.

Bohr C, Kraeck A, Eysholdt U, Ziethe A, Dollinger M (2013). Quantitative analysis of organic vocal fold pathologies in females by high-speed endoscopy. Laryngoscope 123:1686–93.

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40:834–48.

Chollet F (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 1251–58.

Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE (2008). Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. Folia Phoniatr Logop 60:33–44.

Deliyski DD, Hillman RE (2010). State of the art laryngeal imaging: Research and clinical implications. Curr Opin Otolaryngol Head Neck Surg 18:147–52.

Doellinger M, Berry DA (2006). Visualization and quantification of the medial surface dynamics of an excised human vocal fold during phonation. J Voice 20:401–13.

Švec JG, Schutte HK, Svec H (2007). Phonovibrography: The fingerprint of vocal fold vibrations. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing; 949–52.

Döllinger M, Lohscheller J, Svec J, McWhorter A, Kunduk M (2011). Support vector machine classification of vocal fold vibrations based on phonovibrogram features. J Voice 25:435–56.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv abs/2010.11929.

Fehling MK, Grosch F, Schuster ME, Schick B, Lohscheller J (2020). Fully automatic segmentation of glottis and vocal folds in endoscopic laryngeal high-speed videos using a deep convolutional lstm network. PLoS One 15:e0227791.

Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN (2022). Ensemble deep learning: A review. Eng Appl Artif Intell 115:105151.

Gomez P, Kist AM, Schlegel P, Berry DA, Chhetri DK, Durr S, Echternach M, Johnson AM, Kniesburges S, Kunduk M, Maryn Y, Schutzenberger A, Verguts M, Dollinger M (2020). BAGLS, a multihospital benchmark for automatic glottis segmentation. Sci Data 7:186.

Heaton J (2017). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. Genet Program Evolvable Mach 19.

Hu J, Shen L, Albanie S, Sun G, Wu E (2020). Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42:2011–23.

Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition;4700–08.

Ioffe S, Szegedy C (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning; 448–56.

Jiang J, Lin E, Hanson DG (2000). Vocal fold physiology. Otolaryngol Clin North Am 33:699–718.

Kamiloğlu RG, Sauter DA (2021). Voice production and perception. Oxford University Press.

Kist AM, Gomez P, Dubrovskiy D, Schlegel P, Kunduk M, Echternach M, Patel R, Semmler M, Bohr C, Durr S, Schutzenberger A, Dollinger M (2021). A deep learning enhanced novel software tool for laryngeal dynamics analysis. J Speech Lang Hear Res 64:1889–903.

Krogh A, Hertz JA (1991). A simple weight decay can improve generalization. Proceedings of the 5th International Conference on Neural Information Processing Systems; 950–7.

Kunduk M, Döllinger M, McWhorter AJ, Švec JG, Lohscheller J (2012). Vocal fold vibratory behavior changes following surgical treatment of polyps investigated with high-speed videoendoscopy and phonovibrography. Ann Otol Rhinol Laryngol 121:355–63.

Lohscheller J, Eysholdt U (2008). Phonovibrogram visualization of entire vocal fold dynamics. Laryngoscope 118:753–8.

Lohscheller J (2009). Towards evidence based diagnosis of voice disorders using phonovibrograms. International Symposium on Applied Sciences in Biomedical and Communication Technologies; 1–4.

Patidar M, Agrawal J (2016). Which mathematical and physiological formulas are describing voice pathology: An overview. J Gen Pract 4:1–4.

Malinowski J, Pietruszewska W, Kowalczyk M, Niebudek-Bogusz E (2024).Value of high-speed videoendoscopy as an auxiliary tool in differentiation of benign and malignant unilateral vocal lesions.

Murphy KP (2012). Machine learning: A probabilistic perspective.MITPress.

Schlegel P, Kniesburges S, Durr S, Schutzenberger A, Dollinger M (2020). Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings. Sci Rep 10:10517.

Simonyan K, Zisserman A (2014). Very deep convolutional networks for large-scale image recognition. arXiv 14091556.

Spina AL, Maunsell R, Sandalo K, Gusmao R, Crespo A (2009). Correlation between voice and life quality and occupation. Braz J Otorhinolaryngol 75:275–9.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014). Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res15:1929–58.

Stemple JC, Roy N, Klaben B (2020). Clinical voice pathology: Theory and management. Sixth edition. San Diego, CA: Plural Publishing, Inc.

Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence 31.

Voigt D, Dollinger M, Braunschweig T, Yang A, Eysholdt U, Lohscheller J (2010a). Classification of functional voice disorders based on phonovibrograms. Artif Intell Med 49:51–9.

Voigt D, Dollinger M, Yang A, Eysholdt U, Lohscheller J (2010b). Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. Comput Methods Programs Biomed 99:275–88.

Wang X, Girshick RB, Gupta A, He K (2018). Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 7794–7803.