IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 1069-1076
Journal homepage: http://iieta.org/journals/ts

Occlusion Sensitivity Based Morphological Analysis of Mel Spectrograms for Noise Robust Pathological Voice Detection

Sundararajan Mukesh[1], Subbaraj Pravin Kumar[2*]

[1] Department of Biomedical Engineering, Sona College of Technology, Salem 636005, India
[2] Department of Biomedical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai 603110, India

Corresponding Author Email: pravinkumars@ssn.edu.in

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.430240

Received: 8 February 2026
Revised: 15 April 2026
Accepted: 23 April 2026
Available online: 30 April 2026

Keywords:
voice disorder detection, deep learning models, occlusion sensitivity analysis, Mel spectrogram, spatial patterns

ABSTRACT

Automated characterization of pathological voice patterns highly depends on the effective methods of extracting discriminative features. In this paper voice disorder detection is formulated as a 2D pattern recognition problem and Mel spectrograms as the topographical representations of spectral energy distribution. Hierarchical spatial filtering on a set of 296 high-resolution time-frequency maps (ICC > 0.8) from perceptual voice qualities database (PVQD) was performed and classified using a convolutional neural network (CNN). To test the image degradation sensitivity of the model we injected stochastic noise of 10 dB signal-to-noise ratios (SNR) and 20 dB SNR which is equivalent to spectral noise in practice. The overall 5-fold cross validation of the proposed architecture was 98.04% (SD = 1.93) and the accuracy of architecture test was 97.97%. It was shown through occlusion sensitivity mapping based morphological analysis that the classification of the model is largely dependent upon the low-frequency Mel bands, which is associated with the visual manifestations of breathiness and strain within the spectral texture. The system showed better performance of 99.13% accuracy compared to the traditional hybrid convolutional neural network-long short-term memory (CNN-LSTM) and SVD-based systems at high-entropy conditions of 10 dB SNR. Results with the limited dataset available indicate the relevance of using the spatial feature extraction for the objective screening of vocal pathologies.

1. INTRODUCTION

Voice disorders are causing public health concern as they affect 7% of the general population at any time and 30% the professional vocal users such as teachers, singers, and operators of call centers [1].

These disorders are either asymptomatic lesions of the vocal folds or neurological and functional dysphonia [2, 3].

They impair communication, impact quality of life, and place a major financial burden on the economy through productivity loss and health care costs.

Early and accurate diagnosis is of great importance in achieving better treatment outcomes.

Nevertheless, the conventional clinical setup is heavily reliant on subjective perception measures by trained laryngologists or speech-language pathologists through standardized test procedures including:

(i) Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) [4]
(ii) Grade-Roughness-Breathiness -Asthenia-Strain (GRBAS) index [5].

Though these measures are clinically meaningful, they are also susceptible to subjective variations such as inter-rater error, even among expert clinicians (Intraclass Cross Correlation (ICC) is typically 0.80-92) [6]. The subjectivity of perceptual rating processes and the time-intensive nature of these techniques have prompted the necessity of for objective techniques in the detection of voice disorders.

Most of the previously attempted with machine learning methods were based on signal based acoustic features, such as cepstral peak prominence, jitter, harmonics to-noise ratio (HNR), and shimmer [7, 8].

1.1 2D transformation

The developments in digital signal processing have allowed the transformation of 1D acoustic signals to the 2D time-frequency representations.

There is a need to investigate the connection between the severity scores and these visual representations.

One of them is the Mel spectrogram, which is a special topographical map where the frequency bands are weighted according to human hearing.

Once transformed, voice recording is no longer a 1D waveform and it becomes a 2D digital texture that makes the following representations and further analysis based on them possible:

- The harmonic patterns may be regarded as linear and geometrical patterns.
- Pathological changes like breathiness or roughness

can be in the form of stochastic pixel difference or spectral blur.

1.2 Convolutional neural networks as hierarchical spatial filters

In recent years, convolutional neural networks (CNNs) and other deep learning architectures have been shown that they performed better by learningly finding hierarchical representations in time-frequency representations of speech, typically Mel spectrograms [9].

As mentioned earlier, Mel spectrograms transform one dimensional audio signal into two dimensional mappings of spectral energy distribution, making it possible to apply powerful spatial feature extractions techniques, originally developed to operate on image data [10, 11].

Bashir et al. [12] trained a Visual Geometry Group 16 layers (VGG16) and showed an F1-score of 0.97 on mel spectrograms of Arabic Voice Pathology Database (AVPD) and the performance of this approach has in detecting pathological voices, which confirms the capability of deep learning models to predict deviations in the complex acoustic features.

Similarly, it has been shown that hybrid models such as long short-term memory (LSTM) networks along with CNNs show an accuracy level of 92.71, implying that the temporal dependencies lead to a more effective classification performance

1.3 Noise resilience

Despite these developments, a number of challenges still exists. The fact that most of these models are not robust to real-world noise is one of the major constraints of telehealth and mobile health systems, since recordings may contain noisy acoustic conditions [13-18].

The classical models can be shown to be degraded at least to 20 dB signal-to-noise ratios (SNR) which motivates the use of data augmentation techniques [19]. As an example, the noisy SNR (0–15 dB) data have been applied in training to improve the resilience of the model [20].

This approach helps to be consistent with clinical expectations for robust diagnostics under varying conditions.

Another issue is with the distribution of voice pathology data, where pathology samples are usually overrepresented in comparison with healthy controls, which can cause bias in model predictions. Moreover, deep learning models are characterized by a low degree of interpretability, thus, it is difficult to obtain clinical acceptance and reliability [21].

1.4 Morphological validation

The black box approach of deep learning is one of the major concerns in the medical image classification applications.

This paper addresses this through occlusion sensitivity analysis which performs a digital morphometry of the diagnostic process. It displays the local areas of the spectrogram which have the most influence on making the classification decisions.

By gradual masking of spatial regions of the spectrogram, we can determine the specific frequency in neighbourhoods that influences the model to make its decision.

Therefore, by redefining voice disorder detection as a morphological texture analysis problem on 2D spectral images, the work proposes a scalable interpretable image-based pattern recognition system to objectively classify vocal pathological problems.

This can bridge the visual interpretation of voice and quantitative pattern recognition of images.

1.5 Objectives of the study

In the current research, voice disorder detection is viewed as a morphological texture analysis of the 2D spectral images and the following research questions are formulated.

1. How do Mel spectrogram textures visually correlate with GRBAS severity levels?
2. Can occlusion sensitivity maps on Mel spectrograms identify frequency-band-specific regions that align with individual GRBAS attributes?
3. How does noise augmentation impact Mel spectrogram-derived classification performance?

We use a CNN trained on Mel spectrograms based on a perceptual voice qualities database (PVQD) dataset to do binary classification of healthy and pathological voices.

Data augmentation such as random time-domain stretching, pitch shifting, and controlled additive white Gaussian noise at 10 dB and 20 dB SNR has been used in the training pipeline to enhance generalization and noise resilience.

Model performance is tested using 5-fold cross validation, independent test-set metrics, and SNR per class calculations.

Occlusion sensitivity mapping is obtained for the interpretability as it provides mapping of the spatial important features of the spectrogram and indicates the morphological areas that drive the diagnostic decision-making process.

2. METHODOLOGY

The proposed workflow (Figure 1) transforms raw acoustic signals into a high-dimensional spatial representation for binary classification (Healthy vs. Pathological).

Audio samples from the PVQD corpus were preprocessed into fixed-size Mel spectrograms, augmented to enhance generalization and fed into a customized CNN model.

The model used 5-fold cross-validation (CV) for training and optimization, and the predictions were further analyzed for interpretability.

2.1 Preprocessing

The high-fidelity corpus employed in this work was the freely available PVQD [6], which is specifically designed with the perceptual assessment of voice quality.

It contains 296 audio samples of adult speakers with and without voice complaints, and sustained vowels (/a/ and /i/) and standardized sentences.

All the recordings were captured in a quiet clinical setting with the help of a condenser microphone mounted on head at a 44.1 kHz sampling frequency, with minimized background noise and uniform acoustic recordings [6].

Experienced clinicians provided perceptual ratings on two standardized scales: the CAPE-V to indicate overall severity and individual attributes and the GRBAS scale, which uses ordinal ratings (0–3) [4, 5].

The files were rated by a number of raters (a minimum of 2, mostly 4), that shows considerable inter-rater reliability (ICC) (ICC > 0.80 on most features, with the overall ICC ≈ 0.89) [6]. The spreadsheets that accompany these materials have demographic information and voice pathology rating scores.

To classify as binary (healthy vs. pathological) classes, voice samples were considered pathological when the overall average severity score of CAPE-V was above 10 mm on the VAS (visual analog scale), which is a conservative threshold used to differentiate between mild or more severe dysphonia and healthy voice.

This gave a class distribution of 78 healthy and 218 pathological samples.

All audio files were down sampled to 8 kHz to minimise computation costs while maintaining the perceptually relevant frequency contents.

The audio was segmented into fixed-length frames of 25 ms with a 12.5 ms overlap between consecutive frames, and each frame was weighted with a Hamming window.

Mel spectrograms were then calculated with 128 Mel bands represented by 128 × 128 pixel images through zero-padding or truncation. The resultant spectrograms were normalized by mean of zero and unit variance across the whole dataset.

2.2 Data augmentation

Table 1 shows the four data augmentation strategies that were applied to improve generalization and to simulate real-world acoustic variability.

Table 1. Data augmentation methods applied

Time-Domain Stretching	Random Scaling Factor ∈ [0.95, 1.05] Using Phase Vocoder Resampling
Pitch Shifting	Random semitone shift ∈ [−2, +4] semitones.
Additive White Gaussian Noise	Noise added at signal-to-noise ratios (SNR) levels of 10 dB or 20 dB, randomly selected per sample.
SpecAugment-Style Time/Frequency Masking	Random masking of 0–10% of time or frequency bins.

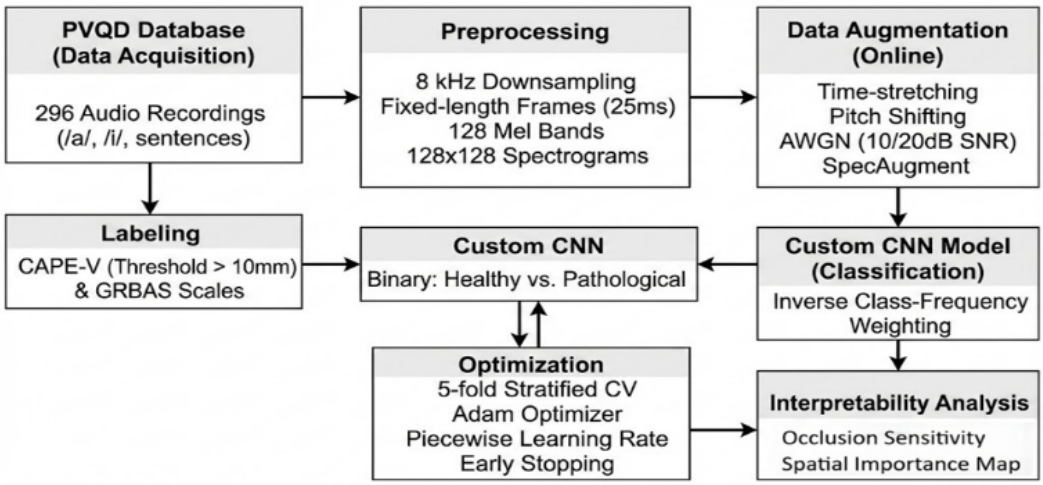

Figure 1. Workflow of the proposed system: Raw acoustic signals are transformed into Mel spectrograms, augmented for robustness, and classified using a custom convolutional neural network (CNN). Interpretability is assessed via occlusion sensitivity mapping

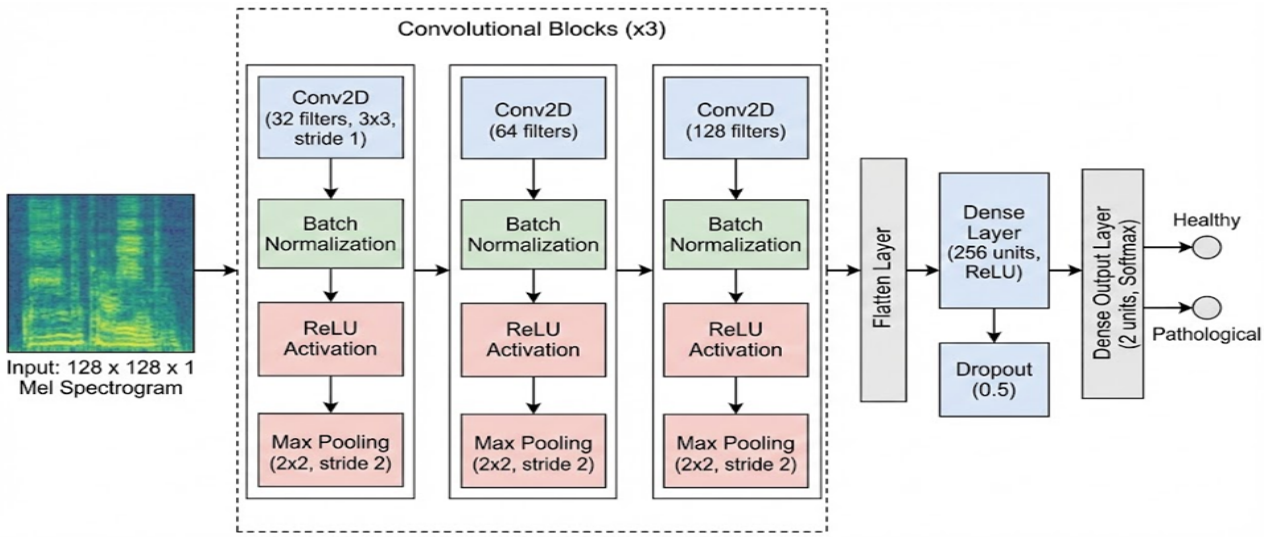

Figure 1. Custom convolutional neural network (CNN) Topology. The raw mel spectrogram topographical mappings are subjected to hierarchical filters, batch normalized and flattened to extract dominant features to categorize them as either Healthy or Pathological

Augmentation was applied during training, and it was aimed at increasing the effective training set size by a factor of 4 while preserving the original class distribution through stratified sampling.

2.3 Convolutional neural network architecture

A custom CNN architecture takes the time-frequency representation of mel spectrogram as a single-channel grayscale topographic map in the form of a 128 × 128 × 1 input tensor and performs hierarchical spatial filtering.

The network is built around three consecutive blocks of convolutional layers intended to capture complicated morphological features of the spectral textures (Figure 2). The description of network components is provided in Table 2.

Table 2. CNN architecture specifications

Layer	Specification	Details
Input	128 × 128 × 1	Mel spectrogram
Conv Block 1	3 × 3, 32 filters	BN + ReLU + 2 × 2 MaxPool
Conv Block 2	3 × 3, 64 filters	BN + ReLU + 2 × 2 MaxPool
Conv Block 3	3 × 3, 128 filters	BN + ReLU + 2 × 2 MaxPool
Flatten + Dense	256 units, ReLU	Dropout (0.5)
Output	2 units, Softmax	Binary classification
Optimizer	Adam	Initial LR = 1e-4
LR Schedule	Piecewise, drop 0.3 every 15 epochs	-
Loss	Weighted Binary Cross-Entropy	Class imbalance correction
Batch Size	16	-
Epochs	30 (with early stopping)	Patience = 20

In order to correct the issues of class imbalance in the PVQD data, inverse class-frequency weighting was incorporated into the cross-entropy loss function. The Adam algorithm (initial learning rate 1×10^{-4}) with piecewise learning rate decay and early stopping criterion were used to optimise the model and achieve the best convergence.

2.4 Occlusion sensitivity analysis

Occlusion sensitivity analysis [22] was performed to identify the spectrogram regions that are most influential for classification decisions. A 4 × 4 pixel occlusion patch was systematically shifted across the 128 × 128 spectrogram (stride 4 pixels).

Each occluded region was replaced with the global mean intensity, and the resulting drop in predicted probability for the correct class was recorded.

The resulting sensitivity map visualizes the spatial importance landscape, highlighting time-frequency regions critical for distinguishing pathological from healthy voice textures.

All experiments were implemented in MATLAB R2023b using the Deep Learning Toolbox and executed on a windows GPU system.

2.5 Training and evaluation protocol

Stratified CV was applied in training with 5 folds and equal balance of classes in each fold. Split up involved 70 percentage of data in training and 30 percentage of validation.

The other hold-out set was used during final testing. The training was performed on NVIDIA RTX 2080 Ti graphic card, mini-batch size of 16 on 30 epochs or until early stopping requirements were achieved.

The model was evaluated based on a number of important measures including accuracy, precision, recall, and F1-score through the use of macro-averaging.

The specific accuracy of signal to noise ratio was tested with the clean audio and at the various levels of noise, including moderate noisy conditions, and at the higher noise levels, tested at 20 dB and 10 dB respectively.

3. RESULTS

Figure 3 shows Mel spectrograms of exemplary voice samples of the PVQD [6] normalized to a standard range of intensity (0–1) to allow direct visual comparison.

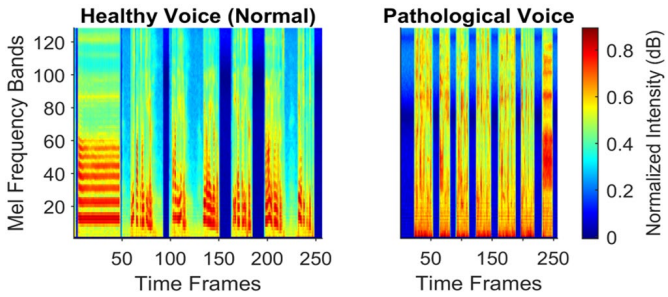

Figure 3. Mel spectrograms of the perceptual voice qualities database (PVQD) dataset of a healthy (left) voice sample that shows regular harmonic striations with distinctive formant structure, and a pathological (right) sample that displays harmonic disruptions

On the left panel, a healthy (normal) voice sample (File Name: NYU1024ENSS.wav, Severity Score of 3), and on the right panel, a pathological voice sample (File Name: PT008ENSS.wav, Severity Score: 98.67) are presented.

The time frames (around 250 frames, or around 2.5 seconds at a hop length of 10 ms), and the Mel frequency bands (0–127, ranging at 0–8 kHz, with a 16 kHz sampling frequency) forms the x and y axes, respectively.

Associated colourscale represents the normalized log-intensity (dB) values, where low, moderate and high energies are coded with blue, yellow and red, respectively.

The healthy voice spectrogram (left panel) has a clear and very structured harmonic pattern. Powerful, sharply cut horizontal striations indicating the base frequency and its harmonies are always apparent on most Mel bands, especially in the lower-to-mid frequency range (bands 20–80).

The clear formant bands are represented by the stable vertical bands of energy with little broadband noise and spectral smearing. The general texture is smooth and periodic, which represents an effective glottal closure, periodic vocal fold vibration, and limited turbulent airflow. These acoustic properties are compatible with perceptually normal voice quality, which consists of stable pitch, appropriate loudness, and the lack of roughness, breathiness, and strain, which is in line with low/absent CAPE-V and GRBAS scores in healthy PVQD samples.

The pathological voice spectrogram (right panel) on the contrary is very irregular and disrupted. In many areas, also in

the mid-to-high Mel bands (40–100) harmonic striations are fragmented, blurred, or completely absent, suggesting aperiodic vibration of the vocal folds and loss of the glottal periodicity. The ubiquitous yellow and red speckles, as well as the energy distribution throughout the spectrum, indicate increased background noise.

The trend is characteristic of turbulent airflow and the vocal folds are tend to be parting but not fully closing when making a sound. There is a lack of formant structure, and the energy is distributed and not concentrated, leading to a chaotic and noisy texture. Another feature of the spectrogram is a noticeable presence of low-frequency noise floor and unevenly shaped vertical striations, which represent perceptual characteristics of roughness (irregular vibration of the vocal folds), breathiness (excessive air leak), and strain (hyperfunction of supraglottis), which are characteristic of the dysphonia with highly rated scores on CAPE-V and GRBAS scales [4-6].

These strong visual contrasts indicate that Mel spectrograms are effective in terms of capturing the morphological and textural signature of voice pathological production through increase in spectral noise, harmonic desynchronization, formant variation and general acoustic chaos and thus are an effective two-dimensional representation of automated detection and quantitative morphological analysis in voice pathology research.

The obvious contrasting nature between the normal phonation with the regular and periodic texture and the pathological phonation with the disrupted and noisy texture emphasizes the usefulness of time-frequency imaging in providing objective, image-based screening of vocal disorders.

The CNN trained on mel spectrograms with PVQD as input data displayed high performance in separating healthy and pathological voices with a 5-fold CV accuracy of 98.04% +-1.93% (mean -standard deviation).

The accuracy of each fold was between 92.83 to 99.32, and in all folds, the model performed well in distinguishing between healthy and pathological samples without any bias in recognition.

To illustrate, Fold 5 had a total accuracy of 99.32, healthy voice had an accuracy of 98.72, and pathological voice had an accuracy of 99.54. The test set accuracy was 97.97 and the precision, recall and F1-score were 0.98, 0.99, and 0.99 respectively.

The test accuracy was found to be 95.73 and 98.77 in healthy and pathological voices, respectively, which indicates that the model is quite effective in identifying the disorder and healthy voices correctly. The summary of these results is presented in Table 3.

Table 3. Model performance metrics

Metric	Value
Cross-Validation (CV) Accuracy (mean ± SD)	98.04% ± 1.93%
Test Accuracy	97.97%
Precision	0.98
Recall	0.99
F1-Score	0.99
Healthy Test Accuracy	95.73%
Pathological Test Accuracy	98.7%

The sensitivity of the model to noise was tested on varying SNR conditions (Table 4). The overall accuracy was also high and was equal to 97.97% in clean audio, 96.73% at 20 dB SNR and 99.13% at 10 dB SNR, which indicated that the augmentation strategy was effective to simulate the noise variations in the real world. Per class SNR performance displayed a slight improvement in pathological detection at lower SNR levels (99.40% at 10 dB) than in healthy samples (98.41% at 10 dB) indicating that the model is able to give more importance to disorder related features such as roughness and breathiness even under noisy conditions.

Table 4. Signal-to-noise ratios (SNR) performance with 95% confidence intervals (mean ±95% CI across 5 folds)

SNR Level	Overall Accuracy (95% CI)	Healthy Accuracy (95% CI)	Pathological Accuracy (95% CI)
Clean	97.97% (96.18–98.93)	95.73% (90.38–98.16)	98.77% (96.89–99.52)
20 dB	96.73% (94.77–98.11)	92.59% (86.03–95.90)	98.12% (96.04–99.15)
10 dB	99.13% (97.70–99.65)	98.41% (93.98–99.53)	99.40% (97.79–99.83)

The occlusion sensitivity map is represented in Figure 4, which reveals the areas of the Mel spectrogram where the occlusion produced the most significant effect on the classification accuracy.

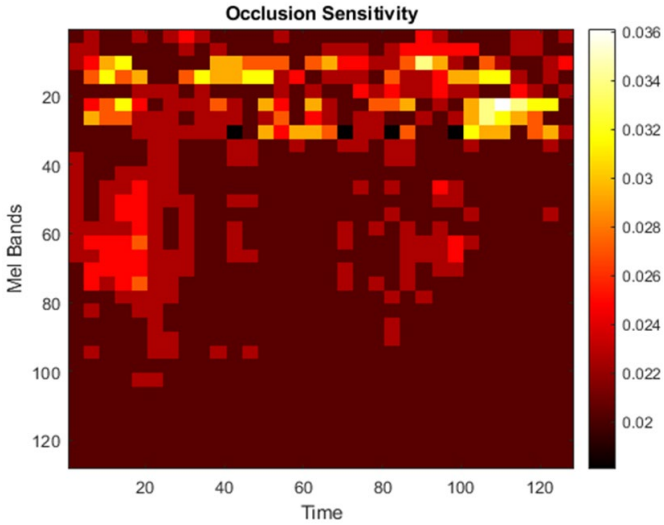

Figure 2. Occlusion Sensitivity Map. Color scale is defined in terms of importance score (0.02–0.036). Darker and lighter colors indicate lower to higher importance, respectively. The discrimination power lies in the lower Mel Bands (0–40 dB)

Lower Mel bands (0–40) across different time segments (0-120) showed a sensitivity to higher importance scores, which probably reflects fundamental frequency and harmonic characteristics of voice disorders. This indicates that the model utilizes low-frequency cues in making the distinction, which is in line with such perceptual attributes as breathiness and strain.

Five example Mel spectrograms in the test set, where the CNN model made incorrect predictions are presented in Figure 5, with the true label and the predicted label shown above the panel (True: 0 = healthy, 1 = pathological; Pred = model prediction). All the spectrograms are adjusted to the same intensity range (0–1) to make a direct comparison.

The x-axis represents time frames (approximately 120 frames, or around 1.2 seconds at 10 ms hop length), and the y-axis represents Mel frequency bands (0–127, ranging 0–8 kHz at 16 kHz sampling rate). Colour encodes normalized log-

intensity (dB), with blue indicating low energy and red/yellow indicating high energy. Interpretation of these results are provided in the Discussion section.

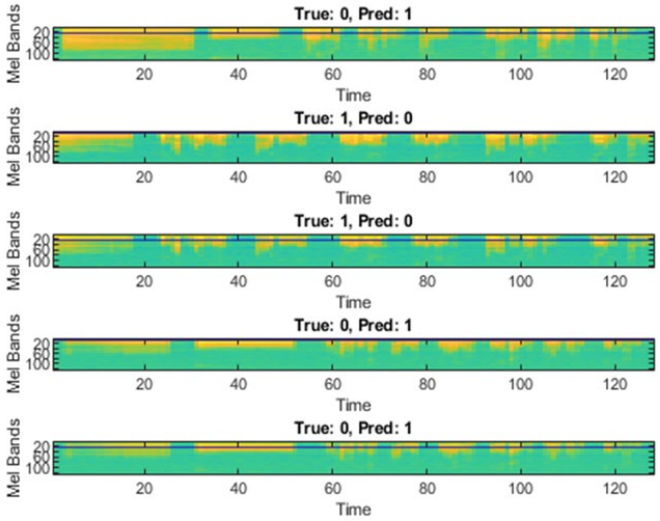

Figure 5. Misclassified spectrograms (True: 0 = healthy, True: 1 = pathological; Pred = predicted label by CNN) with normalized intensity color scale in dB

4. DISCUSSION

The trained CNN model on Mel spectrograms using PVQD dataset had a 5-fold CV accuracy of 98.04% ± 1.93% with fold-specific accuracies of 92.83% to 99.32%. The model performed with an accuracy of 97.97 on the test set with a precision, recall, and F1-score of 0.98, 0.99, and 0.99, respectively.

The findings indicate that the model is very effective in the differentiation between healthy and pathological voices and is even more effective than numerous benchmarks in the detection of voice pathology.

As an example, a CNN-based model that works with Mel spectrograms on the SVD dataset had an F1-score of 0.97 on pathological voices, whereas our model had a high F1-score of 0.99, probably because the PVQD focuses on perceptual characteristics and our training approach is also optimized.

On the same note, hybrid models between CNN and LSTM on comparable datasets have yielded accuracies of 92.71, and this indicates the usefulness of our extraction and augmentation methods.

The noise-resistance of the model is also worth mentioning, as its performance does not decrease significantly between SNR: 97.97% in clean audio, 96.73% and 99.13% at 20 and 10 dB, respectively. This indicates that the SNR percentage addition of 10 and 20 dB in training was an efficient manner of simulating the variability in the real-world environment.

This is consistent with research on SNR robustness in voice classification, where the models tend to deteriorate at levels well under 20 dB.

The per-class SNR analysis showed even-handed results, with pathological performance of 99.40% at 10 dB, which indicates that the model can maintain important perceptual features (e.g., breathiness and strain) even with noise, which is in line with the clinical focus of PVQD which focuses on the robust perceptual rating.

As indicated in the occlusion sensitivity map (Figure 4), the model is dependent on low-frequency Mel bands (0–40) which are fundamental frequency and harmonic structures that are frequently perturbed in pathological voices as mentioned in the PVQD overview. Pathological voices are characterized primarily by degradation, noise or in-periodicity in this low frequency region due to disruptions like incomplete glottal closure, irregular glottal vibration or excessive turbulence in the glottal airflow. This is why these bands were always found to be the most sensitive bands for classification in the occlusion sensitivity map.

Such interpretability is more clinically valuable to the model as it correlates with perceptual scales such as CAPE-V (ICC 0.918 severity) and GRBAS (ICC 0.911 grade).

The falsely classified spectrograms (Figure 5) mostly consist of boundary cases with slight harmonic differences implying that ambiguous perceptual attributes, including mild strain or asthenia (GRBAS ICC -0.85) are handled weakly.

The exemplary misclassifications, as shown in Figure 5 and Table 5, are mostly of a boundary type, in which acoustic characteristics show slight overlap between healthy- and pathological-patterns.

The low-level noise or irregular texture that is consistent in false positives and the preservation of periodicity that is consistent in false negatives gives a clue about the limitations of the model in decision-making and where the model can be improved to ensure a higher degree of robustness and sensitivity to clinical screening results.

Although there are these strengths, there are a number of limitations that are worth discussing. Although the PVQD dataset is reliable (ICC 0.8925), it only has 296 samples, which may be insufficient to generalize it outside of controlled records.

There was a slight loss of data in augmentation errors (e.g., in PT044ENSS.wav and PT099ENSS.wav because of resampling problems), but the effective size (i.e., of the spectrograms, about 1182) was adequate.

Table 5. Interpretation of exemplary misclassification cases

Cases	Observation	Interpretation
True: 0, Pred: 1 (False Positive	These spectrograms show relatively regular harmonic striations in lower Mel bands (20–60) but display mild broadband noise and slight spectral blurring in mid-to-high bands (60–100), producing a texture that the model interpreted as pathological despite the true healthy label.	This suggests the model is sensitive to low-level noise or minor perturbations that may reflect early or subclinical dysphonia not captured by the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) threshold.
True: 1, Pred: 0 (false negative)	These spectrograms exhibit some preserved harmonic structure and formant definition, particularly in lower bands, but contain intermittent aperiodic energy bursts and irregular vertical striations in higher bands, characteristic of pathological voice.	The model's incorrect healthy prediction likely results from over-reliance on preserved low-frequency periodicity, failing to adequately weight the disruptive high-frequency noise and harmonic instability that are perceptually salient in moderate-to-severe dysphonia.

Adjusted weights reduced class imbalance (26.4% healthy) but the smaller healthy recalls in certain folds (e.g., Fold 5: 85.48%) suggest that it may be biased to predict pathological values, as is a frequent problem in voice pathology datasets. Also, the GPU-accelerated training (20–22 seconds/fold) can cause slight numerical variation between CPU-based benchmarks.

These results have significant implications to automated voice disorder screening, which presents a near-clinically reliable tool that can minimize the rater variability (as seen in the multi-rater system of PVQD) and allow telehealth use, particularly in noisy settings.

Future efforts must confirm on larger datasets such as SVD or combine features of GRBAS to learn many tasks, investigate lower SNR rates (e.g., 5 dB), and optimize augmentation to remove errors, further in line with the targets of PVQD of advancing perceptual evaluation with technology.

5. CONCLUSIONS

This paper shows that Mel spectrograms, capturing voice features as a two-dimensional topographic image and classifying them using a CNN with hierarchical spatial filtering approach, form a strong and understandable paradigm of automated pathological voice pattern detection.

The proposed model using high-reliability PVQD [6] results in a 5-fold CV accuracy of 98.04% ± 1.93% and test-set accuracy of 97.97%, and the precision, recall, and F1-score values were 0.98, 0.99, and 0.99, respectively.

The architecture was able to retain strong performance in the simulated noise conditions of real world, achieving 96.73 and 99.13 percent accuracies at 20 and 10 dB SNR, respectively, and this makes one see how the augmentation strategy applied to the architecture worked well to retain the diagnostically relevant spectral features.

Occlusion maps show the model relies mainly on lower Mel bands, associated with pathological features. These regions are directly connected to such features as stable pitch, clear harmonics, and distinct formants that tend to be disrupted in dysphonic voices, particularly rough, breathy, or strained ones [4-6].

These differences were further reflected in the representative spectrograms as near-normal samples showed clean harmonic striations which are periodic and well-defined formant banding, whilst severe pathological samples showed fragmented harmonics, extensive spectral smearing, high noise floors, and chaotic distribution of energy.

The results contribute to objective voice pathology diagnosis, between traditional perceptual assessment and quantified image-based pattern recognition with better performance than previous benchmarks of similar data [23, 24].

Owing to the ease of interpretation, being highly sensitive to pathological features, and working well even in noisy environments, the tool has a high potential of being applied to telehealth applications and clinical decision support as a scalable tool [25].

However, the model should be tested on larger and multi-center data, multi-task-regression of continuous CAPE-V/GRBAS scores should be investigated, and its robustness should be tested to lower SNR and different forms of noise, prior to its application in real-life clinical settings.

ACKNOWLEDGMENT

This work is supported by the SSN Trust (Grant No.: SSN/IFFP/January2019/1-12/08).

REFERENCES

[1] Roy, N., Merrill, R.M., Gray, S.D., Smith, E.M. (2005). Voice disorders in the general population: prevalence, risk factors, and occupational impact. The Laryngoscope, 115(11): 1988-1995. https://doi.org/10.1097/01.mlg.0000179174.32345

[2] Wang, T.V., Song, P.C. (2022). Neurological voice disorders: A review. International Journal of Head and Neck Surgery, 13(1): 32-40. https://doi.org/10.5005/jp-journals-10001-1521

[3] Echternach, M., Döllinger, M., Köberlein, M., Kuranova, L., Gellrich, D., Kainz, M.A. (2020). Vocal fold oscillation pattern changes related to loudness in patients with vocal fold mass lesions. Journal of Otolaryngology-Head & Neck Surgery, 49(1): 80. https://doi.org/10.1186/s40463-020-00481-y

[4] Kempster, G.B., Gerratt, B.R., Abbott, K.V., Barkmeier-Kraemer, J., Hillman, R.E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. American Journal of Speech-Language Pathology, 18(2): 124-132. https://doi.org/10.1044/1058-0360(2008/08-0017)

[5] Hirano, M., McCormick Karen, R. (1981). Clinical Examination of Voice, New York: Springer, 1981.

[6] Walden, P.R. (2022). Perceptual voice qualities database (PVQD): Database characteristics. Journal of Voice, 36(6): 875-e15. https://doi.org/10.1016/j.jvoice.2020.10.001

[7] Arslan, Ö. (2024). A machine learning approach for voice pathology detection using mode decomposition-based acoustic cepstral features. Mathematical Modelling and Numerical Simulation with Applications, 4(4): 469-494. https://doi.org/10.53391/mmnsa.1473574

[8] Ali, Z., Alsulaiman, M., Elamvazuthi, I., Muhammad, G., Mesallam, T.A., Farahat, M., Malki, K.H. (2016). Voice pathology detection based on the modified voice contour and SVM. Biologically Inspired Cognitive Architectures, 15: 10-18. https://doi.org/10.1016/j.bica.2015.10.004

[9] Ilgaz, H., Akkoyun, B., Alpay, Ö., Akcayol, M.A. (2024). CNN based automatic speech recognition: a comparative study. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 13: e29191-e29191. https://doi.org/10.14201/adcaij.29191

[10] Davis, S., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4): 357-366. https://doi.org/10.1109/TASSP.1980.1163420

[11] Stevens, S.S., Volkmann, J., Newman, E.B. (1937). A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America, 8(3): 185-190. https://doi.org/10.1121/1.1915893

[12] Bashir, R.N., Shahid, M.A., Rashid, T., Faheem, M., Saidani, T., Saidani, O., Khan, A.R. (2025). Voice

pathology identification using mel spectrogram features and deep learning. Signal, Image and Video Processing, 19(11): 909. https://doi.org/10.1007/s11760-025-04527-4

[13] Deliyski, D.D., Evans, M.K., Shaw, H.S. (2005). Influence of data acquisition environment on accuracy of acoustic voice quality measurements. Journal of Voice, 19(2): 176-186. https://doi.org/10.1016/j.jvoice.2004.07.012

[14] Deliyski, D.D., Shaw, H.S., Evans, M.K. (2005). Adverse effects of environmental noise on acoustic voice quality measurements. Journal of Voice, 19(1): 15-28. https://doi.org/10.1016/j.jvoice.2004.07.003

[15] Vogel, A.P., Morgan, A.T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. International Journal of Speech-Language Pathology, 11(6): 431-437. https://doi.org/10.3109/17549500902822189

[16] Maryn, Y., Ysenbaert, F., Zarowski, A., Vanspauwen, R. (2017). Mobile communication devices, ambient noise, and acoustic voice measures. Journal of Voice, 31(2): 248-e11. https://doi.org/10.1016/j.jvoice.2016.07.023

[17] Petrizzo, D., Popolo, P.S. (2021). Smartphone use in clinical voice recording and acoustic analysis: A literature review. Journal of Voice, 35(3): 499-e23. https://doi.org/10.1016/j.jvoice.2019.10.006

[18] Van der Woerd, B., Wu, M., Parsa, V., Doyle, P.C., Fung, K. (2020). Evaluation of acoustic analyses of voice in nonoptimized conditions. Journal of Speech, Language, and Hearing Research, 63(12): 3991-3999. https://doi.org/10.1044/2020_JSLHR-20-00212

[19] Javanmardi, F., Kadiri, S.R., Alku, P. (2024). A comparison of data augmentation methods in voice pathology detection. Computer Speech & Language, 83: 101552. https://doi.org/10.1016/j.csl.2023.101552

[20] Kathania, H.K., Kadiri, S.R., Alku, P., Kurimo, M. (2021). Using data augmentation and time-scale modification to improve ASR of children's speech in noisy environments. Applied Sciences, 11(18): 8420. https://doi.org/10.3390/app11188420

[21] Hanif, A.M., Beqiri, S., Keane, P.A., Campbell, J.P. (2021). Applications of interpretability in deep learning models for ophthalmology. Current Opinion in Ophthalmology, 32(5): 452-458. https://doi.org/10.1097/ICU.0000000000000780

[22] Zeiler, M.D., Fergus, R. (2014). Visualizing and understanding convolutional networks. In European Conference on Computer Vision, Zurich, Switzerland, pp. 818-833. https://doi.org/10.1007/978-3-319-10590-1

[23] Islam, R., Tarique, M., Abdel-Raheem, E. (2020). A survey on signal processing based pathological voice detection techniques. IEEE Access, 8: 66749-66776. https://doi.org/10.1109/ACCESS.2020.2985280

[24] Vikram, C.M., Umarani, K. (2013). Text independent classification of normal and pathological voices using MFCCs and GMM-UBM. In 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, India, pp. 1215-1220. https://doi.org/10.1109/CICT.2013.6558286

[25] Oliveira, G., Fava, G., Baglione, M., Pimpinella, M. (2017). Mobile digital recording: adequacy of the iRig and iOS device for acoustic and perceptual analysis of normal voice. Journal of Voice, 31(2): 236-242. https://doi.org/10.1016/j.jvoice.2016.05.023