sciendo

ISSN 1335-8871

MEASUREMENT SCIENCE REVIEW

Journal homepage: https://content.sciendo.com

Detection of Laryngeal Pathologies from Voice using EMD-based Mel-Spectrograms and Scalograms with AlexNet

Sofiane Cherif*, Abdelhafid Kaddour, Abdelmoudjib Benkada, Said Karoui

Signals, Systems, and Data Laboratory (LSSD), Electronics Department, Faculty of Electrical Engineering, University of Sciences and Technology of Oran Mohamed Boudiaf (USTO MB), El Mnaouar, Oran, Algeria, sofiane.cherif@univ-usto.dz

Abstract: In this paper, a novel method for detecting of laryngeal pathologies using deep neural networks and time–frequency signal processing techniques is presented. The proposed approach combines empirical mode decomposition (EMD) and wavelet analysis to extract discriminative features from healthy and pathological voice recordings obtained from the Saarbrücken Voice Database (SVD). Each voice signal is pre-processed and decomposed into intrinsic mode functions (IMFs), from which the most relevant IMF is selected based on a temporal energy criterion. Two sets of features are derived from the selected IMF: Mel-frequency cepstral coefficients (MFCCs) and continuous wavelet transform (CWT) coefficients. These features are converted into Mel-spectrogram and scalogram images, respectively, which serve as inputs to the AlexNet convolutional neural network (AlexNet-CNN) for automatic binary classification. To the best of our knowledge, this is the first study to incorporate scalogram representations with AlexNet-CNN in the context of pathological voice detection. The results show that the proposed method achieves a classification accuracy of 85.66 % when using Mel-spectrograms and 86.4 % when using scalograms, demonstrating its potential for effective and interpretable voice pathology screening.

Keywords: laryngeal pathology detection, voice signal processing, empirical mode decomposition, Mel-spectrogram, scalogram, AlexNet convolutional neural network

1. Introduction

Laryngeal pathologies are disorders that affect the larynx, which houses the vocal folds, leading to various voice problems [1], [2], [3]. Early detection of these pathologies is crucial to prevent permanent damage to the vocal folds and to significantly improve the effectiveness of treatment. The diagnosis of voice disorders usually requires invasive clinical examinations such as laryngoscopy and videostroboscopy. However, vocal signal analysis using the signal processing techniques can be used to extract features that help distinguish between healthy and pathological voices. Therefore, there is a growing need to develop a non-invasive, automated approach based on deep learning to identify pathological voices. A substantial body of related work exists in this domain [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Many studies focus on the extraction of signal processing features such as MFCCs and Wavelet Packet Transform (WPT) [15], [16], [17], [18], [19], [20], as well as the use of deep learning for voice pathology detection [21], [22], [23]. This paper focuses on the detection of laryngeal pathologies using MFCC spectrograms and scalograms — time–frequency representations — derived from the most relevant IMFs. These representations are used as inputs to the AlexNet convolutional neural network (AlexNet-CNN) for automatic classification of normal and pathological voices. This study investigates the use of scalogram representations with AlexNet-CNN for pathological voice detection, an approach that has received little to no attention in the existing literature. The adopted processing workflow is illustrated in the synoptic diagram (Fig. 1). Once the relevant IMFs are extracted from each vocal signal, the signal is segmented, and the MFCCs are calculated for each segment. In parallel, scalograms are generated using continuous wavelet transform (CWT). These MFCC images and scalograms are then used as input to the AlexNet-CNN convolutional neural network for classification. This paper is organized as follows: Section 2 presents the materials and related methods, detailing the methodology and the detection process. Section 3 reports and discusses the results. Finally, the paper concludes with a summary of the results and a comparison of recent studies on pathological voice classification.

2. Materials and Methods

A. Saarbrücken Voice Database (SVD)

The vocal signals used in this study were obtained from the publicly accessible Saarbrücken Voice Database (SVD) [24]. The SVD contains a diverse collection of voice recordings from subjects with various laryngeal pathologies, including both functional and organic disorders. The database contains multiple recordings per speaker, featuring the sustained pronunciation of the vowels /a/, /i/, and /u/ with different intonations: normal, low, high, and low–high–low. This diversity

DOI: 10.2478/msr-2025-0030

*Corresponding author: sofiane.cherif@univ-usto.dz (S. Cherif)

contributes to improved model performance when utilized. For this study, only the sustained /a/ vowels pronounced in normal pitch were selected. This choice was motivated by the fact that the sustained vowel /a/ is a common phonation task found in many voice disorder datasets, and it provides a consistent basis for analysis. All voice recordings in the SVD are sampled at 50 KHz and 16-bit resolution. The subset used in this work consists of 259 healthy voice samples and 50 pathological males samples diagnosed with laryngitis, all corresponding to the neutral vowel /a/. To increase the amount of training data and better capture temporal variations, the most relevant IMFs from the healthy and pathological voice samples were segmented into overlapping frames, thus increasing the input to AlexNet-CNN.

B. *Use of AlexNet-CNN with EMD-based scalograms for pathological voice classification*

In this study, we used AlexNet-CNN, a pre-trained convolutional neural network (CNN), to detect laryngeal pathologies from voice signals. AlexNet-CNN consists of eight layers — five convolutional layers followed by three fully connected layers — and uses the ReLU activation function to improve non-linearity and accelerate training [25]. Scalogram images obtained from the most relevant IMF followed by the CWT were used as input to the network. These time–frequency representations capture rich, multiscale features that are highly relevant for vocal disorder characterization. The model, originally trained on ImageNet, was either fine-tuned for direct classification or used as a deep feature extractor, with the outputs of the penultimate fully connected layer fed into an external classifier, such as a support vector machine (SVM) or Softmax layer. The dataset was split into a training and a validation subsets, with 80 % of the images used for training and the remaining 20 % for validation. The choice of AlexNet-CNN was motivated by its computational efficiency, fast convergence, and demonstrated effectiveness in biomedical imaging tasks, particularly in scenarios with limited datasets. The integration of AlexNet-CNN complements traditional acoustic features such as MFCCs and results in a hybrid, multimodal feature space that improves the robustness and accuracy of pathological voice classification.

C. *Voice signal pre-processing and feature extraction pipeline*

The overall process for detecting laryngeal pathologies from vocal signals is summarized in the synoptic diagram (Fig. 1). It comprises three main phases: signal preprocessing, feature extraction, and classification. The first phase begins with the formation of a matrix containing voice signals from healthy and pathological male subjects (suffering from laryngitis), limited to the sustained neutral vowel /a/. To simulate real-world acoustic conditions, Gaussian noise with a signal-to-noise ratio of SNR = 0 dB and a standard deviation $\sigma = 1$ is added to each signal. Denoising is then applied using wavelet transform-based methods. Next, all signals are normalized and centered to create a zero-mean matrix. To ensure uniformity, the signals are equalized to the same length. Silence segments are removed to focus on the voiced regions, followed by low-pass filtering with a cut-off frequency of 3400 Hz to retain only the relevant spectral content. A Hamming window corresponding to the length of each signal segment is applied to reduce spectral leakage during subsequent analysis.

In the second phase, each pre-processed signal undergoes empirical mode decomposition (EMD) to extract its IMFs. Among these, the IMF with the highest energy is selected as the most relevant component for further analysis. This IMF is then segmented using a sliding window of 23 ms with a 50 % overlap (i.e., half the window length). Feature extraction is performed for each segment to generate two types of representations: Mel-spectrograms and scalograms, which serve as time–frequency descriptors that capture both the spectral and temporal dynamics of the voice signal. Finally, the classification phase is performed using the AlexNet-CNN. Two separate classification paths are considered: one using MFCC images and the other using scalogram images as input. In both cases, the network outputs a binary decision indicating whether the voice is healthy or pathological.

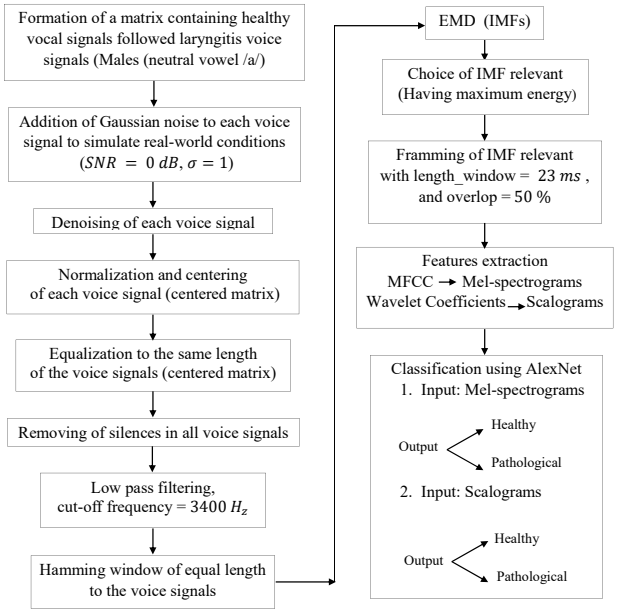

Fig. 1. Synoptic representation of the proposed method: voice signals first undergo signal processing, followed by EMD to extract IMFs. The most energetic IMF is selected to compute Mel-spectrograms and scalograms, which are then fed into a pretrained AlexNet-CNN model for voice pathology classification.

D. *Wavelet-based denoising method*

We applied a denoising technique based on the wavelet transform, which involves decomposing the vocal signal into wavelet coefficients at multiple frequency scales using the Daubechies wavelet of order 4($db4$). Stein's unbiased risk estimate (SURE) method was used to determine the optimal thershold for noise suppression. A hard thresholding approach was applied: coefficients with magnitudes below the estimated threshold were completely discarded (set to zero),

while those above the threshold were kept unchanged. This technique effectively removes the noise while preserving the significant components of the vocal signal. The denoised signal was then reconstructed using the inverse wavelet transform applied to the modified coefficients. To evaluate the effectiveness of the wavelet denoising approach, clean vocal signals were artificially contaminated with Gaussian noise (standard deviation $\sigma = 1$, signal-to-noise ratio SNR = 0 dB) . The results showed that wavelet-based denoising significantly improved signal clarity and preserved diagnostically relevant acoustic features, even under severe noise conditions. To analyze the time–frequency characteristics of the relevant IMF, the CWT was also applied. The CWT of a signal x(t) is calculated by integrating the signal with a family of scaled and shifted wavelets, and is mathematically defined as:

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t-b}{a} \right) dt \qquad (1)$$

where ψ^* denotes the complex conjugate, the wavelet $\psi^* \left(\frac{t-b}{a}\right)$ is obtained by scaling (dilating) and shifting (translating) the mother wavelet $\psi(t)$. Here, t denotes time, $a > 0$ is the scale parameter that controls the frequency resolution, b is the translation parameter that represents the time shift, and $\psi^*(t)$ is the complex conjugate of the mother wavelet $\psi(t)$. This representation captures both the spectral and temporal variations of the signal and is therefore highly suitable for analyzing and classifying pathological voice signals.

E. Normalization and equalization of voice signal lengths

To ensure the consistency of all samples and to enable uniform processing, each voice signal was subjected to normalization and length equalization. Normalization is a critical pre-processing step in which the data are transformed to a standard scale. In this study, we used a combination of two normalization techniques:

- Z-score normalization, defined as $x_{\mathrm{in}} = \frac{x_i - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the signal. This method centers the signal around zero with a unit variance, effectively removing the DC offset and scaling the amplitude distribution.
- Peak amplitude normalization, defined as $X_{\mathrm{in}} = \frac{x_{\mathrm{in}}}{\max(|x_{\mathrm{in}}|)}$, where each sample is divided by the maximum absolute amplitude value to ensure that all signals are within the range [–1, 1].

here, X_{in} denotes the final normalized signal, N is the signal length, and the standard deviation is calculated as:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

After normalization, all signals were adjusted to a uniform length by truncating longer sequences or applying zero-padding to shorter ones. This equalization ensures compatibility with the subsequent stages, particularly during feature extraction and classification using convolutional neural networks (CNNs), which require a fixed-size of the input dimensions.

F. Silence removal using energy thresholding

To improve the signal quality and reduce computational complexity, silent regions were removed from the voice recordings prior to feature extraction. This pre-processing step is particularly important in pathological voice analysis, as non-phonated segments do not contain diagnostically relevant features related to vocal fold behavior. In this study, silence detection was performed using short-term energy analysis. A frame was considered silent when its energy fell below 2 % of the maximum signal energy. This threshold effectively identified low-activity regions while preserving the meaningful voiced segments. Detected silent frames were discarded, which improved the signal-to-noise ratio and ensured that only diagnostically useful components were retained for reliable feature extraction and classification.

G. Low-pass filtering of voice signals

To eliminate high-frequency noise components that are not relevant for speech production, a low-pass filter was applied to all vocal signals. This filtering stage is crucial for preserving the frequency band that is most informative for voice analysis, particularly for pathology detection. We used a low-pass filter with a cut-off frequency of 3400 Hz. This value is typically used in speech processing applications, as the majority of voice energy is below this threshold. Frequencies above 3400 Hz typically contain ambient noise or artifacts irrelevant to the phonatory process. The filtering process contributes to improving the signal-to-noise ratio and increases the reliability of subsequent feature extraction steps, including Mel-spectrograms and scalograms.

H. Windowing of frames using the Hamming function

A Hamming window was applied to each pre-processed voice signal to reduce spectral leakage by truncating the signal at its edges (Fig. 2).

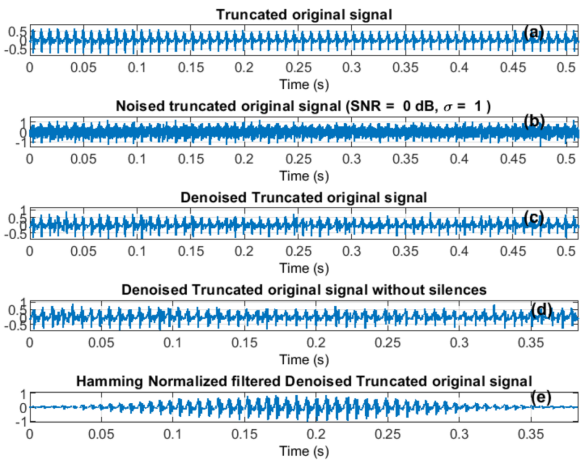

Fig. 2. Voice signals after pre-processing, including noise filtering, amplitude normalization, and segmentation into fixed-length frames. These enhanced signals serve as input for subsequent EMD-based decomposition and feature extraction (e.g., MFCCs and scalograms) to distinguish pathological from healthy voice patterns.

I. EMD algorithm

EMD, introduced by Huang et al. [26], is an adaptive and fully data-driven method designed for analyzing non-linear and non-stationary signals.The EMD method can self-adaptively decompose a complicated multicomponent signal into a finite set of components known as IMFs without any a priori assumptions (Fig. 3). Due to its adaptive nature, EMD is particularly well-suited for the analysis of biomedical and voice signals. It enables for the isolation of the most informative oscillatory modes and thus the extraction of relevant features — such as MFCCs or scalograms — from the most energetically significant IMF. We used a Huang's Empirical Mode Decomposition for the signal analysis. The EMD algorithm is explained next [26], [27], [28]. Let $x(t)$ be a real-valued, non-linear and non-stationary signal. EMD expresses $x(t)$ as the sum of N IMFs and a final residual component $r(t)$:

$$x(t) = \sum_{i=1}^{N} (\mathrm{IMF}_i(t) + r(t)) \tag{2}$$

where $\mathrm{IMF}_i(t)$ is the i-th IMF, $r(t)$ is the final residual and and N is the total number of extracted IMFs. We have calculated the temporal energy for each of the IMFs obtained from the EMD analysis of the voice signal, and only the IMF with the highest energy value is chosen as relevant one, according to the following equation:

$$E = \sum_{i=1}^{K} (\mathrm{IMF}_i(n))^2 \tag{3}$$

where E, K and $\mathrm{IMF}_i(n)$ are the temporal energy, the length of the IMF and the i-th IMF digitized signal, respectively.

Algorithm 1 EMD algorithm (Huang et al.)

Require: Signal $x(t)$
Ensure: A set of intrinsic mode functions (IMFs) $\{IMF_1(t), IMF_2(t), \ldots, IMF_N(t)\}$ and a residual $r(t)$
$r(t) \leftarrow x(t)$
$i \leftarrow 1$
while $r(t)$ has more than two extrema **do**
  $h(t) \leftarrow r(t)$
  repeat
    Identify all local maxima and minima of $h(t)$
    Interpolate maxima to obtain upper envelope $e_{\mathrm{upper}}(t)$
    Interpolate minima to obtain lower envelope $e_{\mathrm{lower}}(t)$
    Compute mean envelope: $m(t) \leftarrow \frac{e_{\mathrm{upper}}(t)+e_{\mathrm{lower}}(t)}{2}$
    Update proto-IMF: $h(t) \leftarrow h(t) - m(t)$
  until $h(t)$ satisfies IMF conditions:
    – Number of extrema and zero crossings differ by at most one
    – Mean envelope is approximately zero
  $IMF_i(t) \leftarrow h(t)$
  $r(t) \leftarrow r(t) - IMF_i(t)$
  $i \leftarrow i + 1$
end while
return $\{IMF_1(t), IMF_2(t), \ldots, IMF_{i-1}(t)\}$ and residual $r(t)$

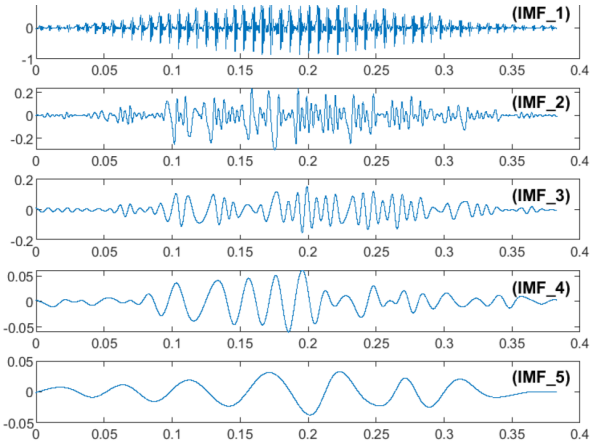

Fig. 3. The voice signal is decomposed into IMFs using EMD. Each IMF represents a distinct oscillatory mode, ordered from high to low frequency content, capturing features relevant to voice characteristics. The most energetic IMF (highlighted) is selected for further analysis, including MFCC-based and scalogram image generation, to support the classification between pathological and healthy voice signals.

J. Extracted features

Mel-spectrogram images

Once the relevant IMF was selected, it was segmented into overlapping frames for Mel-frequency cepstral coefficient (MFCC) extraction. Each IMF was divided into frames of 23 ms duration, with a 50 % overlap between consecutive frames to ensure smooth temporal continuity and to capture transitional acoustic features. For each windowed frame, MFCCs were calculated by transforming the power spectrum into the Mel scale using a bank of triangular filters spaced according to the Mel frequency warping function. This process resulted in a time–frequency representation of the relevant IMF in the perceptually motivated Mel scale. The MFCCs were then aggregated into Mel-spectrogram images, which were subsequently used as inputs for the classification stage (Fig. 4).

Scalogram representation of the relevant IMF using CWT

A scalogram is a two-dimensional time–frequency representation that visually shows how the spectral content of a signal evolves over time. It is particularly well suited for analyzing non-stationary signals such as the human voice, where spectral characteristics change dynamically during phonation. In addition to the MFCC extraction, each frame of the selected relevant IMF was processed using the Morlet CWT filter bank to generate scalograms. This technique enables the identification of localized spectral variations over time, which are crucial for the detection of subtle irregularities associated with vocal fold pathologies. For each frame, the CWT generates a matrix of wavelet coefficients representing the signal's energy distribution over time and frequency. These matrices were then visualized as scalogram images, resized to

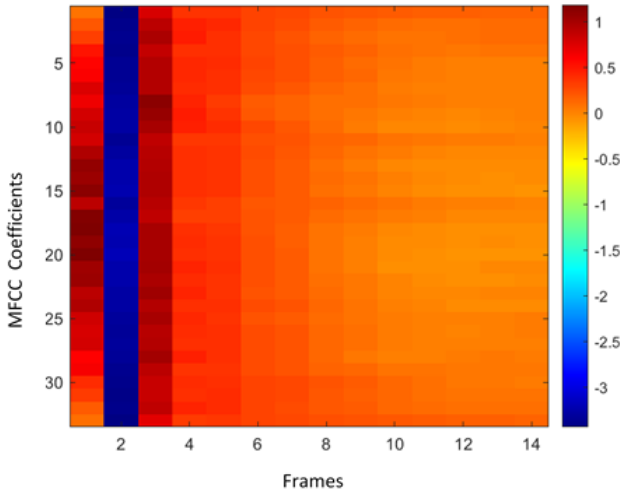

Fig. 4. Example of a Mel spectrogram generated from the most energetic IMF of a pre-processed voice signal. The representation emphasizes perceptually meaningful spectral patterns used to discriminate between healthy and pathological voices in classification tasks.

224×224 pixels, and converted to RGB format to comply with the input specifications of the AlexNet-CNN used in the classification stage (Fig. 5). This frame-level approach not only captures fine-grained, time-localized acoustic features relevant for pathology detection, but also significantly increases the number of training samples. As a result, the combination of scalogram-based representations and MFCCs enriches the feature space and improves the model's robustness in discriminating between healthy and pathological voice signals.

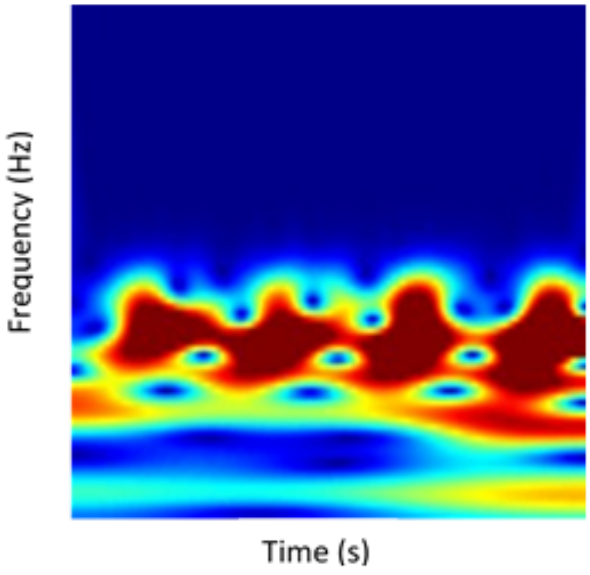

Fig. 5. Scalogram example derived from the most energetic IMF of a pre-processed voice signal using CWT. The representation highlights relevant time-frequency patterns for the subsequent classification of healthy and pathological voices.

3. Results

The effectiveness of the proposed approach was evaluated by training and validating the AlexNet-CNN on two distinct types of input representations: MFCC-based images and CWT-derived scalograms. Each representation was generated from the most relevant IMF extracted from voice signal frames, as described in the previous sections. Classification performance was evaluated using accuracy as the primary metric in the validation datasets. The classifier achieved an accuracy of 85.66 % when using MFCC images, and a slightly higher accuracy of 86.4 % when using scalogram images (Fig. 6). These results suggest that both representations are effective in capturing discriminative features relevant to pathological voice detection. However, the superior performance of scalograms highlights their ability to encode rich time–frequency information that complements, and in some cases, surpasses, conventional cepstral features. The improved performance of the scalograms can be attributed to their fine-grained temporal and spectral resolution, which allows the model to detect subtle irregularities associated with vocal fold dysfunctions. These results confirm the suitability of combining EMD with CWT-based scalograms for robust and accurate voice pathology classification.

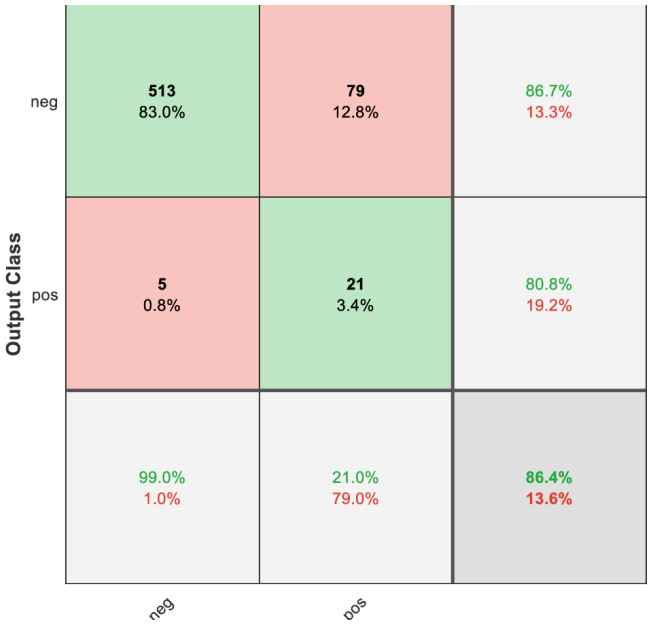

Fig. 6. Confusion matrix showing the performance of AlexNet-CNN on scalograms of the most energetic IMFs obtained with EMD and CWT.

4. Discussion and Comparison with Previous Studies

In this section, we compare the performance of our proposed method with several recent studies on pathological voice classification using different datasets, features, and classification models. Table 1 summarizes the key characteristics and classification accuracies reported in the literature.

Table 1. Comparison of our method with recent studies on pathological voice detection.

Study	Dataset	Features and model	Accuracy [%]
[29]	SVD	Multipeak, Gaussian mixture model (GMM)	91.83
[30]	SVD + HUPA	MFCCs, SVM	71.45–76.19
[31]	MEEI voice disorders	MFCC (500 ms frames, 5 ms shift), SVM	66.4–75.1
[32]	SVD + HUPA	wav2vec, SVM	68.55–83.11
[33]	SVD + HUPA	Mel-spectrogram, SVM	69.45–75
[34]	VOICED	wav2vec 2.0, SVM / KNN	98
[35]	UA-speech + TORGO	MFCCs, SVM	63.13–89.22
This work	SVD	EMD-IMF, Mel-spectrogram + scalogram, AlexNet-CNN	85.66 / 86.4

To evaluate the effectiveness of our proposed method, we compared its performance with several recent studies on pathological voice detection using different datasets, features, and machine learning models (see Table 1). Our approach, based on EMD-derived Mel-spectrograms and scalograms as inputs to the AlexNet-CNN, achieved an accuracy of 85.66 % and 86.4 %, respectively. In comparison, Eskidere et al. [29] achieved a slightly higher accuracy of 91.83 % using a GMM with multipeak features on the same SVD dataset. However, their method did not use deep learning or time–frequency representations, which could limit the generalization. Similarly, Kadiri et al. [30] used MFCCs and SVM on a combined dataset (SVD and HUPA), reporting accuracies ranging from 71.45 % to 76.19 %. More recent studies have explored deep representations, such as wav2vec features in combination with SVM classifiers and achieved accuracies randing from 68.55 % to 83.11 % [32]. Other approaches, including Mel-spectrogram features with SVM [33], and classical MFCC-based systems [31], [35], have generally shown lower or more variable performance, especially when applied to small or heterogeneous datasets. The highest reported accuracy in the literature (98 %) was obtained by Cai et al. [34] using the VOICED database and wav2vec 2.0 features in conbination with SVM and KNN classifiers. While promising, this result is based on a different dataset and may not be directly comparable due to variations in recording conditions, subject demographics, and pathology types.

Our method offers a competitive and robust alternative, particularly because it:

- operates effectively on a publicly available and widely used dataset (SVD),
- integrates the EMD to isolate the most informative IMF component,
- extracts both MFCC images and scalograms, and
- utilizies deep learning through AlexNet-CNN, a well-established architecture for small to medium-sized datasets.

Overall, these results suggest that our hybrid approach, which combines classical signal processing with deep learning, delivers performance that is not only competitive with state-of-the-art methods, but also interpretable and adaptable for non-invasive clinical screening of voice disorders.

5. Conclusion

This study proposed an effective framework for automatic detection of laryngeal pathologies by combining advanced signal processing and deep learning techniques. The voice recordings were pre-processed and decomposed using EMD, and the most relevant IMFs were selected based on temporal energy. From each frame of this IMF, MFCC images and scalograms based on CWT were extracted to capture both spectral and temporal information.

Using a deep CNN, namely AlexNet, we classified the extracted features and achieved promising results: 85.66 % accuracy with MFCC-based spectrograms and 86.4 % with scalograms. These results show the potential of the proposed method to extract and analyze diagnostically relevant information from voice signals for non-invasive and early-stage detection of laryngeal pathologies.

The proposed method provides a novel combination of signal decomposition and time–frequency feature representation that complements recent advances in deep learning-based voice pathology detection. Compared to existing approaches, our framework proves to be not only effective but also interpretable and well-suited for small-scale datasets. Therefore, the method is suitable for an initial screening of pathological voice conditions and can serve as a valuable diagnostic aid. However, further research and large-scale clinical validation are essential to improve its robustness and generalizability for real-world applications.

References

[1] Dhillon, V. K. (2022). *Vocal Cord Disorders*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/vocal-cord-disorders. (Accessed September 2025).

[2] Verdolini, K., Ramig, L. O. (2001). Review: Occupational risks for voice problems. *Logopedics, Phoniatrics, Vocology,* 26 (1), 37–46.

[3] Parsa, V., Jamieson, D. G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research,* 43 (2), 469–485. https://doi.org/10.1044/jslhr.4302.469.

[4] Wang, J., Xu, H., Peng, X., Liu, J., He, C. (2023). Pathological voice detection based on multi-domain features and deep hierarchical extreme learning machine. *The Journal of the Acoustical Society of America,* 153 (1), 423–435. https://doi.org/10.1121/10.0016869.

[5] AL-Dhief, F. T., Latiff, N. M. A. A., Malik, N. N. N. A., Sabri, N., Albadr, M. A. A., Abbas, A. F., Hussein, Y. M., Mohammed, M. A. (2020). Voice pathology detection using machine learning technique. In *2020 IEEE 5th International Symposium on Telecommunication Technologies (ISTT)*. IEEE, 99–104. https://doi.org/10.1109/ISTT50966.2020.9279346.

[6] Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T. A., Farahat, M., Malki, K. H., Bencherif, M. A. (2017)a. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice,* 31 (1), 113.e9–113.e18. https://doi.org/10.1016/j.jvoice.2016.03.019.

[7] Al-Nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z. (2017)b. Investigation of voice pathology detection and classification on different frequency regions using correlation functions. *Journal of Voice,* 31 (1), 3–15. https://doi.org/10.1016/j.jvoice.2016.01.014.

[8] Godino-Llorente, J. I., Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering,* 51 (2), 380–384. https://doi.org/10.1109/TBME.2003.820386.

[9] Mittal., V., Sharma, R. K. (2021). Deep learning approach for voice pathology detection and classification. *International Journal of Healthcare Information Systems and Informatics,* 16 (4), 1–30. https://doi.org/10.4018/IJHISI.20211001.oa28.

[10] Roohum, J., Jayagowri, R. (2020). Voice disorder detection and classification - a review. In *Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics and Cloud in Computational Vision and Bio-Engineering (ISMAC-CVB 2020)*. https://doi.org/10.2139/ssrn.3734762.

[11] Altayeb, M., Al-Ghraibah., A. (2022). Classification of three pathological voices based on specific features groups using support vector machine. *International Journal of Electrical and Computer Engineering (IJECE),* 12 (1), 946–956. https://doi.org/http://doi.org/10.11591/ijece.v12i1.pp946-956.

[12] Hammami, I. (2019). Classification of psychogenic and laryngeal voice diseases based on wavelet transform analysis and teager energy operator. *International Journal of Applied Mathematics, Electronics and Computers,* 7 (3), 49–55. https://doi.org/10.18100/ijamec.458230.

[13] Wu, H., Soraghan, J., Lowit, A., Di Caterina, G. (2018). Convolutional neural networks for pathological voice detection. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–4. https://doi.org/10.1109/EMBC.2018.8513222.

[14] Dibazar, A. A., Narayanan, S., Berger., T. W. (2002). Feature analysis for automatic detection of pathological speech. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Vol. 1. 182–183. https://doi.org/10.1109/IEMBS.2002.1134447.

[15] Arjmandi, M. K., Pooyan., M. (2012). An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomedical Signal Processing and Control,* 7 (1), 3–19.

[16] Hariharan, M., Polat, K., Yaacob, S. (2014). A new feature constituting approach to detection of vocal fold pathology. *International Journal of Systems Science,* 45 (8), 1622–1634. https://doi.org/10.1080/00207721.2013.794905.

[17] Saldanha, J. C., Ananthakrishna, T., Pinto, R. (2014). Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. *Journal of Medical Imaging and Health Informatics,* 4 (2), 168–173. https://doi.org/10.1166/jmihi.2014.1253.

[18] Arias-Londoño, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., Castellanos-Domínguez, G. (2011). Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *IEEE Transactions on Biomedical Engineering,* 58 (2), 370–379. https://doi.org/10.1109/TBME.2010.2089052.

[19] Godino-Llorente, J. I., Aguilera-Navarro, S., Gómez-Vilda, P. (2000). LPC, LPCC and MFCC parameterisation applied to the detection of voice impairments. In *6th International Conference on Spoken Language Processing (ICSLP 2000)*. ISCA, Vol. 3. 965–968. https://doi.org/10.21437/ICSLP.2000-695.

[20] Watts, C. R., Awan., S. N. (2011). Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research,* 54 (6), 1525–1537. https://doi.org/10.1044/1092-4388(2011/10-0209).

[21] Farazi, S., Shekofteh., Y. (2024). Voice pathology detection on spontaneous speech data using deep learning models. *International Journal of Speech Technology,* 27, 739–751. https://doi.org/10.1007/s10772-024-10134-4.

[22] Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Ghani, M. K. A., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., AL-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences,* 10 (11), 3723. https://doi.org/10.3390/app10113723.

[23] Ankışhan, H., İnam, S. Ç. (2021). Voice pathology detection by using the deep network architecture. *Applied Soft Computing* 106, 107310. https://doi.org/10.1016/j.asoc.2021.107310.

[24] Barry, W. J. (2000). *Saarbrücken Voice Database, Version 2.0*. Institute of Phonetics, Saarland University, Germany. https://stimmdb.coli.uni-saarland.de/.

[25] Krizhevsky, A., Sutskever, I., Hinton., G. E. (2017). ImageNet classification with deep convolutional networks. *Communications of the ACM,* 60 (6), 84–90. https://doi.org/10.1145/3065386.

[26] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 454 (1971), 903–995. https://doi.org/10.1098/rspa.1998.0193.

[27] Huang, N. E., Wu, M.-L. C., Long, S. R., Shen, S. S. P., Qu, W., Gloersen, P., Fan, K. L. (2003). A confidence limit for the empirical mode decomposition and hilbert spectral analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,* 459, 2317–2345. https://doi.org/10.1098/rspa.2003.1123.

[28] Wu, Z., Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis,* 1 (1), 1–41. https://doi.org/10.1142/S1793536909000047.

[29] Eskidere, O., Gürhanlı, A. (2015). Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features. *Computational and Mathematical Methods in Medicine* 1–12. https://doi.org/10.1155/2015/956249.

[30] Kadiri, S. R., Alku, P. (2020). Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing,* 14 (2), 367–379. https://doi.org/10.1109/JSTSP.2019.2957988.

[31] Tirronen, S., Kadiri, S. R., Alku, P. (2022). The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection. *Journal of Voice,* 38 (5), 975–982. https://doi.org/10.1016/j.jvoice.2022.03.021.

[32] Tirronen, S., Javanmardi, F., Kodali, M., Kadiri, S. R., Alku, P. (2023). Utilizing Wav2vec in Database-Independent Voice Disorder Detection. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094798.

[33] Javanmardi, F., Kadiri, S. R., Alku, P. (2023). A comparison of data augmentation methods in voice pathology detection. *Computer Speech and Language,* 83, 101552. https://doi.org/10.1016/j.csl.2023.101552.

[34] Cai, J., Song, Y., Wu, J. (2024). Voice disorder classification using Wav2vec 2.0 feature extraction. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2024.09.002.

[35] Javanmardi, F., Tirronen, S., Kodali, M., Kadiri, S. R., Alku, P. (2023). Wav2vec-based detection and severity level classification of dysarthria from speech. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/DOI:10.1109/ICASSP49357.2023.100948577.

Received October 21, 2024
Accepted August 28, 2025