

A Comprehensive Review Of Multimodal Laryngeal Disorder Detection Using Voice And Imaging Data

Shridevi Soma¹, Hafsa Aiman²

^{1,2} Department of Computer Science and Engineering,
Poojya Doddappa Appa College of Engineering (Autonomous),
Kalaburagi, Karnataka, India
Email: Shridevisoma@pdaengg.com, Hafsaaiman10@gmail.com

ABSTRACT

Laryngeal disorders, including vocal cord nodules, polyps, paralysis, and laryngitis, significantly affect voice production and communication quality. Effective treatment and better patient outcomes depend on early and precise diagnosis. Traditional diagnostic procedures, such as laryngoscopy and clinical voice assessment, provide valuable information but often require specialized equipment and expert interpretation. Contemporary developments in artificial intelligence have enabled the development of automated diagnostic systems based on voice recordings, medical imaging, and multimodal data sources. This review presents a comprehensive analysis of deep learning approaches for laryngeal disorder detection using audio signals and laryngeal imaging data. The study examines commonly used acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral features, and spectrogram representations, as well as image-based feature extraction techniques employed in Convolutional Neural Network (CNN) architectures. Furthermore, recent multimodal learning frameworks that combine voice and image information are analyzed to emphasise their potential for development diagnostic performance. The strengths, limitations, and research challenges of existing approaches are critically discussed, including issues related to dataset availability, model generalization capability, data heterogeneity, and real-world deployment. Finally, future research directions are presented, emphasizing the advancement of robust, accurate, and clinically applicable multimodal systems for reliable laryngeal disorder diagnosis.

Keywords: Laryngeal Disorders, Voice Pathology detection, deep learning, convolutional neural networks, MFCC, Medical Imaging, Multimodal Learning.

1. Introduction

Laryngeal disorders are a group of medical conditions that affect the larynx and vocal folds, resulting in abnormalities in voice production and communication. Common disorders include vocal cord nodules, polyps, laryngitis, and vocal cord paralysis. These circumstances could result from excessive voice use, infections, neurological disorders, trauma, or structural abnormalities of the vocal folds. Individuals affected by laryngeal disorders frequently experience symptoms such as hoarseness, vocal fatigue, breathiness, throat discomfort, and reduced speech intelligibility. Since voice is a primary means of human communication, such disorders can significantly impact social interaction, professional activities, and overall quality of life.

For laryngeal diseases to be effectively treated, early and accurate diagnosis is crucial and prevention of long-term vocal impairment. Conventional diagnostic approaches, including laryngoscopy, stroboscopy, and perceptual voice assessment, give insightful clinical data about vocal fold structure and function. But these processes frequently specialized equipment, experienced clinicians, and hospital-based infrastructure, which may limit accessibility in remote and resource-constrained settings. Furthermore, manual interpretation of clinical findings may introduce variability and increase the time required for diagnosis.

Contemporary developments in Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) But these processes frequently developing automated and non-invasive diagnostic systems for voice disorders. Speech signals contain rich acoustic information that reflects the physiological condition of the vocal folds. Features such as pitch, jitter, shimmer, harmonic-to-noise ratio, Mel-Frequency Cepstral Coefficients (MFCCs), and spectrogram representations have been extensively employed for detecting pathological voice characteristics. At the same time, laryngeal imaging techniques provide direct visualization of structural abnormalities, including inflammation, lesions, swelling, and irregular vocal fold movement. The availability of both acoustic and visual information has encouraged researchers to investigate multimodal diagnostic approaches capable of combining complementary sources of information for improved disease detection.

Recent developments in deep learning have significantly improved the automated analysis of speech recordings and medical images. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and combined neural network architectures are capable of identifying complex patterns directly from raw data without extensive manual feature engineering. Numerous studies have reported that these approaches provide better classification performance than conventional machine learning methods in various diagnostic tasks. Furthermore, combining different sources of information, including voice signals and laryngeal imaging data, has enabled more comprehensive assessment of vocal disorders. By utilizing both acoustic and visual characteristics, multimodal systems can capture complementary disease-related information, leading to more reliable and accurate diagnostic outcomes. These advancements highlight the growing potential of multimodal deep learning techniques for supporting automated detection and classification of laryngeal disorders in clinical settings.

The objective of this review is to examine recent developments in automated techniques for identifying laryngeal disorders using speech recordings, medical images, and combined data sources. Particular attention is given to commonly adopted feature extraction methods, neural network architectures, and multimodal integration strategies reported in recent studies. By evaluating the strengths and limitations of existing approaches, this review highlights current challenges and outlines potential directions for future investigation in computer-assisted laryngeal disorder diagnosis.

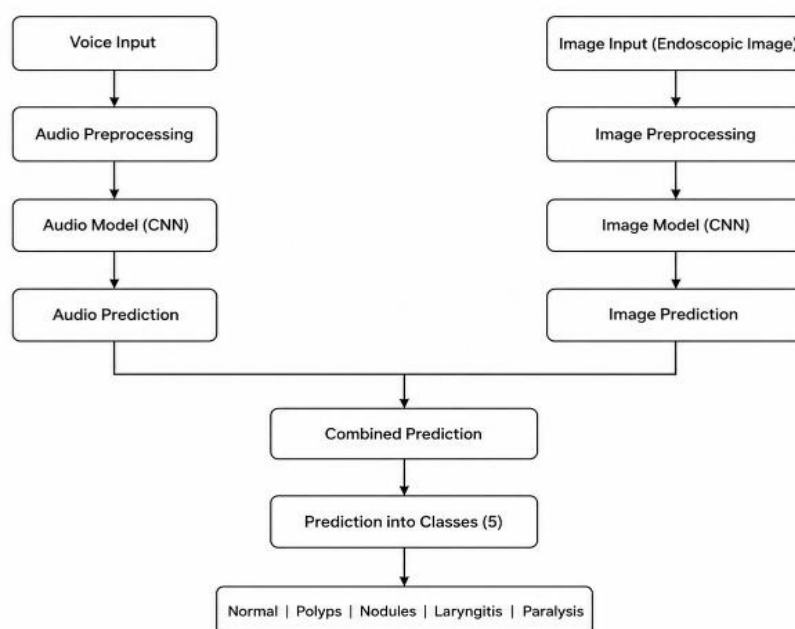


Figure 1. General framework of AI-based laryngeal disorder detection systems

Figure 1 illustrates the workflow of the proposed multimodal vocal disorder detection system using both voice and endoscopic image data. In the first stage, voice signals and endoscopic images are provided as inputs to the system. The audio data undergo preprocessing steps such as spectrogram generation and feature extraction, while the image data are processed through image preprocessing techniques. The processed inputs are then passed into separate CNN-based audio and image models for prediction. After obtaining individual predictions, both outputs are combined in a fusion stage to enhance the overall classification accuracy. Finally, the system predicts one among the five vocal disorder classes: Normal, Polyps, Nodules, Laryngitis, and Paralysis.

2. Review Methodology

This review was conducted through a systematic examination of published literature related to laryngeal disorder detection, voice pathology analysis, deep learning, multimodal learning, and artificial intelligence. Relevant studies were selected from peer-reviewed journals and conference proceedings that addressed automated diagnosis using speech signals and medical imaging. The selected studies were analyzed and categorized according to their methodologies, data modalities, performance characteristics, and limitations to identify current trends, challenges, and future research directions.

3. Methodological Background

3.1 Laryngeal Disorders

Laryngeal disorders include nodules, polyps, vocal cord paralysis, and inflammatory conditions. These disorders affect both structural and functional aspects of voice production.

Table 1. Common Laryngeal Disorders

Disorder	Description	Common Symptoms
Nodules	Benign vocal fold growths	Hoarseness
Polyps	Fluid-filled lesions	Voice fatigue
Laryngitis	Vocal cord inflammation	Sore throat
Paralysis	Restricted vocal fold movement	Breathy voice

3.2 Voice Signal Analysis

Voice signal analysis is one of the most widely used approaches for detecting laryngeal disorders because speech recordings can be obtained non-invasively and at low cost. Acoustic characteristics such as pitch, jitter, shimmer, harmonic-to-noise ratio, and Mel-Frequency Cepstral Coefficients (MFCCs) provide valuable information about vocal fold vibration and voice quality. Variations in these features often indicate the presence of pathological conditions and are commonly used for automated voice disorder detection.

3.3 Image-Based Analysis

Image-based analysis utilizes laryngoscopic and endoscopic images to examine structural abnormalities of the vocal folds. Medical images can reveal important visual indicators such as inflammation, lesions, polyps, nodules, and irregular vocal fold movement. Contemporary developments in image processing and deep learning have enabled the automatic extraction of meaningful visual features, improving the accuracy and efficiency of disease detection.

3.4 Deep Learning

Advances in deep learning have contributed significantly to the development of automated systems for identifying laryngeal disorders. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures have demonstrated the ability to process complex speech and imaging data effectively. Unlike traditional approaches that rely heavily on handcrafted features, these models can learn important patterns directly from the available data. As a result, they have achieved strong performance in classification and diagnostic tasks across various studies. The use of deep learning has also facilitated the analysis of both acoustic and visual information, enabling more comprehensive assessment of vocal fold abnormalities and supporting the development of accurate and reliable diagnostic solutions.

4. Feature Extraction Techniques

4.1 Acoustic Features

Effective analysis of pathological speech depends on the extraction of informative acoustic characteristics. Various signal-processing techniques have been employed to represent changes associated with vocal fold abnormalities. Among these, MFCC-based representations remain widely used because they capture spectral information relevant to human auditory perception. Additional descriptors such as pitch variation, shimmer, jitter, energy measures, and harmonic characteristics are frequently incorporated to improve discrimination between healthy and pathological voices. Advances in deep learning have also reduced dependence on handcrafted descriptors by enabling automatic feature learning directly from spectrograms and raw audio recordings.

Table 2. Common Acoustic Features Used in Voice Analysis

Feature	Purpose
MFCC	Speech representation
Pitch	Vocal frequency analysis
Jitter	Frequency variation
Shimmer	Amplitude variation
Spectrogram	Time-frequency visualization

4.2 Image Features

Medical imaging provides visual evidence of structural changes affecting the larynx and vocal folds. Information derived from laryngoscopic examinations can reveal abnormalities such as lesions, swelling, inflammation, and impaired vocal fold movement. Traditional image analysis methods relied on manually designed descriptors related to texture and shape. More recent approaches employ deep neural networks that automatically identify relevant visual characteristics from image data. These learned representations support the detection and classification of pathological conditions while reducing the need for extensive manual intervention.

5. Deep Learning Approaches for Laryngeal Disorder Detection

The application of deep neural networks has transformed research on automated laryngeal disorder assessment. Instead of relying exclusively on manually selected features, these models learn representations directly from available data. CNN-based architectures are frequently applied to spectrogram analysis and medical imaging because of their ability to recognize spatial patterns. Sequential models such as RNNs and LSTM networks are useful for processing temporal information contained within speech signals. Hybrid architectures that combine multiple learning mechanisms have also been proposed to exploit complementary strengths of different network types. Published studies generally report improvements in predictive performance compared with conventional machine learning methods, although results often depend on dataset size and quality.

6. Literature Review

The selected studies were examined according to their primary source of diagnostic information and methodological approach. For clarity, the literature was organized into categories including speech-based systems, image-driven analysis, multimodal diagnostic frameworks, and survey studies. This organization facilitates comparison of research objectives, computational techniques, reported outcomes, and identified limitations.

1. Ziwei et al. (2014) proposed a multi-parameter voice assessment framework that combined objective acoustic measurements with perceptual evaluation techniques for voice disorder diagnosis. Acoustic parameters such as pitch, jitter, shimmer, and harmonic-to-noise ratio were analyzed to improve diagnostic reliability. The study demonstrated that integrating multiple assessment criteria enhanced clinical evaluation compared with single-parameter methods. However, the framework required considerable manual interpretation and was not suitable for automated large-scale screening applications (**Ziwei et al., 2014**).

2. Muhammad and Alhussein (2021) introduced an Artificial Intelligence (AI) and Internet of Things (IoT)-based healthcare framework for voice pathology detection and remote patient monitoring. Speech signals collected through connected devices were processed using machine learning algorithms to support automated diagnosis. The framework showed potential for telemedicine and continuous health monitoring. However, the study primarily focused on classification performance and lacked extensive validation in diverse clinical environments (**Muhammad & Alhussein, 2021**).

3. Abdulmajeed et al. (2022) presented a comprehensive review of voice pathology detection techniques, diagnostic procedures, and machine learning methods. The authors discussed commonly used feature extraction approaches, including MFCCs and spectral analysis techniques, and highlighted challenges such as limited datasets and variability in recording conditions. Although the review provided valuable insights, it did not include implementation-based evaluation or comparative experimental validation (**Abdulmajeed et al., 2022**).

4. Al-Hussain et al. (2022) conducted a systematic review and meta-analysis of supervised machine learning approaches for screening and diagnosing voice disorders. Their analysis compared several classification algorithms and confirmed the effectiveness of machine learning techniques in improving diagnostic accuracy. The study also identified the need for further investigation of advanced deep learning architectures and multimodal diagnostic systems (**Al-Hussain et al., 2022**).

5. Liu et al. (2023) developed an end-to-end deep learning framework for vocal pathology classification using speech recordings. The model automatically learned discriminative features from spectrogram representations without extensive manual feature engineering. Experimental results demonstrated strong classification performance; however, the evaluation was conducted on relatively limited datasets, which may affect generalizability across broader patient populations (**Liu et al., 2023**).

6. Idrisoglu et al. (2023) performed a systematic review of machine learning techniques for diagnosing voice-related disorders. The study examined feature extraction methods, classification algorithms, and evaluation strategies used in recent research. The authors highlighted the growing adoption of AI in healthcare while emphasizing challenges associated with dataset size, model overfitting, and poor generalization capabilities (**Idrisoglu et al., 2023**).

7. Ahmadian et al. (2024) investigated vocal cord dysfunction using laryngoscopic observations to analyze laryngeal adduction patterns. The research explored whether these visual patterns could predict disease severity and therapeutic response. Findings demonstrated the clinical significance of imaging-based assessment for laryngeal disorders; however, the approach relied heavily on expert interpretation and lacked automated analysis techniques (**Ahmadian et al., 2024**).

8. A recent voice disorder detection study (2024) implemented machine learning algorithms using Mel-Frequency Cepstral Coefficients (MFCCs) extracted from speech recordings. The proposed approach achieved classification accuracies ranging from 85% to 92%, demonstrating the effectiveness of acoustic feature analysis for pathology detection. Nevertheless, model performance was affected by environmental noise and variations in recording quality (**Voice Disorder Detection Study, 2024**).

9. Sindhu and Sainin (2024) reviewed contemporary developments in deep learning methods for automatic speech and voice disorder detection. The study analyzed CNNs, RNNs, LSTMs, and hybrid architectures applied to pathological voice classification. The review concluded that deep learning approaches generally outperform traditional machine learning techniques, although challenges related to dataset availability and clinical deployment remain significant (**Sindhu & Sainin, 2024**).

10. Wang et al. (2024) proposed an AI-based diagnostic framework that integrated voice signals, demographic information, and structured medical records for detecting glottic neoplasms. The multimodal system demonstrated promising diagnostic performance by utilizing complementary information from multiple sources. However, further large-scale clinical validation and integration of additional imaging modalities were recommended (**Wang et al., 2024**).

11. Ma et al. (2025) developed an automated severity assessment system for unilateral vocal cord paralysis using Mel-spectrogram representations and Convolutional Neural Networks (CNNs). The proposed model achieved high classification accuracy and demonstrated the effectiveness of deep learning for objective pathology assessment. Nevertheless, the framework relied exclusively on acoustic information and did not incorporate imaging data (**Ma et al., 2025**).

12. Ozcan (2025) proposed an explainable artificial intelligence framework for differentiating voice disorders using deep learning models. The study improved transparency by identifying important acoustic features contributing to classification decisions while maintaining competitive diagnostic performance. However, validation on larger datasets and integration with multimodal information were identified as important future research directions (**Ozcan, 2025**).

13. Tomaszewska, Kukwa, and Georgakis (2026) developed an audio-based deep learning framework for classifying normal, precancerous, and cancerous laryngeal conditions using Gammatone Cepstral Coefficients (GTCCs). The extracted GTCC features were processed using a Convolutional Neural Network, achieving classification accuracy exceeding 95%. Although the model demonstrated excellent performance, it relied solely

on audio signals and did not incorporate laryngeal imaging information, highlighting the need for multimodal diagnostic systems (Tomaszewska et al., 2026).

7. Comparative Analysis

Table 7.1 : Categorized Analysis of Voice, Image, and Multi-source diagnostic frameworks

A. Voice-Based Approaches

Ref No	Year	Authors & Journal	Methodology	Performance Metrics	Research Gap
1	2024	Ahmadian D et al., <i>Laryngoscope Investigative Otolaryngology</i>	Clinical laryngoscopy analysis	Accuracy: NA, Precision: NA, Recall: NA, Sensitivity: NA, Specificity: NA, F1-Score: NA; Overall therapy response rate = 74.1%	No automation
2	2014	Ziwei Y et al., <i>Journal of Voice</i>	Acoustic + perceptual analysis	Improved diagnostic consistency through combined assessment	Not scalable
3	2025	Ma S et al., <i>BioMedical Engineering OnLine</i>	Mel-spectrogram + CNN	Accuracy: 74.3% (4-class), 84.2% (3-class), 95.4% (binary); Precision: NR; Recall: 74.7%, 83.9%, 95.7%; Sensitivity: 74.7%, 83.9%, 95.7%; Specificity: 91.4%, 92.1%, 95.7%; F1-Score: 74.3%, 84.0%, 95.3%	Only voice-based
4	2024	<i>Engineering Applications of Artificial Intelligence</i>	MFCC + Machine Learning	Accuracy: 85–92% ; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: NR	Noise sensitivity
5	2025	Ozcan F., <i>Scientific Reports</i>	Explainable AI for voice disorder classification	Accuracy: NR; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: NR; Improved model interpretability	Needs multimodal validation
6	2023	Liu GS et al., <i>Laryngoscope Investigative Otolaryngology</i>	End-to-end Deep Learning using stacked vowels	Accuracy: NR; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: 81% (multiclass), 56% (hyperfunctional dysphonia)	Limited dataset
7	2026	Tomaszewska J.Z., Kukwa W., Georgakis A., <i>Biomedical Engineering Advances</i>	GTCC feature extraction + CNN classification	Accuracy: >95% ;	Only audio used; no multimodal integration

8	2020	Harar P. et al., <i>Neural Computing and Applications</i>	Robust voice pathology detection using ML/DL features	Accuracy: NR; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: 73.3%	Limited robustness across datasets
---	------	--	---	---	------------------------------------

B. Multi-source diagnostic frameworks

Ref No	Year	Authors & Journal	Methodology	Performance Metrics	Research Gap
9	2021	Muhammad G. & Alhussein M., <i>IEEE Access</i>	AI + IoT-based voice pathology detection	Accuracy: NR; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: NR; Demonstrated effective remote monitoring	No explainability
10	2024	Wang CT et al., <i>Laryngoscope</i>	Voice signals + demographics + structured medical records + AI	Accuracy: NR; Precision: NR; Recall: NR; Sensitivity: NR; Specificity: NR; F1-Score: NR; Strong classification performance reported	Needs expansion with imaging and larger datasets

c. Review / Survey Papers (Supporting Category)

Ref No	Year	Authors & Journal	Focus	Performance Metrics	Research Gap
11	2022	Abdulmajeed NQ et al., <i>Journal of Intelligent Systems</i>	Voice pathology review	Review paper – No experimental performance metrics reported	No implementation
12	2023	Idrisoglu A et al., <i>JMIR</i>	ML systematic review	Review paper – No experimental performance metrics reported	Dataset limitation
13	2022	Al-Hussain G et al., <i>JMIR</i>	ML meta-analysis	Review paper – No single accuracy value; summarizes multiple studies	Limited DL focus
14	2024	Sindhu I. & Sainin M.S., <i>IEEE Access</i>	Deep learning review	Review paper – No experimental performance metrics reported	Limited explainability discussion
15	2022	Abdul ZKh., Al-Talabani A.K., <i>IEEE Access</i>	MFCC review	Review paper – No experimental performance metrics reported	Review-only study

Table 7.1 compares existing approaches based on methodology, performance, and limitations. It highlights the dominance of voice-based methods and the emerging role of multimodal systems.

Examination of the reviewed studies reveals that speech-oriented diagnostic systems continue to dominate research on laryngeal disorder detection. Most published works employ acoustic representations extracted from voice recordings and use machine learning or deep predictive performance for classification. These approaches have achieved encouraging results across different pathological conditions; however, performance often varies according to recording quality, speaker characteristics, and dataset composition.

Compared with speech-based methods, fewer studies have explored visual analysis of laryngeal abnormalities. Investigations utilizing laryngoscopic images provide direct observation of structural changes affecting the vocal folds but are constrained by limited image datasets and dependence on specialized clinical equipment. Consequently, image-based diagnosis remains less common despite its clinical relevance.

A notable trend identified in recent literature is the gradual movement toward multimodal diagnostic frameworks. By combining acoustic information with imaging and patient-related clinical data, researchers aim to capture multiple aspects of disease manifestation within a single analytical framework. Although these systems frequently report improved predictive capability, widespread adoption remains constrained by data availability, integration challenges, and limited validation across diverse healthcare settings.

The literature also demonstrates increasing reliance on deep predictive performance for feature extraction and classification. Their ability to identify complex patterns from heterogeneous datasets has contributed to measurable improvements in automated diagnosis. Nevertheless, differences in datasets, evaluation strategies, and classification categories make direct comparison among studies difficult. Establishing standardized benchmarks would therefore improve the consistency and reproducibility of future research.

8. Research Gap Analysis

Analysis of the reviewed publications highlights several unresolved issues that continue to influence the development of automated laryngeal disorder detection systems. First, a substantial proportion of existing studies rely exclusively on speech recordings, while relatively few combine acoustic information with visual examination of the vocal folds. Second, many investigations are conducted using limited datasets collected under controlled conditions, making it difficult to assess performance across broader populations and clinical environments. Third, inconsistencies in evaluation procedures and dataset selection complicate direct comparison between published approaches. Finally, only a small number of studies report large-scale clinical validation, creating uncertainty regarding real-world applicability. Addressing these challenges may contribute to the development of more reliable and clinically relevant diagnostic technologies.

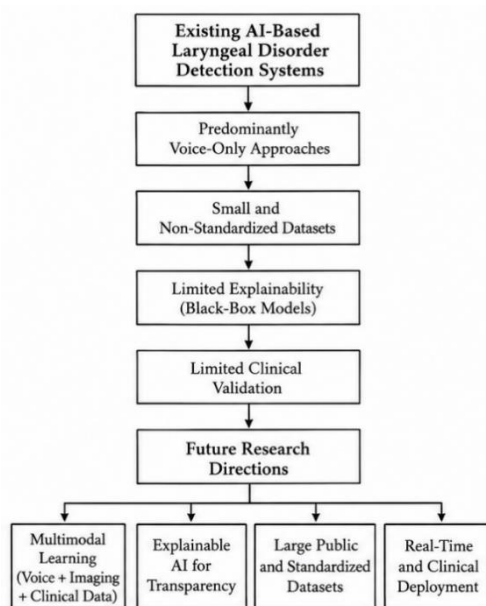


Figure 2. Major research gaps and future directions in AI-based laryngeal disorder detection.

Figure 2 summarizes recurring challenges identified across the reviewed studies. Common issues include limited dataset availability, variability in recording and imaging conditions, insufficient multimodal integration, and restricted evaluation in real clinical environments. These factors continue to influence model reliability and broader healthcare adoption. Future developments are expected to address these limitations through larger datasets, improved integration strategies, and comprehensive validation studies.

9. State-of-the-Art: Voice, Image, and Multi-source diagnostic frameworks for Laryngeal Disorder Detection

9.1 Voice-Based Approaches

Research involving automated assessment of laryngeal disorders has primarily focused on analysis of speech recordings. Voice signals contain acoustic variations associated with pathological changes in vocal fold function, making them suitable for non-invasive screening applications. Traditional studies commonly utilize handcrafted acoustic descriptors, whereas recent investigations increasingly employ predictive performance capable of learning feature representations directly from spectrograms and waveform data. Although many voice-based systems demonstrate encouraging predictive performance, differences in language, speaking style, recording environment, and microphone quality remain important factors influencing reliability outside controlled experimental settings.

9.2 Image-Based Approaches

Visual examination techniques provide valuable information regarding structural abnormalities affecting the larynx and vocal folds. Image-based diagnostic systems use laryngoscopic or endoscopic images to identify pathological characteristics associated with lesions, inflammation, paralysis, and other disorders. Advances in computer vision have enabled automated extraction of visual patterns from medical images using deep neural networks. Despite their clinical importance, image-based approaches are less frequently investigated than speech-based methods because obtaining large annotated image datasets requires specialized equipment, expert labeling, and extensive clinical collaboration.

9.3 Multi-source diagnostic frameworks

Multimodal diagnostic frameworks attempt to integrate information obtained from different clinical sources into a unified analytical model. Most studies combine speech-derived features with imaging information and, in some cases, demographic or medical record data. Such integration allows simultaneous assessment of functional and anatomical characteristics associated with laryngeal disorders. Existing evidence suggests that combining complementary modalities can improve diagnostic consistency; however, the creation of synchronized multimodal datasets remains a major challenge. Additional research is required to evaluate these systems across larger populations and diverse clinical environments.

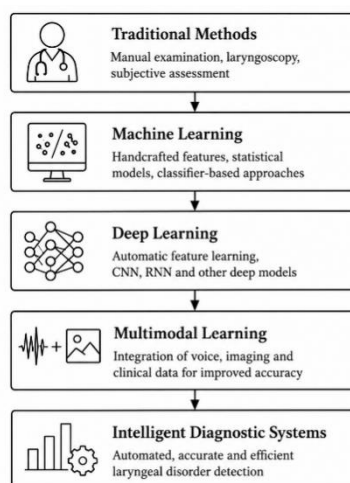


Figure 3. Evolution of laryngeal disorder detection from traditional clinical assessment to intelligent multimodal diagnostic systems.

10. Discussion

Evaluation of the reviewed studies indicates that automated laryngeal disorder detection has evolved considerably during the past decade. Earlier investigations predominantly relied on handcrafted acoustic measurements and conventional machine learning algorithms, whereas recent work increasingly emphasizes representation learning through deep neural networks. Improvements in computational methods have contributed to stronger predictive performance; however, substantial variation exists among datasets, experimental protocols, and evaluation metrics. This variation complicates objective comparison of published findings. Furthermore, research activity remains concentrated on speech-based analysis, while image-based and multimodal investigations represent a

comparatively smaller portion of the literature. Expanding these areas may provide a more comprehensive understanding of pathological vocal conditions and improve diagnostic reliability.

10. Challenges and Open Issues

Several practical and methodological challenges continue to affect progress in this field. Access to large and clinically representative datasets remains limited, particularly for rare pathological conditions. Differences in recording environments, microphone characteristics, and imaging procedures introduce variability that may reduce model robustness. Another challenge involves the lack of standardized evaluation protocols, which makes comparison across studies difficult. In addition, many proposed systems are validated only under controlled research conditions, leaving questions regarding performance in routine healthcare environments. Addressing these issues will be important for translating research outcomes into dependable clinical tools.

11. Future Directions

Future investigations may benefit from collaborative efforts between research institutions and healthcare organizations to establish larger multimodal datasets containing synchronized speech and imaging information. Broader demographic representation would enable more reliable assessment of model generalization. Additional work is also required to develop computationally efficient algorithms suitable for deployment in practical healthcare environments, including portable and mobile platforms. Standardized benchmarking procedures and multi-center validation studies could further strengthen confidence in emerging diagnostic technologies and facilitate their adoption in clinical practice.

12. Conclusion

This review examined contemporary research on automated detection of laryngeal disorders using speech recordings, medical imaging, and multimodal data sources. The findings indicate that speech-based approaches remain the predominant research direction due to their accessibility and non-invasive nature, while image-based and multimodal techniques continue to gain attention because of their ability to capture complementary diagnostic information. Advances in deep learning have improved the extraction of clinically relevant patterns from both acoustic and visual data, contributing to more effective classification systems. Despite these developments, challenges associated with dataset diversity, evaluation consistency, and large-scale clinical validation remain unresolved. Continued progress in multimodal data integration, benchmark dataset development, and collaborative clinical research may support the creation of reliable computer-assisted diagnostic systems for future clinical environments.

References

1. Ahmadian, D., Wehbi, N., Gleadhill, C. M., et al. (2024). *Vocal cord dysfunction: Does laryngeal adduction on laryngoscopy predict disease severity and response to laryngeal retraining therapy?* *Laryngoscope Investigative Otolaryngology*, 9(6), e70039.
2. Ziwei, Y., Zheng, P., & Pin, D. (2014). *Multiparameter voice assessment for voice disorder patients: A correlation analysis between objective and subjective parameters.* *Journal of Voice*, 28(6), 770–774. <https://doi.org/10.1016/j.jvoice.2014.03.006>
3. Ma, S., Liao, W., Zhang, Y., et al. (2025). *Research on automatic assessment of the severity of unilateral vocal cord paralysis based on Mel-spectrogram and convolutional neural networks.* *BioMedical Engineering OnLine*, 24(1), 76. <https://doi.org/10.1186/s12938-025-01401-9>
4. Muhammad, G., & Alhussein, M. (2021). *Convergence of artificial intelligence and internet of things in smart healthcare: A case study of voice pathology detection.* *IEEE Access*, 9, 89198–89209. <https://doi.org/10.1109/ACCESS.2021.3090237>
5. *Voice disorder detection using machine learning algorithms: An application in speech and language pathology.* (2024). *Engineering Applications of Artificial Intelligence*, 133, 108047.
6. Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2022). *A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions.* *Journal of Intelligent Systems*, 31(1), 855–875. <https://doi.org/10.1515/jisys-2022-0137>

7. Idrisoglu, A., Dallora, A. L., Anderberg, P., & Berglund, J. S. (2023). *Applied machine learning techniques to diagnose voice-affecting conditions and disorders: Systematic literature review*. *Journal of Medical Internet Research*, 25, e46105. <https://doi.org/10.2196/46105>
8. Al-Hussain, G., Shuweihdi, F., Alali, H., Househ, M., & Abd-alrazaq, A. (2022). *The effectiveness of supervised machine learning in screening and diagnosing voice disorders: Systematic review and meta-analysis*. *Journal of Medical Internet Research*, 24(10), e38472. <https://doi.org/10.2196/38472>
9. Sindhu, I., & Sainin, M. S. (2024). *Automatic speech and voice disorder detection using deep learning: A systematic literature review*. *IEEE Access*, 12, 49667–49681. <https://doi.org/10.1109/ACCESS.2024.3385327>
10. Ozcan, F. (2025). *Differentiability of voice disorders through explainable artificial intelligence*. *Scientific Reports*, 15(1), 18250. <https://doi.org/10.1038/s41598-025-18250-x>
11. Abdul, Z. K., & Al-Talabani, A. K. (2022). *Mel-frequency cepstral coefficient and its applications: A review*. *IEEE Access*, 10, 122136–122158. <https://doi.org/10.1109/ACCESS.2022.3220325>
12. Bashir, R. N., Shahid, M. A., Rashid, T., et al. (2025). Voice pathology identification using Mel-spectrogram features and deep learning. *Signal, Image and Video Processing*, 19(11), 909.
13. Liu, G. S., Hodges, J. M., Yu, J., Sung, C. K., Erickson-DiRenzo, E., & Doyle, P. C. (2023). End-to-end deep learning classification of vocal pathology using stacked vowels. *Laryngoscope Investigative Otolaryngology*, 8(5), 1312–1318. <https://doi.org/10.1002/lio2.1168>
14. Harar, P., Galaz, Z., Alonso-Hernandez, J. B., Mekyska, J., Burget, R., & Smekal, Z. (2020). Towards robust voice pathology detection. *Neural Computing and Applications*, 32(20), 15747–15757. <https://doi.org/10.1007/s00521-019-04612-4>
15. Wang, C. T., Chen, T. M., Lee, N. T., & Fang, S. H. (2024). *AI detection of glottic neoplasm using voice signals, demographics, and structured medical records*. *Laryngoscope*, 134(11), 4585–4592. <https://doi.org/10.1002/lary.31456>
16. Tomaszewska, J. Z., Kukwa, W., & Georgakis, A. (2026). *Audio-based deep learning classification of laryngeal pathologies using Gammatone Cepstral Coefficients*. *Biomedical Engineering Advances*. (January 2026).