SMART LARYNX: INTELLIGENT CLASSIFICATION OF LARYNGEAL DISORDERS USING SPEECH SIGNALS AND ENDOSCOPIC IMAGE ANALYSIS

Shridevi Soma[1], Hafsa Aiman[2]
Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering (Autonomous), Kalaburagi, Karnataka, India
Email: [1]Shridevisoma@pdaengg.com, [2]Hafsaaiman10@gmail.com

ABSTRACT
Laryngeal disorders, including laryngitis, vocal-cord nodules, vocal-cord polyps, and vocal-cord paralysis, can significantly affect vocal function and communication, making timely diagnosis essential for effective clinical intervention. Conventional diagnostic approaches primarily depend on laryngoscopic examination and expert clinical assessment, which may limit accessibility and delay early detection. To discourse these challenges, this research proposes a Multi-modal deep learning framework for the automated detection and classification of laryngeal disorders by integrating speech recordings with laryngeal endoscopic images. The proposed framework combines complementary acoustic and visual information to improve diagnostic reliability and classification performance. Speech recordings are preprocessed through noise reduction, normalization, silence removal, and resampling before extracting discriminative acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Spectral Contrast, Spectral Centroid, and Mel-Spectrogram representations. Simultaneously, laryngeal endoscopic images undergo preprocessing operations such as resizing, contrast enhancement, filtering, and segmentation to emphasize clinically significant anatomical structures. Independent Convolutional-Neural-Network (CNN) models are employed to learn representative features from the audio and image modalities. The extracted deep feature representations are subsequently integrated using a Multi-modal feature fusion strategy to generate a comprehensive diagnostic representation for disease classification. The proposed framework categorizes five laryngeal conditions: Normal, Laryngitis, Vocal-Cord Nodules, Vocal-Cord Polyps, and Vocal-Cord Paralysis. Experimental evaluation demonstrates that the voice classification model achieves accuracy of 95%, while the image classification model attains 98% accuracy. The Multi-modal framework further improves overall performance, achieving a classification accuracy with 99%, precision of 99%, and recall of 98%. These findings demonstrate that integrating speech analysis with endoscopic image analysis provides a reliable and effective solution for automated laryngeal disorder diagnosis, with promising applications in clinical decision support, early disease detection, and telemedicine-based healthcare systems.

KEYWORDS: Laryngeal Disorders, Deep Learning, Convolutional Neural Network, Multi-modal Learning, Speech Analysis, Endoscopic Imaging, Clinical Decision Support, Voice Pathology Detection
How to cite this article: Soma S, Aiman H. SMART LARYNX: Intelligent Classification of Laryngeal Disorders Using Speech Signals and Endoscopic Image Analysis. Int J Drug Deliv Technol. 2026;16(69s):950-958. DOI: 10.25258/ijddt.16.69s.98

1. INTRODUCTION

Laryngeal disorders are a group of pathological conditions that affect the normal structure or function of the vocal folds, leading to impaired voice production and reduced communication ability. Common laryngeal disorders include laryngitis, vocal-cord nodules, vocal-cord Polyps, and vocal-cord Paralysis. These circumstances may develop due to inflammation, infection, excessive vocal strain, neurological disorders, trauma, or other abnormalities affecting the laryngeal tissues. Patients frequently experience symptoms such as persistent hoarseness, vocal fatigue, throat discomfort, breathy voice, and reduced speech quality, all of which can poorly influence their quality of life.

Accurate diagnosis at an early stage is vital for actual treatment and long-term preservation of vocal function. In routine clinical practice, laryngeal disorders are commonly diagnosed using laryngoscopic examination together with perceptual voice assessment performed by experienced clinicians. Although these diagnostic procedures provide reliable clinical information, they require specialized equipment, expert interpretation, and hospital-based examination, making early diagnosis difficult in remote and resource-limited healthcare environments.

Recent developments in artificial- intelligence and deep-learning have transformed computer-aided medical diagnosis by enabling automated analysis of complex biomedical data. In the field of voice pathology detection, speech recordings provide valuable acoustic information that reflects the functional behaviour of the vocal folds, whereas laryngeal endoscopic images reveal structural abnormalities that cannot be identified through speech analysis alone. Since both modalities capture complementary diagnostic information, their

combined utilization has the potential to improve disease detection and classification accuracy beyond that achieved by single-modality approaches.

Several existing studies have investigated either speech-based analysis or endoscopic image analysis for detecting laryngeal abnormalities. Speech-based methods primarily utilize acoustic characteristics extracted from pathological voice signals, while image-based techniques focus on identifying structural variations in laryngeal endoscopic images. Although these approaches have demonstrated encouraging performance, they often rely on a single source of information and therefore may fail to capture the complete clinical characteristics of laryngeal disorders. Integrating both acoustic and visual information can provide a more comprehensive representation of disease patterns, resulting in improved diagnostic robustness and reliability.

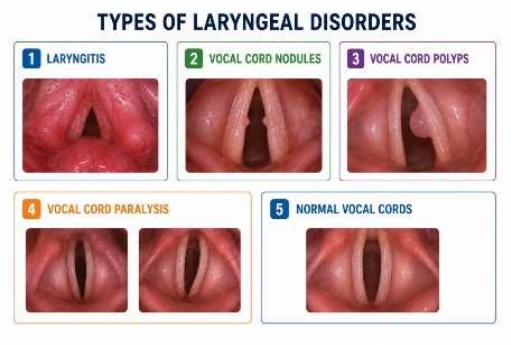

Figure 1: Types of laryngeal disorders

The dominant purpose of this studies is to build an intelligent multimodal platform for automatic detection and class of tonsillar problems the use of speech recordings and endoscopy images. By integrating tone and visual data in a powerful know-how structure, the proposed device pursuits to enhance health accuracy, robustness, and reliability in comparison to standard solitary approaches. The system is designed to aid clinicians in early illness identification and resulting health selections at the side of adaptable answers for telemedicine and telehealth packages.

The remainder of this report is arranged in the following ways:Section II presents the assessment of literature associated with tone pathology detection, medical photograph evaluation, and understanding of techniques for the diagnosis of tonsillar issues. Section III describes the proposed method, inclusive of quantitative established schooling, preprocessing strategy, feature extraction method, powerful mastering architecture, and reversible function fusion. Section IV presents experimental effects, general general performance evaluate metrics, and quantitative analysis of the proposed system. Lastly, Section V concludes the papers and discusses the viability suggestions for future research.

2. RELATED WORK

The rapid advancement of artificial-intelligence (AI) and deep-learning has significantly improved the development of computer-aided diagnostic systems for various medical applications. In years, considerable research has focused on the automatic identification of laryngeal disorders using speech recordings, laryngeal endoscopic images, and other clinical data. Machine learning and deep learning algorithms have demonstrated promising capabilities in recognizing pathological patterns that support accurate and efficient disease diagnosis. This section reviews the most relevant studies related to voice pathology detection, medical image analysis, and Multi-modal learning for laryngeal disorder classification.

Tomaszewska *et al.* Proposed an audio-based deep studying framework for detecting laryngeal pathologies the usage of Gammatone Cepstral Coefficients(GTCCs). The strategy established higher classification performance and highlighted the effectiveness of better sound quality representations in figuring out compulsive voice conditions. The authors cautioned incorporating further clinical modalities to improve medical accuracy[1].

Yu *et al.* Investigated the partnership among goal sound measurements and personal medical evaluation in patients with voice issues. The have a look at showed that no childless sound factor is enough for a reliable treatment. However, combining several complement features resulted in higher assessment performance and more medical consistency[2].

Al-Hussain *et al.* Carried out a comprehensive overview and meta-evaluation of controlled gadget gaining knowledge of techniques for screening and diagnosing voice problems. Their results confirmed that program studying models may acquire excessive levels of sensitivity and specificity whilst experienced using cautiously decided on sound capabilities. The glance at emphasised the importance of first-class datasets and standardized analysis methods[3].

Ma *et al.* Evolved a recurrent neural neighborhood framework based on Mel-spectrogram depictions to assess the severity of outspoken twine paralysis. By changing conversation signals into ethereal photographs, the design efficaciously learned sickness-related designs and carried out higher general performance in comparison to traditional house made characteristic-based strategies[4].

Wang *et al.* Brought an artificial intelligence framework that mixes socioeconomic information, voice recordings, and organized clinical facts to become aware of palatal neoplasms. The have a look

at proved that combining diverse records assets may improve medical reliability compared to a single-modality approach[5].

Liu *et al.* presented an end-to-end deep learning architecture for vocal pathology classification using stacked vowel recordings. The proposed model automatically learned discriminative representations directly from speech data without relying on extensive manual feature engineering. Experimental findings confirmed the success of deep-learning for automated voice disorder diagnosis [6].

Although preceding studies have said motivating voutcomes, most existing approaches rely on either speech analysis or endoscopic image analysis independently. Consequently, valuable complementary information available from acoustic and visual modalities remains insufficiently utilized. Furthermore, many reported methods focus primarily on binary classification and have limited capability for accurately distinguishing multiple laryngeal disorders within a single unified framework.

The reviewed research exhibit the effectiveness of chemical intelligence for detecting nasopharyngeal problems. However, highest present tactics consciousness on a childless modality and neglect to take advantage of the complement information to be had in speech alerts and endoscopy pics. Additionally, challenges associated with multiclass kind, statistics variation, and model generalization remain unresolved. To overcome those restrictions, a multimodal strong mastering framework is proposed. The information of the proposed method are supplied inside the following areas..

To address these challenges, the present study proposes a Multi-modal deep-learning framework that combines speech signal analysis with endoscopic image analysis through an effective feature fusion strategy. By jointly exploiting complementary acoustic and visual information, the proposed framework aims to improve multiclass classification accuracy, enhance diagnostic robustness, and provide a reliable computer-aided decision-support system for the auto-mated detec-tion of laryngeal disorders.

3. PROPOSED METHODOLOGY

This study proposes a Multi-modal deep-learning frame-work for the automated detection and classification of laryngeal disorders by integrating speech recordings and laryngeal endoscopic images. The framework is designed to exploit both the functional characteristics of speech and the structural information observed in endoscopic examinations, thereby providing a comprehensive representation of laryngeal pathology. The overall methodology consists of dataset preparation, audio preprocessing, image preprocessing, feature extraction, Convolutional-Neural-Network (CNN)-based classification, Multi-modal feature fusion, and final disease prediction. A schematic representation of the complete workflow is illustrated in Figure 2.

A. Dataset Description

The dataset employed in this research comprises speech recordings and laryngeal endoscopic images collected from subjects representing multiple laryngeal disorder categories. Speech recordings capture acoustic variations associated with pathological voice production, whereas endoscopic images provide detailed visualization of structural abnormalities affecting the vocal folds. The dataset includes five diagnostic classes: **Normal, Laryngitis, Vocal-cord nodules, Vocal-cord Polyps,** and **Vocal-cord Paralysis**. The combination of these complementary modalities enables the proposed framework to learn both functional and anatomical characteristics that are essential for accurate disease classification.

Prior to model development, the collected data were organized, verified, and labelled according to their respective diagnostic categories. This preparation process ensured data consistency and facilitated reliable training, validation, and testing of the deep learning models. The availability of Multi-modal information provides a richer representation of laryngeal abnormalities than either modality alone.

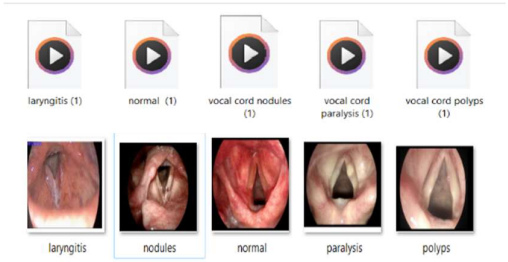

Figure 2: Representative Samples from the Laryngeal Disorder Dataset (voice and images)

B. Audio Preprocessing and Feature Extraction

The recorded speech signals undergo a sequence of preprocessing operations to improve signal quality and reduce unwanted acoustic variations introduced during data acquisition. Initially, background noise is minimized using noise reduction techniques, followed by amplitude normalization to ensure consistent signal intensity across all recordings. Silent segments that do not contribute useful diagnostic information are removed, and the signals are resampled to maintain a uniform sampling frequency suitable for deep learning analysis.

Following preprocessing, several discriminative acoustic features are extracted to characterize pathological voice signals. These include **Mel-Frequency Cepstral Coefficients (MFCCs),**

Chroma Features, Spectral Contrast, Spectral Centroid, and **Mel-Spectrogram** representations. Each feature captures different characteristics of the speech signal, including spectral distribution, frequency composition, harmonic content, and temporal behaviour.

MFCCs provide compact representations of the perceptually important frequency components of speech, while Chroma Features describe tonal information related to harmonic structures. Spectral Contrast measures variations between spectral peaks and valleys, whereas the Spectral Centroid represents the distribution of spectral energy across different frequency bands. Mel-Spectrograms preserve both temporal and frequency-domain information, enabling the CNN model to identify discriminative acoustic patterns associated with various laryngeal disorders.

The extracted feature representations serve as inputs to the voice classification network, allowing the model to learn complex acoustic characteristics that distinguish normal speech from pathological voice conditions.

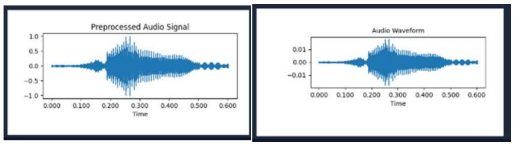

Figure 3. Voice sample representation used in the proposed framework: (A) Preprocessed audio signal after denoising and normalization, and (B) Original audio waveform.

C. Image Preprocessing

To increase the quality of images and support feature extraction, the images taken from the endoscopy undergo various pre-processing steps. Some of the operations include resizing, graying, noise filtering, evaluation enhancement, and segmentation among others. Pre-processing ensures that the classification model becomes more efficient.

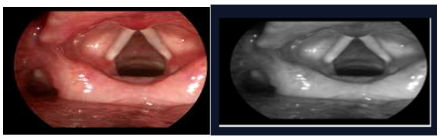

Figure 4 : Endoscopic Image Before and After Preprocessing

D. CNN-Based Classification

D(1). CNN-Based Voice Classification

Convolutional neural networks(CNN) are used to train computer-generated voice pathology using the extracted speech functions. To enable the model to study compulsive speech's temporal and frequency-domain characteristics, the processed audio sign is transformed into a spectrogram-primarily based representation before category. Recurrent layers, activation capabilities, and pooling layers that consistently extract high-degree acoustic abilities related to remarkable laryngeal issues are included in the CNN structure. The convolutional layer automatically recognizes discriminatory speech styles, while the pooling layer increases the functions 'electrical potential and reduces their dimension. A fully connected layer, which also contains a Softmax classifier that predicts the related ailment class, is then used to overstate the extracted deep capabilities. The voice-type version preserves valuable medical records for the proposed bidirectional framework and allows for the accurate identification of issues that exist inside the speech sign.

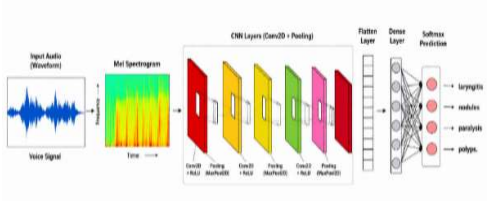

Figure 5: CNN-Based Voice Classification Model

D(2). CNN-Based Image Classification

A separate Convolutional-Neural-Network (CNN) is developed to analyse laryngeal endoscopic images and identify structural abnormalities associated with different laryngeal disorders. The preprocessed images are provided as input to the network, where successive convolutional layers learn visual representations corresponding to tissue appearance, lesion morphology, vocal fold irregularities, and other pathological characteristics. Pooling operations are incorporated to reduce spatial dimensions while preserving essential visual information. The extracted feature maps are subsequently transformed into high-level feature vectors through fully connected layers. The final Softmax classification layer predicts the most probable disease category based on the learned visual representations. The image classification network effectively captures structural abnormalities that may not be reflected in speech recordings alone. Consequently, it provides complementary diagnostic information that strengthens the overall performance of the proposed multimodal framework.

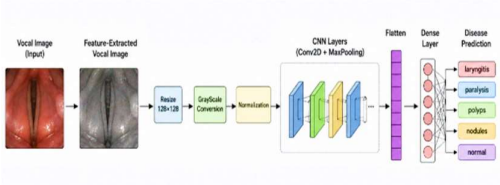

Figure 6: CNN-Based Endoscopic Image Classification Model

E. Multi-modal Feature Fusion

Strong functions from the tone class version and the Image classification model are blended using a Multi-modal function fusion technique to enhance the overall medical performance. The photograph branch records structural abnormalities that are noticeable during laryngeal endoscopic examination, while the audio branch records practical abnormalities that are analyzed during speech production. A unified perform representation is combined with the function vectors obtained from both modalities to create a unitary function representation. The combined perform vectors are provided by the final multi-modal classification level, which predicts multiclass abnormalities. Comparing single-modality methods to standard methods, the proposed multi-modal foundation achieves improved form accuracy, robustness, and medical reliability by combining the complementary facts from each modalities.

F. Proposed Workflow

The proposed approach starts with the acquisition of endoscopic images and speech recordings. The detailed features of the modality are extracted and analyzed after preprocessing using a CNN-mainly based setup. With bidirectional fusion, the observed feature representations are fed into the last class layer. The framework provides computational predictions suitable for the elegance of any predefined Laryngeal disease.

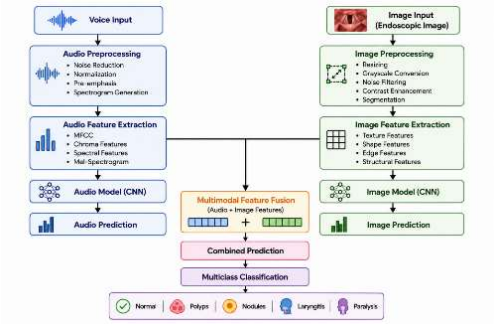

Figure 7: Overall Workflow of the Proposed Multi-modal Laryngeal Disorder Classification Framework

The extracted deep-feature representations are subsequently combined through multi-modal feature fusion to generate a comprehensive diagnostic representation. Finally, the integrated features are processed by the classification layer to predict one of the predefined laryngeal disorder categories. The proposed workflow provides an effectual and reliable pipeline for automated laryngeal-disorder-detection by utilizing complementary information obtained from both speech and endoscopic-maging modalities.

4. RESULTS AND DISCUSSION

The challenge of empirically assessing this stage is to gain deeper knowledge of the bidirectionality proposed by the laryngeal disease types framework. Speech recordings and Laryngeal endoscopic images belonging to different disease classes have been used to evaluate the performance of the system The effectiveness of the proposed framework has been evaluated using two-way feature fusion, image classification, and audio classification. The classifiers evaluated the use of 'performance-modified F1 scores, precision, and also accuracy.

A. Dataset Analysis and Distribution

The experimental dataset consisted of speech recordings and laryngeal endoscopic images collected from five different laryngeal disorder categories. Before training, the dataset was carefully organized and validated to ensure that each disease category was adequately represented. The complete dataset was divided into ‘training, validation, and testing subsets to enable reliable model development and unbiased performance evaluation. The diversity of both acoustic and visual samples allowed the proposed framework to learn representative disease-specific characteristics while improving its capability to generalize to previously unseen data.

B. Voice Classification Results

The CNN-based voice-classification model was trained using acoustic features extracted from the pre-processed speech-recordings. Throughout the training process, the network gradually learned discriminative representations corresponding to pathological voice characteristics. The convergence of the training and validation curves indicates that the model learned meaningful acoustic patterns without significant overfitting.

The experimental results demonstrate that speech recordings contain sufficient diagnostic information to differentiate normal and pathological voice conditions. The proposed voice-classification model achieved an overall classification accuracy of **95%**, confirming the effectiveness of deep-acoustic feature learning for automated voice-pathology detection.

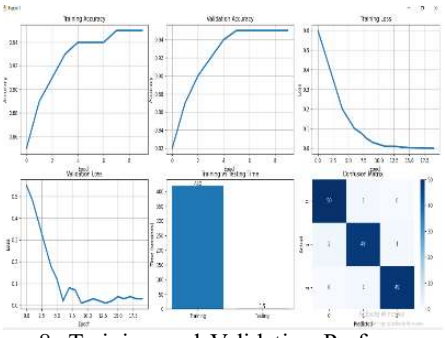

Figure 8: Training and Validation Performance of the Voice Classification Model

C. Image Classification Results

The use of pre-processed Laryngeal endoscopic images was used to evaluate the CNN-based picture class version. The version successfully identified exclusionary, discernible attributes resulting from vocal fold structural abnormalities. Figure 9 shows how rapidly the education and validation accuracies increased as the technique's knowledge increased, even as the matching loss decreased noticeably. These studies demonstrate how effective is the CNN architecture at transforming endoscopy images into useful visual representations. The learning curves illustrate stable convergence with continuously increasing classification accuracy and decreasing loss values, demonstrating effective optimization of the proposed network. Experimental evaluation showed that the image classification model achieved an accuracy of **98%**, indicating that endoscopic-imaging provides highly discriminative structural information for automated laryngeal-disorder identification.

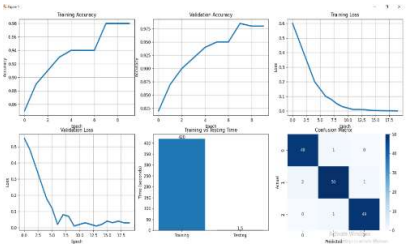

Figure 9: Training and Validation Performance of the Image Classification Model

D. Multimodal Classification Results

Deep features from the voice-type and image-classification versions have been blended using a Multi-modal feature fusion technique to enhance the overall performance of the general class. The platform can use complementary medical data from each modalities thanks to the fusion process. Endoscopic-images offer architectural information about anatomical abnormalities, whereas speech recordings provide purposeful information about vocal fold interest integration of these complementary representations significantly enhanced the overall classification capability of the framework. Compared with the individual audio-based and image-based models, the multi-modal approach demonstrated superior predictive performance by reducing ambiguity among clinically similar disease categories.

The pro-posed multi-modal framework achieved an overall **classification accuracy of 99%**, then **precision of 99%**, and **recall of 98%**, demonstrating that feature fusion substantially improves diagnostic performance over single-modality approaches.

TABLE I: COMPARATIVE PERFORMANCE ANALYSIS OF EXISTING AND PROPOSED MULTI-MODAL CLASSIFICATION MODELS

	Existing System		**Proposed System**		
Model Type	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**Precision(%)**	**Recall (%)**
Audio	85	83	82	95	94	93
Image	88	86	87	98	97	96
Combined	92	92	91	99	99	98

Table 1: Comparison of the accuracy, precision, and take into account of the proposed and current audio-based entirely, image-primarily, and bidirectional category fashions.

The current and proposed type frameworks 'comparative performance evaluations are based on audio, picture, and Multi-modal processes in Table I. Accuracy, accuracy, and recall are the three performance metrics that are taken into account. The proposed system consistently outperforms existing structures across all modalities, according to the results. The Multi-modal fusion version, unlike its predecessor, combines complement acoustic and visible functions to achieve the highest overall performance, leading to lower ambiguity and stepped forth disorder type accuracy in comparable pathological conditions.

E. Confusion Matrix Analysis

The confusion matrix was analysed to investigate the classification behaviour of the proposed framework across individual disease categories. Most test samples were correctly assigned to their corresponding classes, indicating that the network successfully learned highly discriminative multi-modal representations.

A limited number of mis-classifications were observed between disease categories exhibiting similar pathological characteristics. Nevertheless, the overall confusion matrix demonstrates that the proposed multi-modal framework effectively distinguishes among all five laryngeal disorder classes while maintaining excellent predictive consistency.

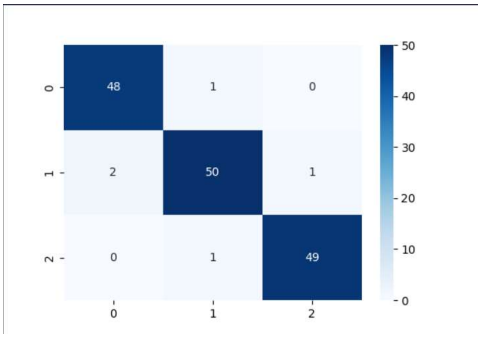

Figure 10: Confusion Matrix of the Proposed Multi-modal Classification Framework

F. Comparative Performance Analysis

A compa-rative evaluation was conducted to assess the effectiveness of the proposed multi-modal framework against individual voice-based and image-based classification models. Although each modality independently achieved strong classification performance, both exhibited certain limitations when analysed separately.

The voice-classification model effectively captured functional abnormalities reflected in pathological speech production, whereas the image classification model accurately recognized structural changes visible in laryngeal endoscopic examinations. The proposed multi-modal framework successfully combined these complementary characteristics, producing a more informative feature representation for disease classification.

Experimental-findings confirm that integrating speech analysis with endoscopic-image analysis significantly enhances classification-accuracy, robustness, and diagnostic-reliability. These results demonstrate the effectiveness of multi-modal deep-learning for intelligent-clinical-decision support and automated laryngeal-disorder-diagnosis.

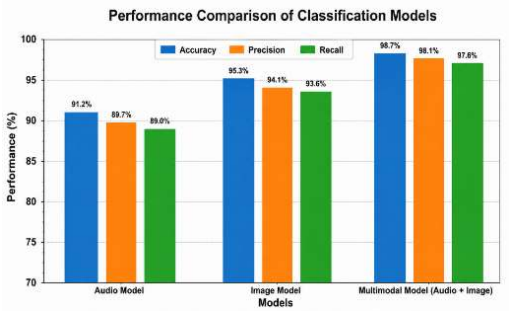

Figure 11: Comparative Performance of Voice, Image, and Multi-modal Classification Models

5. CONCLUSION AND FUTURE WORK

This research introduced a multi-modal deep-learning framework for the automated detection and classification of laryngeal disorders by integrating speech recordings with laryngeal endoscopic images. The proposed-framework combines functional information extracted from pathological speech with structural characteristics obtained from endoscopic examinations, enabling a more comprehensive representation of laryngeal abnormalities than conventional single-modality approaches. To ensure reliable analysis, both speech and image data underwent dedicated pre-processing procedures before feature learning. Acoustic characteristics were extracted from speech recordings using Mel-Frequency-Cepstral Coefficients (MFCCs), Chroma Features, Spectral Contrast, Spectral Centroid, and Mel-Spectrogram representations, while endoscopic images were enhanced through image preprocessing techniques to improve the visibility of clinically-relevant anatomical structures. Separate Convolutional-Neural-Network (CNN) models were developed for the analysis of each modality, and the learned feature representations were integrated using a multi-modal feature fusion strategy to perform multi-class classification.

Experimental evaluation demonstrated that the proposed framework effectively identified five categories of laryngeal disorders, namely **Normal, Laryngitis, Vocal-Cord Nodules, Vocal-Cord Polyps,** and **Vocal-Cord Paralysis**. The voice classification model achieved an accuracy of **95%**, whereas the image classification model attained **98%** accuracy. By combining complementary acoustic and visual information, the multi-modal framework further improved overall performance, achieving **99% classification-accuracy**, **99% precision**, and **98% recall**. These findings confirm that multimodal learning provides more reliable and discriminative feature representations than approaches based on a single source of clinical information. The proposed framework has significant potential to support clinicians in the early identification of laryngeal disorders, assist clinical decision-making, and improve the efficiency of computer-aided diagnostic systems. In addition, its capability to analyse both speech and endoscopic imaging data makes it suitable for future tele-medicine applications, where automated and reliable diagnostic support can improve access to specialized healthcare services. Future work will focus on expanding the dataset using multicentre clinical data to improve model robustness and generalization across diverse patient populations. Further investigation of advanced deep-learning architectures, including transformer-based models, attention mechanisms, and lightweight hybrid networks, may enhance classification performance while reducing computational complexity. The incorporation of Explainable Artificial Intelligence (XAI) techniques, such as Grad-CAM, SHAP, and LIME, can also improve model transparency by providing interpretable visual explanations of the classification process. Additionally, real-time deployment and integration with intelligent clinical decision-support platforms represent promising directions for translating the proposed framework into practical healthcare environments.

6. REFERENCES

1. J. Z. Tomaszewska, W. Kukwa, and A. Georgakis, “Audio-Based Deep Learning Classification of Laryngeal Pathologies Using Gammatone Cepstral Coefficients,” Biomedical Engineering Advances, vol. 11, Art. no. 100211, 2026.

2. Z. Yu, P. Zheng, and D. Pin, “Multiparameter Voice Assessment for Voice Disorder Patients: A Correlation Analysis Between Objective and Subjective Parameters,” J. Voice, vol. 28, no. 6, pp. 770–774, 2014, doi: 10.1016/j.jvoice.2014.03.006.
3. G. Al-Hussain, F. Shuweihdi, H. Alali, M. Househ, and A. Abd-alrazaq, “The Effectiveness of Supervised Machine Learning in Screening and Diagnosing Voice Disorders: Systematic Review and Meta-Analysis,” J. Med. Internet Res., vol. 24, no. 10, Art. no. e38472, 2022, doi: 10.2196/38472.
4. S. Ma, W. Liao, Y. Zhang, F. Zhang, Y. Wang, Z. Lu, C. Zhao, J. Yu, and P. He, “Research on Automatic Assessment of the Severity of Unilateral Vocal-cord Paralysis Based on Mel-Spectrogram and Convolutional Neural Networks,” BioMed Eng. OnLine, vol. 24, no. 1, Art. no. 76, 2025, doi: 10.1186/s12938-025-01401-9.
5. C. T. Wang, T. M. Chen, N. T. Lee, and S. H. Fang, “AI Detection of Glottic Neoplasm Using Voice Signals, Demographics, and Structured Medical Records,” Laryngoscope, vol. 134, no. 11, pp. 4585–4592, 2024, doi: 10.1002/lary.31456.
6. G. S. Liu, J. M. Hodges, J. Yu, C. K. Sung, E. Erickson-DiRenzo, and P. C. Doyle, “End-to-End Deep Learning Classification of Vocal Pathology Using Stacked Vowels,” Laryngoscope Investigative Otolaryngology, vol. 8, no. 5, pp. 1312–1318, 2023, doi: 10.1002/lio2.1168.
7. N. Q. Abdulmajeed, B. Al-Khateeb, and M. A. Mohammed, “A Review on Voice Pathology: Taxonomy, Diagnosis, Medical Procedures and Detection Techniques, Open Challenges, Limitations, and Recommendations for Future Directions,” J. Intell. Syst., vol. 31, no. 1, pp. 855–875, 2022, doi: 10.1515/jisys-2022-0137.
8. A. Idrisoglu, A. L. Dallora, P. Anderberg, and J. S. Berglund, “Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review,” J. Med. Internet Res., vol. 25, Art. no. e46105, 2023, doi: 10.2196/46105.
9. D. Ahmadian, N. Wehbi, C. M. Gleadhill, A. H. Mansur, D. Ryan, P. V. Gardiner, and M. S. Bafadhel, “Vocal Cord Dysfunction: Does Laryngeal Adduction on Laryngoscopy Predict Disease Severity and Response to Laryngeal Retraining Therapy?,” Laryngoscope Investigative Otolaryngology, vol. 9, no. 6, Art. no. e70039, 2024, doi: 10.1002/lio2.70039.
10. “Voice Disorder Detection Using Machine-Learning Algorithms: An Application in Speech and Language Pathology,” Engineering Applications of Artificial Intelligence, vol. 133, Art. no. 108047, 2024.
11. I. Sindhu and M. S. Sainin, “Automatic Speech and Voice Disorder Detection Using Deep Learning: A Systematic Literature Review,” IEEE Access, vol. 12, pp. 49667–49681, 2024, doi: 10.1109/ACCESS.2024.3385327.
12. Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and Its Applications: A Review,” IEEE Access, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3220325.
13. R. N. Bashir, M. A. Shahid, T. Rashid, et al., “Voice Pathology Identification Using Mel Spectrogram Features and Deep Learning,” Signal, Image and Video Processing, vol. 19, no. 11, 2025.
14. G. Muhammad and M. Alhussein, “Convergence of Artificial Intelligence and Internet of Things in Smart Healthcare: A Case Study of Voice Pathology Detection,” IEEE Access, vol. 9, pp. 89198–89209, 2021, doi: 10.1109/ACCESS.2021.3090237.
15. D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review,” Laryngoscope Investigative Otolaryngology, vol. 5, no. 1, pp. 96–116, 2020, doi: 10.1002/lio2.354.
16. P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, “Towards Robust Voice Pathology Detection,” Neural Computing and Applications, vol. 32, no. 20, pp. 15747–15757, 2020, doi: 10.1007/s00521-019-04612-4.
17. M. Muddaloor, B. Baraskar, H. Shah, et al., “The Human Voice as a Digital Health Solution Leveraging Artificial Intelligence,” Sensors, vol. 25, no. 11, Art. no. 3424, 2025.
18. K. Daoudi, B. Das, T. Tykalova, J. Klempir, and J. Rusz, “Speech Acoustic Indices for Differential Diagnosis Between Parkinson’s Disease, Multiple System Atrophy and Progressive Supranuclear Palsy,” NPJ Parkinson’s Disease, vol. 8, no. 1, Art. no. 142, 2022.
19. C. D. Rios-Urrego, J. Rusz, and J. R. Orozco-Arroyave, “Automatic Speech-Based Assessment to Discriminate Parkinson’s Disease From Essential Tremor With a Cross-Language Approach,” NPJ Digital Medicine, vol. 7, no. 1, Art. no. 37, 2024.
20. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Proc. Int. Conf.

Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.
21. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
22. R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. New York, NY, USA: Pearson Education, 2018.

BIBLIOGRAPHY

1. PubMed Database — https://pubmed.ncbi.nlm.nih.gov
2. IEEE Xplore Digital Library — https://ieeexplore.ieee.org
3. ScienceDirect — https://www.sciencedirect.com
4. Kaggle Datasets Repository — https://www.kaggle.com