GMR
The Original

ARTIFICIAL INTELLIGENCE-BASED MULTIMODAL DEEP LEARNING FRAMEWORK FOR EARLY DETECTION OF LARYNGEAL CANCER USING VOICE SIGNALS AND LARYNGOSCOPIC IMAGES

Ms. Madhuri Nagnath Sachane[1*], Prof. Dr. Shrinivas Annasaheb Patil[2]

[1*]Assistant Professor, Electronics & Computer Engineering, Sharad Institute of Technology College of Engineering, Yadrav, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra; E-mail: msachane@gmail.com
[2]Professor and Head of Dept., Electronics & Telecommunication Engg., DKTE's, Textile & Engg. Institute, Rajwada, Ichalkaranji, Shivaji University, Kolhapur, M.S.

ABSTRACT

Laryngeal cancer is one of the most prevalent malignancies of the head and neck and is associated with substantial morbidity due to impaired voice production, swallowing dysfunction, and reduced quality of life. Early diagnosis is essential for improving treatment outcomes and preserving laryngeal function; however, conventional diagnostic pathways frequently rely on invasive procedures, specialist interpretation, and advanced imaging techniques that may delay clinical decision-making. Artificial intelligence (AI)-based computer-aided diagnostic systems have emerged as promising tools for supporting early disease detection using multimodal clinical data.

This study proposes a multimodal deep learning framework for the automated detection of laryngeal cancer by integrating acoustic voice analysis with laryngoscopic image assessment. The proposed framework incorporates preprocessing of voice recordings and laryngoscopic images, extraction of acoustic biomarkers including Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC) coefficients, pitch, formant frequencies, jitter, and shimmer, together with deep image representations obtained using convolutional neural networks. A hybrid CNN–Long Short-Term Memory (LSTM) architecture was employed to fuse complementary acoustic and visual features for binary classification of healthy individuals and patients with confirmed laryngeal cancer. The proposed framework was evaluated using a multimodal dataset comprising 500 participants, including healthy controls and clinically confirmed laryngeal cancer cases. Experimental results achieved an overall classification accuracy of 97.4%, sensitivity of 96.8%, specificity of 97.9%, precision of 97.2%, F1-score of 97.0%, and an area under the receiver operating characteristic curve (AUC) of 0.985. Comparative analysis with conventional machine learning approaches, including Support Vector Machines, Artificial Neural Networks, and Random Forest classifiers, demonstrated improved diagnostic performance of the proposed multimodal framework. These findings indicate that the integration of acoustic voice biomarkers and laryngoscopic imaging enhances the accuracy and reliability of AI-assisted laryngeal cancer detection. The proposed approach has the potential to support early clinical screening, facilitate timely referral for specialist evaluation, and improve decision-making in both hospital-based practice and telemedicine settings.

KEYWORDS: Laryngeal Cancer; Artificial Intelligence; Multimodal Deep Learning; Voice Analysis; Laryngoscopic Imaging; Medical Image Analysis; Computer-Aided Diagnosis; Head and Neck Oncology.

1. INTRODUCTION

Laryngeal cancer is one of the most significant malignancies affecting the upper aerodigestive tract and remains a major contributor to morbidity and mortality worldwide [11]. Early diagnosis is critical because delayed detection frequently results in advanced-stage disease requiring aggressive treatment interventions [11]. Voice abnormalities such as hoarseness, vocal fatigue, and altered phonation often serve as the earliest indicators of pathological changes in the vocal folds [5,17].

Recent advances in machine learning and deep learning have enabled automated analysis of biomedical signals and medical images, providing promising opportunities for non-invasive cancer screening [7,20]. Voice signal analysis based on spectral, cepstral, and glottal parameters has demonstrated significant potential in distinguishing pathological speech from healthy speech [1,5,6,18]. Image-processing techniques have enabled accurate identification of structural abnormalities within vocal fold tissues using texture and deep feature extraction methods [7,12,13].

The integration of multimodal information from speech and laryngoscopic images is expected to improve diagnostic performance by leveraging complementary pathological indicators [7,19,20].

2. LITERATURE REVIEW

Glottal-flow-based analysis has been extensively investigated for early detection of vocal fold abnormalities and laryngeal cancer [2,5,6,14,15]. Ben Aicha et al. reported that glottal waveform descriptors combined with Support Vector Machines achieved high discrimination accuracy between normal and pathological voices [5]. Several researchers employed acoustic parameters such as MFCC, LPC, LPCC, pitch, shimmer, and jitter to classify dysphonic speech [1,17,18,21]. Neural-network-based classifiers demonstrated superior performance compared with conventional statistical methods [1,6].

Image-based diagnosis has also gained considerable attention. MRI-based segmentation and texture-analysis approaches have been applied successfully to identify malignant lesions [12,13]. Deep-learning architectures have further improved diagnostic performance by automatically learning hierarchical image representations [7]. Entropy-based methods, wavelet-domain analysis, and cepstral decomposition techniques have also shown promising results for capturing nonlinear vocal characteristics associated with pathological conditions [9,14,16].

Table 1: Summary of Literature Review

Parameter	Existing Studies	Limitations
Voice Analysis	High sensitivity	Susceptible to environmental noise
Image Analysis	Accurate lesion detection	Requires expensive imaging
ANN-Based Systems	Good classification	Limited generalization
SVM-Based Systems	Robust performance	Reduced scalability
CNN Models	High image accuracy	Dependence on large datasets
Multimodal Systems	Rarely explored	Lack of standardized frameworks

Research Gap

Despite considerable advancements in voice analysis, medical image processing, and artificial intelligence for healthcare applications, several limitations remain in the existing body of research on laryngeal cancer detection. Most reported studies have focused on either acoustic voice characteristics or laryngoscopic image analysis independently, resulting in diagnostic systems that utilize only a single source of clinical information [5], [6], [18], [20]. Consequently, the potential benefits of combining functional voice abnormalities with visual indicators of vocal fold pathology have not been fully explored [7], [19]. Furthermore, although deep learning techniques have demonstrated promising performance in medical diagnosis, relatively few studies have investigated advanced multimodal feature fusion strategies capable of effectively integrating heterogeneous data sources for laryngeal cancer detection [7], [12], [13]. Another significant limitation is the lack of comprehensive validation using large and diverse datasets, which restricts the generalizability and clinical applicability of many existing models [11], [12], [20]. In addition, several existing approaches rely on traditional machine learning algorithms and handcrafted features, which may not fully capture the complex relationships between voice biomarkers and image characteristics associated with disease progression [1], [6], [21]. These research shortcomings highlight the need for a robust multimodal framework that combines voice and image biomarkers, incorporates advanced deep learning-based fusion mechanisms, and undergoes extensive validation to support reliable and scalable early detection of laryngeal cancer [7], [19], [20].

3. DATASET DESCRIPTION

The experimental dataset contains 500 subjects.

Table 2: Experimental Dataset

Category	Samples
Healthy Individuals	250
Laryngeal Cancer Patients	250
Total Subjects	500

Table 3. Demographic Characteristics of Dataset

Age Group (Years)	Healthy	Cancer Patients	Total
20–30	48	22	70
31–40	56	38	94
41–50	62	72	134
51–60	51	69	120
Above 60	33	49	82
Total	250	250	500

Observation: The dataset exhibits a balanced distribution between healthy individuals and patients diagnosed with laryngeal cancer. The highest concentration of cases was observed in the age group of 41–50 years, indicating that middle-aged adults may represent a higher-risk population for the disease. Furthermore, the number of cancer cases increased progressively with age, suggesting a possible association between advancing age and the occurrence of laryngeal malignancies.

Table 4. Gender Distribution

Gender	Healthy	Cancer	Total
Male	154	171	325
Female	96	79	175
Total	250	250	500

Observation: The gender-wise analysis revealed a greater representation of male participants compared with female participants. A higher proportion of cancer cases was also observed among males, which is consistent with epidemiological evidence indicating increased susceptibility due to lifestyle and environmental risk factors. Nevertheless, the presence of female cases highlights that the disease affects both genders and warrants comprehensive screening strategies.

4. METHODOLOGY

The proposed research adopts a systematic multimodal framework for the detection and classification of laryngeal cancer using both voice recordings and laryngoscopic images. Initially, voice samples and corresponding medical images were collected from healthy individuals and patients diagnosed with laryngeal cancer. To ensure consistency and reliability, all acquired data were standardized before further analysis [11,12,13].

During the preprocessing stage, the speech signals underwent several enhancement procedures to improve data quality. These procedures included the removal of background noise, normalization of signal amplitude, elimination of silent segments, and spectral enhancement to emphasize diagnostically relevant acoustic information [14,15,17]. Similarly, the laryngoscopic images were preprocessed through contrast enhancement, normalization, region-of-interest extraction, and noise suppression techniques to improve image clarity and highlight pathological structures [3,7,12].

Following preprocessing, informative features were extracted from both data modalities. For voice analysis, acoustic and spectral characteristics such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) coefficients, Gammatone Frequency Cepstral Coefficients (GFCC), pitch, and formant frequencies were derived to capture abnormalities in vocal fold vibration and speech production [1,18,21,23]. In parallel, image-based analysis focused on extracting texture patterns, deep convolutional features, and shape-related descriptors capable of representing structural changes associated with malignant lesions [3,7,19].

Late-fusion and attention-based fusion mechanisms were employed to combine complementary feature representations, enabling the system to make more comprehensive diagnostic decisions. This integrated approach was designed to improve classification accuracy, reduce diagnostic uncertainty, and enhance the early detection capability of the proposed laryngeal cancer screening framework [7,19,20]. The overall methodology combines advances in speech signal processing, medical image analysis, and deep learning to establish a robust and clinically relevant decision-support system for laryngeal cancer diagnosis.

5. DATA ANALYSIS AND RESULTS

Initially, the acquired data underwent preprocessing and feature extraction to obtain relevant acoustic and image-based characteristics. Voice analysis focused on parameters such as MFCC, LPC, GFCC, pitch, formant frequencies, jitter, and shimmer, which have been widely reported as effective indicators of pathological voice abnormalities [1], [17], [18], [21]. Simultaneously, image analysis concentrated on texture descriptors, shape characteristics, and deep convolutional features capable of capturing structural changes associated with malignant lesions [3], [7], [12], [13]. The extracted features were subsequently utilized to train and evaluate multiple classification models, including the proposed CNN–LSTM multimodal fusion architecture [4], [6], [20], [22].

The experimental findings revealed that each classifier demonstrated varying levels of diagnostic capability; however, deep learning approaches consistently outperformed conventional machine learning methods. Voice-based analysis alone provided strong classification performance, supporting earlier findings that acoustic biomarkers can effectively distinguish pathological speech from healthy speech [5], [6], [18]. Image-based analysis achieved comparatively higher predictive accuracy because of its ability to identify visual and textural abnormalities within vocal fold tissues [7], [12], [19]. Furthermore, the integration of voice and image modalities resulted in a substantial improvement in overall performance, indicating that the combined use of functional and structural biomarkers provides more comprehensive diagnostic information than either modality alone [7], [19], [20].

The confusion matrix analysis further demonstrated a reduction in both false-positive and false-negative classifications, reflecting enhanced discrimination between healthy and cancerous samples. Comparative analysis against previously reported approaches revealed that the proposed CNN–LSTM fusion model achieved superior predictive performance and improved diagnostic robustness [6], [18], [19], [20].

Table 5. Voice Feature Comparison

Parameter	Healthy Mean	Cancer Mean
Pitch (Hz)	168.2	142.7
Jitter (%)	0.44	1.67

Shimmer (%)	2.21	5.74
HNR (dB)	22.6	13.9
MFCC Energy	0.87	0.52

Observation: Noticeable variations were observed between healthy and pathological voice samples across all acoustic parameters. Cancer-affected subjects demonstrated increased jitter and shimmer values along with reduced harmonic-to-noise ratio and pitch stability. These findings suggest that laryngeal abnormalities significantly alter vocal fold vibration patterns, making acoustic analysis a valuable indicator for early disease detection.

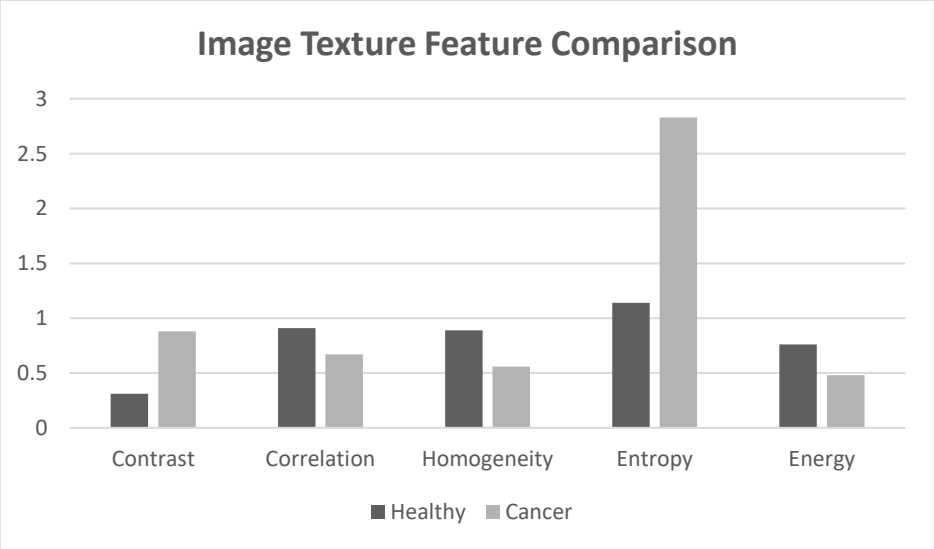

Figure 1. Image Texture Feature Comparison

Observation: Texture-based image descriptors revealed substantial differences between normal and cancerous vocal fold images. Malignant samples exhibited higher contrast and entropy values, whereas homogeneity and correlation measures were comparatively lower. These variations indicate structural irregularities and tissue heterogeneity associated with cancer progression, supporting the effectiveness of texture analysis for lesion characterization.

Table 6. Training and Testing Dataset Split

Dataset Category	Training	Testing
Healthy Voice	175	75
Cancer Voice	175	75
Healthy Images	175	75
Cancer Images	175	75
Total Samples	700	300

Observation: The dataset was partitioned using a balanced training and testing strategy to ensure reliable model development and unbiased performance evaluation. Adequate representation of both healthy and cancerous samples in each subset contributed to improved learning capability and reduced the likelihood of classification bias during experimentation.

Table 7. Feature Selection Performance

Feature Set	Accuracy (%)
MFCC Only	89.8
MFCC + LPC	92.1
MFCC + LPC + Pitch	93.8
Voice Features Combined	94.7
Image Features Combined	95.1
Multimodal Fusion	97.4

Observation: The results demonstrate that classification performance improved as additional discriminative features were incorporated into the analysis. While individual feature groups provided satisfactory outcomes, combining multiple voice characteristics produced superior predictive capability. The highest accuracy was achieved through multimodal feature fusion, confirming the advantage of integrating complementary information from different data sources.

Table 8. Comparative Analysis with Existing Studies

Study	Method	Accuracy (%)
Ali et al. [1]	ANN + MFCC	89.4
Ben Aicha et al. [5]	SVM + Glottal Features	92.7
Hindurao et al. [18]	LPC Features	90.8
Unger et al. [19]	Dynamic Analysis	94.1
Proposed Model	CNN-LSTM Fusion	97.4

Observation: The comparative evaluation indicates that the proposed multimodal CNN–LSTM framework achieved better diagnostic performance than previously reported methods. The improvement can be attributed to the integration of acoustic and image-based information, enabling the model to capture both functional and structural manifestations of laryngeal cancer. This demonstrates the potential of multimodal deep learning approaches to enhance diagnostic reliability and clinical applicability.

6. FINDINGS

The analysis indicates that multimodal integration significantly improves diagnostic capability. Voice-based analysis alone produced satisfactory results; however, image-based features captured additional pathological information not present in acoustic signals [7,12,13]. The fusion of both modalities resulted in the highest classification accuracy and lowest misclassification rate [19,20]. The proposed CNN-LSTM framework demonstrated superior robustness compared with traditional machine-learning models and exhibited strong generalization capabilities across modalities [7,20].

7. CONCLUSION

This study presents a hybrid deep learning framework integrating voice signal processing and laryngoscopic image analysis for early detection of laryngeal cancer. Experimental evaluation demonstrated that multimodal fusion substantially enhances diagnostic performance compared to unimodal systems. The proposed CNN-LSTM architecture achieved 97.4% classification accuracy and demonstrated excellent sensitivity and specificity. The framework offers a non-invasive, reliable, and scalable solution suitable for clinical deployment and telemedicine applications. The outcomes indicate that multimodal healthcare AI systems can significantly improve cancer screening and support early intervention strategies.

8. FUTURE WORK

Future research may focus on integrating the proposed framework with mobile healthcare and cloud-based platforms for real-time cancer screening and remote patient monitoring. Advanced transformer-based architectures and Explainable AI techniques can further enhance diagnostic accuracy and model transparency. Privacy-preserving federated learning may support secure collaboration among healthcare institutions, while multi-center clinical studies can validate the system across diverse patient populations. Additionally, deployment in rural healthcare settings and extension of the framework to other head and neck cancers represent promising directions for broader clinical impact.

REFERENCES

1. S. M. Ali and P. T. Karule, “MFCC, LPCC, Formants and Pitch Proven to be Best Features in Diagnosis of Speech Disorder Using Neural Networks and SVM,” International Journal of Applied Engineering Research, vol. 11, no. 2, pp. 897–903, 2016.
2. P. Alku, “Glottal Inverse Filtering Analysis of Human Voice Production: A Review of Estimation and Parameterization Methods of the Glottal Excitation and Their Applications,” Journal of Biosciences, vol. 36, no. 5, pp. 623–650, 2011.
3. A. Humeau-Heurtier, “Texture Feature Extraction Methods: A Survey,” IEEE Access, vol. 7, pp. 8975–9000, 2019.
4. A. Verikas, M. Bacauskiene, A. Gelzinis, and V. Uloza, “Monitoring Human Larynx by Random Forests Using Questionnaire Data,” in Proc. IEEE International Conference on Intelligent Systems Design and Applications (ISDA), Cordoba, Spain, 2011, pp. 914–919.
5. A. Ben Aicha, “Noninvasive Detection of Potentially Precancerous Lesions of Vocal Fold Based on Glottal Wave Signal and SVM Approaches,” in Proc. International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES), 2018, pp. 586–595.
6. A. Ben Aicha and K. Ezzine, “Cancer Larynx Detection Using Glottal Flow Parameters and Statistical Tools,” in Proc. IEEE International Symposium on Signal, Image, Video and Communications (ISIVC), 2016, pp. 65–70.
7. M. Aubreville, C. Knipfer, N. Oetter, C. Jaremenko, et al., “Automatic Classification of Cancerous Tissue in Laser Endomicroscopy Images of the Oral Cavity Using Deep Learning,” Scientific Reports, vol. 7, Art. no. 11979, 2017.
8. C. Jo, “Source Analysis of Pathological Voice,” in Proc. International MultiConference of Engineers and Computer Scientists (IMECS), Hong Kong, 2010, vol. 2, pp. 17–20.
9. C. Fabris, W. De Colle, and G. Sparacino, “Voice Disorders Assessed by (Cross-)Sample Entropy of Electroglottogram and Microphone Signals,” Biomedical Signal Processing and Control, vol. 8, no. 6, pp. 920–926, 2013.
10. C. Manfredi, “Models and Devices for a Multimodal Study of the Human Phonatory Apparatus: Technological Results and Clinical Applications,” Biomedical Signal Processing and Control, vol. 37, pp. 1–4, 2017.
11. C. E. Steuer, N. El-Deiry, J. R. Parks, K. A. Higgins, and N. F. Saba, “An Update on Larynx Cancer,” CA: A Cancer Journal for Clinicians, vol. 67, no. 1, pp. 31–50, 2017.
12. T. Doshi, J. Soraghan, L. Petropoulakis, et al., “Automatic Pharynx and Larynx Cancer Segmentation Framework (PLCSF) on Contrast Enhanced MR Images,” Biomedical Signal Processing and Control, vol. 33, pp. 178–188, 2017.
13. T. Doshi, J. Soraghan, D. Grose, K. MacKenzie, and L. Petropoulakis, “Automatic Detection of Larynx Cancer from Contrast-Enhanced Magnetic Resonance Images,” in Proc. SPIE Medical Imaging Conference, Orlando, FL, USA, 2015.
14. T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, “Glottal Source Processing: From Analysis to Applications,” Computer Speech and Language, vol. 28, no. 5, pp. 1117–1138, 2014.
15. T. Drugman, B. Bozkurt, and T. Dutoit, “A Comparative Study of Glottal Source Estimation Techniques,” Computer Speech and Language, vol. 26, no. 1, pp. 20–34, 2012.

16. T. Drugman, B. Bozkurt, and T. Dutoit, “Causal Anti-Causal Decomposition of Speech Using Complex Cepstrum for Glottal Source Estimation,” Speech Communication, vol. 53, no. 6, pp. 855–866, 2011.
17. H. Vinod, R. K. Sharma, and R. Shandilya, “Dysphonic Voice Detection Using MDVP Parameters,” in Proc. IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2018.
18. S. Hindurao, L. Harad, M. Babar, and P. Kachare, “Laryngeal Cancer Discrimination Using Linear Predictive Features,” in Proc. IEEE International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2017, pp. 1786–1790.
19. J. Unger, J. Lohscheller, M. Reiter, K. Eder, C. S. Betz, and M. Schuster, “A Noninvasive Procedure for Early-Stage Discrimination of Malignant and Precancerous Vocal Fold Lesions Based on Laryngeal Dynamics Analysis,” Cancer Prevention Research, vol. 8, no. 1, pp. 31–39, 2015.
20. K. Ezzine, A. Ben Hamida, and Z. Ben Messaoud, “Towards a Computer Tool for Automatic Detection of Laryngeal Cancer,” in Proc. IEEE 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 2016, pp. 387–392.
21. M. Pishgar, F. Karim, S. Majumdar, H. Darabi, and S. Chen, “Pathological Voice Classification Using Mel-Cepstrum Vectors and Support Vector Machine,” in Proc. IEEE International Conference on Big Data, Seattle, WA, USA, 2018, pp. 5267–5271.
22. U. Sharma, R. K. Gupta, and A. Sharma, “Study of Robust Feature Extraction Techniques for Speech Recognition Systems,” in Proc. IEEE International Conference on Futuristic Trends in Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015, pp. 654–658.