Research Paper

Explainable Multi-Backbone Deep Learning with Attention-Based Feature Fusion Framework for Laryngeal Carcinoma Classification using Medical Imaging

Muthulakshmi Lakshmanan[1*], Shobana Govindaswamy[2], Meghana Solanki[3], Vidhya Venkatesan[4], Sandhiya Balamurugan[5], Prabakaran Perumal[6]

1. *MCA, M.Kumarasamy College of Engineering (Autonomous), Thalavapalayam, Tamil Nadu, India.*
2. *Department of Computer Science & Engineering (Cyber Security), Madras Engineering College, Vellarai, Chennai, Tamil Nadu, India.*
3. *CSE - Cyber Security, Dayanand Sagar University, Bangalore, India.*
4. *Computer Science and Engineering (Data Science), Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India.*
5. *Department of Computer Science and Engineering, School of Engineering and Technology, Christ University, Bangalore, India.*
6. *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vadeswaram, India.*

Citation Lakshmanan M, Govindaswamy S, Solanki M, Venkatesan V, Balamurugan S, Perumal P. Explainable Multi-Backbone Deep Learning with Attention-Based Feature Fusion Framework for Laryngeal Carcinoma Classification using Medical Imaging International Journal of Medical Toxicology and Forensic Medicine. 2026; 16:E53090.

https://doi.org/10.22037/ijmtfm.v16.53090

Article info:

Received: 04 August, 2026

Accepted: 26 August, 2026

Published: 28 September, 2026

Keywords:

Laryngeal Carcinoma Classification, Medical Image Analysis, Deep Learning,

Attention-Based Feature Fusion, Kolmogorov-Arnold Networks, GradCAM++

ABSTRACT

Background: Automated diagnostic classification of laryngeal carcinoma using laryngoscopic endoscopic images can ease the workload of otolaryngologists, thereby enhancing their efficiency in practice and supporting young clinicians in achieving accurate diagnoses. Deep Learning has become an effective approach for medical image classification and relies on large labeled datasets. However, several existing methods for laryngeal carcinoma analysis often struggle to capture local and global dependencies, thereby affecting classification accuracy and reliability.

Methods: This study presents an Explainable Multi-Backbone Deep Learning Framework for Laryngeal Carcinoma Detection (XMBDL-LCD) in Medical Imaging. The proposed technique utilizes ConvNeXt V2, Swin Transformer V2, and DINOv2 models to extract complementary feature representations from laryngeal images. Besides, the proposed method uses a Multi-Head Cross-Attention Fusion mechanism to combine the extracted features into a unified representation. A Kolmogorov-Arnold Networks-based classifier is applied to perform the final classification of laryngeal carcinoma. To enhance the performance of the proposed method, Explainable AI using Grad-CAM++ is incorporated to visualize the regions that contribute to the model's decision.

Results: The proposed XMBDL-LCD method can be evaluated on a laryngeal endoscopic dataset, and the results demonstrate its efficiency compared with other recent medical image classification approaches, with a maximum accuracy of 94.57%.

Conclusion: Therefore, the proposed model is an assistive tool for laryngeal carcinoma using laryngoscopic endoscopic images.

*** Corresponding Author:**
Muthulakshmi Lakshmanan, PhD
MCA, M.Kumarasamy College of Engineering (Autonomous), Thalavapalayam, Tamil Nadu, India.
E-mail: *muthulakshmiassociateprofessor@gmail.com*

Introduction

Laryngeal Carcinoma (LC) is one of the most common and clinically significant conditions, a malignant tumor that develops in the mucous membrane of the larynx and is a common type of Head and Neck cancer, accounting for nearly one-third of cases [1, 2]. Recent advances in machine learning (ML) and deep learning (DL) have created new opportunities for clinicians to overcome this limitation and deliver improved diagnostic outcomes for LC in a timely manner [3]. Though these methods offer better outcomes, most DL models are considered black boxes, delivering results without clear explanations of how they arrive at their decisions. Based on this requirement, the proposed work presents an explainable multi-backbone deep learning system for laryngeal carcinoma classification, combining advanced deep learning techniques with explainable AI to enhance diagnostic accuracy and clinical understanding. The major contributions of this study are as follows:

- The proposed approach uses a multi-backbone feature extraction strategy using ConvNeXt V2, Swin Transformer V2, and a self-supervised model (DINOv2) for capturing local texture information and global contextual details from the images.

- A Multi-Head Cross-Attention Fusion (MHCAF) mechanism is designed to effectively combine features from different backbones into a unified and more discriminative representation.

- A Kolmogorov-Arnold Networks (KANs) based classifier is employed for final classification, which helps in modeling complex nonlinear relationships in the fused feature space.

- Explainable AI using Grad-CAM++ is incorporated into the proposed work to highlight important regions in the endoscopic images and improve interpretability for clinical understanding.

- The proposed XMBDL-LCD model is evaluated on a benchmark laryngeal endoscopic dataset, and the outcomes are examined using different performance metrics to assess its effectiveness.

Chen et al. [4] discussed comparing the outcomes of the DL and radiomics algorithms for predicting cartilage invasion in LSCC patients and estimating the optimal prognostic approach. The authors [5] presented a Fusion of efficient transfer learning methods with the Pelican optimizer for a laryngeal cancer recognition model. Jiang et al. [6] developed a comprehensive prognostic strategy that integrates clinical factors with DL and radiomics derived from medical imaging to predict recurrence-free survival (RFS) in LSCC patients. Gu et al. [7] investigated the clinical value of Surface-enhanced Raman spectroscopy (SERS) combined with ML for the initial screening of laryngeal cancer in serum, focusing on a 1D-CNN algorithm. Nie et al. [8] designed an enhanced YOLOv8 strategy, MSEC-YOLO, specifically for identifying laryngeal cancer in endoscopic images. A new, improved convolutional model has been designed to enhance the method's feature-extraction capabilities for small-sized goals. Srivastava et al. [9] examined a new technique to categorize laryngeal lesions into 9 morphological classes, including Squamous Cell Carcinoma (SCC) and non-cancerous lesions, and to analyze endoscopic images using an innovative CNN-based approach with image processing and DL. The authors [10] proposed a reliable LC identification employing chaotic metaheuristics incorporated with the DL strategy. Sachane and Patil [11] investigated a hybrid quantum dilated CNN-deep neuro-fuzzy network for laryngeal cancer recognition, utilizing both image and voice data in federated learning.

Materials and Methods

The proposed methodology comprises several key phases: preprocessing, feature representation, feature fusion, classification, and explainable AI.

Overview of XMBDL-LCD Framework

The overall framework of the proposed model is shown in Figure 1. The presented model uses ConvNeXt V2, Swin Transformer V2, and DINOv2 approaches to capture complementary feature representations from laryngeal images.

Preprocessing

Noise reduction was applied via bilateral filtering to remove imaging artifacts, and CLAHE-based contrast enhancement is used to enhance tissue structure visibility in laryngeal endoscopic images. Finally, all images were normalized to maintain uniform intensity distribution and resized to 224 × 224 pixels.

Feature Extraction

This stage is used to extract meaningful features from laryngeal endoscopic images using DL models such as ConvNeXt V2, Swin Transformer V2, and DINOv2.

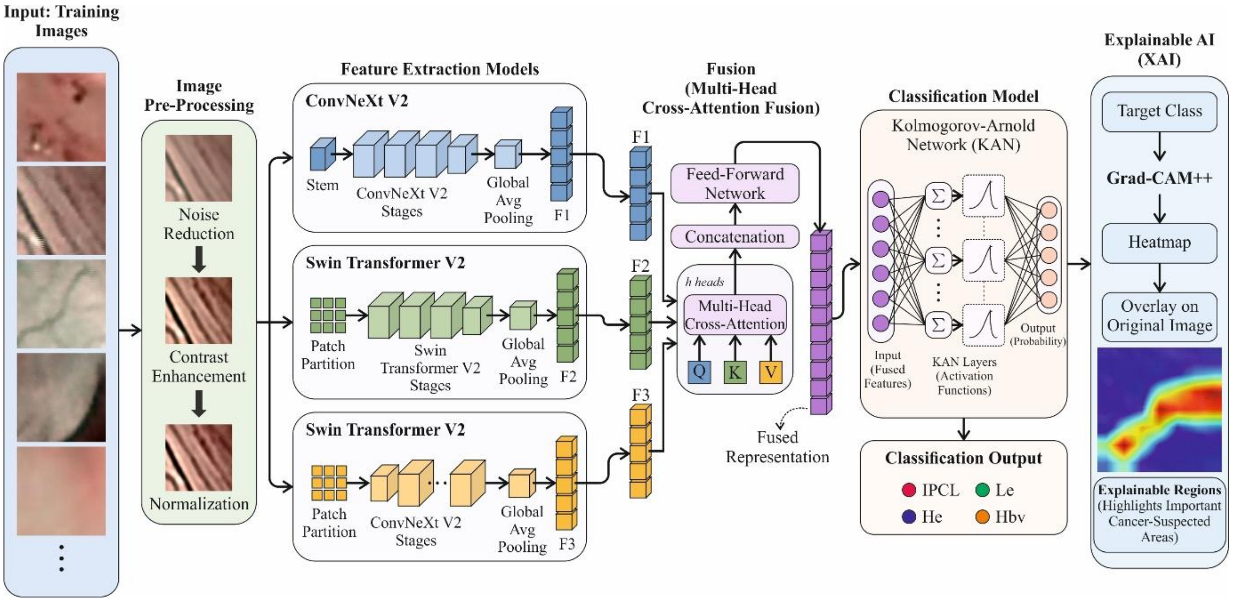

Figure 1. Overall workflow of the XMBDL-LCD model.

ConvNeXt V2

ConvNeXt V2 is an innovative CNN method that extends ConvNeXt V1 by integrating structural optimization and self-supervised learning to advance laryngeal carcinoma diagnosis and medical imaging. ConvNeXt V2 improves feature learning by using fully convolutional masked autoencoder (FCMAE) pretraining and a global response normalizer (GRN), enabling better representation of laryngeal tissue images. ConvNeXt V2 utilizes sparse convolutions to handle masked images, highlighting visible regions and improving pretraining effectiveness.

Swin Transformer V2

Swin Transformer V2 is used to extract rich local and global features from laryngeal endoscopic images to improve classification. It also has the capability to efficiently handle larger-size images and the potential to successfully learn intricate features, thus supporting it in accomplishing greater classification precision in subtle tissue differentiation processes, like differentiating healthy and cancerous laryngeal regions. The Swin Transformer V2 utilizes the standard self-attention formula. The mathematical form of transformer-based models can be given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In this expression, the parameter d_k represents a scaling factor for dimensionality, the term V points to the value matrix, K points to the key matrix, and Q indicates the query matrix

DINOv2

DINOv2 was used as the feature-extraction backbone for the analysis of laryngeal tissue. DINOv2 is an advanced computer vision foundation model developed by Meta AI. It operates on processing images to extract rich, universal feature vectors (embeddings) that can instantly perform various visual tasks. Unlike convolutional neural networks (CNNs), which restrict feature extraction to nearby regions, the self-attention mechanism (SAM) in the Vision Transformer (ViT) captures dependencies across the entire image. This broad contextual awareness is crucial in laryngeal endoscopic images, especially when examining tissues with subtle or uniform surface patterns. While local tissue regions show weak or subtly different patterns owing to visually similar healthy and cancerous mucosal structures, the model can utilize structural and global morphological settings throughout all tissue regions to aid discrimination, thus possibly considering the basic benefit over a locally operated CNN framework.

Multi-Head Cross-Attention Fusion

The DINOv2, Swin Transformer V2, and ConvNeXt V2 modules are employed by extracting the complementary feature representations from the laryngeal endoscopic imaging, and the multi-modal correlation amongst these feature embeddings is obtained for efficient fusion in the laryngeal carcinoma classification process [12]. The MHCAF module incorporates content information obtained by DINOv2, Swin Transformer V2, and ConvNeXt V2 using a multi-level adaptive method. The mathematical

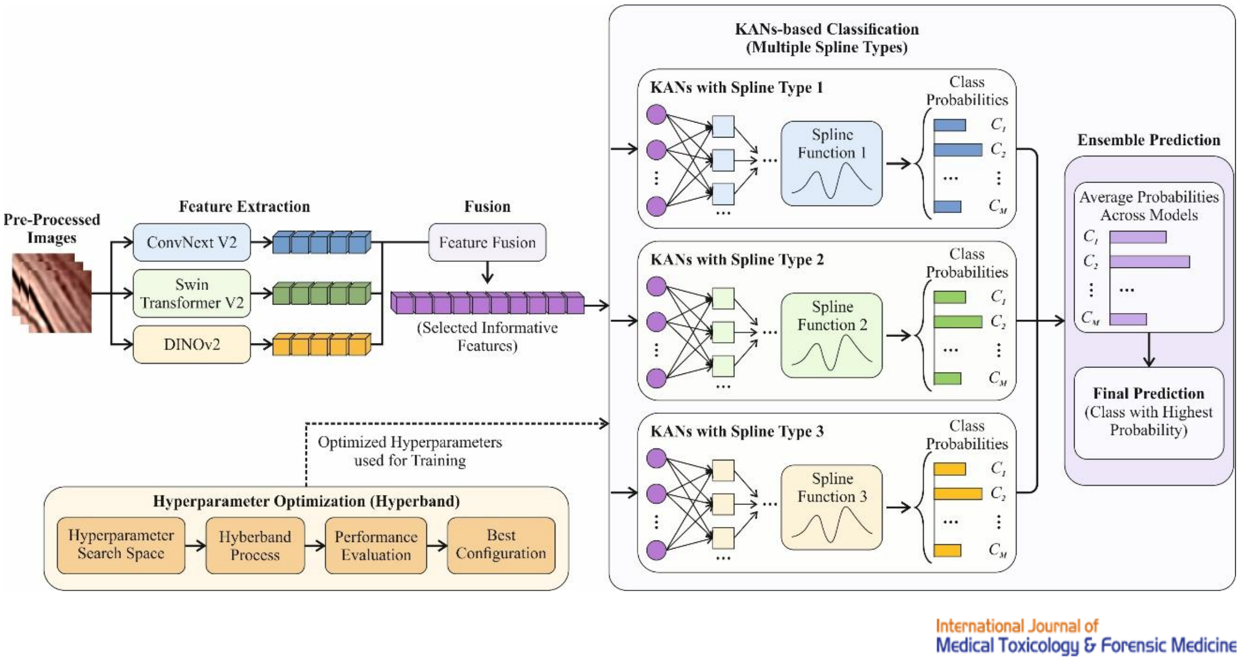

Figure 2. Architecture of KAN model for classification.

operation is given by:

$$R = F_{att}(Q, K, V) \qquad (2)$$

From this mathematical expression, the value $V = W_v Y$ query $Q = W_q Y$, and the key $K = W_k Y$. The fusion representation Y is the input of the MHCAF model that can be described as:

$$Y = \gamma Z_{CN} + (1 - \gamma) Z_S + \delta Z_D \qquad (3)$$

Here, the parameters Z_D, Z_S, and Z_{CN} indicate the output feature representations of DINOv2, Swin Transformer V2, and ConvNeXt V2, respectively.

KANs-based Laryngeal Carcinoma Detection

This stage is responsible for classifying the fused features into categories of laryngeal carcinoma. The Kolmogorov-Arnold Networks (KANs) framework can achieve higher interpretability and accuracy in real-time laryngeal carcinoma classification [13]. KANs use learnable spline functions for modeling feature relationships and perform classification effectively. The calculation of the KANs can be determined by:

$$F(x) = \sum_{K=1}^{2n+1} \theta_K \left(\sum_{P=1}^{n} \theta_{K,P} \left(x_p \right) \right) \qquad (4)$$

Here, the parameters $\theta_K$and $\theta_{K,P}$ point to the univariate functions with respect to the factors of K and P, the term x_p represents the parameter counts, and n refers to the total features extracted from DINOv2, Swin Transformer V2, and ConvNeXt V2. Figure 2 showcases the architecture of the KAN-based classification approach.

The Hyperband optimization method is used to fine-tune the KAN framework. Hyperband is an adjustable resource-allocation and early-stopping method-driven optimization technique. This can efficiently distribute resources to various parameter configurations. The parameters, comprising activation functions, dropout rates, learning rate, layer counts, weight initialization, and batch size, have been fine-tuned by applying the Hyperband optimizer.

Explainable AI Model

The proposed system applies the Grad-CAM++ technique to improve interpretability, building on the capabilities of Grad-CAM [14]. Grad-CAM++ is highly effective at detecting subtle patterns in healthy and cancerous laryngeal endoscopic images. To improve class-discriminative areas, Grad-CAM++ enhances standard Grad-CAM.

Results

The proposed model is evaluated on the benchmark Laryngeal dataset [15], which comprises 1320 image patches of healthy and early-stage cancerous laryngeal tissues. Figure 3 shows the representative sample images.

Figure 4 illustrates the confusion matrix evaluation for training (80%) and testing (20%) on the proposed model across 4 classes. In both phases, most predicted

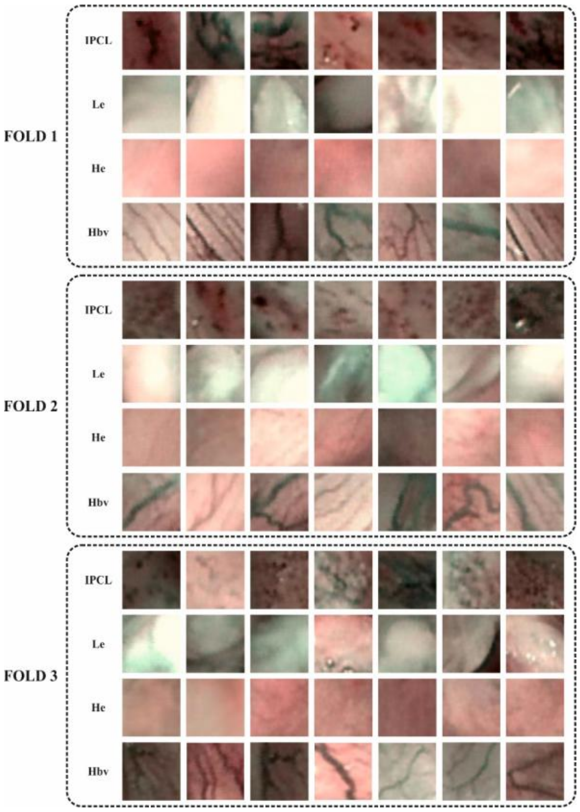

Figure 3. Representative sample images.

values align with the actual class labels, indicating the model's strong learning capability. Misclassifications are minimal in both phases, which reflects the model's strong predictive performance. Therefore, the proposed XMBDL-LCD model demonstrates strong and accurate classification performance in both phases and is well suited for the laryngeal tissue classification task.

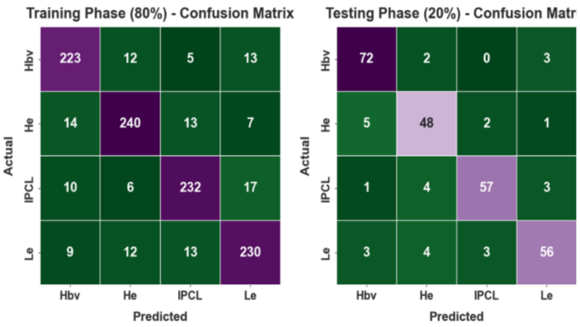

Figure 4. Confusion matrix evaluation of the XMBDL-LCD approach using 80:20 train-test splits.

Figure 5 presents the confusion matrix evaluation for the XMBDL-LCD model on the training (70%) and testing (30%) sets across 4 classes. The results show strong diagonal dominance and very few classification errors across all classes in both phases. The model achieves consistently high correct predictions across all classes with minimal misclassifications, which demonstrates the stability and reliability of the model. Therefore, the confusion matrix analysis validates the XMBDL-LCD approach's efficiency in achieving accurate classification performance.

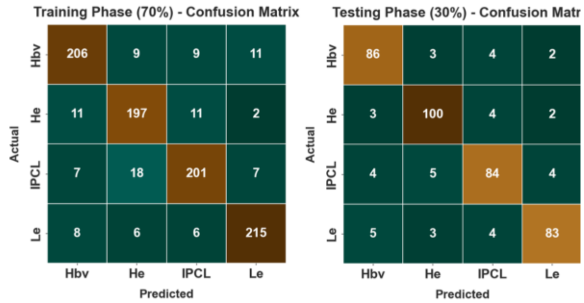

Figure 5. Confusion matrix evaluation of the XMBDL-LCD approach using 70:30 train-test splits.

Discussion

Table 1 exhibits the performance comparative analysis of the XMBDL-LCD model with existing approaches [16, 17]. Among the compared methods, the traditional approaches such as SqueezeNet1, MobileNet, and AlexNet achieved moderate performance, with accuracy values ranging from 78% to 82.60%. Advanced deep learning models, including Xception, MobileNetV2, DenseNet121, and NasNetMobile, delivered notable improvements across all performance metrics but had comparatively lower performance than the proposed model. The proposed XMBDL-LCD model outperforms all compared methods by achieving 89.14% precision, 89.06% recall, 89.08% F-score, and 94.57% accuracy.

Table 1. Comparison study of the XMBDL-LCD method with existing approaches.

Methods	Precision	Recall	F-Score	Accuracy
Xception	88.06	88.66	87.81	90.71
MobileNetV2	86.37	85.17	87.23	90.85
DenseNet121	85.95	85.61	85.62	86.79
NasNetMobile	84.89	84.73	84.80	86.23
Alexnet	80.89	81.00	79.81	82.60
Mnasnet	78.89	82.20	75.53	78.00
Squeezenet1	74.89	78.30	76.62	79.00
XMBDL-LCD	89.14	89.06	89.08	94.57

Figure 6 compares the visual explanation approaches using Layer-CAM, Grad-CAM, and Grad-CAM++ on similar input images. The heatmaps show how all approaches highlight the significant areas that

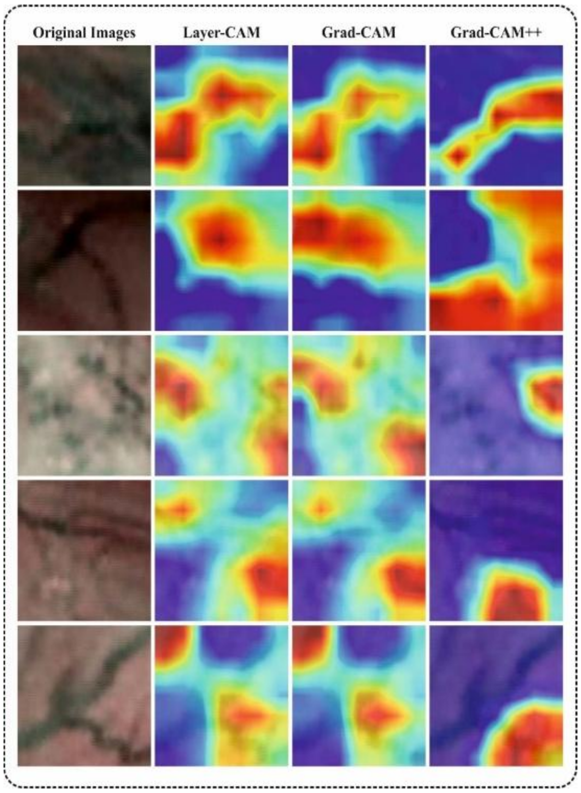

Figure 6. Grad-CAM++ comparison of visual explanation methods.

contribute to the method's prediction.

Table 2 presents an ablation study of the proposed XMBDL-LCD framework, highlighting the contribution of each component. The integration of effective pre-processing and multi-head cross-Attention Fusion with KAN enables the presented system to achieve an accuracy of 94.57%. Therefore, the XMBDL-LCD framework is highly effective in laryngeal carcinoma classification.

Table 2. Ablation study of the proposed XMBDL-LCD model.

Methods	Accuracy	Precision	Recall	F-Score
KAN Only	90.74	85.10	84.13	84.82
ConvNeXt V2 + KAN	91.86	85.98	86.15	86.02
ConvNeXt V2+Swin Transformer V2+KAN	92.78	87.1	87.24	87.14
ConvNeXt V2 + Swin Transformer V2 + DINOv2 + KAN (Without Fusion)	93.61	88.11	88.05	88.07
XMBDL-LCD (Preprocessing + Multi-Head Cross-Attention Fusion + KAN)	94.57	89.14	89.06	89.08

International Journal of Medical Toxicology & Forensic Medicine

Conclusion

In this article, the XMBDL-LCD model is presented for the early and accurate detection of laryngeal carcinoma in endoscopic images. The presented model used ConvNeXt V2, Swin Transformer V2, and DINOv2 approaches to capture complementary feature representations from laryngeal images. In addition, the XMBDL-LCD technique employed an MHCAF mechanism to integrate the extracted features into a unified representation. A KAN-based classifier is used to perform the final classification of laryngeal carcinoma. To enhance the effectiveness of the presented system, Grad-CAM++ is integrated to highlight the regions that contribute to the model's decision. The experimental assessment of the presented XMBDL-LCD technique can be evaluated using a laryngeal endoscopic dataset, and the results demonstrate the model's proficiency compared to existing medical image classification techniques.

Ethical Approval

Not applicable.

Acknowledgment

None.

Funding

No external funding was received for this research.

Conflicts of Interest

The authors report there are no competing interests to declare.

References

[1] Wang SS, Cheng WJ, Wang CT, Hung CH, Fang SH. Improved l. Expert Syst Appl a critical history-based multimodal approach. Expert Syst Appl. 2026:132643. [DOI: 10.1016/j.eswa.2026.132643]

[2] Kim HB, Song J, Park S, Lee YO. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. Sci Rep. 2024;14(1):9297. [DOI: 10.1038/s41598-024-58817-x]

[3] Kang YF, Yang L, Hu YF, Xu K, Cai LJ, Hu

BB, et al. Self-attention mechanisms-based laryngoscopy image classification technique for laryngeal cancer detection. Head Neck. 2025;47(3):944-55. [DOI: 10.1002/hed.27999]

[4] Chen X, Xia S, Jiang H, Lv F, Yu Q, Ning Y, et al. CT-based radiomics and deep learning models for predicting thyroid cartilage invasion and patient prognosis in laryngeal carcinoma. Sci Rep. 2025;15(1):40193. [DOI: 10.1038/s41598-025-23809-y]

[5] Al Khulayf AMF, Alamgeer M, Obayya M, Alqahtani M, Ebad SA, Aljehane NO, et al. Mathematical fusion of efficient transfer learning with fine-tuning model for laryngeal cancer detection and classification in the throat region using biomedical images. Alexandria Eng J. 2025;129:658-71. [DOI: 10.1016/j.aej.2025.06.044]

[6] Jiang H, Xie K, Chen X, Ning Y, Yu Q, Lv F, et al. A prognostic model integrating radiomics and deep learning based on CT for survival prediction in laryngeal squamous cell carcinoma. Sci Rep. 2025;15(1):29990. [DOI: 10.1038/s41598-025-15166-7]

[7] Gu M, Yu B, Liu H, Kong H. Multivariate feature analysis of early-stage laryngeal cancer serum components using surface-enhanced Raman spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc. 2026:127625. [DOI: 10.1016/j.saa.2026.127625]

[8] Nie X, Zhang X, Wang D, Liu Y, Xing L, Liu W. Laryngeal cancer diagnosis based on improved YOLOv8 algorithm. Mach Learn Sci Technol. 2025;6(1):015011. [DOI: 10.1088/2632-2153/ada2d9]

[9] Srivastava R, Kumar N, Sandhan T. Binary classification of laryngeal images utilising ResNet-50 CNN architecture. Indian J Otolaryngol Head Neck Surg. 2025;77(2):644-51. [DOI: 10.1007/s12070-024-05202-9]

[10] Alazwari S, Maashi M, Alsamri J, Alamgeer M, Ebad SA, Alotaibi SS, et al. Improving laryngeal cancer detection using chaotic metaheuristics integration with squeeze-and-excitation ResNet model. Health Inf Sci Syst. 2024;12(1):38. [DOI: 10.1007/s13755-024-00296-5]

[11] Sachane MN, Patil SA. Hybrid QDCNN-DNFN for laryngeal cancer detection using image and voice analysis in federated learning. Sens Imaging. 2024;26(1):1. [DOI: 10.1007/s11220-024-00531-z]

[12] Huo G, Zhang Y, Gao J, Wang B, Hu Y, Yin B. CaEGCN: cross-attention fusion based enhanced graph convolutional network for clustering. IEEE Trans Knowl Data Eng. 2023;35(4):3471-83. [DOI: 10.1109/TKDE.2021.3125020]

[13] Sait ARW, AlBalawi E, Nagaraj R. Ensemble learning driven Kolmogorov-Arnold networks-based lung cancer classification. PLoS One. 2024;19(12):e0313386. [DOI: 10.1371/journal.pone.0313386]

[14] BR B. Comparative analysis of Grad-CAM variants for enhancing explainability in PCOS detection using ultrasound imaging. 2025. [DOI: 10.12785/ijcds/1571152580]

[15] Laryngeal dataset [dataset on the Internet]. Kaggle; [date unknown] [cited 2026 Sep 21]. [Link]

[16] Emre E, Cetintas D, Yildirim M, Emre S. Vocal fold disorders classification and optimization of a custom video laryngoscopy dataset through structural similarity index and a deep learning-based approach. J Clin Med. 202 structural similarity index and a deep learning-based approach. J Clin Med. [DOI: 2023390/jcm14196899]

[17] Xu ZH, Fan DG, Huang JQ, Wang JW, Wang Y, Li YZ. Computer-aided diagnosis of laryngeal cancer based on deep learning with laryngoscopic images. Diagnostics. 2023;13(24):3669. [DOI: 10.3390/diagnostics13243669]