A Multimodal Model for Laryngeal and Hypopharyngeal Lesions Diagnosis: A Multicenter Retrospective Study

Jiahong Zhang* *Zhejiang University

Article is in preparation

Abstract

Background: Diagnosing laryngeal cancer (LCA) is a challenging yet highly valuable task. In recent years, an increasing number of studies have attempted to leverage the capabilities of deep learning algorithms to extract and integrate patient information to improve diagnostic accuracy.

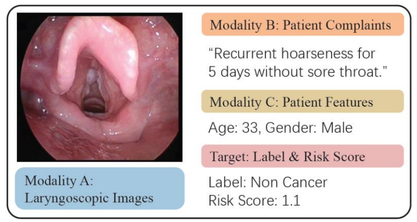
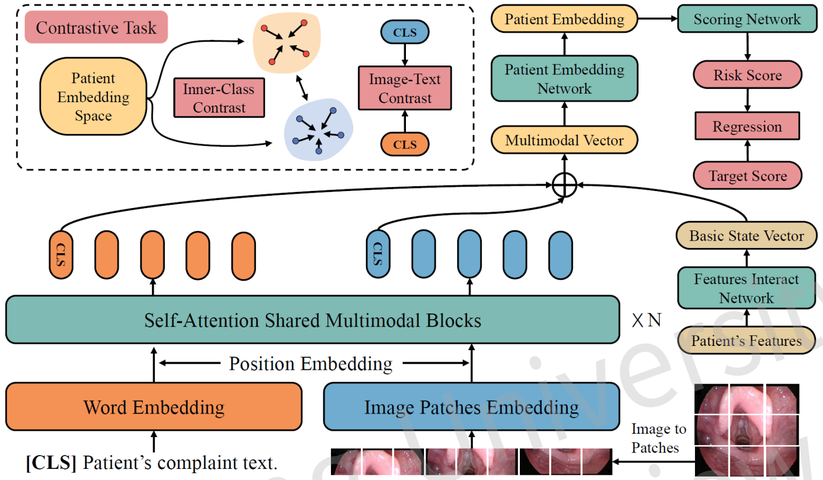


Figure 1: Multimodal patient sample example. Three modalities are involved, which are patient features, patient complaints, and laryngoscopic images.

Methods: This paper proposes a deep learning nerwork architecture based on the fusion of multimodal information and feature space contrastive interaction, called Multimodal Fusion Diagnostic Network (MFDN, Info-Dig), to explore this problem.

Result: The MFDN model achieves an accuracy of 95.42%, a precision of 95.42%, a recall of 95.40% and a F1-score of 95.38% in overall performance. To our best knowledge, this article is the **first work** to conbine aryngoscopy, outpatient history data, and clinicopathological findings to construct a tripartite model in diagnosing laryngeal cancer. Analysis of the experimental results reveals that the multimodal diagnostic model demonstrates promising performance and generalization, making it a highly potential research direction.

Framework Overview & Experiments



Modality	Train Set				Validation Set			
	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.
Image	95.39	94.36	95.65	94.96	84.17	84.17	85.53	84.46
Text	88.78	87.60	89.54	88.42	65.02	65.00	67.37	65.74
Image+Feat	98.64	98.45	98.35	98.39	87.92	88.38	87.92	87.92
Image+Text	98.43	98.19	98.30	98.24	87.50	89.26	87.50	87.76
Text+Feat	97.17	96.64	97.44	96.96	67.50	67.50	71.14	67.41
Image+Text+Feat	95.91	94.67	96.58	95.53	**92.50**	**93.11**	**92.50**	**92.52**

	Internal Validation Set				External Validation Set			
	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.
Info-Dig	95.42	95.42	95.40	95.38	84.95	84.91	86.16	85.29
Expert	86.15	86.14	87.43	86.41	58.06	55.07	58.25	56.20
Senior	82.92	82.92	84.51	82.77	56.99	54.33	55.78	54.76
Competent	79.44	79.44	83.04	79.45	50.90	48.60	50.65	49.26

Model	Train Set				Validation Set			
	Acc.	Pre.	Rec.	F1.	Acc.	Pre.	Rec.	F1.
Vision Backbones								
Resnet50	96.86	95.95	97.23	96.55	79.17	80.38	79.17	79.05
Resnet101	95.28	94.25	95.86	94.96	81.25	82.67	81.25	81.48
ViT-Base	96.02	95.32	96.00	95.60	81.34	82.45	81.25	81.52
ViT-Large	96.12	95.30	95.93	95.59	82.50	83.05	82.50	82.65
SwinT	95.60	94.84	95.33	95.06	84.17	86.27	84.17	84.32
CLIP-Vision	95.39	94.86	94.65	94.75	84.48	84.46	84.37	84.12
Vision-Language Aggregation								
Resnet50 + BERT	96.65	95.89	97.02	96.42	81.27	84.96	81.25	81.49
Resnet101 + BERT	99.79	99.76	99.76	99.77	82.08	83.32	82.08	82.21
ViT-Large + BERT	98.32	97.79	98.48	98.13	84.58	86.49	84.58	84.85
SwinT + BERT	98.89	96.35	98.43	98.02	85.83	86.26	85.83	85.96
CLIP	95.81	95.22	95.58	94.89	86.67	86.91	86.67	86.74
Visual-Textual interaction and Patient information								
Info-Dig (ours)	98.74	98.25	98.93	98.57	**95.42**	**95.42**	**95.40**	**95.38**