

후두내시경 영상에서의 라디오믹스에 의한 병변 분류 연구

박준하¹ · 김영재^{1,2,3,4} · 우주현^{5**} · 김광기^{1,2,3,4*}

¹가천대학교 보건과학대학 의용생체공학과, ²가천대길병원 의료기기 R&D센터,

³가천대학교 의과대학 의공학교실, ⁴가천융합의과학원 의용생체공학과,

⁵가천대길병원 이비인후과

Research on the Lesion Classification by Radiomics in Laryngoscopy Image

Jun Ha Park¹, Young Jae Kim^{1,2,3,4}, Joo Hyun Woo^{5**} and Kwang Gi Kim^{1,2,3,4*}

¹Department of Biomedical Eng., College of Health & Sci., Gachon University

²Medical Devices R&D Center, Gachon University Gil Medical Center

³Department of Biomedical Engineering, Pre-medical Course, College of Medicine, Gachon University

⁴Department of Health Sci. & Technol., Gachon Advanced Institute for Health Sci. & Technology (GAIHST), Gachon University, Lee Gil Ya Cancer and Diabetes Institute

⁵Department of Otorhinolaryngology-Head and Neck Surgery, Gachon University School of Medicine, Gil Medical Center

(Manuscript received 22 August 2022 ; revised 18 October 2022 ; accepted 22 October 2022)

353

Abstract: Laryngeal disease harms quality of life, and laryngoscopy is critical in identifying causative lesions. This study extracts and analyzes using radiomics quantitative features from the lesion in laryngoscopy images and will fit and validate a classifier for finding meaningful features. Searching the region of interest for lesions not classified by the YOLOv5 model, features are extracted with radiomics. Selected the extracted features are through a combination of three feature selectors, and three estimator models. Through the selected features, trained and verified two classification models, Random Forest and Gradient Boosting, and found meaningful features. The combination of SFS, LASSO, and RF shows the highest performance with an accuracy of 0.90 and AUROC 0.96. Model using features to select by SFM, or RIDGE was low lower performance than other things. Classification of larynx lesions through radiomics looks effective. But it should use various feature selection methods and minimize data loss as losing color data.

Key words: Radiomics, Machine learning, Laryngoscopy, Laryngeal disease, Quantitative

*Corresponding Author : Kwang Gi Kim

Gil Medical Center, Gachon University, College of Medicine,
38-13, Dokjeom-ro 3beon-gil, Namdong-gu, Incheon, Republic
of Korea

Tel: +82-32-458-2770

E-mail: kimkg@gachon.ac.kr

**Corresponding author: Joo Hyun Woo

ENT office, Namdong-daero 774 beon-gil 21, Namdong-gu, Incheon,
Republic of Korea

Tel: +82-32-460-3765

E-mail: woojh@gilhospital.com

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT
연구센터육성지원사업(IITP-2022-2017-0-01630)과 경기도의 경기도
지역협력연구센터 사업의 일환으로 수행하였음[GRRC-가천2020(B01),
AI기반 의료영상분석].

I. 서 론

후두 질환은, 신체적, 사회, 직업적, 의사소통에 큰 영향을
끼치며, 환자의 삶의 질을 떨어뜨린다. 음성장애지수(Voice
Handicap Index, VHI)와 음성과 관련된 삶의 질(Voice-
Related Quality Of Life, V-RQOL)과 같은 음성과 관련된
심리측정적 평가 도구들의 존재는 삶의 질에 대한 후두 질환의
영향을 보여준다[2]. 후두내시경은 후두 질환 진단에 대표적
으로 사용되는 검사 방법이며, 표면 점막의 이상을 보여주고,
점막의 움직임이나 병변에 대한 다양한 정보를 제공한다는
장점이 있다[3,4]. 컴퓨터 단층촬영(Computed Tomography,

CT) 및 자기공명영상(Magnetic Resonance Imaging, MRI)은 후두내시경으로 보기 힘든 해부학적 공간의 객관적인 평가를 가능하게 하지만, 보조적인 검사에 국한된다[5]. 내시경 검사는 후두 질환 진단에 중요한 역할을 하지만, 전문의의 충분한 훈련과 경험을 요구하고, 의료진의 주관적인 진단을 해야 하는 어려움이 있다[6,7].

기존의 후두내시경 검사 방법의 어려움을 보완하기 위해 합성곱 신경망(Convolution Neural Network, CNN)을 이용한 연구사례가 발표되고 있다[8-13]. 2019년, Hao Xiong 등은 후두 내시경 영상 10,892장을 기반으로 GoogLeNet Inception v3를 백본 네트워크로 이용한 심층 합성곱 신경망(Deep Convolutional Neural network, DCNN) 모델을 통해 후두암 진단을 시도하였으며, 제안된 모델은 정확도 0.87의 정확도를 보였다[8]. 2021년, Muhammad Adeel Azam 등은 624장의 백색광 및 협대역 영상을 기반으로 YOLOv5 모델을 통해 후두 편평 세포 암종의 검출을 시도하였고, 제안된 모델은 25fps 속도에서 mAP 0.63의 검출 성능을 보였다[9]. 2021년, Gen Ye 등은 114개의 사례를 기반으로 ResNet50 모델을 통해 인후두 역류 질환의 진단을 시도했으며, 제안된 모델은 최대 73.4%의 정확도로 진단하였다[10].

내시경 영상에선 합성곱 신경망을 이용한 연구만이 아니라, 이미지의 히스토그램 분포, 모양 및 텍스처 특징을 이용한 연구들도 보고되고 있다[11-13]. 2011년, Keiichi Onji 등은 165명의 결장직장 내시경 영상을 Gray-level difference matrix, spatial Gray-level dependence matrix, Gray-level histogram moments로 분석하여, 정량화된 특징과 종양의 조직학적 특징의 유의미한 관계와 이를 통한 직장 병변 유형의 구분의 가능성을 보여주었다[11]. 2013년 Sonu Sainju 등은 캡슐 내시경에서 히스토그램 분석을 통해 얻은 특징을 인공신경으로 분석해 장내 출혈 검출하는 모델을 제안하였고, 최대 89%의 정확도로 검출하였다[12]. 2021년, Danh-Huy Vu 등은 227개의 내시경 영상을 기반으로 색상 히스토그램 확률 및 텍스처 특징을 서포트 벡터 머신(Support Vector Machine, SVM)을 이용해 역류성 식도염 경계선의 병변 분류를 시도하였으며, 제안된 모델은 최대 100%의 정확도로 분류하였다[13].

라디오믹스(Radiomics)는 의료 영상에서 히스토그램 분포, 모양 및 텍스처 특징 등을 대량으로 추출, 분석하는 기법이다[14]. 추출된 특징들은 병변 내의 생물학적인 정보를 포함하며, 유의한 특징을 찾아 비침습적으로 바이오마커로 사용할 수 있다[15,16]. 이러한 바이오마커는 병변의 진단만이 아니라 암의 병기, 전이, 예후를 예측하는 인자로 사용할 수 있다[17]. CT 영상을 라디오믹스 특징으로 대장암의 병기를 수신자 응답 곡선 아래 영역의 넓이(Area Under the ROC Curve, AUC) 0.71로 구분하는 성능을 보여줌으로써, 라디

오믹스 특징이 병기의 예측이 지표가 될 수 있음을 보였다[18]. 폐암의 예후나 전이를 예측하는 데에 지표로 사용하였으며[19,20], MRI 영상을 기반으로 뇌종양의 전이성 종양 유형을 예측하는 지표로 사용할 수 있다[21].

후두내시경에서 라디오믹스 특징을 추출하여 정량적 판단이 가능한 바이오마커를 선정할 수 있다. 이러한 바이오마커를 통한 정량적인 진단은 전문의의 경험에 기반한 주관적인 판단의 한계를 극복할 수 있다. 또한 이러한 바이오 마커는 인공지능 모델의 진단 결과를 설명할 수 있는 근거가 될 수 있는, 설명가능한 모델을 설계 가능하다는 장점이 있다. 하지만 후두내시경에서는 라디오믹스를 통해 유의한 특징을 찾고 바이오 마커를 선정하는 연구는 미비하다. 따라서 본 논문에서는 후두내시경 영상에서 라디오믹스 특징들을 추출하고 특징들을 인공지능으로 분석, 선별할 것이다. 선별된 특징을 이용하여 분류기를 학습시키고, 평가하여 라디오믹스 특징과 후두 병변 사이의 유의성을 평가할 것이다.

II. 연구 방법

1. 후두내시경 영상 및 병변의 관심영역 수집

기관생명윤리위원회(IRB)의 승인을 거쳐(GDIRB2021-165), PACS 데이터베이스로부터 정상(Normal), 악성(Malignant), 양성(Benign) 세 종류의 후두내시경 영상을 수집하였다. 수동으로 후두 병변을 감싸는 직사각형의 관심 영역(Region Of Interest, ROI)을 전문의의 도움으로 수집하였다. 병변 탐지에서는 학습(Train) 데이터 900장, 검증(Validation) 데이터 300장, 평가(Test) 데이터 300장의 데이터를 사용하였다. 후두 병변의 분류에서는 학습시킨 병변 탐지 모델에서 검출한 데이터를 사용하였다. 병변 탐지 모델은 악성과 양성으로 구성된 400장의 영상을 총 346장의 데이터를 검출하였다. 검출한 데이터를 약 3:1비율인 학습 데이터 259장 평가 데이터 87장으로 나누어 병변 분류 모델 학습에 사용하였다.

2. 후두내시경 영상 속 병변 탐지

후두 병변의 위치를 검출하기 위한 딥러닝 모델을 학습하기 위해 YOLOv5 아키텍처를 사용하였다. 후두 병변이 보이는 악성(Malignant), 양성(Benign)을 묶어 비정상(Abnormal) 클래스로 라벨링하고, 나머지 영상을 정상으로 라벨링 하였다. YOLOv5sl에서 YOLOv5xl까지 5개로 구분되는 모델의 사이즈 중 대형(YOLOv5l)을 사용하였다. 이미지 크기 640에 Epochs 200, Batch 16으로 학습을 진행하였다.

3. 라디오믹스 기반 특징 추출

후두 병변 분류를 위한 라디오믹스 특징 추출은 오픈 소스 파이썬 패키지인 파이라디오믹스(Pyradiomics, version 3.0.1,

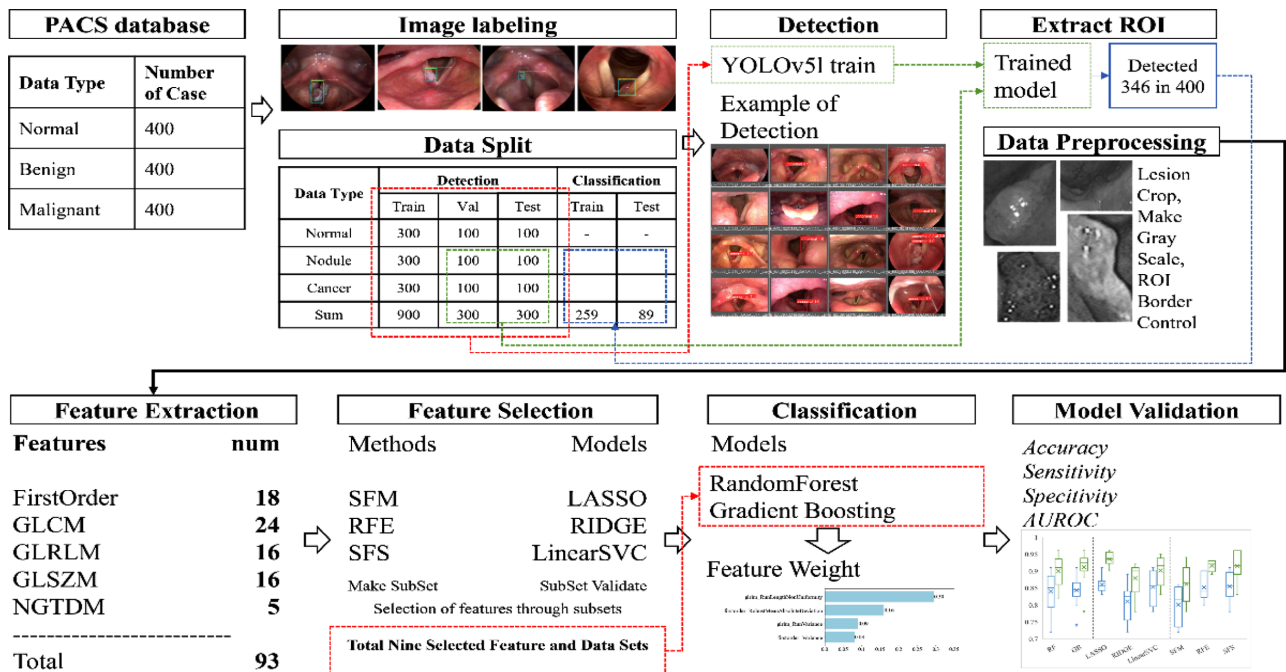


그림 1. 후두내시경 영상 속 라디오믹스를 통한 병변 분류 연구 흐름도

Fig. 1. Study of lesion classification by radiomics in laryngoscopy image flowchart

표 1. 데이터 세트의 구성, a: 병변 탐지 모델 학습에 사용된 데이터 세트 구성 b: 병변 분류 모델 학습에 사용된 데이터 세트 구성

Table 1. Data set composition, a: Composition of data set used to train detect lesions model, b: Composition of data set used to train lesions classification model

	Detection			Classification	
	Train	Validation	Test	Train	Test
Normal	300	100	100	(-)	(-)
Nodule	300	100	100		
Cancer	300	100	100		
Total	900	300	300	259	89

<https://github.com/AIM-Harvard/pyradiomics.git>[15]를 이용하였다. 해당 패키지의 경우 그레이 컬러의 이미지를 요구하므로, 먼저 병변 탐지 모델을 통해 얻은 ROI를 잘라내고, 이미지를 흑백 이미지로 만들어 사용했다. 추출된 라디오믹스 특징은 평균, 분산과 같은 1차 히스토그램 통계(First Order)[22] 18개와 텍스처 분석의 통계적 접근 방식인 GLCM(Gray Level Co-occurrence Matrix)[23] 24개, GLDM(Gray Level Dependence Matrix)[24] 14개, GLRLM(Gray Level Run Length Matrix)[25] 16개, GLSZM(Gray Level Size Zone Matrix)[26] 16개, NGTDM(Neighbouring Gray Tone Difference Matrix)[27] 5개로, 총 93개의 특징을 추출하였다.

4. 특징 선택

추출된 특징을 선택하고 분류하기 위해 파이썬 기반 오픈

소스 기계학습 라이브러리인 사이킷런(Scikit-learn, version 1.0.2, <https://scikit-learn.org/>)[28]을 사용하였다. 추출된 라디오믹스 특징은 그 수가 많으며, 이렇게 많은 양의 특징은 모델의 학습 시간을 늘리고 성능을 떨어뜨릴 수 있다[29]. 따라서 병변 분류에 유용한 특징을 선택해야 한다.

특징 선택에서는 Select From Model(SFM), Recursive Feature Elimination(RFE), Sequential Feature Selector(SFS) 세 가지 기법을 사용하였다. SFM은 모델이 평가한 특징의 가중치를 가지고, 가중치가 임계값(Threshold)을 넘지 않는 것을 제거하는 방법이다[30]. RFE는 점점 더 작은 서브 세트(SubSet)를 재귀적으로 고려하여 특징을 선택하는 방법으로 초기 특징 집합을 학습하고, 가중치가 가장 낮은 특징을 제거하는 방법이다[30]. SFS는 0개의 특징 집합에서 시작하여, 서브 세트의 교차 검증(Cross Validation) 점수를 최대화하는 기능을 찾고 추가하는 방법이다[30].

각 기법은 특징의 가중치를 제공하거나(SFM, RFE), 서브 세트의 교차 검증 점수를 제공할(SFS) 모델이 필요하다. 이를 위해 사용된 모델을 이하 보조 모델이라 정의한다. 보조 모델은 선형 회귀 모델인 LASSO[31]와 RIDGE[32], 사이킷런에서 선형 커널로 구현된 서포트 벡터 머신(Support Vector Machine, SVM)인 LinearSVC[33] 세 가지를 이용하였다.

5. 병변 분류 및 특징 가중치

랜덤 포레스트(Random Forest, RF), 그래디언트 부스팅(Gradient Boosting, GB) 두 가지의 분류기를 선택된 특징들을 통해 학습시키고, 병변을 분류하였다. 학습한 모델의 성능과 특징 가중치를 분석하여 어떤 라디오믹스 특징이 바이오마커로 사용 가능한지 도출할 것이다. 랜덤 포레스트는 무작위의 의사 결정 트리를 독립적으로 학습시킨 후, 평균화를 사용하여 예측함으로써 정확도를 개선하고 과적합을 제어하는 배깅(Bagging) 방식의 앙상블 학습 방법이다[34]. 그래디언트 부스팅은 순차적으로 의사 결정 트리가 이전 트리의 오차를 보정하는 방식으로 추가되어 손실 함수를 최적화하는 부스팅(Boosting) 방식의 앙상블 학습 방법이다[35].

6. 모델 평가 지표

모델을 평가하기 위한 지표로는 먼저 실제 값과 모델의 예측 값을 비교하여 얻은 참 긍정(true positive, TP), 참 부정(true negative, TN), 거짓 긍정(false positive, FP), 거짓 부정(false negative, FN)값을 이용해, 병변 검출 모델의 평가를 위해 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall)을 $f1$ 스코어($f1$ score)를, 병변 분류 모델의 평가를 위해 정확도(Accuracy)와 민감도(Sensitivity), 특이도(Specificity)를 구하였으며, 계산 방법은 식 (1), (2), (3), (4), (5), (6)과 같다. 병변의 양성, 악성의 분류 성능을 평가하기 위해 수신자 조작 특성(Receiver Operating Characteristics, ROC) 분석을 수행하였고, ROC 곡선의 아래 면적에 해당하는 AUROC(Area

Under ROC curve)를 구하였다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$f1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

III. 연구 결과

라디오믹스 기법을 통해 분석할 후두 병변을 찾아내기 위해 YOLOv5를 학습하였다. YOLOv5모델이 효과적으로 후두 병변을 검출하는지 알아보기 위해 민감도, 특이도 정확도로 분석한 결과이다. 후두 내시경 영상 내에 병변 유무에 따라, 비정상과 정상 두가지의 클래스로 객체를 지정하였고, 모델이 이를 탐지하는 결과를 분석하였다. 후두 병변이 있는 경우는 참으로, 후두 병변이 없는 경우는 거짓으로 정의하였다. 예측 영역과 정답 영역과의 IOU값이 0.5이상일 경우 예측에 성공한 것으로 판단하였다. YOLOv5는 비정상 200건 중 159건을 비정상으로 판정하였으며(TP), 41건을 정상으로 판정하였다(FN). 정상 100건 중 100건 모두 정상으로 판정하였으며(TN) 정상을 비정상으로 판정하는 경우는 없었다(FP). 이미지 검출의 예시를 그림 2에 나타냈다. YOLOv5는 0.86의 정확도로 이미지를 검출하였으며, 민감도는 0.80, 특이도는 1이었다. 지표를 표 2에 나타냈다.

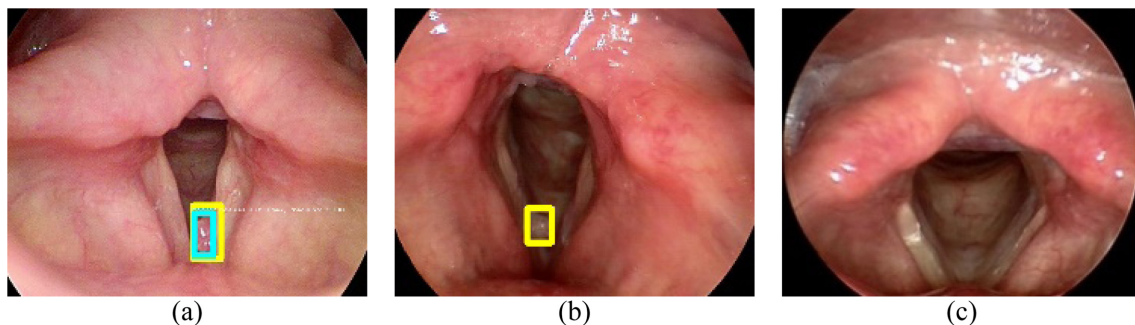


그림 2. YOLOv5 병변 검출 결과 예시 이미지, (a) TP, (b) FN, (c) TN, 노란색 경계 상자: 정답 ROI, 청록색 경계 상자: 학습시킨 YOLOv5가 찾은 ROI 경계

Fig. 2. YOLOv5 Examples of lesion detection results, (a) TP, (b) FN, (c) TN, Yellow bounding box: ground truth ROI, Cyan bounding box: ROI found by YOLOv5

표 2. YOLOv5의 병변 검출 정밀도, 재현율, F1-점수, 정확도 결과
Table 2. Precision, recall, F1-score, accuracy of lesion detection by YOLOv5

Class	Precision	Recall	F1 score	Accuracy
Abnormal	1	0.79	0.89	0.86

본 논문에서는 3가지의 특징 선택 기법과 3개의 보조 모델로 2개의 분류기가 학습할 특징을 선별하였다. 학습된 모델이 예측한 결과와 전문의의 진단 결과를 비교하여 후두 병변의 악성(Malignant) 및 양성(Benign)의 분류 성능을 평가하였다. 악성, 양성 두 가지 클래스의 이진 분류의 결과로 얻어진 정확도(Accuracy)와 AUROC(Area Under ROC)를 성능 평가의 지표로 사용하였다. 모든 지표는 이진 분류에서 양성을 참(True), 악성을 거짓(False)으로 정의하였다.

표 3에서 분류기 18개의 정확도와 AUROC를 결과를 제시하였다. SFS와 LASSO와, RF의 조합이 Accuracy 0.91, AUROC 0.96으로 가장 성능이 높았다. 반면 SFM과 RIDGE, GB의 조합은 Accuracy 0.74, AUROC 0.78로 성능이 낮았다.

특징 선택 기법, 보조 모델, 분류기가 최종 성능에 어떤 영향을 끼치는지 알기 위해, 각 요소를 포함하는 모델들의 정확도와 AUROC의 값들의 분포를 그림 3에 나타냈다. 특징 선택 기법에서 SFM를 포함하는 모델은 RFE와 SFS를 포함하는 모델보다 성능이 떨어지는 모습을 보인다. RFE와 SFS 간에는 큰 성능 차이를 보이지 않는다. 보조 모델에서

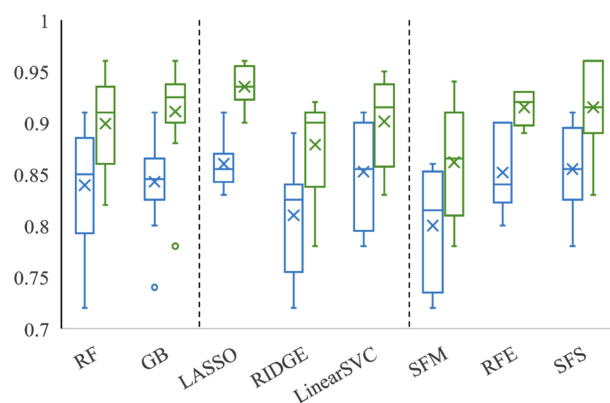


그림 3. 각 요소를 포함하는 모든 모델의 정확도, AUROC 점수의 분포를 나타냈다. 파란색: 정확도, 초록색: AUROC, x: 평균
Fig. 3. The accuracy and AUROC scores distribution of all models including each element are shown. Blue: Accuracy, Green: AUROC, x: mean

RIDGE는 LASSO와 LinearSVC보다 성능이 떨어지는 모습을 보인다. LASSO와 LinearSVC 간에는 큰 성능 차이를 보이지 않는다. 분류기에서 RF와 GB는 성능에 큰 영향을 끼치지 않았다.

가장 성능이 높았던 SFS와 LASSO, RF 조합이 효과적으로 후두 병변을 분류하는지 알아보기 위해, 모델의 예측 확률에 대해 t 검정을 진행하였다[36]. 양성의 예측 확률($M=0.83$, $SD=0.18$)과 악성의 예측 확률($M=0.25$, $SD=0.23$)은 $t(85)=$

표 3. 후두 병변의 분류에 유용한 라디오믹스 특징을 찾기 위해, 18가지의 방법으로 특징을 선택, 분류한 결과이다. 제시된 지표는 정확도와 AUROC이다. 각지표의 최댓값을 굵은 글꼴 처리하였다.

Table 3. The result of selecting and classifying features using 18 methods to find radiomics features useful for lesion classification. accuracy and AUROC presented. The maximum value was bold.

RF: Random Forest, GB: Gradient Boosting, SFM: Select From Model, RFE: Recursive Feature Elimination, SFS: SequentialFeatureSelector

		SFM		RFE		SFS	
		Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
RF	LASSO	0.86	0.90	0.83	0.92	0.91	0.96
	RIDGE	0.72	0.82	0.84	0.89	0.89	0.91
	LinearSVC	0.78	0.85	0.90	0.92	0.78	0.83
GB	LASSO	0.85	0.94	0.84	0.93	0.87	0.96
	RIDGE	0.74	0.78	0.80	0.90	0.84	0.92
	LinearSVC	0.85	0.88	0.90	0.93	0.84	0.91

표 4. SFS와 LASSO, RF 조합의 양성 및 악성의 예측 확률 t 검정 결과 N: 관측 수, M: 평균, SD: 표준편차

Table 4. t-test result of predict probability of benign and malignant in SFS, LASSO, and RF N: The sample size, M: mean, SD: Standard deviation

Malignant			Benign			t	P
N	M	SD	N	M	SD		
44	0.27	0.06	43	0.84	0.03	-12.80	0.00

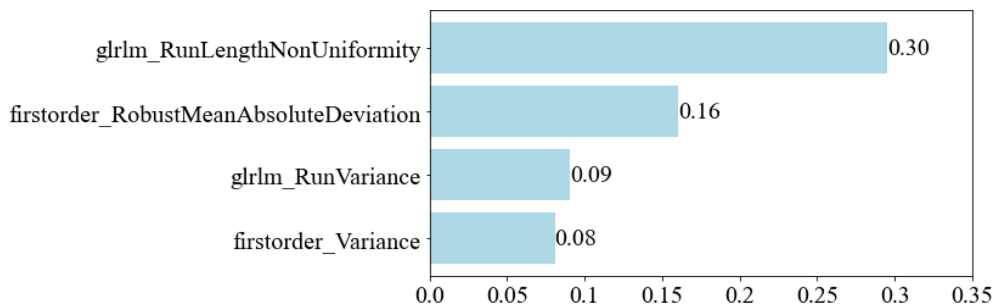


그림 4. 후두 병변 분류에 유의한 특징을 찾기 위해 SFS와 LASSO, RF 조합의 특징 가중치를 분석한 결과. 유의한 가중치를 가지는 상위 4개의 특징을 보여준다

Fig. 4. The result of feature weights in SFS, LASSO, and RF to find Significant features in the classification of laryngeal lesions. It shows the top 4 features having significant weight. GLRLM: Gray Level Run Length Metrix, Run Length Non-Uniformity - RLN, Robust Mean Absolute Deviation - rMAD

-12.89, $p < 0.05$ 로 유의한 차이를 나타냈다. 따라서 본 모델은 효과적으로 후두 병변을 분류하였다. 모델의 민감도(Sensitivity)는 0.93, 특이도(Specificity)는 0.86이다.

후두 병변 진단에 유의한 특징을 찾기 위해, 가장 성능이 높았던 SFS와 LASSO, RF 조합의 특징 가중치를 분석하여, 유의한 가중치를 가지는 상위 4개의 특징을 그림 4에 나타냈다. 가장 높은 가중치를 가지는 특징은 GLRLM의 Run Length Non-Uniformity이다. 같은 그레이 레벨 값을 가진 연속 픽셀의 수로 정의되는 그레이 레벨 런 령스의 유사성을 이미지 전체에서 측정한 값이며, 값이 낮을수록 균일함을 나타낸다[37]. 이어서 First Order의 Robust Mean Absolute Deviation, GLRLM의 Run Variance, First Order의 Variance가 유의한 가중치를 가지며, 각각 10번째 백분위 수와 90번째 백분위 수 사이 값들의 절대 편차, 그레이 레벨 런 령스의 분산, 이미지에 픽셀 강도의 분산을 나타낸다[37].

IV. 고찰 및 결론

라디오믹스 기법을 이용해 학습한 모델은 결과에 대한 설명과 분석을 할 수 있다는 장점이 있다. 이러한 점은 유의한 결과를 보이는 라디오믹스 특징을 병변 진단에 유용한 바이오마커로 사용할 수 있게 한다. 본 논문에서는 전문의의 경험에 기반한 주관적인 판단의 한계를 극복하기 위해 라디오믹스 기법을 통한 병변의 정량적 분석을 진행하였다. 분석을 진행할 병변을 찾기 위해 YOLOv5 모델을 학습시켜 후두 병변을 검출하게 하였다. YOLOv5는 정확도 0.86의 성능으로 후두내시경 영상에서 병변을 찾았다. 검출한 병변에서 라디오믹스 기법을 통해 정량적 특징이 추출되었다. 정량적인 특징을 분석할 18개의 모델을 제시하였고, SFS와 LASSO와, RF의 조합을 통해 정확도 0.91과 AUROC 0.96의 최대 분류 성능을 얻을 수 있었다. 각 모델의 결과와 모델들이 사용한

특징들을 분석하여, 후두 병변의 진단에 유용한 라디오믹스 특징을 알아보았다.

라디오믹스로 추출된 93개의 특징은 다양한 정보를 포함하고 있다. 그러나 어떤 특징이 후두 병변 진단에 유의한 정보를 포함하는지 알아보는 것은 어렵다. 본 논문에서는 후두 병변 진단에 유의한 정보를 포함하는 특징을 알아보기 위해 3가지 특징 선택 기법과 3가지의 보조 모델을 이용해 특징을 선별하고 이를 2가지의 분류기가 학습하였다 학습한 분류기의 성능 평가함으로써 선별된 특징들이 후두 병변 진단에 유용한지 알아보았다. 또한 분류기의 특징 가중치를 분석함으로써, 후두 병변 진단에 유의한 정보를 포함하는 특징을 알아보고자 하였다.

특징들의 값은 양성과 악성을 특정 임계값을 가지고 명확하게 나누지 않으며, 두개의 클래스를 분류하지 못하는 영역이 존재한다. 이는 해당 특징을 사용하여 병변을 분류할 때 성능의 저하를 야기한다. 만약 이를 보완하는 다른 특징이 상호적으로 사용된다면 모델의 분류 성능 향상을 기대할 수 있다. 따라서 특징 선택에 있어 서로 다른 분류 성능의 한계를 보완할 수 있는 특징을 선택하는 기법이 최종 성능 향상에 유리할 것으로 시사된다.

특징 선택 기법에서 SFM를 사용하는 모델들은 다른 기법들을 사용하는 모델들보다 성능이 떨어지는 모습을 보인다. SFM은 보조 모델을 통해 특징들의 가중치를 얻고, 가중치가 임계값 보다 낮은 특징을 한 번에 제거하므로 다른 기법에 비해 속도가 빠르다. 하지만 특징 간의 관계성을 전혀 고려하지 않으므로, 분류 성능의 한계를 보완하지 못하는 특징들이 선택될 수 있다. 반면 RFE와 SFS는 특징을 추가하며 정해진 개수만큼의 특징이 선택될 때까지 서브 세트를 반복적으로 평가해 특징들을 선택한다. 이 때 먼저 선택된 특징의 한계를 보완해 줄 수 있는 특징이 추가된 서브 세트가 그렇지 않은 서브 세트보다 성능이 높을 것이며, 각 단계에서 선택된 가

능성이 크다. 결과적으로 RFE와 SFS는 각 특징의 한계를 보완하는 특징들을 선택할 가능성을 크게 하며, 최종 성능을 상승시키는 요인으로 시사된다.

가장 성능이 높았던 조합은 SFS와 LASSO, RF를 조합한 모델이었으며, 이 경우 유의한 특징은 GLRLM의 Run Length Non-Uniformity(RLN)와 First Order의 Robust Mean Absolute Deviation(rMAD)임을 확인할 수 있었다. RLN은 GLRLM에서 그레이 런 렱스의 유사성을 측정하며, 낮은 값은 이미지의 런 렱스의 균일함을 시사한다. rMAD는 이미지 그레이 레벨에서 10번째 백분위 수와 90번째 백분위 수 사이 값들의 절대 편차이다. 악성 병변이 양성 병변보다 두 특징의 값이 큰 경향을 보인다. 두 특징의 결과는 악성 병변이 양성 병변보다 불균일함을 시사하며, 이는 종양 내 이질성 (Intra-tumor heterogeneity)에서 기인한 것으로 보인다 [38,39].

색상 정보는 MRI, CT와 같은 다른 의료 영상에서 얻을 수 없는 내시경 영상의 장점이다. 하지만 파이라디옉믹스를 이용하기 위해 이미지를 흑백 이미지로 변환하였고, 이로 인한 색상 정보의 손실이 발생하였다. 따라서 RGB의 컬러 영상을 각 채널로 나누어 특징을 추출하는 방식과 같은 영상의 색상 정보 또한 활용하는 보완이 필요하다. 또한 특징의 선택에 있어, 특징 간의 상관관계를 고려하는 기법은 사용되지 않았으며, 유사한 결과를 보이는 특징을 제외하는 선행 작업이 이루어지지 않았다. 이는 서로의 한계를 보완하는 특징들이 선택되기 어렵게 만들며, 앞선 SFM의 성능 저하의 가장 큰 요인으로 추정된다. 정확한 결과의 해석 및 효과적인 특징 선택을 위한 특징들의 전처리 과정이 추가적으로 필요하다. 본 논문에서 사용된 YOLOv5의 정확도 0.86(수치 변경)이라는 수치는 객체 검출에 있어 낮은 성능을 보여준다. 이것은 전체 연구에 신뢰도를 낮추는 결과를 가져온다. 이후 연구에서 다양한 객체 검출 모델, 혹은 병변 검출을 위한 추가적인 데이터 전처리에 관한 연구를 통해 성능을 높일 필요가 있다.

라디오믹스 기법을 통하여 내시경 영상 속 후두 병변의 정량적인 특징을 추출할 수 있었고, 이를 총 24가지의 방법으로 선별, 학습하였다. 학습시킨 모델의 평가를 통해 후두 병변의 진단을 보조하기 위한, 정량적인 특징을 분석하여 내시경 영상에서 사용할 수 있는 바이오마커로 제시하며, 이를 이용한 후두 병변의 분류가 효과적임을 보였다. 최근 인공지능을 통한 의료 영상 내의 병변 분류는 합성곱 신경망과 같은 딥러닝을 활용한 경우가 주류이다. 하지만 딥러닝을 활용하여 진단을 내리는 경우 진단 결과에 대한 근거를 제시하기 어렵다는 한계가 있다. 라디오믹스를 이용한 병변의 진단은, 특징 값에 기반하여 이루어 지므로, 모델 진단의 결과를 근거를 들어 설명할 수 있다는 장점이 있다. 본 연구에서는 후두내시경 영상에서 라디오믹스를 통한 병변 분류를 시도하였으며 그

유용성을 보였다. 하지만, 라디오 믹스 기법을 활용하기 위해 사용한 객체 검출 모델의 성능 향상을 위한 노력이 추가로 필요해 보이며, 이미지를 흑백으로 전환하는 과정에서의 정보 손실을 최소화하는 방법의 적용과 더 다양하고 효과적인 특징 선택 기법 및 분류 기법의 적용을 통한 추가적인 연구가 필요해 보인다.

Reference

- [1] Cohen SM, Dupont WD, Courey MS. Quality-of-life impact of non-neoplastic voice disorders: a meta-analysis. *Annals of Otolaryngology, Rhinology & Laryngology*. 2006;115(2):128-134.
- [2] Kim JO, Lim SE, Park SY, Choi SH, Choi JN, Choi HS. Validity and reliability of Korean-version of voice handicap index and voice-related quality of life. *Speech Sciences*. 2007;14(3):111-125.
- [3] Ni XG, He S, Xu ZG, Gao L, Lu N, Yuan Z, Lai S-Q, Zhang T-M, Yi J-L, Wang X-L. Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. *The Journal of Laryngology & Otology*. 2010;125(3):288-296.
- [4] Lee SH. Diagnostic Role of Stroboscopy. *Journal of the Korean Society of Laryngology. Phoniatrics and Logopedics*. 2010;21(1):13-16.
- [5] Chu EA, Kim YJ. "Laryngeal cancer: diagnosis and preoperative work-up". *Otolaryngologic Clinics of North America*. 2008;41(4):673-695.
- [6] Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, Ma L, Sun Y, Xu W, Yang N, Zou J, Zheng Y, Chen M, Gan W, Xiang T, An J, Liu R, Lv C, Lin K, Zheng X, Lou F, Rao Y, Yang Hui, Liu Kai, Liu G, Lu T, Zheng X, Zhao Y. Automatic recognition of laryngoscopic images using a deep-learning technique. *The Laryngoscope*. 2020;130(11):E686-E693.
- [7] Cho WK, Choi SH. Comparison of convolutional neural network models for determination of vocal fold normality in laryngoscopic images. *Journal of Voice*. 2020;36(5):590-598.
- [8] Xiong H, Lin P, Yu JG, Ye J, Xiao L, Tao Y, Jiang Z, Lin W, Liu M, Xu J, Hu W, Lu Y, Liu H, Li Y, Zheng Y, Yang H. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*. 2019;48:92-99.
- [9] Azam MA, Sampieri C, Loppi A, Africano S, Vallin A, Mocellin D, Fragale M, Guastini L, Moccia S, Piazza C, Mattos L, Peretti G. Deep Learning Applied to White Light and Narrow Band Imaging Videolaryngoscopy: Toward Real-Time Laryngeal Cancer Detection. *The Laryngoscope*. 2021;132(9):1798-1806.
- [10] Ye G, Du C, Lin T, Yan Y, Jiang J. Deep Learning for Laryngopharyngeal Reflux Diagnosis. *Applied Sciences*. 2021;11(11):4753.
- [11] Onji K, Yoshida S, Tanaka S, Kawase R, Takemura Y, Oka S, Tamaki T, Raytchev B, Kaneda K, Yoshihara M, Chayama K. Quantitative analysis of colorectal lesions observed on magnified endoscopy images. *J Gastroenterol*. 2011;46:1382-1390.
- [12] Sainju S, Bui FM, Wahid K. Bleeding detection in wireless capsule endoscopy based on color features from histogram probability. 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) IEEE. 2013;1-4

- [13] Vu DH, Nguyen LT, Nguyen VT, Tran TH, Dao VH, Vu H. Boundary delineation of reflux esophagitis lesions from endoscopic images using color and texture. 2021 International Conference on Multimedia Analysis and Pattern. 2021;1-6.
- [14] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Foster K, Aerts HJWL, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*. 2012;30(9):1234-1248.
- [15] Griethuysen JJMV, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research*. 2017;77(21):e104-e107.
- [16] Machicado JD, Koay EJ, Krishna SG. Radiomics for the Diagnosis and Differentiation of Pancreatic Cystic Lesions. *Diagnostics*. 2020;10(7):505.
- [17] Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Physics in Medicine & Biology*. 2016;61(13):R150.
- [18] Liang C, Huang Y, He L, Chen X, Ma Z, Dong D, Tian J, Liang C, Liu Z. The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget*. 2016;7(21):31401-31412.
- [19] Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based prognosis analysis for non-small cell lung cancer. *Scientific reports*. 2017;7(1):1-8.
- [20] Coroller TP, Grossmann P, Hou Y, Velazquez ER, Leijenaar RTH, Hermann G, Lambin P, Haibe-Kains B, Mak RH, Aerts HJWL. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*. 2015;114(3):345-350.
- [21] Kniep HC, Madesta F, Schneider T, Hanning U, Schonfeld MH, Schon G, Fiehler J, Gauer T, Werner R, Gellissen S. Radiomics of brain MRI: utility in prediction of metastatic tumor type. *Radiology*. 2019;290(2):479-487.
- [22] Kandemirli SG, Chopra S, Pyiya S, Ward C, Locke T, Soni N, Srivastava S, Jones K, Bathla G. Presurgical detection of brain invasion status in meningiomas based on first-order histogram-based texture analysis of contrast enhanced imaging. *Clinical neurology and neurosurgery*. 2020;198:106205.
- [23] Mohanaiah P, Sathyanarayana P, GuruKumar I. Image texture feature extraction using GLCM approach. *International journal of scientific and research publications*. 2013;3(5):1-5.
- [24] Ahmadi N, Akbarizadeh G. Iris tissue recognition based on GLDM feature extraction and hybrid MLPNN-ICA classifier. *Neural Computing and Applications*. 2020;32(7):2267-2281.
- [25] Sohail ASM, Bhattacharya P, Mudur SP, Krishnamurthy S. Local relative GLRLM-based texture feature extraction for classifying ultrasound medical images. 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE). 2011;001092-001095.
- [26] Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification. 2011 18th IEEE International Conference on Image Processing. 2011;53-56.
- [27] Amadasun M, King R. Textural features corresponding to textural properties. in *IEEE Transactions on Systems, Man, and Cybernetics*. 1989;19(5):1264-1274.
- [28] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, Vanderplas J, Joly A, Varoquaux G. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv*. 2013;1309.0238.
- [29] Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 40.1 (2014): 16-28.
- [30] https://scikit-learn.org/stable/modules/feature_selection.html, Accessed on 29 Jun 2022.
- [31] https://scikit-learn.org/stable/modules/linear_model.html. Accessed on 1 Aug 2022.
- [32] McDonald GC. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2009;1(1):93-100.
- [33] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC>. Accessed on 30 Jun 2022.
- [34] Biau G, Scornet E. A random forest guided tour. *TEST*. 2016; 25:197-227.
- [35] Natekin A, and Knoll A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*. 2013;7:21.
- [36] Antoine AA. The student's t-test: a brief description. *Research & Reviews: Journal of Hospital and Clinical Pharmacy*. 2019;5(1):1.
- [37] <https://pyradiomics.readthedocs.io/en/latest/features.html>. Accessed on 4 Jul 2022.
- [38] Mazon T, et al. Intratumoral heterogeneity of the epigenome. *Cancer cell*. 2016;29(4):440-451.
- [39] Aerts HJWL, Pankov A, Song JS, Costello JF. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*. 2014;5(1): 1-9.