



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



비소세포 폐암에서 기계학습을 이용한
비침습적 PD-L1 발현 수치 분류기 개발
및 성능평가

연세대학교 대학원

융합의학과

장병한

비소세포 폐암에서 기계학습을 이용
한 비침습적 PD-L1 발현 수치 분류기
개발 및 성능평가

지도교수 김 성 준, 김 휘 영

이 논문을 석사 학위논문으로 제출함

2022년 12월 23 일

연세대학교 대학원

융합의학과

장병한

장병한의 석사 학위논문을 인준함

심사위원 김 성 준 인

심사위원 김 휘 영 인

심사위원 서 영 주 인

연세대학교 대학원

2022년 12월 23 일

감사의 글

우선, 지도교수님 들인 김성준 교수님과 김휘영 교수님께 감사드립니다. 덕분에 대학원 생활을 마무리할 수 있었습니다. 또한, 교수님들 외에 대학원 생활에 함께한 동료들과 친구들에게 감사의 인사를 하자면,

우선, 상욱이랑 민규야 정말 고마워. 처음에 입학했을 때 아무것도 몰랐던 내가 그래도 방장 일이든 프로그래밍이든 할 수 있었던 것은 모두 상욱이 덕이다. 그리고 항상 모르는 것 물어보면 성의껏 설명해주고 새벽까지 같이 논문 찾아봐 준 민규야 정말 고맙다. 너희 둘이 거의 과외선생님처럼 많이 알려줘서 나도 많이 배우고 성장할 수 있었어. 고맙다.

그리고 TaiLab 후배들에게도 고맙다. 항상 잘 따라주고 응원해준 점 모두에게 고맙고 미안한 마음이 있어. 이제는 고마움만 남기도록 할 게. CCIDS에 같이 생활했던 진희, 기성이, 현재야. 너희 모습에서 동기부여도 많이 받았던 것 같다. 거기 같이 있을 때가 제일 재미있었어.

대학원 시절 응원해 준 의공 친구들 및 대학원 같이 다니며 여러 도움을 의공 후배들도 고마워. 대학원 다니다가 가끔 만나서 이야기할 때면 동기부여가 많이 되었다. 특히 백규형, 영명이 내가 힘들다고 할 때 항상 이야기 들어주고 밥 사줘서 고맙다. 이제 내가 사주도록 할 게. 윤호랑 도현이는 학교 다닐 때 생각 많이 나더라. 학부 때 너희랑 수업 들을 때가 좋아서 대학원을 가볼까 생각했지. 결국 여기까지 왔네. 그 밖에 모두들 고마워, 다 적지 못해 미안해.

은식, 현유, 영덕이, 동혁이 힘들다고 할 때 이야기 들어줘서 너무 고맙다. 특히 이번 학기 내 멘탈에 많이 도움됐다. 마지막으로 승학이 윤서, 아직 학생이라고 배려해주고, 힘들 때 이야기 들어주고 응원해줘서 고맙다. 어릴 때부터 너희가 있어서 든든했어.

대학원 생활 뿐 아니라, 살면서 많은 사람들의 도움을 받아 여기까지 올 수 있었던 것 같습니다. 다들 감사합니다.

차례

그림 차례	iii
표 차례	iv
국문 요약	v
제 1 장. 서론	1
1.1 연구 배경	1
1.1.1 폐암	1
1.1.2 Programmed cell Death Ligand – 1	1
1.2 선행 연구	2
1.3 연구 목적	4
제 2 장. 연구 방법	6
2.1 데이터 세트	6
2.2 분류 모델	7
2.2.1 3D image analysis	9
2.2.1.1 Handcrafted radiomics feature	9
2.2.1.2 Radiomics 모델	12
2.2.2 2D Patch based analysis	14
2.2.2.1 Label smoothing	15
2.2.2.2 Attention based deep multiple instance learning	16
2.3 평가 방법	19



제 3 장. 결과.....	22
3.1 예측 성능	22
3.2 Feature importance	24
3.3 Patch attention map.....	25
제 4 장. 고찰.....	28
제 5 장. 결론.....	34
참고문헌.....	36
Abstract	42

그림 차례

그림 1 PD-L1 분류기 ROC 커브	22
그림 2 Machine learning PD-L1 분류기의 feature importance	24
그림 3 PD-L1 수치 음성인 병변에 대한 patch attention map	25
그림 4 PD-L1 1%인 병변에 대한 patch attention map	26
그림 5 PD-L1 90%인 병변에 대한 patch attention map	27



표 차례

표 1 비침습적 PD-L1 수치 예측 선행 연구	3
표 2 데이터 세트.....	7
표 3 PD-L1 분류기 예측 성능	23
표 4 선행연구 및 본 연구 데이터 세트 비교	33

국문요약

비소세포 폐암에서 기계학습을 이용한 비침습적 PD-L1 발현 수치 분류기 개발 및 성능평가

Programmed cell Death Ligand 1(PD-L1)은 막 횡단 단백질로서 T세포의 Programmed cell death 1(PD-1)과 결합하여 해당 세포를 정상 세포로 판단하는 기능을 한다. 하지만 일부 암세포들은 이러한 PD-L1을 발현함으로써 T세포가 그 암세포를 정상세포로 판단하도록 한다. 이러한 PD-L1의 기전을 역이용한 면역 관문 억제제의 사용으로 인하여 최근 바이오 마커로서 PD-L1을 이용한 연구가 활발히 진행되고 있으며 항암 치료에서 각광받고 있다. PD-L1 수치를 판단하기 위해서 조직 검사를 이용하는데, 조직 검사의 경우 침습적인 방법을 이용하는 검사이기 때문에 모든 환자에게 적용할 수 없다. 이러한 점 때문에 비침습적인 방법을 이용하여 PD-L1 수치를 판단할 수 있는 방법은 중요하다.

본 연구에서는 CT 영상을 이용하여 병변의 PD-L1 수치를 분류하는 모델을 제안한다. PD-L1 양성인지 음성인지를 분류하는 이진 분류 모델과 PD-L1 수치 50%를 이상인지 50%미만인지를 분류하는 이진 분류 모델을 만들었다. 전체 폐암 환자 중 PD-L1 수치 50%이상의 환자들은 병변의 초기일수록 적기 때문에 불균형한 데이터 세트를 이용하고도 분류 성능이 높은 모델을 만드는 것을 목적으로 patch based analysis를 이용하여 multiple instance learning(MIL) 기법을 사용하였고, MIL 모델의 분류 성능을 보완하기 위하여 3D image analysis 방법으로서 handcrafted radiomics feature를 이용한 전통적인 기계학습(machine learning) 분류기를 radiomics 모델로 함께 사용하여 최종적으로 PD-L1 양성 분류기에서 AUC 0.81, PD-L1 수치 50% 분류기에서 AUC 0.93을 얻었다.

본 연구에서는 각각의 분류기가 어떻게 병변을 분류하게 되었는지에 대



한 해석을 위하여 patch attention map과 feature importance를 이용하였다. 각각의 MIL 모델들은 PD-L1 수치에 따른 attention map의 차이를 보였으며, radiomics 모델도 MIL 모델의 성능을 보완할 수 있는 특징들을 중요도 높게 학습했다는 것을 feature importance를 통해 확인할 수 있었다.

핵심 되는 말 : PD-L1, 비침습. MIL, handcrafted radiomics feature

제 1 장. 서론

1.1 연구 배경

1.1.1 폐암

폐암은 대한민국에서 암으로 인한 사망률이 가장 높은 질병으로, 2000년대 이후 인구 10만명 당 폐암으로 인한 사망자가 계속 증가하는 추세이다[1]. 폐암은 조직학적인 하위분류에 따라 비소세포폐암과 소세포폐암으로 나뉘게 되는데, 이 중 85%가 비소세포 폐암이며 나머지 15% 정도가 소세포폐암이다.[2] 소세포폐암은 암이 성장하는 시간이 비소세포 폐암에 비해 빠르고, 전이가 빨리 진행됨으로 빠른 화학요법이나 방사선 치료가 요구된다[3]. 비소세포 폐암은 소세포폐암에 비해 성장 속도와 전이가 진행되는 속도가 느리기 때문에 초기에는 수술을 통해 완치될 수 있다. 비소세포폐암 치료에서 수술이 최적의 치료법으로 여겨지지만, 25~30% 환자들만 수술에 적합하다. 또한, 최적의 수술적 관리에도 불구하고 5년 생존율은 TNM 분류체계가 IA 인 환자들에게서 73%, IIIA 인 환자들에게서 25%에 불과하다[2]. 현재 여러 임상 실험에서 다양한 adjuvant therapy나 neo adjuvant therapy를 추가하여 이러한 결과를 개선하려고 시도중이다 [4].

1.1.2 Programmed cell Death Ligand – 1(PD-L1)

비소세포 폐암 치료를 위하여 암세포 표면에 PD-L1 표현 수치를 바이오 마커로 사용하는 면역 관문 억제제가 승인되었다. PD-L1은 원래 정상 세포에 존재하는 막 횡단 단백질로서 T 세포의 Programmed cell death protein 1(PD-1)과 결합하여 해당 세포를 정상 세포로 판별하는 기능을 한다. 하지만 일부 암세포들은 PD-L1을 발현하여 T세포의 PD-1과 결합을 함으로 T세포가 그 암세포를 암세포로 구분하지 못하도록 한다. 면역

관문 억제제는 이런 PD-1/PD-L1 결합을 억제하는 기능을 한다. 면역 관문 억제제가 T 세포의 PD-1과 먼저 결합을 함으로써 T 세포가 PD-L1을 가진 암세포를 암세포로 인식할 수 있게 한다[5]. 이러한 면역 관문 억제에 기반한 치료는 기존의 세포독성 화학요법에 비해 높은 반응률 및 생존율과 내약성을 보였다[4]. National comprehensive cancer network(NCCN)에서 발표한 가이드라인에 따르면, PD-L1 수치가 특히 50% 이상인 경우가 치료 결정에 중요한 지표로 사용되며, 경우에 따라 1차 치료 약제로 쓰이도록 권장하고 있고[6,7] 면역요법과 화학용법을 결합한 치료의 경우 모든 PD-L1 수치에서 향상된 생존율을 보였다.[8]

전이성 폐암에서의 PD-1 및 PD-L1 바이오 마커의 성공은 종양 재발 방지 및 치료율 개선이 주된 목표인 초기 단계의 비소세포폐암에의 사용에 대한 관심 또한 증가되었다. PD-L1 수치가 수술이 가능한 초기 단계의 폐암 환자들에 있어서도 예후의 차이가 있다고 보고되었고[9], 한 시범 연구에서는 20명의 환자들을 대상으로 PD-1 억제제인 Nivolumab을 수술 전 2회 투여한 후 예후를 관찰하였는데, 수술 후 추적 관찰 12개월에서 수술 절제 후 80%의 환자가 살아 있었고 재발이 없었다. 18개월에서는 73%의 환자가 재발 없이 살아 있었다[10]. 초기 단계의(I-III) 비소세포폐암에서 최초 치료를 받았던 환자들이 60%까지 재발을 겪기 때문에 임상적인 필요성이 있다고 한다[4].

이러한 PD-L1 수치를 판별하기 위한 방법은 조직 검사이다. 병리과에서 암세포의 조직을 추출하여 면역화학염색을 한 후 현미경으로 확인하여 PD-L1 발현이 있는지, 얼마나 있는지를 확인한다[11]. 하지만 조직검사는 침습적 검사 방법이라는 단점이 있다. 다시 말해서, 암세포의 위치나 형태의 따라 검사 자체가 어려울 수도 있다.

1.2 선행 연구

현재 PD-L1 수치를 판별할 때는 조직 검사를 하게 된다. 하지만 조직

검사는 몸에 바늘을 넣어 병변을 추출하는 과정이 필수적이다. 암 세포가 존재하는 위치에 따라 이러한 과정은 어려울 수 있다. 또한, 조직 검사를 하기 이전에 CT 등의 방사선 의학 영상을 통한 진단은 선행된다. 따라서 방사선 의학 영상을 통해서 조직 검사를 진행해야만 알 수 있는 PD-L1 수치를 알 수 있게 된다면 시간적, 비용적으로 효율적이다. Chengdi Wang 외 연구진[12]은 1135 명의 TNM 병기가 골고루 분포된 비소세포 폐암 환자들을 가지고 PD-L1 음성인 그룹, PD-L1 수치 1~49%인 그룹, 50% 이상인 그룹에 대하여 CT radiomics feature와 Deep Learning feature 와 Clinical feature 들을 가지고 다중 클래스 분류를 수행하였고, 각각의 그룹에 대하여 AUC 0.95, 0.934, 0.946이라는 수치를 얻었다. Panwen Tian 외 연구진[13]은 939 명의 폐암 환자들을 대상으로 CT radiomics feature와 Deep Learning feature와 Clinical feature를 이용하여 PD-L1 50% 이상 그룹에 대하여 이진 분류를 하였고 AUC 0.76을 얻었다. Ying Zhu 외 연구진[14]은 120명의 진행성 비소세포폐암 환자들로 PD-L1 양성그룹과 50%그룹에 대하여 각각 이진분류를 하였고 각각 AUC 0.78과 0.77을 얻었다. Qiang Wen 외 연구진[15]은 120명의 진행성 비소세포폐암 환자들로 PD-L1 50% 그룹에 대하여 이진 분류를 하였고 AUC 0.79를 얻었다. Zekun Jiang 외 연구진[16]은 수술한 후의 비소세포 폐암 환자들로 PD-L1 양성 그룹에 대하여 AUC 0.85를 얻었다. Stefano Bracci 외 연구진[17]은 72명의 진행성 비소세포폐암 환자들로 PD-L1 50% 이상 그룹에서 AUC 0.79를 얻었다.

Author	Cut-off	Population
Chengdi wang(2022)[12]	1%, 50%	1,135 NSCLC
Panwen Tian(2021)[13]	50%	939 NSCLC
Ying Zhu(2020)[14]	1%, 50%	127 Metastatic NSCLC
Qiang Wen(2021)[15]	1%	120 Advanced NSCLC
Zekun Jiang(2021)[16]	1%	125 Surgically resectable NSCLC
Stefano Bracci[17]	50%	72 NSCLC

표 1 비침습적 PD-L1 수치 예측 선행 연구

각각의 선행연구들은 연구에 사용된 데이터들의 병기의 진행상황, 분류의 기준이 되는 cut-off, 각각의 병기의 진행 상황에 따른 환자들의 분포가 모두 다르다.

1.3 연구 목적

기준의 비침습적 PD-L1 수치 예측 연구에는 두 가지 문제점이 있다. 첫 째로 데이터 량이 한정되어 있다는 점이 있다. 이는 모든 기계학습 분야에서 공통적으로 가지고 있는 문제점이긴 하지만 의학 분야에 적용할 때에는 그 한계점이 좀 더 크게 적용한다. 그 이유는 의료 데이터는 중요한 개인 정보이고 병원들 간의 협업을 통하여 큰 데이터 세트를 구축하기에는 제약이 있다. 따라서 데이터 량이 적은 상황에서 좋은 성능을 내는 것이 중요한 점이라고 할 수 있다.

두 번째는 PD-L1 수치 데이터 세트 간의 불균형성을 가진다는 점이다. 가령, treatment decision에서 중요하게 취급되는 50% 이상의 PD-L1 수치를 가진 환자의 경우, 이러한 수치를 가지는 환자들의 수는 전체 비소세포폐암 환자들의 50%에 한참 모자란다. 이는 모든 비소세포폐암 병변의 세포막에서 PD-L1 이 발현되지 않기 때문인데, 이런 데이터 불균형 문제가 심하면 심할수록 데이터가 많은 방향으로 예측할 가능성이 커진다. 예를 들어, PD-L1 미 발현인 환자들의 데이터가 99%이고 PD-L1 수치가 50% 이상인 환자가 전체의 1%인 데이터 불균형 문제가 심해진다면 목적함수가 어떤 데이터들에 대해서도 PD-L1 미발현으로 분류하게 되는, $y=0$ 인 함수로 정의된다고 하더라도 전체 학습 데이터에 대한 정확도는 99%가 된다. 물론 이는 아주 극단적인 예시이지만 입력 데이터의 편향이 심해질수록 예측 모델의 성능도 편향을 가질 수 있게 된다. 특히, 초기 병변에서는 PD-L1 수치가 50% 이상인 병변의 7.4% 정도[18]로 매우 낮다. 초기 상태의 병변의 경우에도 보조적



항암화학요법이 의미가 있다는 점과 폐암 초기 병변이라도 CT 는 찍는다는 것을 고려하면 수술 가능한 초기 병변의 환자들에게 있어서 데이터세트의 불균형성을 고려할 수 있는 비침습적 PD-L1 수치 예측은 유의미하다고 할 수 있다.

본 연구에서는 이러한 문제점들에 대해 많지 않은 데이터와 불균형한 데이터세트를 이용하고도 분류 성능을 높이고 해석 가능한 예측 모델을 만드는 것이 목적이다.

제 2 장. 연구 방법

2.1 데이터 세트

본 연구에서는 2019년부터 2020년 신촌 세브란스 병원에서 절제 수술을 받은 NSCLC 환자들 238명이 촬영한 CT 영상을 사용하였다. 238명 중 1명의 환자는 병변이 3개 있었고 8명의 환자는 병변이 2개씩 있었다. 따라서 총 248개의 개별적인 병변에 대하여 연구가 진행되었다. 병변을 표시한 mask는 연구자 본인이 직접 라벨링 하였으며 후에 수술을 집도하신 교수님께 해당 부위를 확인받았다. 훈련 데이터 세트에는 PD-L1 양성인 병변 67개, 음성인 병변 142개로 총 209개의 병변이 사용되었고, 테스트 데이터 세트에는 PD-L1 양성인 병변 19개 음성인 병변 20개로 총 39개의 병변을 사용하였다. 각 데이터 세트에서 PD-L1 수치가 50% 이상인 병변은 훈련 데이터 세트에는 23개, 테스트 데이터 세트에는 8개를 사용하였다. PD-L1 양성 분류기와 PD-L1 수치 50% 이상의 분류기의 성능 평가 지표를 계산할 때는 각각 PD-L1 양성과 PD-L1 수치 50% 이상을 참인 클래스로 정의하였다.

데이터 세트를 구성하고 있는 데이터들의 분포는 표 2에 정리되었다. 적은 수의 데이터를 이용하여 많은 데이터를 예측하기 위하여 PD-L1 수치 50% 이상의 병변을 테스트 세트에 많이 포함하였다.

본 연구에서 측정된 병변들의 부피는 volume rendering을 하여 계산되었다. 병변이 라벨링된 mask 영역 내의 픽셀과 CT 영상의 pixel spacing 및 slice thickness attribute를 각각 곱하여 각각의 CT마다 1 pixel이 가지는 병변의 부피를 바탕으로 전체 병변 영역이 계산되었다. 본 데이터 세트에서 제시된 볼륨은 간유리 음영을 포함한 mask 전체 영역이다.

		Train	Test	Total
PD-L1	Negative	142(67.6%)	20(51.2%)	162(65.1%)
	0~49	44(21.1%)	11(28.2%)	55(22.1%)
	50 or higher	23(11.0%)	8(20.5%)	31(12.4%)
T – stage	total	209	39	248
	is	2(0.9%)	–	2(0.8%)
(Pathological)	1	145(68.9%)	32(82.1%)	177(71.3%)
	2	40(19.0%)	5(12.8%)	45(18.1%)
	3	15(7.1%)	2(5.1%)	17(6.8%)
	4	7(3.3%)	–	7(2.8%)
	N – stage	0	184(88.0%)	35(89.7%)
(Pathological)	1	8(3.8%)	1(2.6%)	9(3.6%)
	2	17(8.1%)	3(7.7%)	20(8.0%)
	3	–	–	–
Average Size(cm^3)		21.58	5.31	19.04
Standard deviation of Size(cm^3)		57.28	6.84	53.03
95% range of size(cm^3)		0.40~159.06	0.83~95.27	0.40~172.10

표 2. 데이터 세트

2.2 분류 모델

본 연구에서 병변이 라벨링된 mask는 병변이 존재하는 부위는 1, 존재하지 않는 부분은 0으로 코딩되었다. 따라서 원본 CT 영상과 mask를 곱하면 병변 영역만 남게 되는데, 이를 바탕으로 2가지 모델을 만들었다. 첫 째로 3D 영상에서 그대로 feature를 추출한 3D image analysis와 2D 영상의 patch들로 분할한 2D patch based analysis이다. 3D image analysis에서는 handcrafted radiomics feature와 전통적인 기계학습 모델을, 2D patch based analysis에서는 딥 러닝 기반의 multiple instance learning(MIL)이 사용되었다.

3D image analysis에서는 각각의 병변에서의 크기 차이에서 오는 편차로 인한 학습 오류를 줄이기 위하여 handcrafted radiomics feature 기반의 전통적인 기계학습 모델을 사용하였다. 간유리 음영을 포함한 mask를 학습하기 때문에, 병변마다 크기의 차이가 많이 나는데, 이 경우 원본 이미지를 사용하게 되면 작은 병변에서는 intensity의 값이 0인 부분이 많아지게 되므로 정상적인 학습이 어렵다. 3D interpolation을 이용하여 해상도를 늘리는 경우에는 원본 정보의 손실이 크게 일어나게 되므로, 수학적으로 잘 정의된 handcrafted radiomics feature를 이용하게 되었다.

데이터 불균형 문제를 해결하기 위한 핵심적인 방법으로 multiple instance learning 모델과 2D patch based analysis를 사용하였다. MIL 모델은 bag 내에 하나의 instance 만 참인 클래스에 속해도 bag 전체를 참으로 분류하게끔 설계되어 있다. 따라서, 훈련 데이터 세트의 patch들은 꼭 하나의 병변을 하나의 bag으로 정의하지 않아도 학습이 가능하다. 본 연구에서는 전체 209개의 훈련 데이터 세트의 병변들의 32x32 픽셀 크기의 patch들로 bag을 구성하고, 그 중 절반의 bag에는 참일 확률이 있는 instance를 랜덤하게 포함시키고 나머지 절반의 bag은 PD-L1 음성에서 나온 patch들만으로 구성하는 방법을 통해 데이터 불균형 문제를 해결하고자 한다.

3D image 기반 모델과 2D patch 기반 모델의 성능 평가 및 결과를 해석

해보고 최종적으로는 두 모델의 softmax 출력을 평균을 내어 3D 및 2D 기반 모델을 만들었다.

2.2.1 3D image analysis

2.2.1.1 Handcrafted radiomics feature

Radiomics는 이미지를 정량화된 수치들로 나타내는 과정을 의미한다[19]. 이 용어에 대하여 엄격하게 어떻게 수치화를 해야 한다는 명확한 정의는 없지만[20], 크게 Filtered, Intensity, shape, texture feature의 4 가지 feature 그룹들로 나눈다[21]. 이 중, Filtered feature란 원본 이미지에 특정한 filtering 한 후에 나온 이미지를 수치화한 feature들을 의미한다. 어떤 이미지에 filtering을 거치지 않은 상태에서 나온 feature들을 Original feature들이라고 정의하고, Original 대신에 filter의 이름을 기재하면서 원본 이미지에 대한 feature와 구분한다. Intensity feature는 가장 쉽게 정의할 수 있는 feature 그룹으로서, 이미지의 관심 영역 내의 이미지 Intensity들을 하나의 값으로 수치화한 feature 들을 의미한다. 관심 영역 내의 최솟값, 최댓값, 평균, 분산부터, 편포도, 첨도 등으로 관심 영역 내의 Intensity 분포에 대한 정보를 나타낼 수 있는 어떤 정량화 방법들은 Intensity feature라고 정의할 수 있다. Shape feature란 관심 영역 내에 크기를 수치화한 feature 들을 의미한다. 면적이나 부피, 길이, 곡률, 곡률의 왜곡된 정도 등을 ROI 내에서 나타낸 수치들을 의미한다. 이러한 Shape feature의 경우 영역 내의 Intensity와는 관계가 없기 때문에 관심영역을 표시한 마스크로만 수치화 할 수 있으며 filtering해서 관심 영역이 변화하지 않으므로 역시 수치가 변하지 않는다. 마지막으로 texture feature는 ROI 내의 Intensity 들이 변화하는 정도를 수치화한 feature들을 의미한다. Greyscale 이미지에서 이를 수치화 할 때 대표적으로 사용되는 방법으로 원본 이미지를 Grey-Level Cooccurrence Matrix(GLCM), Grey-Level Run-length Matrix(GLRLM), Grey-Level Size Zone Matrix(GLSZM), Neighborhood Grey-Tone Difference Matrix(NGTDM), Grey-Level

Dependence Matrix(GLDM) 등의 행렬로 변환하는 방법이다. 각각의 행렬들은 각각의 정의에 따라 이미지의 Intensity의 수치가 변화하는 정도가 공간적인 정보와 상관없이 어떻게 분포되어 있는지를 나타낸다 [22,23,24,25,26].

GLCM 행렬은 정해진 방향과 크기에 위치한 Intensity의 순서쌍(i,j)가 GLCM 위의 점(i,j)에 몇 개가 분포되었는지를 나타내는 행렬이다[22]. 2D image에서는 3개의 방향(수평선, 수직선, 대각선)이 정의될 수 있고 3D에서는 13개(인접한 26개의 점으로의 벡터 중 한 직선 위에 있는 두 방향을 제외한 13가지 방향)의 방향이 정의될 수 있으며 일반적으로는 거리가 인접한 영역에서 GLCM 행렬을 정의하게 된다. 이러한 GLCM 행렬은 이미지의 불균일성과 랜덤성을 나타내는 행렬로 정의된다.

GLRLM 행렬은 특정 방향에 대하여 어떤 Intensity i의 길이가 j만큼 동시에 있는 부분이 이미지 내에 몇 개가 있는지를 나타내는 행렬이다 [23]. GLRLM의 방향은 수학적으로는 임의의 모든 각도에 대하여 정의할 수 있으나, 디지털 이미지에서 쉽게 얻을 수 있는 행렬은 GLCM과 마찬가지로 2차원에서 3가지 방향, 3차원에서는 13가지 방향이다.

GLSZM 행렬은 특정 Intensity i가 인접한 영역에서 몇 개 존재하는지를 나타내는 행렬이[24]. 이 때, 인접한 대각선의 길이는 1로 간주한다. 즉, 2차원 이미지의 경우 9개의 열을 가지고 있고 3차원 이미지의 경우 27개의 열을 가지고 있다. GLSZM 행렬의 (i, j) 성분은 intensity i가 j개 존재하는 영역의 개수로 정의되며 큰 영역의 부분집합은 행렬에 성분으로 세어지지 않는다. 즉, 어떤 2차원 이미지가 임의의 intensity i에 대하여 인접한 9개의 픽셀에서만 intensity i를 가지고, 나머지는 intensity가 0이라고 한다면, GLSZM 행렬에서의 (i,9)의 값은 1이다. 하지만 그 영역의 부분집합인 j=4인 4개의 영역들은 (i, 4)에 포함되지 않는다. 즉, 그 2차원 이미지에서의 GLSZM matrix는 (i,9)에서만 1의 수치를 가지고 나머지는 0을 갖

는 행렬이 된다. GLSZM의 특징은 바로 인접한 영역들의 불균형성만 정의 된다는 점과 방향에 따라 다르게 정의되는 행렬이 아니라는 점이다.

NGTDM 행렬은 관심 영역 내의 intensity를 첫 번째 열로 가지고, 2 번째 열은 관심 영역 내의 intensity가 존재하는 개수, 3번째 열은 관심 영역 내에서 해당 intensity가 존재하는 확률, 4번째 열은 관심영역 내의 각각의 intensity를 가지는 모든 위치에 대하여 해당 위치를 제외한 인접한 부분의 평균 intensity와 i의 차이의 크기를 모두 더한 값을 가진다[25]. 다시 말해서 NGTDM 행렬에서 순서쌍($i, 1$)의 값은 관심 영역 내에서 i 번째로 큰 intensity의 크기를 나타낸다.

GLDM 행렬은 임의의 거리 δ 와 임의의 크기 α 에 대하여 Intensity 간의 의존성을 정의한 행렬이다. 이 때, δ 는 보통 1로 정의가 되며, 이 경우 바로 인접한 영역의 의존성만을 수치화 하고 δ 가 1보다 클 경우 2 픽셀 차이까지의 의존성을 수치화 하며 인접한 대각선의 길이는 1로 간주한다. 이미지의 어느 한 점 a 에서 δ 안에 위치한 각각의 점들의 intensity들과 a 의 intensity의 차이가 α 이하로 차이가 날 경우, 의존성이 있다고 정의하게 된다[26]. 이 때, 통상적으로 α 는 0을 사용하며 이 경우에는 점 a 와 그 인접한 영역에서 intensity가 같은 점들의 개수만을 의존성이 있는 점들이라고 간주하게 된다. GLDM 행렬은 이미지 내의 모든 픽셀들에 대하여 인접한 위치의 intensity들과 비교하고 의존성이 있는 픽셀들의 개수를 수치화 하게 된다. 즉, GLDM 행렬의 (i, j) 에서의 원소는 intensity i 가 j 개 만큼 인접하게 있는 개수를 의미한다. 이 때, GLSZM 행렬과의 차이점은 어떤 2차원 이미지가 임의의 intensity i 에 대하여 인접한 9개의 픽셀에서만 intensity i 를 가지고, 나머지는 intensity가 0이라고 한다면, GLDM 행렬도 GLSZM과 마찬가지로 $(i, 9)$ 에서 1을 가지지만, 그 영역의 부분집합들도 행렬에 포함이 된다. 즉, 이 이미지가 GLSZM 행렬에서는 $(i, 9) = 1$ 에서만 1을 가지고 나머지는 0을 갖는 행렬이 되는 것과 다르게 GLDM 행렬은 $(i, 9)$ 에서 1을 가지고, $(i, 4) = 4$, $(i, 6) = 4$ 의 값을 가지는 행렬이 된다. 이

령게 정의된 texture 행렬에서 정의할 수 있는 수치들이 texture feature 들이다.

CNN은 convolution을 이용하여 feature를 추출하는데, 이 convolution 과정은 결국 인접한 영역의 가중치의 합을 얻게 되는 과정이고, pooling layer는 convolution 이후에 얻은 벡터에서 대푯값을 추출하는 과정이다. 즉, 바로 인접한 영역에서의 정보를 잃게 되기 쉬운데, texture 행렬로 얻을 수 있는 정보들에는 바로 인접한 영역에서 얻을 수 있는 정보들이 잘 반영되어 있다. 본 연구에서는 기계학습 모델의 feature들로 filtering 되지 않은 원본 이미지에 대한 intensity, shape, texture feature 들의 총 110개의 handcrafted radiomics feature를 사용하였다.

2.2.1.2 Radiomics 모델

전통적인 기계학습(machine learning)은 모델의 기능을 프로그래밍 하지 않고 데이터를 이용하여 어떤 작업을 할 수 있는 프로그램을 만드는 인공지능의 한 분야이다[27]. 본 연구에서는 PD-L1 수치에 따라 병변을 분류하는 분류기로서 지도 학습 기반의 전통적인 기계학습 분류기를 사용하여 3D radiomics model을 만들었다. 지도 학습이란 훈련 데이터 세트의 데이터들의 정답이 기재되어 되어있는 상황에서 훈련 데이터세트를 학습하는 방법이다. 기계학습 모델은 훈련 데이터 세트 내에 입력 데이터에 대하여 예측 값을 출력하게 되는데, 이 때 해당 데이터는 정답이 기재되어 있으므로 정답이 틀렸는지 맞았는지에 대한 수치화가 가능하고, 이를 점점 맞출 수 있게 되는 방향으로 모델을 학습시키는 방법이다. 본 연구에서 사용된 기계학습 모델은 딥 러닝 기반의 attention based deep multiple instance learning과 전통적인 radiomics 모델이 있는데, 본 장에서는 전통적인 Radiomics 모델들을 설명하고자 한다. 본 연구에서 사용된 radiomics 모델이 딥 러닝 모델과 구분되는 가장 큰 특징은 모델이 데이터의 클래스를 예측하기 위한 입력 데이터의 특징(feature)를 스스로 만들어 내는지 여부이다. 이 때 사용된 radiomics 모델들의 feature들은 handcrafted

radiomics feature들을 min-max로 정규화 해서 사용하였다. 본 연구에서는 PD-L1 발현 여부에 대한 radiomics 모델로 로지스틱 회귀 분류기를 사용하였으며 PD-L1 수치 50% 이상을 분류하는 radiomics 모델로 XGBoost를 선정하였다

1% radiomics 모델에는 로지스틱 회귀 모델이 선정되었다. 로지스틱 회귀 모델은 선형 회귀 모델에 sigmoid 함수를 취한 것으로 정의된다. 로지스틱 회귀모델은 N개의 변수 x 에 대한 종속 변수 y 의 관한 식으로 정의할 수 있다.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

$$\hat{y} = \sigma\left(\sum_{i=1}^N (\beta_i x_i) + \beta_0\right) \quad (2)$$

식 (1)은 시그모이드 함수를 의미하며, (2)는 로지스틱 회귀 모델의 정의식을 의미한다. 시그모이드 내부의 값은 선형 회귀 모델의 정의식이다. 로지스틱 회귀 모델은 범주형 데이터를 대상으로 사용하기 때문에 분류 모델로 사용할 수 있다. 로지스틱 회귀 분석의 손실 함수는 negative log likelihood로서 정의될 수 있으며 이진 분류에서 이는 binary cross entropy와 같다[28].

XGBoost는 의사결정 나무 기반의 boosting 모델이다. Boosting 기법이란 기계학습에서 양상을 기법 중 하나이다. 양상을이란 weak learner라고 하는 개별적인 모델을 조합하여 성능을 향상시키는 방법을 의미한다. 이 중 boosting 기법은 weak learner 들이 한 단계씩 학습을 진행해 나가면서 적절한 가중치를 부여하는 방법으로 학습이 진행되는 양상을 기법이다. XGBoost에서 사용된 weak learner 인 의사결정 나무는 입력 변수들이 상위 루트에서 하위루트로 내려가는 과정에서 불순도를 줄이는

방법으로 학습된다. 의사결정 나무 기반 앙상블 모델은 각각의 의사결정 나무 f_k 에 대하여 다음과 같이 표현할 수 있다.

$$\dot{y} = \Phi(x_i) = \sum_{k=1}^K f_k(x_i) \quad (4)$$

이 때, XGBoost는 다음과 같은 손실함수 $L(\Phi)$ 를 학습한다.

$$L(\Phi) = \sum_i l(\dot{y}_i, y_i) + \sum_k \Omega(f_k) \quad (4)$$

$$\text{where, } \Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (5)$$

여기서 l 은 예측 값 \dot{y} 와 실제 값 y 에 대하여 임의의 미분가능한 convex 형 손실 함수를 의미하고 Ω 의 T 는 각 tree의 leaf 개수를 의미하고 w 는 leaf의 가중치를 의미한다. 이 추가된 정규화 부분 Ω 로 인하여 XGBoost는 오버 피팅을 방지하는 방향으로 학습하게 된다[29].

2.2.2 2D Patch based analysis

Patch based analysis는 병리 영상의 Whole Slide Image(WSI)를 computer vision 분야에서 분석할 때 많이 사용되는 기법이다. Patch based analysis의 장점은 이미지의 세부적인 부분을 봄으로서 국소적인 특징을 추출해내기 쉽다는 점에 있다. 본 연구에 사용된 병변들은 간유리 음영이 포함된 영역이 관심 영역으로 라벨링 되어있고, Handcrafted radiomics feature에서 병변의 전체적인 특징을 추출하였기 때문에 좀 더 부분적인 영역에서의 특징을 추출하기 위하여 병변을 32x32 픽셀 크기를 갖는 2D patch로 나누어 patch based analysis를 적용하였다. 통상적인 3D CNN을 적용한다면, 입력 데이터의 크기를 맞추어야 하는데 이 때 작은 크기의 병변을 보간 하거나 큰 크기의 병변을 다운사이징 해야 한다. 이 과정에서 정보의 손실이 일어나게 되는데 PD-L1이 본래 조직 검사를 통해 얻을 수 있는 정보이고, 조직 검사에 사용되는 검체의 크기는 CT 상의 1 픽셀 정도와 비슷하기 때문에 병변의 크기를 변화시킴으로 인해 일어나는 정보의 손실은 치명적이다. 만약 원본 이미지의 크기를 그대로 사

용하게 될 경우에는 0의 값을 갖는 여백이 많아지기 때문에 학습에 비효율적이다. 따라서, 원본 이미지의 사이즈를 수정하지 않고 handcrafted radiomics feature에 보완할 수 있는 특징을 추출할 수 있는 방법으로 원본 이미지를 32x32 픽셀의 각각의 patch로 나누어 학습시켰다.

2.2.2.1 Label smoothing

정답이 1, 나머지는 0으로 구성된 벡터를 원 핫 인코딩 벡터라고 하며, 이러한 라벨링 방식을 hard label이라고 한다. ImageNet과 같은 큰 데이터 세트에는 라벨링 오류가 존재한다. 이러한 라벨링 오류가 있을 때 hard labeling된 데이터들은 오답에 대한 손실함수를 그대로 완전히 학습하게 된다. 이러한 점을 보완하기 위한 soft labeling 및 그 때 손실함수를 smoothing 하는 방법을 label smoothing이라고 한다[30]. Label smoothing은 $K+1$ 개의 클래스가 있을 때, smoothing parameter α ($0 < \alpha < 1$)에 대하여 원래 정답이라고 알려진 클래스를 α 로 코딩하고, 나머지 클래스를 $\frac{\alpha}{K}$ 로 코딩한 soft label을 적용한다. $K = 1$ 인 이진 분류에서는 hard labeling된 벡터 $[1 \ 0]$ 에 대하여 $[\alpha \ (1 - \alpha)]$ 로 코딩하게 되고, 원래 정답이라고 알려졌던 클래스의 softmax 출력 값을 S 라고 할 때 binary cross entropy loss를 적용하게 되면,

$$L = -(\alpha \log(S) + (1 - \alpha) \log(1 - S)) \quad (6)$$

를 학습하게 된다. Hard labeling된 경우 $(1 - \alpha)$ 가 0의 값을 가지게 되는 반면에 원래 정답이 아니라고 알려졌던 클래스에 대하여 $(1 - \alpha) \log(1 - S)$ 만큼의 보정을 얻는다. 클래스에서 보통 label smoothing을 적용하는 경우에 label이 틀렸을 확률을 예측할 수 있거나 임의로 상수로 정의해서 사용하게 된다. 하지만, 본 연구에서는 상수를 smoothing parameter를 정의할 수 없다. 어느 병변의 조직검사로 획득한 PD-L1 수치가 P 라고 할 때 병변에 있는 하나의 patch의 PD-L1 수치 또한 P 라고 할 수 없다. 그 이유는 폐암의 이질성 때문인데, 하나의 폐암 병변은 동일한 세포로 이루어진 것이 아니기 때문에 조직 검사의 수치가 병변의 위치에 따라 다를 수 있다. Patch의 PD-L1 수치를 결정하기 위하여 이항분포

를 적용하였다. 이항분포는 이산적으로 분포된 확률을 나타내는 방법이다. 이항 분포의 확률 함수 $P(k)$ 는 다음과 같다.

$$C_k^n = \frac{n!}{k!(n-k)!} \quad (7)$$

$$P(k) = C_k^n p^k (1-p)^{n-k} \quad (8)$$

C_k^n 은 전체 n개 중 k개의 원하는 대상이 있을 경우의 수이다. p 는 개별 대상이 가지는 확률 값으로 $P(k)$ 는 전체 n개의 데이터 중 내가 원하는 데이터일 확률이 각각 p 일 경우, 그러한 데이터가 k개 있을 확률을 의미한다 [31]. 본 연구에서는 p 를 조직 검사로 얻어진 PD-L1 수치의 백분율로 정의하였고 n 은 patch 내에 있는 병변의 픽셀 개수로 정의하였다. 의료 영상은 픽셀마다 서로 다른 intensity를 가지게 되므로 이항분포를 적용하였다. 따라서 각각의 병변에 대한 PD-L1 수치의 smoothing parameter α 는,

$$\sum_{k=1}^n P(k) \quad (9)$$

$$\sum_{k=(n/2)}^n P(k) \quad (10)$$

$$\sum_{k=(n/2)+1}^n P(k) \quad (11)$$

로 정의하였다. (9)은 PD-L1 positive 분류기에 대한 smoothing parameter이고 (10),(11)은 각각 patch 내의 병변의 픽셀 수가 짹수, 홀수일 때의 PD-L1 50% 분류기에 대한 smoothing parameter이다.

2.2.2.2 Attention based deep multiple instance learning

Multiple instance learning(MIL)의 개념에는 bag과 instance라는 개념

념이 포함된다. 여기서 instance라는 것은 실제로 구분해야 하는 객체를 의미하고, bag은 여러 instance의 집합이다. instance의 클래스가 몇 개인지는 상관없이 원하는 클래스를 target instance, 그렇지 않은 클래스들을 non-target instance로 나눈다. 그 후 임의의 instance들의 집합인 bag들 중에서 target instance가 포함된 bag과 그렇지 않은 bag을 분류하는 것을 목적으로 고안된 이진 분류 방법을 MIL이라고 한다[32].

이러한 MIL의 개념은 전통적인 machine learning 모델을 포함한 여러 가지 방법으로 시도될 수 있는데, 본 연구에서는 deep learning 기반의 MIL 모델이 사용되었다. Deep learning 기반의 MIL 모델의 구성요소를 3가지로 나누면 instance level feature extractor, MIL pooling, classifier 3가지로 나눌 수 있다. 이 때, 일반적인 deep learning 기반의 분류 모델과 비교했을 때의 차이점은 MIL pooling의 존재 여부이다.

본 모델에서 instance level feature extractor는 convolutional neural network(CNN)을 사용한다. CNN은 시각적 이미지를 분석하는 데 가장 일반적으로 사용되는 모델이다[33]. CNN의 feature extraction은 convolution layer와 pooling layer로 구성된다. Convolution layer에서는 kernel과 입력 데이터의 합성곱(convolution)을 통해 feature map을 추출하게 된다. Kernel과 입력 데이터의 convolution은 공간 주파수 영역에서의 입력 데이터와 kernel의 콤볼드가 되기 때문에 convolutional kernel에 대하여 filtering된 효과를 가진다[34]. 이 때, convolution kernel은 모델의 출력 값이 손실함수에서 각각의 layer로 편미분을 하는 과정인 역전파를 통해 최적화된다. 입력 데이터의 크기와 같은 차원의 kernel의 개수만큼 feature map의 개수가 많아지게 되고, 이 때 feature map의 개수를 channel 수라고 한다. Pooling layer는 feature map에서 근처의 값들을 하나로 통합하는 과정을 의미한다. Pooling layer를 거치면 feature map의 크기는 작아지게 된다. 이러한 과정을 통해 1차원 벡터 형태로 feature map이 형성되게 되고 이를 deep learning feature라고 하며, 본 연구에서는 이렇게 형성된 feature map이 instance level feature가 된다. 이렇게 형성된

instance level feature들은 MIL pooling 과정을 통해 bag level feature가 되고, bag level feature는 fully connected layer(FCN)를 통과하면서 PD-L1 수치에 따른 class로 분류가 되게 된다. 이러한 FCN이 본 모델에서의 classifier 역할을 한다.

일반적인 CNN 분류 모델은 feature extraction과 classification 과정을 거치면서 instance level에서 분류 작업을 하게 된다. 하지만, 본 모델에서는 bag level classification이 목적이므로 instance level feature를 bag level feature로 통합하는 과정이 필요하다. 이러한 역할을 하는 과정이 MIL pooling이다. MIL pooling layer는 permutation invariance 조건을 만족해야 한다. Permutation invariance는 입력 벡터의 순서가 바뀌어도 결과가 변하지 않는 함수의 특성을 말한다[35]. Permutation invariance 조건은 입력의 순서를 바꾸는 permutation matrix P 에 대하여 입력 벡터 X 는 다음과 같은 식이 성립한다.

$$f(PXP^T) = f(X) \quad (12)$$

본 모델 이전에 deep learning 기반 MIL 모델에서는 instance-level approach와 embedding-level approach가 있었다. Instance level approach는 모델의 instance level feature로 분류를 한 후 MIL pooling을 적용하는 방법이다. 이 방법의 장점은 MIL 모델 내에서 bag level classification에 영향을 준 instance가 무엇인지 확인하기 용이하다는 단점이 있지만, 성능이 떨어진다는 단점이 있었다. Embedding level approach는 각각의 feature map마다의 instance level feature의 값들에 MIL pooling을 적용하는 것이다. 이 경우 instance level approach에 비하여 성능은 좋았으나 어떤 instance가 높은 영향을 주었는지를 확인할 수 있는 방법이 없었다. 이러한 두 가지 접근방법을 상호 보완하기 위하여 고안된 방법이 MIL pooling에 attention mechanism을 적용한 attention based deep multiple instance learning이다[36]. bag내의 k번째 instance의 Attention weight는 다음과 같이 정의된다.

$$a_k = \frac{w^T \tanh(V \vec{h}_k^T)}{\sum_{j=1}^N w^T \tanh(V \vec{h}_j^T)} \quad (13)$$

Instance level feature vector의 길이가 L 일 때, N은 instance의 개수, \vec{h}_j 는 j 번째 instance level feature를 의미하며 w와 V는 각각 모델의 parameter로 (Lx1), (LxN)의 크기를 가지고 있다. 위와 같이 정의된 attention weight는 instance level feature의 값에 따라 정의되므로 permutation invariance를 만족하며, 각각의 instance에 따라 정의되므로 MIL 모델이 어떤 instance를 볼 때 가중치를 높게 판단했는지를 확인할 수 있다.

2.3 평가 방법

본 연구에서 모델의 평가 방법으로 모델의 분류성능 지표와 radiomics 모델의 feature importance, MIL 모델의 patch attention map을 사용하였다.

모델의 분류 성능 지표에는 첫 째로 ROC 커브와 그 면적인 AUC가 있다. ROC 커브란 분류기의 공역 내에서 정해진 클래스를 참으로 출력하는 역치 값을 조정하면서 나타나는 위양성률(1-특이도)과 진양성률(민감도)의 관계를 나타내는 그래프이다. 이진 분류기이므로 역치에 대한 기준은 각각 PD-L1 양성과 PD-L1 수치 50% 이상을 기준으로 하며 각 분류기의 출력 결과에는 2가지 클래스에 대한 softmax 함수의 결과값으로 정의되기 때문에 공역은 [0,1]을 가진다.

위양성률은 낮을수록 좋고 진양성률은 높을수록 좋으므로 이상적인 분류기가 최적의 역치 값은 ROC 그래프 상에서 (0,1)인 점을 ROC 커브가 지날 때 (0,1)에서의 역치 값이다. 역치 값이 공역의 최대값일 때는 모두 음성으로 분류하므로 ROC 커브는 (0,0)의 역치 값이 공역의 최소값일 때는 모두 양성으로 분류하므로 ROC 커브는 (1,1)을 갖는다. 역치 값이 감소하면서 위양성률과 위음성률이 동등하게 증가한다면 ROC 커브는 $y = x$

위에 존재하게 되며, 이러한 분류기는 어떠한 역치 값을 가져도 데이터를 제대로 분류할 수 없게 된다. 이 때 ROC 커브의 면적은 $\frac{1}{2}$ 이 된다. 이상적인 분류기는 ROC 커브가 (0,1)을 지나게 되고 모든 위양성률에 대하여 진양성률이 1이므로 ROC 그래프의 면적이 1이된다. 즉, ROC 커브의 면적으로 분류기의 예측 성능을 나타낼 수 있는데 1에 가까울수록 분류 성능이 높고 $\frac{1}{2}$ 에 가까울수록 전혀 분류를 할 수 없다. $\frac{1}{2}$ 보다 작은 경우에는 0에 가까울수록 반대로 분류하게 된다. ROC 커브 상에서 최적의 분류 성능을 나타내는 역치 값을 optimal operating point라고 한다[37].

또 다른 예측성능 지표로서 Accuracy, sensitivity, positive predict value(PPV), negative predict value(NPV), specificity를 사용하였다.

각각의 성능평가 지표들은 다음과 같은 혼동행렬에 대하여

$$\begin{bmatrix} \text{True positive}(TP) & \text{False positive}(FP) \\ \text{False negative}(FN) & \text{True negative}(TN) \end{bmatrix}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$Specificity = \frac{TN}{FN + TN}$$

로 정의되며 Accuracy는 전체 데이터 중 분류기가 맞춘 데이터, sensitivity는 참인 데이터 중 분류기가 맞춘 데이터, PPV은 분류기가 참으로 분류한 데이터 중 정말 참인 데이터를 의미하며 NPV는 분류기가 거짓으로 분류한 데이터 중 정말 거짓인 데이터, specificity는 거짓인 데이터 중 분류기가 맞춘 데이터를 의미한다. 이 때, 참은 PD-L1 양성 분류기와 PD-L1 수치 50% 분류기에 대하여 PD-L1 양성일 때와 PD-L1 수치 50% 이상일 때를 의미한다.

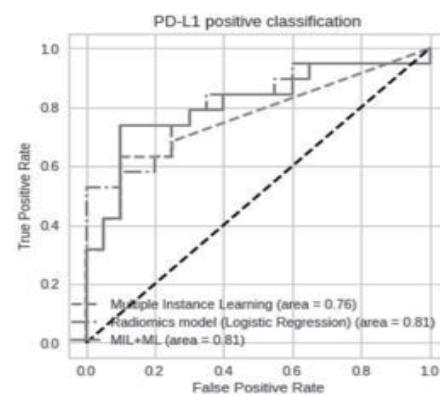
두 번째로 radiomics 모델에서 중요한 feature를 확인하기 위하여 feature importance를 이용하였다. Feature importance란, 전통적인 기계 학습 모델 중 feature들에 대한 가중치를 수치화 할 수 있는 모델들에서 출력할 수 있으며 어떤 feature가 분류에 높은 영향력을 끼쳤는지를 의미하는 지표이다.

세 번째로 MIL 모델의 성능평가를 위하여 patch attention map을 사용하였다. Attention map이란 모델이 어떤 영역을 높은 가중치로 분류하였는지를 의미하는 시각적 지표이다. 본 연구에서 사용된 attention based deep multiple instance learning은 instance level feature에서 bag level feature로 aggregation 되는 과정에서 각각의 가중치를 계산할 수 있게 되는데, 이렇게 계산된 수치가 각각의 instance의 가중치로서 표현된다. 본 연구에서 instance는 각각의 patch를 의미하므로, 계산된 가중치를 통하여 3D 모델링을 통해 시각화하였다.

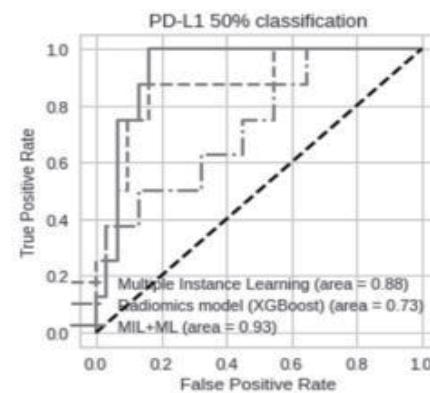
제 3 장. 결과

이번 장에서는 PD-L1 각 PD-L1 각 분류기들의 예측 성능과 radiomics 모델의 feature importance와 MIL 모델의 patch attention map을 제시한다.

3.1 예측 성능



(a) PD-L1 양성 분류기 ROC 커브



(b) PD-L1 수치 50% 분류기 ROC 커브

그림 1 PD-L1 분류기 ROC 커브

그림 1은 PD-L1 양성 분류기와 PD-L1 수치 50%분류기의 ROC 커브이다. PD-L1 양성 분류기와 PD-L1 50% 분류기 모두 MIL 모델과 radiomics 모델의 softmax 출력 값을 평균했을 때 AUC 값이 가장 높았다. PD-L1 양성 분류기에서는 radiomics 모델이 MIL 모델보다 AUC가 높았다.

	Accuracy	Sensitivity	PPV	NPV	Specificity
MIL model	0.77	0.63	0.86	0.72	0,9
Radiomics model	0.72	0.52	0.83	0.67	0.9
MIL + Radiomics	0.82	0.74	0.88	0.78	0.9

(a) PD-L1 양성 분류기 예측 성능

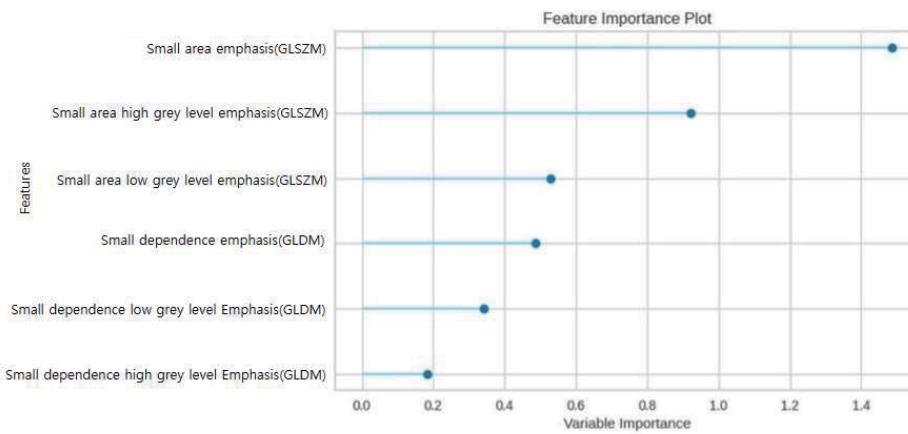
	Accuracy	Sensitivity	PPV	NPV	Specificity
MIL model	0.85	0.875	0.58	0.96	0.84
Radiomics model	0.79	0.50	0.5	0.87	0.87
MIL + Radiomics	0.87	1.00	0.62	1.00	0.84

(b) PD-L1 수치 50% 분류기 예측 성능

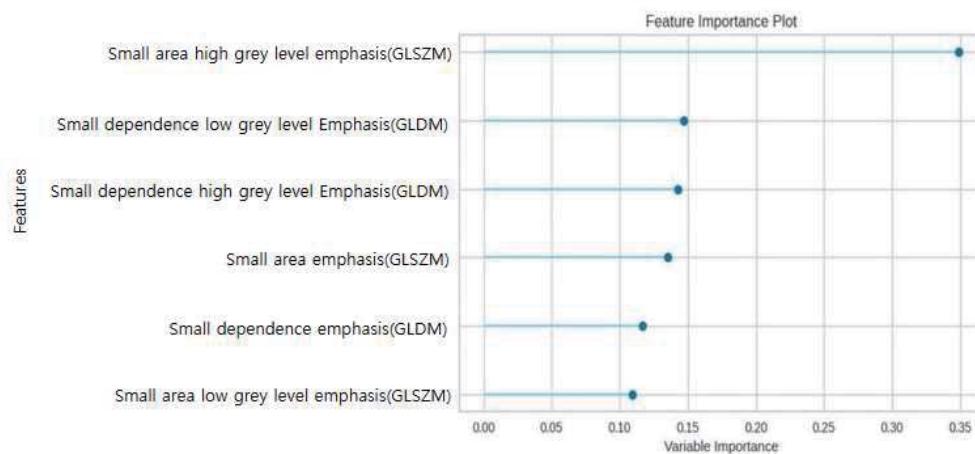
표 3 PD-L1 분류기 예측 성능

표 3.1은 분류기의 예측 성능을 나타낸다. PD-L1 양성 분류기에서 AUC는 MIL 모델보다 radiomics 모델에서 높았지만 다른 예측 평가 지표는 MIL 모델이 높았으며 MIL 모델과 radiomics 모델을 평균했을 때 가장 높았다. PD-L1 수치 50% 이상의 모델에서는 specificity를 제외하고는 모든 성능평가 지표가 MIL 모델이 radiomics 모델보다 높았으며 MIL 모델과 radiomics 모델을 평균했을 때 가장 높았다.

3.2 Feature importance



(a) Machine learning PD-L1 양성 분류기의 feature important



(b) Machine learning PD-L1 수치 50% 분류기의 feature importance

그림 2 Machine learning PD-L1 분류기의 feature importance

그림 2는 각각의 radiomics 모델들에 대한 feature importance를 나타

낸다. XGBoost 인 PD-L1 50% 분류기에서는 가중치 기반에 feature importance를 나타냈고 Logistic regression 분류기인 PD-L1 양성 분류기에서는 계수 기반의 feature importance를 top 6 feature들에 대하여 나타냈다.

3.3 Patch attention map

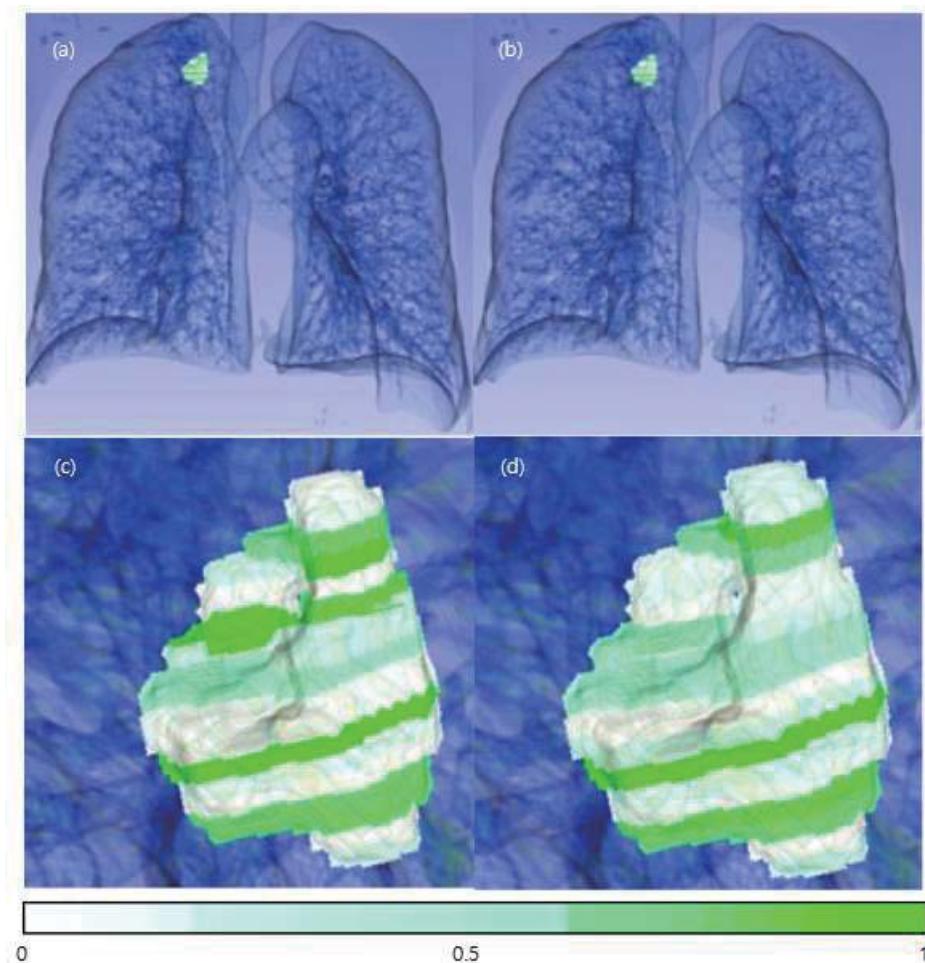


그림 3 Multiple instance learning 모델의 PD-L1 수치 음성인 병변에 대한 patch attention map (a), (c) PD-L1 양성 분류기의 patch attention map (b), (d) PD-L1 50% 수치 분류기의 patch attention map

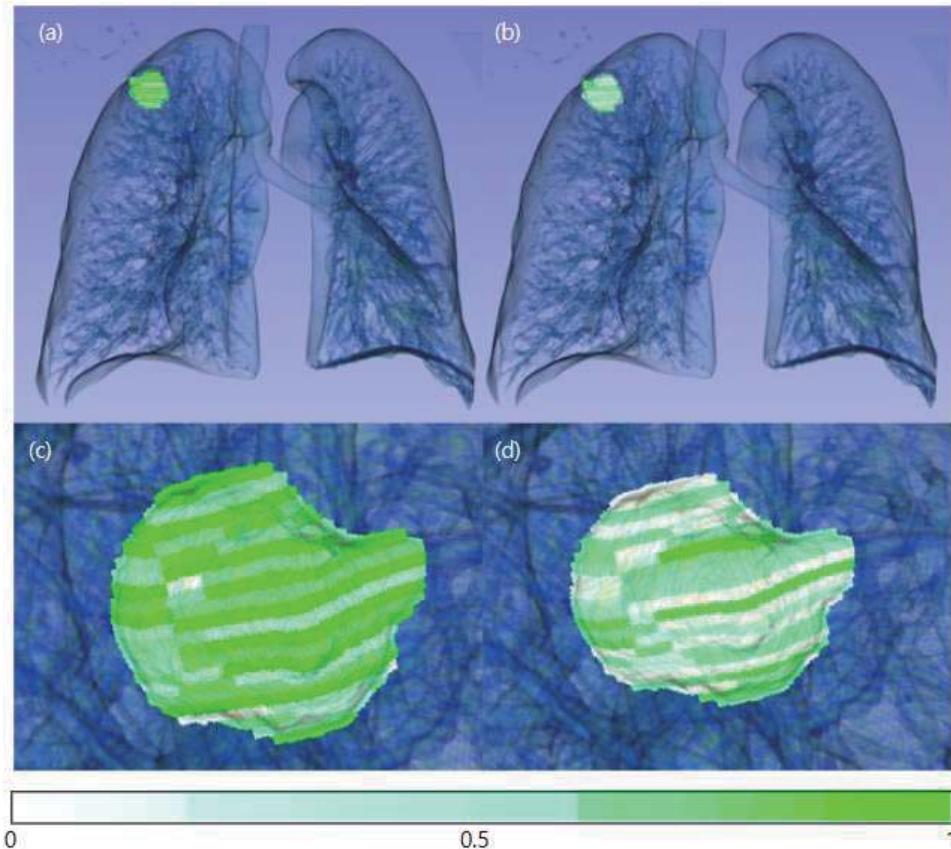


그림 4 Multiple instance learning 모델의 PD-L1 1%인 병변에 대한
patch attention map (a), (c) PD-L1 양성 분류기의 patch attention map
(b), (d) PD-L1 50% 수치 분류기의 patch attention map

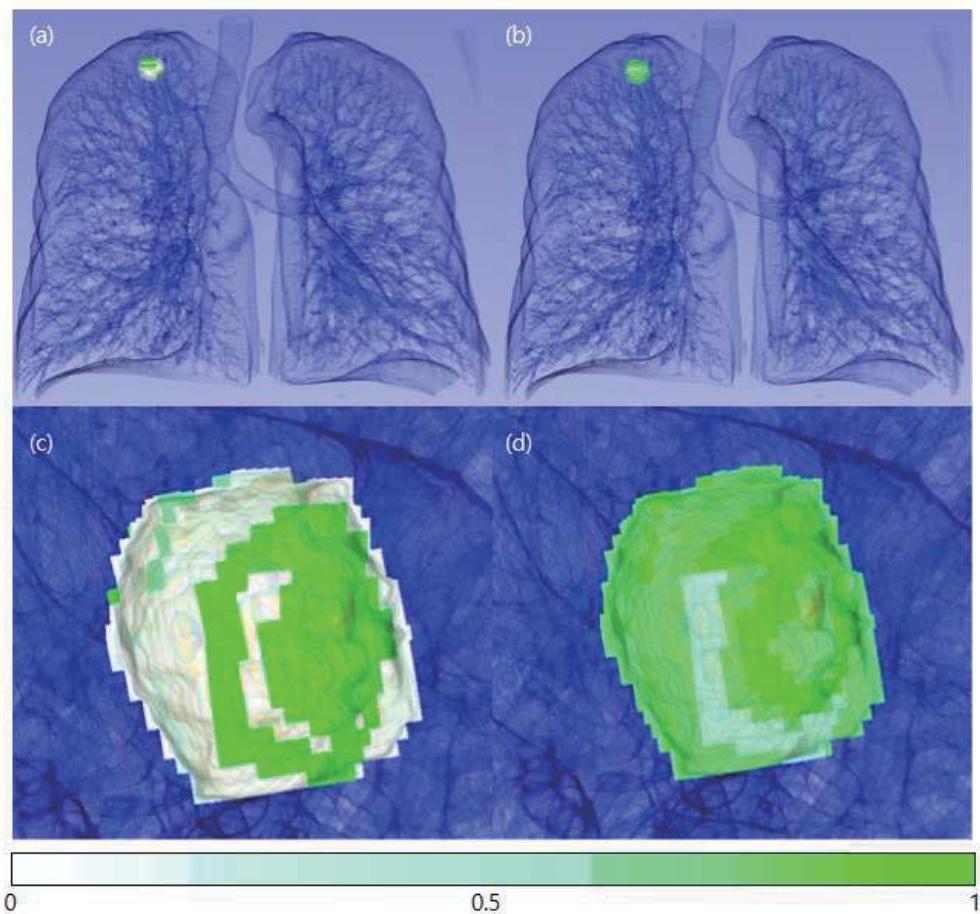


그림 5 Multiple instance learning 모델의 PD-L1 90%인 병변에 대한 patch attention map (a),(c) PD-L1 양성 분류기의 patch attention map (b), (d) PD-L1 50% 수치 분류기의 patch attention map

그림 3부터 5는 각각 PD-L1 음성, PD-L1 수치 1%, PD-L1 수치 90%인 병변에 대하여 patch attention map을 원본 데이터와 겹쳐서 출력하였다. Patch attention map은 MIL 모델에서 attention weight들을 각각의 병변에 대하여 min-max로 normalize한 후 각 patch들이 가지는 상대적인 수치들을 시각화한 것이다.

제 4 장. 고찰

표 4은 선행 연구들과 본 연구의 데이터 세트 및 성능 비교이다. Changdi wang 외 연구진[12]은 1135개의 데이터로 PD-L1 수치 예측에 양성과 수치 50% 모두에서 아주 높은 AUC를 얻었다. Panwen Tian 외 연구진[13]은 939개의 많은 데이터로 전이성 폐암에 대하여 AUC 0.76을 얻었다. 전이성 폐암만을 가지고 연구했기 때문에 본 연구와는 데이터 세트 자체가 다르다고 볼 수 있다. Ying zhu 외 연구진[14]은 127개의 적은 수의 데이터로 0.78 및 0.77의 다소 높은 수치를 얻었으나 본 연구의 AUC 보다 조금 떨어졌으며 데이터의 불균형이 본 연구만큼 심하지는 않았다. Qing wen 외 연구진[15]은 T3와 T4의 병변만 사용하였다. 적은 데이터로 0.79라는 다소 높은 수치를 얻었으나 본 연구의 결과보다 다소 떨어졌다. Zekun Jiang 외 연구진[16]은 적은 수의 데이터로 PD-L1 양성 예측에서 0.85라는 높은 수치를 얻었다. Qing wen 연구진 및 Zekun Jinag 연구진은 PD-L1 양성 및 음성인 데이터에 대해서는 본 연구만큼 불균형이 심하였으나 PD-L1 양성인 환자들의 데이터가 음성인 환자들의 데이터보다 많았다는 것이 본 연구와는 반대되는 점이다. 또한 PD-L1 수치 50% 이상에 대한 분류는 시행하지 않았다. 다른 연구들과 비교해서 PD-L1 양성 반응을 가진 병변들의 숫자가 많았다. Stefano Bracci 외 연구진[17]은 아주 적은 수의 데이터를 이용하여 PD-L1 수치 50% 이상을 분류하는 연구에서 AUC 0.79로 다소 높은 성능을 냈으나 데이터 세트가 상당히 균형적이었다. 본 연구에서는 PD-L1 수치 50% 이상인 병변들에 대한 분류를 데이터 불균형이 심한 데이터 세트를 가지고 높은 분류성능을 가진 분류기를 만들었다. 또한 PD-L1 음성 및 양성을 분류하는 분류에서도 상당히 높은 분류 성능을 냈다.

선행 연구들과의 또 다른 차이점으로는 본 연구는 수술이 가능했던 병변들에 대하여 연구를 진행하였다. 선행 연구 중 수술이 가능한 병변들을 사용한 연구는 Zekun Jinag 외 연구진 밖에 없었다. 수술이 가능한 환자라

도 adjuvant therapy나 neo adjuvant therapy를 받았을 때 예후가 좋다는 점을 고려한다면[4] 수술이 가능한 환자들에 대한 연구도 의미가 있다.

본 연구에서 PD-L1 양성 분류기의 AUC와 PD-L1 수치 50% 이상 분류기의 특이도를 제외하고는 MIL 기반 모델이 Radiomics 기반 모델보다 예측 성능이 잘 나왔다. 또한 MIL과 Radiomics 모델의 softmax 값의 평균으로 aggregation된 MIL + Radiomics 모델이 PD-L1 수치 50% 이상의 특이도를 제외하고 모든 예측성능이 가장 좋았다. PD-L1 수치 50%의 radiomics 기반 모델은 데이터 불균형에 대한 보정이 없었기 때문에 특이도가 높은 것은 일종의 오버 피팅으로 볼 수도 있을 듯하다. 여기서 MIL + Radiomics 모델의 softmax 값의 평균이 각각의 분류기들보다 성능이 좋다는 점의 의미는 MIL 모델이 잘 분류하지 못한 병변과 Radiomics 모델이 잘 분류하지 못한 병변들에 대하여 상호 보완적인 효과를 가지고 있다는 점을 의미한다. 이를 분석하기 위하여 Radiomics 모델의 feature importance를 출력해 보았는데 GLDM 및 GLSZM 행렬에서의 small area의 수치들을 중요한 feature로 학습한 분류기들이었다는 점을 발견하였다. 이는 MIL 모델의 feature extraction 과정에서 손실될 수 있는 바로 인접한 영역의 값의 차이가 특정 병변들의 PD-L1 수치를 예측하는데 유의미한 정보를 가지고 있었다는 것을 의미하며, 이러한 특징이 CNN 기반 feature로 학습된 모델과 상호 보완하여 예측 성능에 긍정적인 영향을 끼쳤다고 보인다.

본 연구에서는 병변의 크기를 계산하였다. 병변의 크기는 종양의 크기가 클수록 불균질성이 높다고 알려져 있으며 영상에서 불균질성을 가늠할 때 계산되는 요소이다[38]. 훈련 데이터 세트에 병변들은 테스트 데이터 세트의 병변들보다 큰 편차를 가지며, 평균적인 병변의 크기 또한 더 크다. PD-L1의 QIF는 병변의 크기와 상관성이 없다고 보고된 연구가 있다[39]. 이에 따르면 본 연구의 모델들은 병변의 불균질성에 대하여 노이즈가 심한 데이터로 훈련하였으나 PD-L1을 잘 예측했다고 해석할 수 있다.

그림 3은 PD-L1 음성인 병변으로 테스트 데이터 중 두 분류기가 모두

negative 클래스로 예측한 병변이다. 그럼 3.4는 PD-L1 수치 1%인 병변으로 PD-L1 양성 분류기는 양성으로 PD-L1 50% 분류기는 50% 이하로 분류한 병변이며, 그럼 3.5는 PD-L1 수치 90%인 병변에 대하여 두 모델 모두 positive 클래스로 분류한 병변이다. 각각의 patch attention map을 나타냈다. 그럼 3.3의 경우 두 모델에서 모두 비슷한 attention weight의 분포를 가졌고, 그럼 3.4는 PD-L1 양성 분류기가 PD-L1 수치 50% 분류기에 비해 골고루 높은 attention weight를 가졌으며 그럼 3.5는 그 반대의 attention weight를 가졌음을 알 수 있다. Patch attention map이 attention weight의 상대적 분포를 나타낸다는 점과 MIL 모델이 전체 bag 내에서 하나의 instance만 target class에 속해도 참으로 분류하게끔 설계되었다는 점을 고려한다면 다음과 같은 해석을 할 수 있다.

우선, attention weight가 1에 가까운 patch들의 feature 들이 높은 가중치로 MIL 모델의 분류단에 입력된다. 그럼 3.5의 PD-L1 수치 90%의 병변에 대하여 (a),(c)는 가중치가 높은 patch가 몇 개 없음에도 PD-L1 양성으로 분류되었고, 그럼 3.4의 PD-L1 수치 1%의 병변에 대하여 (a),(c)는 전체 patch들을 골고루 높은 가중치로 판단하여 PD-L1 양성으로 분류되었다는 점은 PD-L1 수치가 90%에 가까운 patch를 포함한 병변과 PD-L1 수치가 1%에 가까운 patch들에서 형성되는 weight의 차이가 존재한다는 것을 의미한다. 그 이유는 식 (13)에서 전체 병변의 정보를 가지고 있는 벡터 V 에 의해 각 patch의 attention weight가 부여되게 되기 때문이다. 또한, 그럼 5의 (b), (d)와 그럼 3.4의 (b), (d)에서의 attention weight 분포가 차이가 있고 이 차이가 그럼 4와 5의 (a), (c)에서의 차이와 다르다는 점은 각각의 분류기가 참인 클래스로 학습한 patch들에서 차이가 있음을 뒷받침한다고 할 수 있다. 마지막으로 그럼 3의 (c)와 (d)를 보면 attention weight의 분포가 유사함을 알 수 있는데, 이는 조직검사 결과로 PD-L1 음성을 얻은 해당 병변이 정말 모든 patch에서 PD-L1 음성이라는 가설이 참이라면 PD-L1 음성인 patch들은 PD-L1 수치가 양성인 patch들과 구분되는 영상적 특징이 존재함을 뒷받침할 수 있는 증거가 된다. 음성인 patch들은 PD-L1 양성 분류기와 PD-L1 수치 50% 분류기에



서 모두 거짓인 클래스로 학습이 되었기 때문이다.

본 연구와 같이 수술이 가능한 PD-L1 병변들의 경우, 절개된 병변들을 이용하여 원하는 부위의 조직 검사를 다시 해볼 수 있다. 따라서, 해당 연구의 후속 연구로서 절개된 병변들의 patch attention에 따른 조직 검사 결과와 PD-L1과 연관성이 있는 다른 유전자 정보를 결합해 본다면 PD-L1 수치에 따른 병변 분류 연구의 해석을 정량화 할 수 있을 듯하다.

		Chengdi Wang [12]	Panwen Tian [13]	Yigh zhu [14]	Qing wen [15]	Zekun Jiang [16]	Stefano Bracci [17]	This research
	Negative	722(63.6%)	611(65.1%)	81(63.8%)	32(for training)	36(28.8%)	23(0.32)	162(65.1%)
PD-L1	0~49	50(4.4%)	363(32.0%)	328(34.9%)	10(7.9%)	58(for training)	25(0.35)	55(22.1%)
	50 or higher	1135	939	38(30.0%)	38(30.0%)	120	24(0.33)	31(12.4%)
	total	2(0.2%)	—	—	—	125	72	248
	is	440(38.8%)	440(38.8%)	22(17.3%)	—	4(3.2%)	—	2(0.8%)
T – stage	1	368(32.4%)	368(32.4%)	33(26.0%)	33(26.0%)	106(84.8%)	177(71.3%)	177(71.3%)
	2	100(8.8%)	100(8.8%)	23(18.1%)	23(18.1%)	80(66.6%)	45(18.1%)	45(18.1%)
	3	171(15.1%)	171(15.1%)	49(38.6%)	49(38.6%)	40(33.3%)	7(2.8%)	17(6.8%)
	4	54(4.7%)	54(4.7%)	—	—	—	—	—
T _x	0	603(53.1%)	metastatic NSCLC	19(15.0%)	—	—	219(88.3%)	—
	1	83(7.3%)	C	13(10.2%)	—	—	9(3.6%)	—
	2	240(21.1%)		46(36.2%)	—	—	20(8.0%)	—
	3	121(00.7%)		49(38.6%)	—	—	—	—
N – stage	0	88(7.8%)		—	—	—	—	—
	1	781(68.8%)		—	—	—	—	—
	2	300(26.4%)		—	—	—	—	—
	3	54(4.8%)		—	—	—	—	—
M – stage	0	—		—	—	—	—	—
	1	—		—	—	—	—	—
	2	—		—	—	—	—	—
TNM Stage	IIIA	—		—	—	—	22(30.6%)	—
	IIIB	—		—	—	—	20(27.8%)	—
	IIIC	—		—	—	—	6(8.7%)	—
	IV	—		—	—	—	24(33.3%)	—
AUC	Positive	0.95	—	0.78	0.79	0.85	—	0.81
	50%	0.95	0.76	0.77	—	—	—	0.93
Method	DL+ Radiomics	DL	DL	Radiomics	Radiomics	Radiomics	DL+ Radiomics	19.04(cm ²)
Average size(95% range)	—	—	—	—	—	—	(0.40~172.10)	—



Average train size(95% range)	-	-	-	-	-	-	21.58(cm^2)
Average test size(95% range)	-	-	-	-	-	-	0.40~159.06
							5.31(cm^2)
							0.83~95.27

표 4 선형연구 및 분 연구 테이터 세트 비교

제 5 장. 결론

본 논문은 CT 영상을 이용하여 비침습적인 방법으로 PD-L1 양성과 PD-L1 수치 50%의 병변을 분류하는 모델을 제안하였다. 본 연구에서는 3D image에서 3차원 handcrafted radiomics feature를 출력하여 radiomics 모델을 이용한 분석과 2D patch based analysis를 이용한 multiple instance learning(MIL) 모델을 이용한 분석을 함께 진행하고 분류 결과에 대하여 설명하였다. 예측 성능 평가 지표는 대체로 MIL 모델에서 우수하였으나 radiomics 모델과 MIL 모델을 결합하였을 때 가장 우수한 성능을 보였다.

본 연구에서 사용된 데이터세트는 PD-L1 수치 50% 이상의 병변이 전체의 약 12%로 데이터 불균형이 아주 심했다. 이러한 문제를 해결하기 위하여 병변을 patch로 분할하여 MIL 모델을 적용함으로써 불균형한 데이터 세트를 균형 있는 bag 데이터 세트로 만들어서 분류 성능을 향상시켰다.

MIL 모델에서 예측하지 못했던 병변들 중 radiomics 모델과 결합하였을 때 분류가 가능한 병변들이 있었다. 선정된 radiomics 모델들은 바로 인접한 픽셀들의 intensity 차이가 수치화 된 feature들을 중요하게 학습한 모델들이었다. MIL 모델의 feature extraction 과정에서 바로 인접한 영역의 intensity 차이는 학습할 수 없게 되지만 radiomics 모델에서 학습한 feature 들이 이를 보완할 수 있었다고 생각한다. 또한 patch attention

map을 이용하여 MIL 모델에서 분류한 병변의 patch 들의 가중치를 시각화 하였다. PD-L1 양성인 병변들의 patch를 참인 클래스로 학습한 분류기와 PD-L1 수치 50% 이상인 병변을 참인 클래스로 학습한 분류기가 PD-L1 음성인 병변은 가중치의 분포가 비슷하나 PD-L1 양성이면서 수치가 다른 병변들에 대한 가중치의 분포는 다르다는 점에서 PD-L1 수치에 따라 병변의 intensity에 영향을 줄 수 있음을 확인하였다.

선행 연구들과 비교하여 데이터양이 적음에도 PD-L1 50% 이상의 병변들을 잘 분류하였으며, PD-L1 양성 분류도 준수한 성능을 보였다. 특히 PD-L1 수치 50% 이상의 병변들은 초기 단계의 병변들이었기 때문에 데이터 불균형이 선행 연구들에 비해 아주 심한 편이었으나 성능이 높았다. 또한, patch attention map과 feature importance를 이용하여 해석 가능한 모델을 만들었다.

참고 문헌

- [1] 통계청. (2022) 사망원인통계
- [2] Le Chevalier T. Adjuvant chemotherapy for resectable non-small-cell lung cancer: where is it going? *Ann Oncol.* 2010 Oct;21 Suppl 7:vii196-8. doi: 10.1093/annonc/mdq376. PMID: 20943614.
- [3] Elias AD. Small cell lung cancer: state-of-the-art therapy in 1996. *Chest.* 1997 Oct;112(4 Suppl):251S-258S. doi: 10.1378/chest.112.4_supplement.251s. PMID: 9337299.
- [4] Vansteenkiste J, Wauters E, Reymen B, Ackermann CJ, Peters S, De Ruysscher D. Current status of immune checkpoint inhibition in early-stage NSCLC. *Ann Oncol.* 2019 Aug 1;30(8):1244-1253. doi: 10.1093/annonc/mdz175. PMID: 31143921.
- [5] Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. *Immunity.* 2013 Jul 25;39(1):1-10. doi: 10.1016/j.jimmuni.2013.07.012. PMID: 23890059.
- [6] Ettinger DS, Aisner DL, Wood DE, Akerley W, Bauman J, Chang JY, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 5.2018. *J Natl Compr Canc Netw.* 2018 Jul;16(7):807-821. doi: 10.6004/jnccn.2018.0062. PMID: 30006423.
- [7] Martinez P, Peters S, Stammers T, Soria JC. Immunotherapy for the First-Line Treatment of Patients with Metastatic Non-Small Cell Lung Cancer. *Clin Cancer Res.* 2019 May 1;25(9):2691-2698. doi: 10.1158/1078-0432.CCR-18-3904. Epub 2019 Jan 14. PMID: 30642913.
- [8] European society for medical oncology, Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment

and follow-up, Ann Oncol 2018;29(Suppl 4): iv192–iv237

[9] Mu CY, Huang JA, Chen Y, Chen C, Zhang XG. High expression of PD-L1 in lung cancer may contribute to poor prognosis and tumor cells immune escape through suppressing tumor infiltrating dendritic cells maturation. Med Oncol. 2011 Sep;28(3):682–8. doi: 10.1007/s12032-010-9515-2. Epub 2010 Apr 6. PMID: 20373055.

[10] Forde PM, Chafft JE, Smith KN, Anagnostou V, Cottrell TR, Hellmann MD, et al. Neoadjuvant PD-1 Blockade in Resectable Lung Cancer. N Engl J Med. 2018 May 24;378(21):1976–1986. doi: 10.1056/NEJMoa1716078. Epub 2018 Apr 16. Erratum in: N Engl J Med. 2018 Nov 29;379(22):2185. PMID: 29658848; PMCID: PMC6223617.

[11] Tsunoda A, Morikawa K, Inoue T, Miyazawa T, Hoshikawa M, Takagi M, et al. A prospective observational study to assess PD-L1 expression in small biopsy samples for non-small-cell lung cancer. BMC Cancer. 2019 Jun 7;19(1):546. doi: 10.1186/s12885-019-5773-3. PMID: 31174496; PMCID: PMC6555021.

[12] Wang C, Ma J, Shao J, Zhang S, Li J, Yan J, et al. Non-Invasive Measurement Using Deep Learning Algorithm Based on Multi-Source Features Fusion to Predict PD-L1 Expression and Survival in NSCLC. Front Immunol. 2022 Apr 7;13:828560. doi: 10.3389/fimmu.2022.828560. PMID: 35464416; PMCID: PMC9022118.

[13] Tian P, He B, Mu W, Liu K, Liu L, Zeng H et al. Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. Theranostics. 2021 Jan 1;11(5):2098–2107. doi: 10.7150/thno.48027. PMID: 33500713; PMCID: PMC7797686.

[14] Zhu Y, Liu YL, Feng Y, Yang XY, Zhang J, Chang DD, et al. A CT-

derived deep neural network predicts for programmed death ligand-1 expression status in advanced lung adenocarcinomas. Ann Transl Med. 2020 Aug;8(15):930. doi: 10.21037/atm-19-4690. PMID: 32953730; PMCID: PMC7475404.

[15] Wen Q, Yang Z, Dai H, Feng A, Li Q. Radiomics Study for Predicting the Expression of PD-L1 and Tumor Mutation Burden in Non-Small Cell Lung Cancer Based on CT Images and Clinicopathological Features. Front Oncol. 2021 Aug 6;11:620246. doi: 10.3389/fonc.2021.620246. PMID: 34422625; PMCID: PMC8377473.

[16] Jiang Z, Dong Y, Yang L, Lv Y, Dong S, Yuan S, et al. CT-Based Hand-crafted Radiomic Signatures Can Predict PD-L1 Expression Levels in Non-small Cell Lung Cancer: a Two-Center Study. J Digit Imaging. 2021 Oct;34(5):1073-1085. doi: 10.1007/s10278-021-00484-9. Epub 2021 Jul 29. PMID: 34327623; PMCID: PMC8554954.

[17] Bracci S, Dolciami M, Trobiani C, Izzo A, Pernazza A, D'Amati G, et al. Quantitative CT texture analysis in predicting PD-L1 expression in locally advanced or metastatic NSCLC patients. Radiol Med. 2021 Nov;126(11):1425-1433. doi: 10.1007/s11547-021-01399-9. Epub 2021 Aug 9. PMID: 34373989; PMCID: PMC8558266.

[18] Cooper WA, Tran T, Vilain RE, Madore J, Selinger CI, Kohonen-Corish M et al. PD-L1 expression is a favorable prognostic factor in early stage non-small cell carcinoma. Lung Cancer. 2015 Aug;89(2):181-8. doi: 10.1016/j.lungcan.2015.05.007. Epub 2015 May 18. PMID: 26024796.

[19] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563-77. doi: 10.1148/radiol.2015151169. Epub 2015 Nov 18. PMID: 26579733;

PMCID: PMC4734157

- [20] Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypliński P, Gibbs P, et al. Introduction to Radiomics. *J Nucl Med.* 2020 Apr;61(4):488–495. doi: 10.2967/jnumed.118.222893. Epub 2020 Feb 14. PMID: 32060219; PMCID: PMC9374044.
- [21] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imaging.* 2020 Aug 12;11(1):91. doi: 10.1186/s13244-020-00887-2. PMID: 32785796; PMCID: PMC7423816.
- [22] Haralick, R. M., Shanmugam, K., & Dinstein, I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics,* 1973; *SMC-3*(6):610–621. doi:10.1109/TSMC.1973.4309314
- [23] Galloway, M. M. Texture classification using gray level run length. *Comput. Graph. Image Process.* 4.2 1975; 172–179
- [24] Thibault, G., Angulo, J., & Meyer, F. Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification. *IEEE Transactions on Biomedical Engineering,* 2014;61(3):630–637. doi:10.1109/TBME.2013.2284600
- [25] Amadasun, M., & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics,* 1989; 19(5) : 1264–1274. doi:10.1109/21.44046
- [26] Sun, C. Wee, W.G. Neighboring gray level dependence matrix for texture classification. *Comput. Vis. Graph. Image Process.* 1982; 23, 341–352.
- [27] Tom Mitchell, Machine learning, McGraw Hill(NY); 1997

- [28] DeMaris, A.. A Tutorial in Logistic Regression. *Journal of Marriage and Family*, 1995; 57(4), 956–968. <https://doi.org/10.2307/353415>
- [29] Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785–794. New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>
- [30] Müller R, Kornblith S, Hinton GE. When Does Label Smoothing Help? *CoRR*. 2019;abs/1906.02629. <http://arxiv.org/abs/1906.02629>
- [31] Henry S, John W. Probability, Statistics, and Random Processes for Engineers. 4th ed. PEARSON; 2012
- [32] Marc-André C, Veronika C, Eric G, Ghyslain G. Multiple instance learning: A survey of problem characteristics and applications, *Pattern Recognition*, Volume 77, 2018,
- [33] M.V. Valueva, N.N. Nagornov, P.A. Lyakhov, G.V. Valuev, N.I. Chervyakov. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation, *Mathematics and Computers in Simulation*, Volume 177, 2020,
- [34] Rafael C, Richard E. Digital Image Processing 3rd ed. Prentice-hall; 2006
- [35] Zaheer M, Kottur S, Ravanbakhsh S, Póczos B, Salakhutdinov R, Smola AJ. Deep Sets. *CoRR*. 2017;abs/1703.06114. <http://arxiv.org/abs/1703.06114>
- [36] Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. *CoRR*. 2018;abs/1802.04712. <http://arxiv.org/abs/1802.04712>

- [37] Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8(4):283–298. doi:10.1016/s0001-2998(78)80014-2
- [38] Davnall F, Yip CS, Ljungqvist G, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?. *Insights Imaging.* 2012;3(6):573–589. doi:10.1007/s13244-012-0196-6
- [39] McLaughlin J, Han G, Schalper KA, et al. Quantitative Assessment of the Heterogeneity of PD-L1 Expression in Non-Small-Cell Lung Cancer [published correction appears in *JAMA Oncol.* 2016 Jan;2(1):146]. *JAMA Oncol.* 2016;2(1):46–54. doi:10.1001/jamaoncol.2015.3638

Abstract

Development and Performance Evaluation of classifiers through Non-invasive method to classify PD-L1 expression level using machine learning in Non-Small Cell Lung Cancer

Jang, Byoung Han

Dept. of Integrative Medicine

The Graduate School

Yonsei University

Programmed cell death ligand 1 (PD-L1) is a transmembrane protein that combines with Programmed cell death 1 (PD-1) of T cells to determine the cells as normal cells. However, some cancer cells express such PD-L1 so that T cells judge the cancer cells as normal cells. Due to the use of immune checkpoint inhibitors using the mechanism of PD-L1, research using PD-L1 as a biomarker has recently been actively conducted and has been in the spotlight in chemotherapy. A biopsy is used to determine PD-L1 levels, but in the case of a biopsy, it cannot be applied to all patients because it is an invasive method. Because of that, it is important to determine PD-L1 level using a non-invasive method.

In this study, I propose a model for classifying PD-L1 levels using CT images. I propose a binary classification model that classifies whether PD-L1 is positive or negative, and a binary classification model that classifies whether PD-L1 level above 50% or not. Since there are few patients with PD-L1 levels above 50% of all patients in the early stages of the lesion, multiple instance learning (MIL) technique was used to make classification models with high classification performance even with unbalanced datasets. Machine learning classifiers trained with handcrafted radiomics features as radiomics model were used to extract features from 3D images. The final models were defined as the average of the softmax outputs of the MIL model and radiomics model for each positive classifier and PD-L1 level 50% classifier. Finally, AUC 0.81 in the PD-L1 positive classifier and AUC 0.93 in the PD-L1 level 50% classifier were obtained.

In this study, patch attention map and feature importance were used to interpret how each classifier classified PD-L1 level. MIL models showed a difference in attention maps according to PD-L1 level, and it was confirmed through feature importance that the radiomics classifier also learned features that could supplement the performance of the MIL models



Key words : PD-L1, Non-invasive. MIL, Handcrafted radiomics feature