



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 원저작자를 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 권리는 위의 내용에 의하여 영향을 받지 않습니다.

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

[Disclaimer](#)

비소세포폐암 수술 후 예후 예측을 위한 설명 가능한 머신러닝 모델의 적용

- 진단 시 체질량지수의 영향 분석 -

연세대학교 보건대학원

보건정보통계학과 보건정보통계전공

장 보 윤

비소세포폐암 수술 후 예후 예측을 위한 설명 가능한 머신러닝 모델의 적용

– 진단 시 체질량지수의 영향 분석 –

지도 박 소 희 교수

이 논문을 보건학 석사학위 논문으로 제출함.

2024년 6월 19일

연세대학교 보건대학원

보건정보통계학과 보건정보통계전공

장 보 윤

장보윤의 보건학 석사학위 논문을 인준함.

심사위원 박 소 희 인

심사위원 남 정 모 인

심사위원 하 민 진 인

연세대학교 보건대학원

2024년 6월 19일

감사의 말씀

학부를 졸업한 지 15년 만에 석사과정을 시작하며 학문의 세계에 발을 내딛었습니다. 학업은 일과 육아에 더해져 저에게 치열한 도전이었으며 무한한 성장과 발전을 이루는 시간이었습니다. 배움과 연구의 시간 동안 잘 버텨낼 수 있도록 함께 해주시고 응원해 주신 모든 분께 감사의 인사를 전합니다.

보건정보통계학의 길로 이끌어 주시고, 일과 육아 그리고 학업까지 하는 학생들이 더 잘한다며 부족함이 많은 저에게 자신감을 심어주시고 애정 어린 관심과 지도로 많은 가르침을 주시려고 노력해 주신 박소희 교수님 진심으로 감사드립니다. 덕분에 석사과정 동안 크게 성장할 수 있었습니다. 논문의 완성도를 높이기 위해 세심하고 깊이 있는 조언을 해주신 남정모 교수님과 SAS, R 통계 프로그램을 다룰 수 있게 가르침을 주시고 따뜻한 격려를 해주신 하민진 교수님께도 감사의 말씀을 드립니다.

2022년 8월 OT에서 처음 만나 2년 넘게 희노애락을 함께한 우리 동기들 고맙습니다. 각자의 자리에서 최선을 다하는 동기들을 보며 자극받으며 성장하였고, 너무나 소중한 인연들과의 시간이었기에 석사과정이 외롭지 않고 풍성했습니다. 특히 우유빛깔 박소희 교수님 지도하에 함께한 유혜연, 오유진, 주동진, 김승호 선생님, 그대들이 있어 중간에 포기하지 않고 더욱 힘을 받아 논문을 마무리하고 졸업할 수 있었습니다. 고맙습니다.

매일 함께했던 국립암센터 의료정보관리실 직원들과 어려운 일을 맡아 함께 고생하고 있는 어벤져스팀, 차세대 실무지원팀 분들에게도 감사의 인사들 드립니다. 특히, 존경하는 직장 상사이자 친근한 옆집 큰언니처럼 아낌없이 지원해 주시고 어려움에 부딪힐 때마다 지혜롭게 조언해 주신 박은혜 실장님 진심으로 감사드립니다. 또한, 석사과정을 시작할 수 있는 발판을 마련해 주시고 응원해 주신 암빅데이터센터의 최귀선 센터장님과 차효성, 김현진, 채희정 팀장님께도 감사의 마음을 전합니다. 처음 접하는 통계로 어려워할 때마다 알기 쉬운 설명으로 이해력을 높여주고, 처음 시도해 보는 머신러닝 연구에도 많은 도움을

주신 이상원 선생님과 갑작스러운 연락에도 세심한 조언을 해주셨던 아주대병원 변상일 선생님 감사합니다. 워킹맘 힐링추진단의 최강미모 김지연, 배연민 파트장님, 힘들어할 때마다 할 수 있다며 격려해 주신 말들 많은 힘이 되었습니다. 감사드립니다.

늦게 시작한 학위과정을 응원해 주시고 수업이 있는 날과 시험 기간이면 한걸음에 달려오시어 육아와 가사를 도와주신 부모님과 시부모님께 감사의 말씀을 드립니다. 앞으로 더 잘하며, 보답하겠습니다. 오래도록 건강하세요. 사랑합니다. 마지막으로 묵묵히 옆에서 지원해 주고 힘들 때마다 버팀목이 되어준 든든한 나의 남편, 그리고 엄마의 빈자리가 많았음에도 내색 없이 알아서 척척 자기 일을 해내던 세상에서 가장 소중하고 대견한 나의 천사 로아, 너무나도 사랑합니다.

저 혼자서는 못해냈을 것 같은 2년 반의 시간,
함께해 주시고 격려해 주신 모든 분께 다시 한번 감사와 존경을 표합니다.

2024년 6월

장 보 운 올림

차 례

국문요약

I . 서론	1
1. 연구 배경 및 필요성	1
2. 선행연구 고찰	5
3. 연구 목적	9
II . 연구 방법	10
1. 연구 자료 및 연구 대상자	10
2. 변수 정의	14
3. 통계분석 방법	16
III . 연구 결과	26
1. 연구대상자의 일반적 특성	26
1.1. 사망 여부에 따른 대상자의 일반적 특성	26
1.2. 재발 여부에 따른 대상자의 일반적 특성	31
1.3. 진단 시 체질량지수에 따른 대상자의 일반적 특성	35
1.4. 단순 및 다중 로지스틱 회귀분석	40
2. 예후 예측 모델 개발	45
3. 머신러닝 분류 모델 성능 비교	50
3.1. 릿지 & 라쏘 회귀분석(Ridge & Lasso regression)	50
3.2. 랜덤 포레스트(Random Forest)	55
3.3. 익스트림 그래디언트 부스팅(XGBoost)	58
3.4. 모델 결과	61

4. 변수 중요도(Feature importance) 비교	63
4.1. 릿지 & 라쏘 회귀분석(Ridge & Lasso regression)	64
4.2. 랜덤 포레스트(Random Forest)	69
4.3. 익스트림 그래디언트 부스팅(XGBoost)	72
5. 주요 변수에 따른 예후 예측	76
 IV. 고 찰	 79
 V. 결 론	 82
 참고문헌	 83
 ABSTRACT	 89

표 차례

Table 1. Treatment methods according to cancer stage	3
Table 2. Systematic Review Search Term Steps and Counts(Survival) ..	5
Table 3. Systematic Review Search Term Steps and Counts(Recur)	7
Table 4. Korea Lung Cancer Registry collection variation	12
Table 5. Variables used in models	13
Table 6. Pathological morphology classification criteria	15
Table 7. Tuned hyperparameters for each ML algorithm	20
Table 8. Confusion Matrix	23
Table 9. Evaluation of AUC	25
Table 10. Baseline characteristics by groups of live and death	28
Table 11. Baseline characteristics by groups of Recurrence	32
Table 12. Baseline characteristics by groups of BMI	36
Table 13. Odds Ratio Estimates for BMI on Simple Logistic Regression for survival	40
Table 14. Odds Ratio Estimates for BMI on Simple Logistic Regression for recurrence	41
Table 15. Odds ratio of variables associated with survival	42
Table 16. Odds ratio of variables associated with recurrence	43
Table 17. Baseline characteristics by training and testing set	47
Table 18. Best Hyperparameters of Ridge and Lasso regression	50
Table 19. Performance of Ridge and Lasso regression	51
Table 20. Best Hyperparameters of Random Forest	55
Table 21. Performance of Random Forest	55
Table 22. Best Hyperparameters of XGBoost	58
Table 23. Performance of XGBoost	59
Table 24. Performance comparison of Model	61
Table 25. Features Importance Top 10 rankings by Ridge regression	65
Table 26. Features Importance Top 10 rankings by Lasso regression	66

Table 27. Features Importance Top 10 rankings by Random Forest ...	70
Table 28. Features Importance Top 10 rankings by XGBoost	73

그림 차례

Fig. 1 Trends in 5-year relative survival rates for major cancer types	2
Fig. 2 Study population	10
Fig. 3 Research model	18
Fig. 4 K-fold Cross Validation	21
Fig. 5 Odds Ratios Plot	44
Fig. 6 Correlation Plot	46
Fig. 7 3-year survival ROC curve using Ridge regression	52
Fig. 8 3-year survival ROC curve using Lasso regression	52
Fig. 9 5-year survival ROC curve using Ridge regression	53
Fig. 10 5-year survival ROC curve using Lasso regression	53
Fig. 11 2-year Recurrence ROC curve using Ridge regression	54
Fig. 12 2-year Recurrence ROC curve using Lasso regression	54
Fig. 13 3-year survival ROC curve using Random Forest	56
Fig. 14 5-year survival ROC curve using Random Forest	56
Fig. 15 2-year recur ROC curve using Random Forest	56
Fig. 16 3-year survival ROC curve using XGBoost	59
Fig. 17 5-year survival ROC curve using XGBoost	59
Fig. 18 2-year recur ROC curve using XGBoost	60
Fig. 19 Model comparison on Various metrics of 5 year survival	62
Fig. 20 Model comparison on Various metrics of 2 year recur	62
Fig. 21 Feature Importance of 5-year survival using Ridge and Lasso	67
Fig. 22 Feature Importance of 2-year recur using Ridge and Lasso	68
Fig. 23 Feature Importance using Random Forest	71
Fig. 24 Feature importance of 5-year survival using XGBoost	74
Fig. 25 SHAP summary plot of 5-year survival using XGBoost	74
Fig. 26 Feature importance of 2-year recur using XGBoost	75
Fig. 27 SHAP summary plot of 2-year recur using XGBoost	75
Fig. 28 Box plot of predict survival by BMI group	77

Fig. 29 Box plot of predict survival by recur group	77
Fig. 30 Box plot of predict recurrence by BMI group	78
Fig. 31 Box plot of predict recurrence by radiation group	78

국 문 요 약

비소세포폐암 수술환자 예후 예측을 위한 설명 가능한 머신러닝 모델 적용

- 진단 시 체질량지수의 영향 분석 -

연구 배경:

세계보건기구(WHO)에 따르면 폐암은 세계적으로 약 220만 명이 매년 발생하고, 180만 명 정도가 사망하는 사망률 1위의 암이다. 우리나라 2022년 주요 암 사망 통계 자료를 보면, 전체 사망자(37만 2,939명) 가운데 22.4%(8만 3,378명)가 암으로 사망했으며 폐암 사망자는 22.3%(1만8,548명)로 국내 암 사망자 수에서 1위를 차지하고 있다. 폐암은 조직학적인 형태에 따라 소세포암과 비소세포암으로 구분되며, 보통 폐암의 85% 이상은 비소세포성 폐암으로 소세포폐암에 비해 상대적으로 성장 속도가 느리며, 주변 조직으로 침범한 후 전신으로 전이되는 경향이 있다. 따라서 수술이 가능한 초기 단계에 발견하는 것이 무엇보다 중요하며, 수술 후 2년 내 재발도 많아 지속적인 관리가 필요한 질환이다.

본 연구의 목적은 전통적인 통계기법 및 다양한 머신러닝 방법론을 활용하여 모델의 성능을 평가하고, 완전 절제 수술을 받은 비소세포폐암 환자의 생존 및 재발을 예측하는데 있어 중요한 변수, 특히 최근 주목받고 있는 예측인자인 체질량지수(BMI) 변수의 중요도와 영향력을 파악하는 것이다. 이를 통해, 체질량지수가 비소세포폐암 환자의 생존 및 재발 예측 모델에서 어떤 역할을 하는지, 그리고 이를 기반으로 한 임상적 의사결정 및 환자 관리 전략에 대한 깊이 있는 통찰을 제공하고자 한다.

연구 방법:

폐암병기조사데이터 2014년부터 2017년까지 비소세포폐암(NSCLC)으로 완전 절제(R0)를 받은 환자를 후향적으로 분석하였다. 머신러닝 모델을 구축하기 위해 규제 선형모델인 능형(Ridge) 회귀분석, 라쏘(LASSO) 회귀분석, 앙상블 모

델의 랜덤 포레스트(RF)와 엑스트림 그라디언트 부스팅(XGBoost) 분석을 수행하였다. 모형의 성능 평가는 민감도, 특이도, 정확도, 정밀도, F1 점수, AUC를 기반으로 5겹 교차검증을 통해 수행하였다. SHAP(Shapley Additive Explanation) 방법을 통해 변수들의 중요도를 평가하였고, 각 모델은 그리드 검색을 사용하여 하이퍼 매개변수를 최적화한 다음 설명 가능한 방법을 사용하여 결정하였다.

연구 결과:

모든 모델에서 AUC 0.8 이상의 우수한 성능을 보였으며, 3년, 5년으로 나누어 비교한 생존 예측에서는 5년 생존 예측 성능이 가장 높게 나타났다. 모델 간 성능의 큰 차이는 없었으나 전통적인 통계적 방법에서는 릿지(Ridge) 회귀분석이 머신러닝 모델에서는 XGBoost가 높은 성능을 보였다. 연구에서 확인된 생존 예측에 가장 큰 영향력을 준 주요 변수는 재발 여부였으며, 연령, 조직학적 T 병기 및 치료, 진단 시 체질량지수를 수술 후 생존에 영향을 미치는 주요 예후 인자로 식별하였다. 특히, 진단 시 체질량지수에서 저체중 그룹은 낮은 5년 생존율을 보였고 과체중/비만 그룹은 높은 5년 생존율을 보였다. 재발 예측에 있어서 체질량지수의 영향이 높지는 않았으나 다른 그룹에 비해 저체중 그룹에서의 수술 후 2년 재발률이 다소 높게 나타났다. 재발 예측의 주요 인자로는 방사선 치료 및 보조 항암 치료, 조직학적 T 병기, 흉막 침범임을 확인하였다.

결론:

본 연구는 비소세포폐암 환자의 수술 후 생존과 재발 가능성을 우수한 성능으로 예측하는 머신러닝 모델을 구축하였으며, 릿지(Ridge)와 라쏘(Lasso) 로지스틱 회귀분석과 같은 전통적인 통계적 방법과 랜덤 포레스트(RF)나 XGBoost 같은 더 복잡한 머신러닝 모델 간의 성능에는 큰 차이는 없었다. 또한, 생존 및 재발 예측에 있어 변수 중요도를 SHAP 방법을 결합하여 설명 가능한 머신러닝 모델을 적용하였으며 폐암에서의 비만역설 가능성을 확인하였다.

핵심어: 폐암, 비소세포폐암, 머신러닝, 생존예측, 재발예측, 체질량지수

I. 서론

1. 연구 배경 및 필요성

빅데이터 분야가 발전하면서 헬스케어는 개별 환자의 특성과 참여에 초점을 맞춘 정밀의료, 맞춤의학, 참여의학으로 패러다임을 전환하고 있다. 특히 건강증진, 질병예방, 건강관리, 맞춤형 의학을 활용한 효과적인 치료와 개입을 위해서는 신뢰할 수 있는 헬스케어 빅데이터를 기반으로 한 정확한 예측 모델이 중요하다. 최근에는 암의 조기진단, 중재, 치료 분야에서 의료 빅데이터의 활용이 활발히 이루어지고 있으며, 암 환자의 종합적인 맞춤형 치료를 위해서는 치료 후 환자의 예후를 예측하는 모델을 개발하는 것이 중요하다(Hung Viet Nguyen, 2023).

의료 환경에서 머신러닝 기반 접근 방식의 실제 적용 및 결과 해석에 대한 회의론은 여전히 남아있음에도 불구하고 의료분야의 다양한 영역에 걸쳐 머신러닝은 필수적인 기술로 인식되고 있으며, 의료 혁신과 품질 향상을 위한 핵심 요소로 자리 잡아 가며 빠른 속도로 증가하고 있다.

세계보건기구(WHO)에 따르면 폐암은 매년 220만 명의 신규 환자가 발생 되고 있으며, 이 중 180만 명 정도가 사망한다는 점에서 매우 치명률이 높은 암이다. 국내 상황도 다르지 않다. 통계청의 통계정보에 따르면, 암은 우리나라 사망원인 1위의 질환으로 2022년 기준 전체 사망자 372,939명 중 22.4%(83,378명)가 암으로 사망하였다(통계청, 2022). 특히, 폐암은 전체 암 사망자의 22.3%인 18,584명이 사망할 정도로 사망률이 가장 높았으며, 간암(12.2%), 대장암(11.0%), 췌장암(8.8%), 위암(8.6%)이 그 뒤를 따랐다. 암종별 최근 생존율을 살펴보면, 남녀 전체에서 폐암은 38.5%로 상대적으로 낮은 생존율을 보였다(Fig. 1). 이렇듯 폐암은 우리나라 전체 암종 중 사망률 1위, 발생률 3위에 해당하는 높은 질병 부담을 가진 암종이다(중앙암등록본부, 2023).

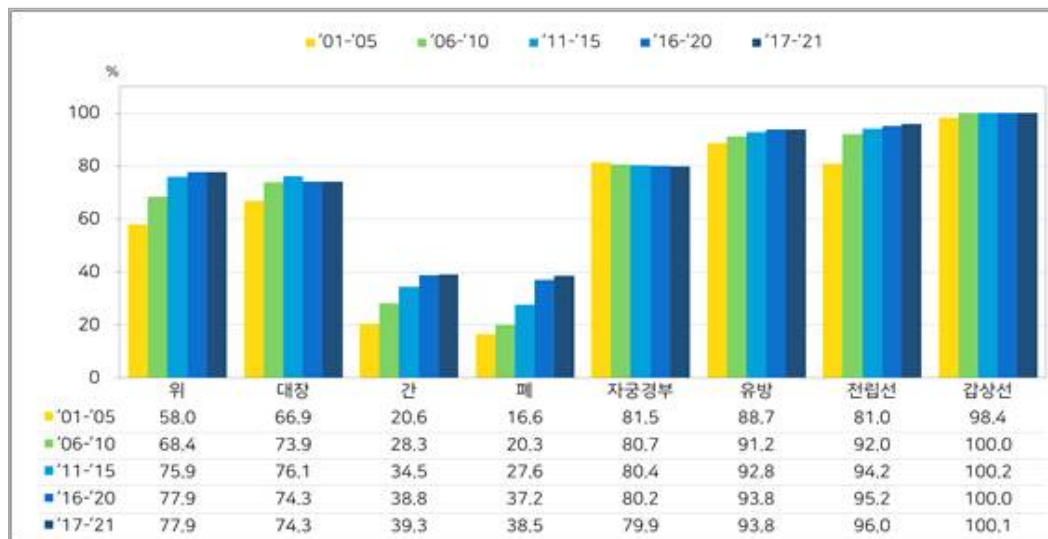


Fig. 1 Trends in 5-year relative survival rates for major cancer types

원발성 폐암은 암세포의 크기와 형태를 기준으로 비소세포폐암(Non-Small Cell Lung Cancer, NSCLC)과 소세포폐암(Small Cell Lung Cancer, SCLC)으로 구분한다. 폐암 가운데 80~85%는 비소세포폐암이며 악성 종양 관련 사망의 중요한 원인을 차지하고 있는 질환으로 비교적 천천히 진행하여 조기에 발견됐을 때는 수술을 통한 완치를 기대할 수 있다. 폐암의 치료 방법으로는 수술과 항암치료 그리고 방사선치료가 포함되며, 환자의 병기 및 치료 적응도에 따라 치료 방법과 조합이 달라질 수는 있으나 비소세포폐암 환자에게는 수술이 가장 효과적인 치료 방법이다. 주요 병기(Stage)에 따른 치료방법은 Table1과 같다. 일반적으로 사용되는 병기 분류 기준은 미국 공동 암 위원회(American Joint Committee on Cancer; AJCC)에서 보고한 TNM 병기 분류로써, 암의 크기 및 위치(T), 주변 림프절 전이 여부(N), 다른 장기로의 전이 여부(M)를 나타낸다. 병기에 따라 1기, 2기, 그리고 3A기 일부에서 절제술을 시행하고 보조적 항암 치료를 하기도 하며, 3B기부터는 수술적 치료가 어려워 항암, 방사선치료를 시행하고, 4기는 수술 없이 항암치료로만 진행한다.

Table 1. Treatment methods according to cancer stage

병기	기본적 치료방법
STAGE I	수술
STAGE II	수술 후 보조항암화학치료
STAGE IIIA	수술 또는 항암-방사선 동시치료
STAGE IIIB	항암화학치료 또는 항암-방사선 동시치료
STAGE IV	항암화학치료

폐암은 다른 암종에 비해 재발이나 전이가 빈번하다. 특히, 비소세포폐암 환자의 55~80%는 처음 진단을 받을 때 상당히 진행되었거나 주변 장기로의 전이를 동반하며, 근치적 절제 수술을 받은 환자의 20~50%가 재발을 보이는 것으로 보고되고 있다. 비소세포폐암은 수술 후 2년 내 재발도 많고, 세포의 형태나 침범 정도에 따라 차이가 존재하나 국소 재발보다는 다른 장기로의 원격전이가 많다고 최근 보고되어 초기 치료가 성공적으로 끝나도 재발과 새로운 암의 발생을 예방하고 치료 부작용을 극복하기 위한 추후 관리가 매우 중요하다. 따라서 수술 후 비소세포폐암 환자의 생존 및 재발 결과와 예후 인자를 연구하는 것은 여전히 중요하다 할 수 있으며, 특히 수술 후 예후를 예측함으로써, 적절한 사후 관리를 통한 재발을 방지하여 환자의 예후를 개선하는 것이 필요하다.

또한, 최근 연구들은 다양한 임상적, 병리학적, 그리고 분자생물학적 요인들이 비소세포폐암 환자의 생존에 영향을 미칠 수 있음을 시사하고 있다. 이 중 체질량지수(Body Mass Index, BMI)는 특히 주목받는 예측인자로, 환자의 영양 상태와 전반적인 건강 상태를 반영하는 중요한 지표로 여겨진다. 체질량지수는 다양한 암 종류의 발병 위험과 진행, 그리고 환자의 예후와 관련이 있음이 보고되었으나, 비소세포폐암 환자의 생존 예측에 있어서 체질량지수의 역할에 대해서는 여전히 명확하지 않은 부분이 많다. 특히, 비만 역설(obesity paradox)이라고 불리는 현상, 즉 비만이 특정 질병의 위험 요소로 간주 되는 동시에, 일부 질환의 예후에 있어서는 비만 환자가 더 나은 생존율을 보이는 상반된 결과에 관심이 높아지고 있다.

본 연구에서는 전통적 통계 방법 및 다양한 머신러닝 모델을 통해 임상적 및 인구학적 변수들을 포함하여, 체질량지수의 상대적 중요도를 정량화하고, 비소세포폐암 환자의 생존 및 재발에 미치는 영향을 평가할 예정이다. 이러한 분석을 통해, 체질량지수와 폐암 생존 및 재발 간의 관계를 이해하고, 체질량지수와 같은 요인이 환자의 예후 예측 모델에서 어떻게 활용될 수 있는지에 대한 이해를 심화시키며, 최종적으로는 개별 환자에 대한 맞춤형 치료 전략의 개발을 지원하고자 한다.

2. 선행 연구

2.1 생존 예측

관련 주제인 ‘폐암’과 ‘생존’ 그리고 ‘머신러닝’과 ‘예후인자’를 나타내는 용어들로 PubMed를 이용하여 검색하였다.

Table 2. Systematic Review Search Term Steps and Counts(Survival)

Search Terms	Counts
((lung cancer[Title/Abstract]) OR (lung neoplasms[Title/Abstract]) OR (lung tumor[Title/Abstract])) AND ((survival[Title/Abstract]) OR (survivors[Title/Abstract])) AND ((predict[Title/Abstract]) OR (prognostic factors[Title/Abstract]) OR (predictor[Title/Abstract])) AND ((machine learning[Title/Abstract]) OR (model[Title/Abstract]) OR (algorithm[Title/Abstract])) ((resection[Title/Abstract]) OR (resected[Title/Abstract]) OR (surgical[Title/Abstract]) OR (lobectomy[Title/Abstract]) OR (pneumonectomy[Title/Abstract])) ((NSCLC[Title/Abstract]) OR (non-small cell lung cancer[Title/Abstract]))	2,319
	534
	379

Zhi-Hui Wang 등(2022)은 SEER(Surveillance, Epidemiology, and End Results) 데이터베이스에서 2010년부터 2015년까지 수술 후 1기 비소세포폐암 환자 10,218명의 연구 대상을 추출하여 콕스 회귀 분석으로 확인된 독립적인 예후 인자를 노모그램에 통합하여 이들 개인의 3년 및 5년 전체 생존 기간을 예측하였다. 다변량 분석을 통해 연령, 성별, 조직학적 유형, 분화 등급, 수술 유형, T 병기 및 치료가 1기 비소세포폐암의 생존에 대한 중요한 예측 인자인 것으로 확인되었다. 훈련 세트에서 노모그램은 해외 적용이 가능한 TNM 병기

설정 시스템, 특히 3년 및 5년 생존에 대한 C-index(0.669 vs 0.580) 및 AUC(ROC 곡선 아래 면적)에서 탁월한 우수성을 나타냈다.

Zhenyu Yang 등(2020)은 2007년 1월부터 2016년 12월까지 쓰촨대학교 서중국병원 흉부외과에서 근치 절제술을 받은 비소세포폐암 N2기 환자 773명을 포함하여 임상병리학적 특징의 예후 가치를 탐색하기 위해 콕스 비례 위험 회귀 모델을 사용했으며, 종양-림프절 전이(TNM) 단계, 침범된 림프절 부위의 수, 전이, 림프 또는 혈관 침범 및 보조 화학요법은 독립적인 예후 인자였다.

Haike Lei 등(2022)은 진행성 비소세포폐암 환자 6,586명을 대상으로 전체 생존율을 예측하기 위해 노모그램을 모델링하고 검증하였으며, 이후 5가지 머신러닝모델(로지스틱 회귀분석, 랜덤 포레스트, XGBoost, 의사결정 트리, Light Gradient Boosting Machine)을 사용하여 생존상태를 예측하였다. 5가지 기계 학습 모델 중 랜덤 포레스트 모델의 정확도가 다른 모델보다 우수했다.

Long Jin 등(2023)은 2012-2017년까지 SEER(Institute Surveillance, Epidemiology, and End Results) 데이터베이스 내 절제 수술을 받은 3기 비소세포폐암 환자 4,617명을 대상으로 3개의 머신러닝 생존모델인 딥러닝 신경망, 랜덤 포레스트, 콕스 비례위험 모델을 훈련하고 검증하였다. 생존 예측 성능은 일치지수(c-index)를 사용하여 평가되었으며, 딥러닝 알고리즘의 C-index 평균은 0.843으로 랜덤 포레스트(0.678)와 콕스 비례위험 모델(0.678)보다 훨씬 높았다.

Jiahui Zhang 등(2019)은 병리학적 1기 및 2기 비소세포폐암으로 진단되고 신보조 화학요법이나 방사선요법 없이 근치적 절제술을 받은 총 443명의 환자를 모집하여 분석했다. 전체 생존율을 예측하기 위한 최종 노모그램에서 가장 유의미한 예측 변수를 선택하기 위해 로그 순위 테스트와 다변량 콕스 회귀분석을 사용했다. 또한 모델은 부트스트랩 방법으로 검증되었으며 C-index 및 보정 플롯으로 측정되었다. 전체 생존에 대한 종양 분화도, 샘플링된 림프절의 위치, 병리학적 T 및 병리학적 N과 같이 4개의 독립적인 예측인자가 확인되어 노모그램 묘사에 포함되었으며, 비교적 안정적인 차별(부트스트랩 보정 C-index =0.622,

95% CI: 0.572-0.672)과 우수한 보정을 보여주었다.

Koichi Fukumoto 등(2020) 일본 폐암 레지스트리에 추적된 비소세포폐암 데이터베이스에서 R0 절제된 비소세포폐암 환자 16,509명을 대상으로 로지스틱 회귀분석 및 콕스 비례 위험 모델을 사용하였으며, 연령, 성별 및 임상적 병기 외에도 수술 전 체질량지수가 완전히 절제된 비소세포폐암 환자의 독립적인 예후 인자로 확인하였다. 수술 전 저체중은 낮은 생존율과 관련되었다.

2.2 재발 예측

관련 주제인 ‘폐암’과 ‘재발’ 그리고 ‘머신러닝’과 ‘예후인자’를 나타내는 용어들로 PubMed를 이용하여 검색하였다.

Table 3. Systematic Review Search Term Steps and Counts(Recur)

Search Terms	Counts
((lung cancer[Title/Abstract]) OR (lung neoplasms[Title/Abstract]) OR (lung tumor[Title/Abstract])) AND ((recur[Title/Abstract]) OR (recurrence[Title/Abstract])) AND ((predict[Title/Abstract]) OR (prognostic factors[Title/Abstract]) OR (predictor[Title/Abstract])) AND ((machine learning[Title/Abstract]) OR (model[Title/Abstract]) OR (algorithm[Title/Abstract])) ((resection[Title/Abstract]) OR (resected[Title/Abstract]) OR (surgical[Title/Abstract]) OR (lobectomy[Title/Abstract]) OR (pneumonectomy[Title/Abstract])) ((NSCLC[Title/Abstract]) OR (non-small cell lung cancer[Title/Abstract]))	322
	162
	121

Kar í 등(2023)은 2008년 1월부터 2018년 6월까지 수술을 받은 1기 비소세포 폐암 환자 268명을 대상으로 예후 요인(병리학적 단계, 조직학적 유형, 샘플링 된 종격동 림프절 위치 수, 절제 유형, 병변의 SUVmax)을 분석하였다. 콕스 비례 위험 모델, 랜덤 서바이벌 포레스트(RSF), DeepSurv의 세 가지 다른 방법을 사용하여 재발에 영향을 줄 수 있는 방법을 연구하고 이러한 방법의 성능을 Harrell의 C-index로 비교한 결과 훈련 세트에서는 DeepSurv가 세 가지 모델 중 가장 좋은 성능을 보였으며 훈련용 세트의 C-index는 0.832로 RSF(0.675), CoxPH(0.672)가 그 뒤를 이었다. 시험용 세트에서는 RSF가 세 모델 중 가장 좋은 성능을 보였으며 DeepSurv가 0.677, CoxPH 방법이 0.625로 나타났다.

이현주 등(2009)은 완전 절제술을 받은 비소세포폐암 환자 1,578명 중 조기 재발군과 3년 비 재발군을 비교하여 수술 후 1년 이내의 조기 재발에 영향을 미치는 요인을 평가하였다. 선암종과 편평 세포 암종을 독립적으로 분석했으며, 다중 로지스틱 회귀 분석을 사용하여 재발에 대한 독립적인 임상 예측 변수를 식별하고 콕스의 비례 위험 회귀 방법을 사용하여 임상 예측 모델을 개발하였다. 무작위로 환자를 훈련 및 테스트 하위 집합으로 나누어 분석하였으며, 병리학적 병기, 종양 세포 유형, 종양의 분화도, 보조 항암요법 및 연령은 다변량 분석에 중요한 요소로 확인되었다.

Dariusz Adam Dziedzic 등(2016)은 2009년부터 2014년 1월까지 여러 기관에서 비소세포폐암으로 완전 절제술을 받은 환자를 대상으로 재발 빈도에 대한 영향에 대해 임상병리학적 변수를 평가하고, 콕스 비례 회귀 위험 모델 분석을 통해 비소세포폐암의 재발 연관성을 조사하였다. 분석에서는 재발 위험에 병리학적 병기의 증가, 고령, 폐절제술, 폐엽하절제술, 림프 및 혈관 침범, 내장 흉막 침범이 독립적으로 국소 및 원격 재발과 관련이 있는 것으로 나타났다.

3. 연구 목적

본 연구의 목적은 폐암병기조사데이터를 전통적 통계기법 및 다양한 머신러닝 기법으로 분석하여 비소세포폐암 수술 환자의 생존 및 재발 예측 모형을 개발하는 데 있다. 이를 통해 예후를 예측할 수 있는 주요 변수를 파악한 후 고위험군을 판별하고, 주관심 변수인 진단 시 체질량지수의 중요도와 영향력을 파악하고자 한다. 구체적인 연구 목적은 다음과 같다.

첫째, 비소세포폐암 수술환자의 생존과 재발에 영향을 미치는 주요 요인을 문헌 검토를 통해 선정하고, 주요 요인을 식별하는 근거를 제시하고자 한다.

둘째, 전통적 통계기법 및 다양한 머신러닝 기법을 활용하여 개발한 예측 모델의 성능을 비교 평가함으로써 최적의 모델을 수립하고자 한다.

셋째, 예측 모델의 결과와 의사결정 과정에 대한 설명을 제공할 수 있도록 변수 중요도(Feature importance)를 확인하여, 사용한 설명변수 중 수술을 받은 비소세포폐암 환자의 생존 및 재발 예측에 영향을 미치는 중요 요인들을 선별하고, 주관심 변수인 진단 시 체질량지수(BMI) 변수의 중요도와 영향력을 파악한다.

넷째, 구축한 모델을 통해 진단 시 체질량지수(BMI) 및 주요 변수에 대한 예측 생존율 및 재발률을 확인한다.

II. 연구 방법

1. 연구 자료 및 연구 대상자

본 연구는 중앙암등록본부에 등록된 전체 폐암 발생자를 대상으로 10%를 확률 표본 추출하여 후향적 의무기록조사를 통해 수집된 폐암병기조사데이터를 사용하였다. 2014년부터 2017년까지 발생 연도별 폐암 원시자료 데이터베이스에 2021년까지 사망 추적된 통계청 사망 데이터(사망일자, 사망원인)가 연계된 폐암 환자를 대상으로 하였다. 추출된 10,943건 중 소세포폐암 환자, 수술받지 않은 환자 및 결측치를 제외한 2,659건을 최종 분석 대상으로 진행하였다(Fig. 2).

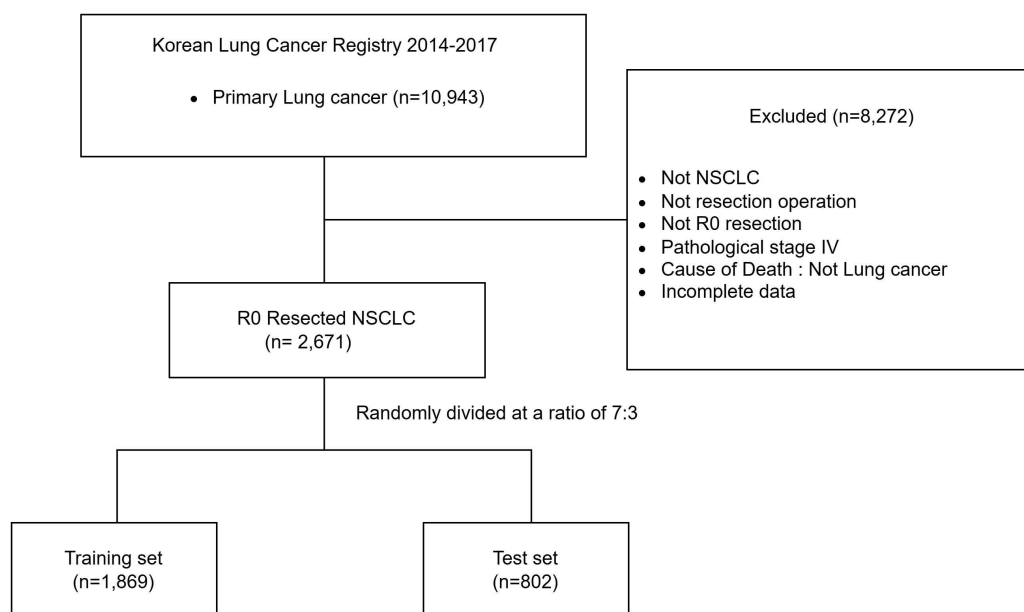


Fig. 2 Study population

1.1 폐암병기조사데이터

우리나라의 암등록통계사업은 주요 암 통계 생산을 통해 국가의 암 발생 및 생존, 유병 현황과 국가암관리 사업 계획 및 평가를 위한 주요 근거자료를 생산하고 있으나, 암이 발생한 시점의 기본적인 항목만을 수집하고 있어 자료의 활용적 측면에서는 제한이 있을 수 밖에 없다. 이에 중앙암등록본부는 암등록통계에서 암감시(Cancer Surveillance)로의 확대를 위해 협력병기조사사업을 2011년 시범사업을 거쳐 매년 조사를 수행하고 있다. 협력병기조사사업은 모든 암종 병기분류 시스템 간 호환이 가능하고, 암 병기와 관련된 정보뿐만 아니라 치료 및 예후와 관련된 정보를 제공하는 협력병기(Collaborative Staging)를 우리나라의 암등록체계에 도입하기 위한 사업으로 중앙암등록본부에 등록된 주요 호발암의 암종에 대해 확률표본 추출하여 후향적 의무기록 조사를 통해 수행한다. 이를 시작으로 2013년부터는 간암병기조사사업을, 2017년부터는 우리나라 전체 암종 중 사망률 1위, 발생률 3위에 해당하는 높은 질병 부담을 가진 암종인 폐암에 대한 폐암병기조사사업이 시작되었다. 폐암병기조사사업은 중앙암등록본부에 등록된 폐암 발생자를 대상으로 확률표본 추출하여 후향적 의무기록조사를 수행한 자료이다. 폐암병기조사사업의 목적은 국가암관리정책의 기준마련을 위하여 폐암환자의 위험요인, 세부 진단내용, 치료기술 및 행태와 관련된 정보 수집이 필요함에 따라 우리나라에서 발생하는 폐암의 발생 양상을 정확히 파악하여 향후 폐암 연구의 기본 자료로 활용하며, 나아가 폐암의 치료와 사망 및 생존에 관한 체계적인 통계자료를 구축하고자 하는데 있다. 암 환자수 기준 75% 이상을 포함하는 높은 순위의 39개 병원 및 13개 국가 지정 암센터 병원을 조사대상 병원으로 선정하여 중앙암등록본부에 등록된 전체 폐암 발생자를 계통추출 방법으로 10% 추출하여 조사가 수행되었으며, 수집하는 주요 변수로는 Table 4와 같이 일련번호, 등록연도, 나이, 성별 등의 기본정보와 키, 체중, 내원동기, 증상과 같은 문진사항과 흡연력, ECOG PS, 폐기능 검사, 조직학적 진단정보, 수술정보, TNM Stage, 치료정보, 재발/전이, 치료의 완전성 등이 있다(중앙암등록본부, 자료제공서비스).

Table 4. Korea Lung Cancer Registry collection variation

구분	변수 종류
기본정보	일련번호, 등록년도, 나이, 성별
문진사항	키, 체중, 내원동기, 증상
흡연력	흡연, 흡연량, 흡연기간, 금연 시작 연도
Performance	
Status (ECOG PS)	
폐기능검사	FVC(L), FEV1(L), DLco (mL/min/mmHg), DLco (%)
조직/병리	조직학적 진단연월, 조직학적 진단방법, 조직진단명, EGFR mutation, ALK IHC, ALK FISH
근치적 수술 유무	
TNM stage	영상진단소견(Clinical Stage) - TNM 정보, 종양최대직경, 종양의 위치, Main bronchus 침범유무, 폐쇄성 폐렴 유무, Separate tumor nodule 여부, 종양 침범 정도, 국소림프절 침범 여부(N1, N2, N3), 원격전이 여부(M1a, M1b), 전이개수, SCLC stage 조직학적 진단소견(Pathological Stage) - TNM 정보, 종양최대직경, 흉막침범, Main bronchus 침범유무, 폐쇄성 폐렴 유무, Separate tumor nodule 여부, 종양 침범 정도, 국소림프절 침범 여부(N1, N2, N3)
수술진단정보	수술조직진단명, 종양의 위치, 종양의 개수
치료정보	수술 - 수술 시행 연월, 수술 목적, 수술명, 수술접근법, 부가 절제술, Mediastinal LN resection, 완전절제술, 합병증 방사선치료 - 방사선치료 시행/종료 연월, 치료목적, 치료부위, 치료방법 항암치료 - 치료 시행/종료 연월, 투여 횟수, 항암치료 종류, 항암치료 약물
재발/전이	재발/전이 유무, 재발/전이 확인 연월, 최종 내원 연월
치료완전성	

2. 변수 정의

비소세포폐암 수술환자의 예후에 영향을 주는 요인은 선행연구 고찰을 통하여 위험 요인으로 알려진 변수들을 포함하였으며, 사용된 변수는 Table 5와 같다.

Table 5. Variables used in models

Variables	Value
인구사회학적 변수	
나이	>=65, <65
성별	Female, Male
생활습관 변수	
진단 시 체질량지수(BMI)	Normal, Overweight/Obese, Underweight
흡연여부	Non smoker, Current smoker, Past smoker
임상적 변수	
진단경로	Without symptoms, With symptoms
ECOG PS	0-1, 2-4, Unknown
FEV1/FVC(%)	>=70, <70
EGFR mutation	Positive, Negative, Not tested
흉막침범 여부	Negative, Positive
수술명	Lobectomy, Bilobectomy, Pneumonectomy, Segmentectomy, Sleeve lobectomy, Wedge resection
조직학적 종양크기	1-2, 2-3, 3-4, 4-5, >=5, <=1
종양개수	Single, Multiple
수술조직 진단명	Adenocarcinoma, Other, Squamous
조직학적 병기	I, II, III
조직학적 T	T1, T2, T3, T4
조직학적 N	N0, N1-3
방사선 치료여부	None, Unknown, Yes
전보조적 치료여부	None, Unknown, Yes
재발여부	No recur, Recur
사망여부	Live, Death

1) 인구학적 변수

연속형 변수인 나이는 65세 미만과 65세 이상으로 그룹을 나누어 범주형으로 변환하여 분석에 포함하였으며 성별은 남자와 여자로 나누어 구분하였다.

2) 생활 습관 변수

진단 시 측정된 신장과 몸무게 변수는 진단 시 체질량지수(BMI)로 산출하여 포함하였으며, 아시아-태평양 비만 진단 기준을 참고하여 $bmi < 18.5$ 는 저체중(Underweight), $bmi < 25$ 는 정상체중(Normal weight), $bmi < 30$ 은 과체중(Overweight), $bmi \geq 30$ 은 비만(Obey)의 구분 중 비만 그룹의 대상자가 적어 과체중과 비만 그룹을 합하여 분석하였다. 흡연 그룹은 비흡연, 과거흡연(1년 이상 금연자), 현재 흡연 중으로 구분하였다.

3) 임상적 변수

내원동기는 증상발현과 우연한 발견으로 나누어지는데, 우연한 발견은 증상이 없는 그룹으로 정의하여 사용하였다. ECOG PS의 경우 0~1과 2~5, 9(불명) 세 그룹으로 나누어 구분하였다. PEV1(%)는 FEV1을 FVC에 100을 곱하여 나눈 값으로 70 이상인 그룹과 70 미만인 그룹으로 나누어 구분하였으며, EGFR mutation은 세포의 성장, 분화 및 생존에 중요한 역할을 하는 수용체 단백질로 음성(-)과 양성(+), 시행안함 세 그룹으로 나누어 구분하였다. 병기(Stage)는 수술환자만을 모집단으로 하여 분석하므로, 조직학적 병기를 사용하였으며, IA, IB, IIA, IIB, IIIA, IIIB, IV로 구분 되어지나 IIIB와 IV를 분석대상에서 제외한 뒤, 최종 I, II, III로 구분하였다. 조직학적 T병기는 T1, T2, T3, T4로 구분하였으며, 조직학적 N병기는 림프절 전이가 없는 N0와 림프절 전이가 있는 N1-3 두 그룹으로 나누어 구분하였다. 폐암 병기 기준은 American Joint Committee on Cancer(AJCC) 7th edition을 따라 분류되었다. 최근 연구에서 생존과 재발에 중요 인자로 보고되고 있는 흉막 침범 여부는 침범이 있는 군과 없는 군으로 나누어 구분하였으며, 종양 개수의 경우 종양의 개수가 1개인 경우, Single로 2개 이상인 경우 multi로 나누어 구분하였다. 수술명은 근치적 목적으로 시행한 Lobectomy와 Segmentectomy, Wedge resection, Sleeve lobectomy, Pneumonectomy, Bilobectomy로 나누었으며, 조직학적 형태학명은 단일값이 아닌 다중값도 존재하였기에 다중값에 대해 Table 6과 같이 정리

하여 사용하였다.

Table 6. Pathological morphology classification criteria

Criteria	Morphology
Adenocarcinoma	Adenocarcinoma + NSCLC_NOS
Squamous carcinoma	Squamous carcinoma + NSCLC_NOS
	Adenocarcinoma + Squamous carcinoma
	Adenocarcinoma + Squamous carcinoma + NSCLC_NOS
	Large cell + NSCLC_NOS
Other	Adenocarcinoma + Large cell
	Adenocarcinoma + carcinoid
	Large cell + carcinoid
	Adenocarcinoma + Squamous carcinoma + NSCLC_NOS

항암치료는 항암 보조적 치료를 추가로 시행한 그룹과 시행하지 않은 그룹으로 구분하였고, 방사선치료는 방사선치료를 시행한 그룹과 시행하지 않은 그룹으로 나누어 구분하였으며 각 치료그룹에는 알 수 없음도 포함하여 분석하였다.

재발 여부는 폐암 진단 후 조사 시점인 3년까지의 추적 관찰된 데이터이며, 사망 여부는 통계청 사망 정보가 업데이트된 기준으로 최종 날짜인 2021.12.31.로 적용하여 Follow time을 계산하였다.

3. 통계 분석 방법

머신러닝은 의료분야에서 생존시간을 예측하는데 주로 활용한다. 본 연구에서는 비소세포폐암 수술환자의 생존율을 예측하기 위해 기계학습을 활용하는 연구를 검토하였다. 체계적인 문헌검토를 통해 Pubmed 데이터베이스를 검색하여 최종 생존과 관련된 연구는 379개, 재발과 관련된 연구는 121개의 연구를 확인할 수 있었다. 이 중 생존과 관련된 주요 위험 인자는 연령, 성별, 병리학적 N 병기가 가장 많이 선택되었으며, 수술 전 체질량지수(BMI)도 완전히 절제된 비소세포폐암 환자의 독립적인 예후 인자로 확인할 수 있었다(Koichi Fukumoto, 2020). 재발과 관련된 위험 인자로 연령, 조직학적 병기 및 분화도, 림프 전이와 흉막 전이가 가장 많이 선택되었다. 예측 모델로는 고식적 통계 방법인 콕스 비례 위험 모델이, 머신러닝 모델로는 랜덤 포레스트 등 앙상블 모델이 가장 많이 사용된 주요 모델이었다.

비소세포폐암 수술 후 사망한 환자군과 생존군의 특성을 비교하기 위해 선행연구를 통해 확인한 주요 변수를 인구사회학적 변수(성별, 나이), 생활 습관 변수(진단 시 체질량지수), 임상학적 변수(흡연력, 입원 시 환자상태, 폐기능검사, 조직학적 진단명, EGFR mutation, 조직학적 진단조건, 종양크기, 재발/전이 유무, 수술명, 방사선치료여부, 보조적 치료여부)로 나누어 통계적으로 차이가 있는지 Chi-square test를 수행하였으며, 단순 로지스틱 회귀분석을 통해 진단 시 체질량지수와 사망 및 재발과의 연관성을 확인하였다. 이후, 주요 임상변수들을 포함한 다중 로지스틱 회귀분석을 시행하여 유의미한 임상 특징을 확인할 수 있었다.

폐암병기조사데이터를 다양한 머신러닝 알고리즘을 활용하여 훈련용 데이터(Training data)로 머신러닝 모델을 생성하고 검증용 데이터(Test data)로 비소세포폐암 수술 후 3년, 5년으로 나누어 생존율을 예측하고, 생존율에 영향을 미치는 요인을 분석하는 연구모형을 Fig. 3과 같이 설계하였다. 또한, 비소세포폐암의 재발은 절제 후 첫 2년 동안 가장 빈번하며, 재발의 80% 이상이 이 기간 내에 발생하는 것으로 보고되어 수술 후 2년 재발 예측 모형을 추가적으로 분

석하였다(Maeda R, 2010).

머신러닝 분석 시에는 데이터 전처리, 데이터 분할, 하이퍼파라미터의 튜닝과 교차검증, 훈련데이터의 학습과 검증데이터 적용 및 평가의 순으로 진행하였다. 범주형 데이터의 경우 범주값을 숫자형으로 변경하여 사용하였으며, 그대로 사용 시 숫자의 연속성으로 인한 문제가 발생하기 때문에 연속성을 제외하기 위해 더미변수를 생성하여 적용하였다.

비소세포폐암 수술환자의 예후 예측 모형 개발에 있어 전체 대상자의 70%를 훈련용 데이터(Training data), 30%를 시험용 데이터(Test data)로 나누어 선형모델인 능형 회귀분석(Ridge)과 Lasso 회귀분석, 앙상블 모델인 랜덤 포레스트(RF)와 익스트림 그래디언트 부스팅(XGBoost) 알고리즘을 적용하여 분석하였으며 각 모델의 최적의 하이퍼파라미터를 조정하고 예후 예측을 위한 최상의 모델을 선택하기 위해 5겹 교차검증(5-fold CV)을 수행하였다.

생존과 재발 예측 모델에서의 변수중요도를 확인하였으며, 설명 가능한 방법인 SHAP(Shapley Additive Explanations)를 사용하여 각 변수가 모델의 예측에 기여하는 영향도를 평가하였다.

분석에 사용된 통계분석 프로그램은 R4.3.2이며, 능형 회귀분석(Ridge)과 라쏘(Lasso) 회귀분석에는 'glmnet' 패키지를 사용하였으며, 익스트림 그래디언트 부스팅 모델은 'xgboost' 패키지를, 랜덤 포레스트 모델은 'RandomForest' 패키지를 사용하여 분석하였다. 성능평가에는 'caret', 'pROC' 패키지를, 시각화를 위해서는 'fmsb', 'ggplot2', 'survminer' 패키지를 활용하였다.

이 연구는 연세의료원 연구심의위원회(Institutional Review Board, 이하 IRB) 심의면제 승인을 받아 진행되었다(과제승인번호 : 4-2024-0571).

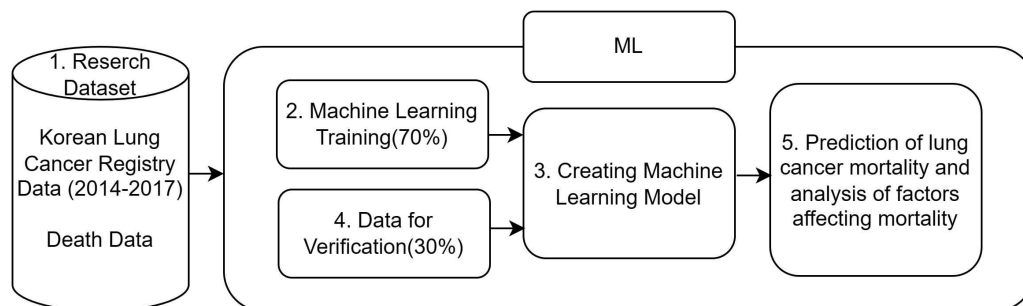


Fig. 3 Research model

3.1 머신러닝 알고리즘

3.1.1 로지스틱 회귀분석(Logistic regression)

로지스틱 회귀 분석은 특정 대상이 어떤 집단(범주)에 포함되는지 여부를 알 수 있게 해주는 분류를 진행하는 지도학습이다. 딥러닝과 머신러닝에서 자주 사용되는 함수 중 하나로 주로 이진 분류 문제에서 사용되며, 입력 변수의 값을 0과 1 사이로 제한하는 역할을 한다. 실수 범위의 입력을 확률로 매핑하는 함수로 사용되기 때문에, 일반적으로 0과 1 사이의 값으로 출력하며 이는 이진 분류 문제에서 어떠한 데이터가 어느 클래스에 속할 확률이 더 높은지를 예측하기 위해 사용된다.

3.1.2 Ridge & Lasso regression

머신러닝을 활용한 로지스틱 회귀 모형에서의 Penalty 정규화 방법은 회귀 계수의 절댓값을 계산하는 방식이며 L1-정규화인 Lasso와 L2-정규화인 Ridge로 나뉜다. Lasso는 필요하지 않은 변수의 계수를 0으로 만들 수 있으며 이는 특징 선택(feature selection)으로 이어져 모델이 더 단순하고 해석하기 쉬워진다. 결과적으로는 변수의 수를 줄여 중요한 변수만을 남기는 방법이다. Ridge는

변수들의 영향력을 줄이지만 완전히 제거하지는 않으며 모든 변수가 어느 정도의 정보를 제공한다고 가정한다. 변수들 사이에 상관관계가 있거나, 데이터에 다중공선성 문제가 있을 때 유용하다.

3.1.3 Random Forest (RF)

랜덤 포레스트는 분류나 회귀 문제를 해결하기 위한 강력하고 유연한 머신러닝 모델 중 하나이다. 이 모델은 앙상블 학습 방법의 일종이며, 여러 개의 결정 트리(Decision Trees)를 결합하여 사용한다. Breiman(2001)에 의해서 개발되었으며 이론적 설명이나 최종 결과에 대한 해석이 어렵다는 단점이 존재하지만, 예측력은 매우 높은 방법으로 알려져 있다. 모델 축소에 유용한 변수 중요도 정보를 제공하며, OOB 검증으로 추가적인 검증 데이터셋을 사용하지 않고도 모델의 일반화 성능을 추정하고 모델이 과적합 되었는지 여부를 판단하는데 도움이 된다.

3.1.4 eXtreme Gradient Boosting (XGBoost)

XGBoost는 eXtreme Gradient Boosting의 줄임말로, 고성능 그라디언트 부스팅(Gradient Boosting) 알고리즘을 구현한 라이브러리이다. 그라디언트 부스팅은 여러 개의 약한 학습기(weak learner), 특히 결정 트리(decision tree)를 결합하여 강력한 예측 모델을 만드는 앙상블 방법인 부스팅 알고리즘 중 하나로 회귀(Regression)와 분류(Classification) 문제를 모두 지원하며 기존 그라디언트 부스팅 방법을 발전시켜 속도와 성능을 크게 향상시켰다. 과적합(Overfitting)을 방지하기 위한 정규화(regularization) 기능을 갖추고 있으며 데이터의 크기에 상관없이 확장성이 뛰어나 대규모 데이터셋에 적합하다.

3.2 하이퍼파라미터 조정(Hyperparameter tuning) : Grid search

주요 머신러닝 알고리즘의 성능을 향상시키고 과적합을 방지하기 위해 각 머신러닝 알고리즘의 하이퍼파라미터를 조정한다. 하이퍼파라미터는 학습 과정을 제

어하는 데 사용되는 매개변수로, 각 모델의 모델링에 사용된 하이퍼파라미터 중 조정이 필요한 하이퍼파라미터를 선택하고 선택된 하이퍼파라미터의 가능한 값의 범위를 지정한다. 그리드 탐색 방법을 사용하여 최종 모델에서 사용되는 최적의 하이퍼파라미터 조합을 찾는다. 그리드 탐색 방법은 하이퍼파라미터 값의 그리드로 검색 공간을 정의하며, 그리드의 모든 위치를 평가하고 모든 하이퍼파라미터의 조합을 확인하여 최적의 조합을 식별한다. 각 머신러닝 모델의 하이퍼파라미터는 Table 7과 같다.

Table 7. Tuned hyperparameters for each ML algorithm

ML algorithm	Tuned hyperparameters
Ridge Regression Lasso Regression	lambda
Random Forest	mtry
XGBoost	Max depth / eta Minimum child weight / Subsample colsample_bytree

3.3 교차검증

교차검증(Cross Validation)이란 모델의 예측성능을 평가하기 위한 기법으로 학습용 데이터(Training data)와 시험용 데이터(Test data)를 나누는 기법이다. 교차검증에는 여러 종류가 있으나 보통은 학습용 데이터로 모델을 훈련하고 시험용 데이터로 모델을 검증한다. 하지만, 고정된 시험용 데이터를 사용하여 모델의 성능을 검증하고 수정하는 과정을 반복할 경우, 최종적으로 구축한 모델이 시험용 데이터에만 특화되어 잘 작동되는 문제가 발생할 수 있다. 즉 모델이 시험용 데이터에 과적합(Overfitting)되어, 다른 실제 데이터로 예측을 수행하면 정확도가 감소하는 결과가 나타난다. 이 문제를 해결하기 위해 사용되는 방법이

바로 교차검증이며, K-fold 교차검증은 가장 일반적으로 사용되는 교차검증이다. 주로 회귀 모델에 사용되며, 데이터가 독립적이고 동일한 분포를 가질 때 적용된다. K-fold 교차검증은 훈련용 데이터를 계속 변경하며 모델을 훈련시키는 방법으로 자세한 과정은 다음과 같다. 먼저 전체 데이터셋을 훈련용 셋과 시험용 셋으로 나눈다. 훈련용 셋을 훈련용 셋과 검증용 셋으로 사용하기 위해 k 개의 폴드로 나눈다. 첫 번째 폴드를 검증용 셋으로 사용하고 나머지 폴드들을 훈련용 셋으로 사용한다. 모델을 훈련시킨 뒤 첫 번째 검증용 셋으로 평가한다. 차례대로 다음 폴드를 검증용 셋으로 사용하며 3번을 반복한다. 총 K개의 성능 결과가 도출되며, 이들의 평균이 해당 학습 모델의 성능이다(Fig. 4). K는 직접 지정할 수 있으며, 일반적으로 5 또는 10을 많이 사용한다. K가 적어질수록 Bias는 커지고, K가 커질수록 Variance가 커지며 시간도 많이 소요하게 된다.

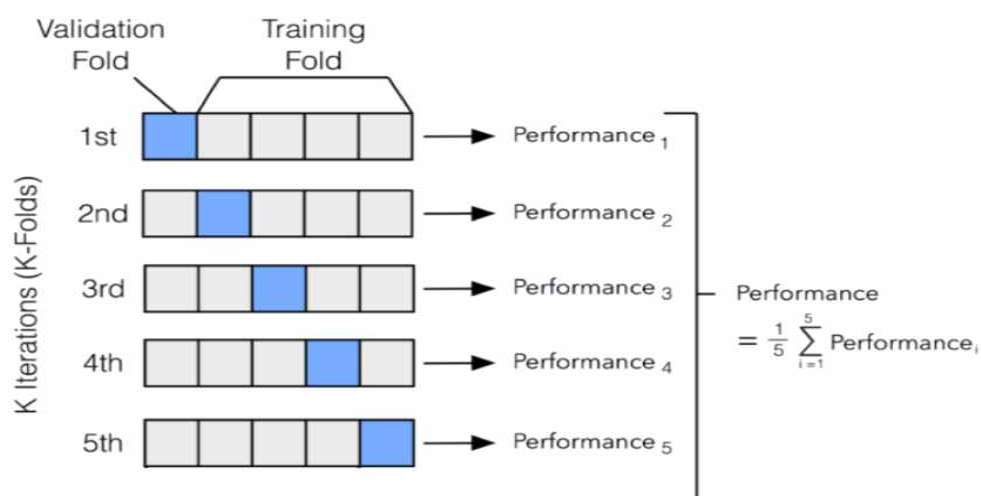


Fig. 4 K-fold Cross Validation

(출처: https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-validation/featured.png)

3.4 분류 모델 성능 평가

분류 모델의 평가 방법도 회귀 모형과 비슷하게 실제 데이터와 예측 결과 데이터가 얼마나 서로 비슷한지에 기반하지만, 단순히 이것만 가지고 판단하기에는 어려워 분류 모델에 대한 대표적인 성능 평가 지표로 혼동행렬(confusion matrix)을 사용한다. 혼동행렬(Confusion Matrix)은 분류의 예측 범주와 실제 데이터의 분류 범주를 교차표(Cross Table) 형태로 정리한 행렬로 이진 분류의 예측 오류가 얼마인지와 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타낸다(Table 8). 이를 통해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), 민감도(Sensitivity), 특이도(Specificity) 등 여러 가지 지표를 같이 볼 수 있다.

Table 8. Confusion Matrix

		예측	
		Positive	Negative
실제	Positive	TP	FN
	Negative	FP	TN

TP(True Positive)는 실제 값(Actual Class)이 True이며 예측된 값(Predicted Class) 또한 True인 경우를 말하며, FP(False Positive)는 실제 값(Actual Class)이 False이며 예측된 값(Predicted Class) 또한 False인 경우를 말한다. FN(False Negative)은 실제 값(Actual Class)이 False이며 예측된 값(Predicted Class)은 True인 경우이며, 제 1종 오류이다. TN(Ture Negative)은 실제 값(Actual Class)이 True이며 예측된 값(Predicted Class)은 False인 경우이며, 제 2종 오류이다.

3.4.1 민감도(Sensitivity)

민감도(Sensitivity)는 재현율(Recall)과 동일한 지표로, 실제 양성인 데이터 중 모델이 양성으로 예측한 데이터의 비율을 의미한다. 양성 결과를 정확히 예측하는 능력으로, 모델의 완전성을 평가하는 지표로 사용된다. 실제 Positive인데 Negative로 예측했을 때 상대적으로 더 문제가 되는 경우 중요성을 가진다.

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

3.4.2 특이도(Specificity)

특이도(Specificity)는 민감도와 정반대의 개념으로 실제 음성인 데이터 중 모델이 음성으로 예측한 데이터의 비율을 의미한다.

$$\frac{TN}{TN+FP}$$

3.4.3 정밀도(Precision)

정밀도(Precision)는 양성 예측의 정확도를 강조하는 지표로써, 모델이 양성으로 예측한 데이터 중 실제 양성인 데이터의 비율을 의미한다.

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

3.4.4 재현율(Recall)

재현율(Recall)은 실제 양성인 데이터 중 모델이 양성으로 예측한 데이터의 비율이다. 거짓양성(FP)을 최소화하고자 할 때 중요한 지표이다.

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

3.4.5 F1 score

F1 score는 정밀도와 재현율을 결합한 성능 지표로 정밀도와 재현율 두 지표 모두 어느 한쪽으로 치우치지 않는 수치를 나타내고 조화를 이룰 때 상대적으로

F1 Score는 높은 값을 가지게 된다. 즉, 정밀도와 재현율의 균형을 강조하는 지표이다.

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.4.6 정확도(Accuracy)

정확도(Accuracy)는 전체 데이터 중 올바르게 예측한 비율로, 예측 데이터가 실제 데이터와 얼마나 같은지 비교하여 산출된다. 일반적으로 많이 보지만 정확도만으로 판단하기는 어려우며, 예측하려고 하는 종속변수의 비율이 불균형할 때 정확도의 신뢰도는 낮아진다.

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

3.4.7 AUC(Area Under the Curve)

AUC(Area Under the Curve)는 여러 절단값에서의 민감도와 특이도의 관계를 나타내는 그래프인 ROC 곡선 아래의 면적을 의미하며, 그 값이 클수록 예측력이 우수하다고 평가할 수 있어, 모델의 성능을 종합적으로 평가하는 지표로 사용된다. Hosmer 등.[3]에 의하면 AUC 0.5~0.7은 Poor, 0.7~0.8은 Acceptable, 0.8~0.9는 Excellent로 간주된다(Table 8). 정확도, 민감도, 특이도, 정밀도, F1 Score, AUC는 0~1 사이의 값으로 표현되며, 이 값들이 1에 가까워질수록 모델의 성능이 높다는 것을 나타낸다.

Table 9. Evaluation of AUC

AUC range	0.5 ~ 0.7	0.7 ~ 0.8	0.8 ~ 0.9	0.9 ~ 1.0
Evaluation	Poor	Acceptable	Excellent	Outstanding

3.5 SHAP(Shapley Additive Explanation)

모델의 예측을 해석하기 위해 게임 이론의 샤플리 값에 기반한 방법으로 각 특성이 모델의 예측에 기여하는 정도를 설명함으로써 모델의 결정 과정을 투명하게 만든다. SAHP 값은 각 특성의 기여도를 수치화하여, 예측값에 대한 각 특성의 영향력을 평가할 수 있도록 한다. SHAP 값은 아래와 같이 특성 i 에 대한 샤플리 값 ϕ_i 를 계산하여 정의된다.

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup i) - f_x(S)]$$

ϕ_i : i 데이터에 대한 Shapley Value

N : 전체 집합

S : 전체 집합 중 i 번째 데이터를 뺀 나머지의 전체 부분 집합

$f_x(S \cup i)$: i 번째 데이터를 포함한 기여도

$f_x(S)$: i 번째 데이터를 뺀, 나머지 부분 집합의 기여도

Ⅲ. 연구 결과

1. 연구 대상자의 일반적 특성

1.1. 사망유무에 따른 대상자의 일반적 특성

연구 대상자 2,659명의 비소세포폐암 수술환자의 사망여부에 따른 일반적인 특성을 확인하였다(Table 10). 사망하지 않은 그룹이 2,107명이었고, 사망한 그룹은 552명이었다.

나이 그룹을 나누어 비교할 시, 65세 미만 그룹은 사망여부에 따라 사망하지 않은 그룹이 1,125명(53.4%), 사망한 그룹이 201명(36.4%)이었으며, 통계적으로 유의한 차이를 보였다($p<0.001$). 성별은 남자의 경우, 사망하지 않은 그룹이 1,173명(55.7%), 사망한 그룹이 434명(21.4%)였으며, 여자의 경우 사망하지 않은 그룹이 934(44.3%), 사망한 그룹이 118명(21.4%)으로 남성의 비소세포폐암 수술 후 사망 비율이 여성에 비해 더 높았으며 통계적으로 유의한 차이를 보였다($p<0.001$).

특히 주 관심 변수인 진단 시 체질량지수에서 정상체중 그룹에서의 사망은 359명(65.0%)이었으며, 과체중/비만 그룹에서의 사망은 162명(29.3%), 저체중 그룹에서의 사망은 31명(5.62%)으로 그룹 간 통계적으로 유의한 차이가 있음을 확인하였다($p<0.001$).

흡연 그룹을 나누어 비교할 시, 비흡연자는 사망하지 않은 그룹이 1,034명(49.1%), 사망한 그룹이 148명(26.8%), 현재 흡연자는 사망하지 않은 그룹이 533명(25.3%), 사망한 그룹이 214명(38.8%)이며, 과거 흡연자의 경우 사망하지 않은 그룹이 540명(25.6%), 사망한 그룹이 190명(34.4%)로 그룹 간 통계적으로 유의한 차이가 있음을 확인하였다($p<0.001$). 증상이 없는 그룹에서의 사망은 316명(57.2%)에 비해 사망하지 않은 그룹은 1,594(75.7%)이며, 증상이 있는 그룹에서

의 사망은 236명(42.8%)에 비해 사망하지 않은 그룹은 513명(24.3%)으로, 그룹 간 통계적으로 유의한 차이가 있음을 확인하였다($p<0.001$).

형태학적 분류의 경우에는 사망하지 않은 그룹에서는 Adenocarcinoma가 1,618명(76.8%), 사망한 그룹에서는 316명(56.2%)으로 반을 넘어서는 분포였다. Squamous cell carcinoma는 사망하지 않은 그룹에서 382명(18.1%), 사망한 그룹에서는 185명(33.5%)으로 통계적으로 유의한 차이를 보인다($p<0.001$). 조직학적 병기에 따른 그룹의 차이를 비교해 보았을 때, Stage I의 경우 사망하지 않은 군이 1,637명(77.7%)으로 가장 많은 분포를 차지하였으며, 사망한 군에서는 Stage I의 분포가 217명(39.3%), Stage II의 분포가 166명(30.1%), Stage III의 분포가 169명(30.6%)으로 병기 간 사망한 군의 분포는 비슷함을 알 수 있다. 이는 그룹 간에 통계적으로 유의한 차이가 있었다($p<0.001$).

수술 이외 치료와 관련하여, 보조적 항암치료의 경우 항암 치료를 받은 그룹에서의 사망은 191명(34.6%), 받지 않은 그룹에서의 사망은 276명(50.0%)으로 그룹 간 통계적으로 유의한 차이가 있음을 확인하였다($P<0.001$). 방사선 치료의 경우 방사선 치료를 받은 그룹에서의 사망은 209명(37.9%), 치료를 받지 않은 그룹에서의 사망은 276명(50.0%)로 받지 않은 그룹에서의 사망분포가 더 높았으며 그룹 간에 통계적으로 유의한 차이가 있었다($P<0.001$).

재발 여부에 따라서는 재발이 없는 경우 사망하지 않은 그룹에서 1,915명(90.9%)으로 많은 분포를 차지하고 있었으며, 재발이 있는 경우 사망한 그룹이 269명(48.7%), 재발이 없는 사망 그룹이 283명(51.3%)로 두 그룹 간에 통계적으로 유의한 차이가 있음을 확인하였다($P<0.001$).

Table 10. Baseline characteristics by groups of live and death

	Live N=2,107	Death N=552	p.overall
Age:			<0.001
>=65	982 (46.6%)	351 (63.6%)	
< 65	1,125 (53.4%)	201 (36.4%)	
Sex:			<0.001
Female	934 (44.3%)	118 (21.4%)	
Male	1,173 (55.7%)	434 (78.6%)	
BMI at diagnosis:			<0.001
Normal weight	1,317 (62.5%)	359 (65.0%)	
Overweight/Obese	744 (35.3%)	162 (29.3%)	
Underweight	46 (2.18%)	31 (5.62%)	
Smoke:			<0.001
Non smoker	1,034 (49.1%)	148 (26.8%)	
Current smoker	533 (25.3%)	214 (38.8%)	
Past smoker	540 (25.6%)	190 (34.4%)	
Pathway:			<0.001
Without symptoms	1,594 (75.7%)	316 (57.2%)	
With symptoms	513 (24.3%)	236 (42.8%)	
ECOG PS:			0.005
0-1	1,766 (83.8%)	439 (79.5%)	
2-4	34 (1.61%)	19 (3.44%)	
Unknown	307 (14.6%)	94 (17.0%)	
FEV1(%):			<0.001
>=70	1,508 (71.6%)	334 (60.5%)	
< 70	599 (28.4%)	218 (39.5%)	
EGFR Mutation:			<0.001
Positive	644 (30.6%)	112 (20.3%)	
Negative	791 (37.5%)	290 (52.5%)	
Not tested	672 (31.9%)	150 (27.2%)	

Table 10. Baseline characteristics by groups of live and death
(continued)

	Live N=2,107	Death N=552	p.overall
Operation name:			<0.001
Lobectomy	1,666 (79.1%)	441 (79.9%)	
Bilobectomy	47 (2.23%)	22 (3.99%)	
Pneumonectomy	26 (1.23%)	23 (4.17%)	
Segmentectomy	174 (8.26%)	20 (3.62%)	
Sleeve lobectomy	16 (0.76%)	16 (2.90%)	
Wedge resection	178 (8.45%)	30 (5.43%)	
Tumor size(cm):			<0.001
1-2	615 (29.2%)	82 (14.9%)	
2-3	633 (30.0%)	126 (22.8%)	
3-4	335 (15.9%)	122 (22.1%)	
4-5	175 (8.31%)	83 (15.0%)	
>=5	165 (7.83%)	131 (23.7%)	
<=1	184 (8.73%)	8 (1.45%)	
Tumor number:			<0.001
Single	2,060 (97.8%)	519 (94.0%)	
Multiple	47 (2.23%)	33 (5.98%)	
Pleural invasion:			<0.001
Negative	1,698 (80.6%)	328 (59.4%)	
Positive	409 (19.4%)	224 (40.6%)	
Pathological morphology:			<0.001
Adenocarcinoma	1,618 (76.8%)	316 (57.2%)	
Other	107 (5.08%)	51 (9.24%)	
Squamous	382 (18.1%)	185 (33.5%)	
Pathological stage:			<0.001
I	1,637 (77.7%)	217 (39.3%)	
II	295 (14.0%)	166 (30.1%)	
III	175 (8.31%)	169 (30.6%)	

Table 10. Baseline characteristics by groups of live and death
(continued)

	Live N=2,107	Death N=552	p.overall
Pathological T:			<0.001
T1	1,242 (58.9%)	138 (25.0%)	
T2	751 (35.6%)	307 (55.6%)	
T3	99 (4.70%)	94 (17.0%)	
T4	15 (0.71%)	13 (2.36%)	
Pathological N:			<0.001
N0	1,794 (85.1%)	302 (54.7%)	
N1-3	313 (14.9%)	250 (45.3%)	
Adjuvant CTx:			<0.001
None	1,220 (57.9%)	319 (57.8%)	
Unknown	471 (22.4%)	42 (7.61%)	
Yes	416 (19.7%)	191 (34.6%)	
Radiation Tx:			<0.001
None	1,310 (62.2%)	276 (50.0%)	
Unknown	617 (29.3%)	67 (12.1%)	
Yes	180 (8.54%)	209 (37.9%)	
Recurrence:			<0.001
No recur	1,915 (90.9%)	283 (51.3%)	
Recur	192 (9.11%)	269 (48.7%)	

1.2. 재발 여부에 따른 대상자의 일반적 특성

연구 대상자 2,659명의 비소세포폐암 수술환자의 재발 여부에 따른 일반적인 특성은 Table 11과 같다. 재발 그룹이 461명이었고, 재발이 안된 그룹은 2,198으로 연령, 진단 시 체질량지수 및 ECOG PS에서는 재발 그룹과 재발이 안된 그룹 사이의 유의미한 차이가 없음을 확인하였다.

사망여부($p<0.001$), 성별($p<0.001$), 흡연($p<0.001$), 진단경로($p<0.001$), 흉막침범여부($p<0.001$), EGFR mutation($p<0.001$), 종양크기($p<0.001$), 종양개수($p<0.001$), 수술명($p<0.001$), 조직학적 진단명($p=0.001$), 조직학적 진단소견($p<0.001$), 전보조적 치료여부($p<0.001$), 조직학적 T병기($p<0.001$), 조직학적 N병기($p<0.001$), FEV1비율 ($p=0.011$), 방사선 치료여부($p<0.001$)가 재발 그룹과 재발이 안된 그룹 사이의 유의미한 차이를 보였다.

Table 11. Baseline characteristics by groups of Recurrence

	No recur N=2,198	Recur N=461	p.overall
Age:			0.806
>=65	1,099 (50.0%)	234 (50.8%)	
< 65	1,099 (50.0%)	227 (49.2%)	
Sex:			<0.001
Female	909 (41.4%)	143 (31.0%)	
Male	1,289 (58.6%)	318 (69.0%)	
BMI at diagnosis:			0.106
Normal weight	1,374 (62.5%)	302 (65.5%)	
Overweight/Obese	686 (31.2%)	125 (27.1%)	
Underweight	59 (2.68%)	18 (3.90%)	
Smoke:			<0.001
Non smoker	1,374 (62.5%)	302 (65.5%)	
Current smoker	765 (34.8%)	141 (30.6%)	
Past smoker	59 (2.68%)	18 (3.90%)	
Pathway:			<0.001
Without symptoms	1,619 (73.7%)	291 (63.1%)	
With symptoms	579 (26.3%)	170 (36.9%)	
ECOG PS:			0.345
0-1	1,812 (82.4%)	393 (85.2%)	
2-4	45 (2.05%)	8 (1.74%)	
Unknown	341 (15.5%)	60 (13.0%)	
FEV1(%):			0.011
>=70	1,546 (70.3%)	296 (64.2%)	
< 70	652 (29.7%)	165 (35.8%)	
EGFR Mutation:			<0.001
Positive	619 (28.2%)	137 (29.7%)	
Negative	848 (38.6%)	233 (50.5%)	
Not tested	731 (33.3%)	91 (19.7%)	

Table 11. Baseline characteristics by groups of Recurrence(Continued)

	No recur N=2,198	Recur N=461	p.overall
Operation name:			<0.001
Lobectomy	1,734 (78.9%)	373 (80.9%)	
Bilobectomy	60 (2.73%)	9 (1.95%)	
Pneumonectomy	30 (1.36%)	19 (4.12%)	
Segmentectomy	173 (7.87%)	21 (4.56%)	
Sleeve lobectomy	19 (0.86%)	13 (2.82%)	
Wedge resection	182 (8.28%)	26 (5.64%)	
Tumor size(cm):			<0.001
1-2	627 (28.5%)	70 (15.2%)	
2-3	626 (28.5%)	133 (28.9%)	
3-4	369 (16.8%)	88 (19.1%)	
4-5	195 (8.87%)	63 (13.7%)	
>=5	196 (8.92%)	100 (21.7%)	
<=1	185 (8.42%)	7 (1.52%)	
Tumor number:			<0.001
Single	2,147 (97.7%)	432 (93.7%)	
Multiple	51 (2.32%)	29 (6.29%)	
Pleural invasion:			<0.001
Negative	1,758 (80.0%)	268 (58.1%)	
Positive	440 (20.0%)	193 (41.9%)	
Pathological morphology:			0.001
Adenocarcinoma	1,625 (73.9%)	309 (67.0%)	
Other	116 (5.28%)	42 (9.11%)	
Squamous	457 (20.8%)	110 (23.9%)	
Pathological stage:			<0.001
I	1,658 (75.4%)	196 (42.5%)	
II	327 (14.9%)	134 (29.1%)	
III	213 (9.69%)	131 (28.4%)	

Table 11. Baseline characteristics by groups of Recurrence(Continued)

	No recur N=2,198	Recur N=461	p.overall
Pathological T:			<0.001
T1	1,241 (56.5%)	139 (30.2%)	
T2	821 (37.4%)	237 (51.4%)	
T3	120 (5.46%)	73 (15.8%)	
T4	16 (0.73%)	12 (2.60%)	
Pathological N:			<0.001
N0	1,825 (83.0%)	271 (58.8%)	
N1-3	373 (17.0%)	190 (41.2%)	
Adjuvant CTx:			<0.001
None	1,293 (58.8%)	246 (53.4%)	
Unknown	490 (22.3%)	23 (4.99%)	
Yes	415 (18.9%)	192 (41.6%)	
Radiation Tx:			<0.001
None	1,428 (65.0%)	158 (34.3%)	
Unknown	629 (28.6%)	55 (11.9%)	
Yes	141 (6.41%)	248 (53.8%)	
Event:			<0.001
Live	1,915 (87.1%)	192 (41.6%)	
Death	283 (12.9%)	269 (58.4%)	

1.3. 체질량지수(BMI)에 따른 대상자의 일반적 특성

연구 대상자 2,659명의 비소세포폐암 수술환자의 주관심 변수인 진단 시 체질량지수에 따른 일반적인 특성을 알아보았다. 정상체중 그룹은 1,676명으로 가장 많았으며, 저체중 그룹이 77명, 과체중/비만 그룹이 906명이었다.

사망 여부로 나누어 비교할 시, 정상체중 그룹에서의 사망은 359명(21.4%), 생존은 1,317(78.6%), 저체중 그룹에서의 사망은 31명(40.3%) 생존은 46명(59.7%), 과체중/비만 그룹에서의 사망은 162명(17.9%), 생존은 744명(82.1%)였으며, 과체중/비만 그룹에서의 생존율이 82.1%로 가장 높게 확인되었다. 사망 여부별 체질량지수에 따른 특성은 통계적으로 유의한 차이를 보였다($p<0.001$).

재발 여부로 나누어 비교할 시, 정상체중 그룹에서의 재발은 302명(18.0%), 재발안함은 1,374명(82.0%), 저체중 그룹에서의 재발은 18명(23.4%) 재발안함은 59명(76.6%), 과체중/비만 그룹에서의 재발은 141명(15.6%), 재발안함은 765명(84.4%)로 재발 여부별 체질량지수에 따른 특성은 통계적으로 유의한 차이를 보이지는 않았다($p=0.106$).

그 외 흡연($p<0.012$), 진단경로($p=0.004$)가 체질량지수 그룹 사이의 유의미한 차이를 보였으며, 각 변수별 자세한 결과는 Table 12와 같다.

Table 12. Baseline characteristics by groups of BMI

	Normal weight N=1,676	Over weight/ Obese N=906	Under weight N=77	p.overall
Age:				0.711
>=65	846 (50.5%)	442 (48.8%)	38 (49.4%)	
< 65	830 (49.5%)	464 (51.2%)	39 (50.6%)	
Sex:				0.402
Female	656 (39.1%)	370 (40.8%)	26 (33.8%)	
Male	1020 (60.9%)	536 (59.2%)	51 (66.2%)	
Smoke:				0.012
Non smoker	750 (44.7%)	404 (44.6%)	28 (36.4%)	
Current smoker	469 (28.0%)	243 (26.8%)	35 (45.5%)	
Past smoker	457 (27.3%)	259 (28.6%)	14 (18.2%)	
Pathway:				0.004
Without symptoms	1,200 (71.6%)	667 (73.6%)	43 (55.8%)	
With symptoms	476 (28.4%)	239 (26.4%)	34 (44.2%)	
ECOG PS:				.
0-1	1,397 (83.4%)	742 (81.9%)	66 (85.7%)	
2-4	32 (1.91%)	20 (2.21%)	1 (1.30%)	
Unknown	247 (14.7%)	144 (15.9%)	10 (13.0%)	
FEV1(%):				0.585
>=70	523 (31.2%)	268 (29.6%)	26 (33.8%)	
< 70	1,153 (68.8%)	638 (70.4%)	51 (66.2%)	

Table 12. Baseline characteristics by groups of BMI(continued)

	Normal weight N=1,676	Over weight/ Obese N=906	Under weight N=77	p.overall
EGFR Mutation:				0.77
Positive	475 (28.3%)	258 (28.5%)	23 (29.9%)	
Negative	685 (40.9%)	361 (39.8%)	35 (45.5%)	
Not tested	516 (30.8%)	287 (31.7%)	19 (24.7%)	
Operation name:				.
Lobectomy	1,334 (79.6%)	717 (79.1%)	56 (72.7%)	
Bilobectomy	48 (2.86%)	20 (2.21%)	1 (1.30%)	
Pneumonectomy	33 (1.97%)	14 (1.55%)	2 (2.60%)	
Segmentectomy	116 (6.92%)	71 (7.84%)	7 (9.09%)	
Sleeve lobectomy	20 (1.19%)	10 (1.10%)	2 (2.60%)	
Wedge resection	125 (7.46%)	74 (8.17%)	9 (11.7%)	
Tumor size(cm):				0.07
1-2	434 (25.9%)	241 (26.6%)	22 (28.6%)	
2-3	482 (28.8%)	265 (29.2%)	12 (15.6%)	
3-4	294 (17.5%)	149 (16.4%)	14 (18.2%)	
4-5	157 (9.37%)	93 (10.3%)	8 (10.4%)	
>=5	200 (11.9%)	82 (9.05%)	14 (18.2%)	
<=1	109 (6.50%)	76 (8.39%)	7 (9.09%)	

Table 12. Baseline characteristics by groups of BMI(continued)

	Normal weight	Over weight/ Obese	Under weight	p.overall
	N=1,676	N=906	N=77	
Tumor number:				0.694
Single	1,624 (96.9%)	881 (97.2%)	74 (96.1%)	
Multiple	52 (3.10%)	25 (2.76%)	3 (3.90%)	
Pleural invasion:				0.448
Negative	1,280 (76.4%)	692 (76.4%)	54 (70.1%)	
Positive	396 (23.6%)	214 (23.6%)	23 (29.9%)	
Pathological morphology:				.
Adenocarcinoma	1,213 (72.4%)	669 (73.8%)	52 (67.5%)	
Other	106 (6.32%)	49 (5.41%)	3 (3.90%)	
Squamous	357 (21.3%)	188 (20.8%)	22 (28.6%)	
Pathological stage:				0.143
I	1,149 (68.6%)	654 (72.2%)	51 (66.2%)	
II	311 (18.6%)	138 (15.2%)	12 (15.6%)	
III	216 (12.9%)	114 (12.6%)	14 (18.2%)	
Pathological T:				.
T1	869 (51.8%)	478 (52.8%)	33 (42.9%)	
T2	654 (39.0%)	372 (41.1%)	32 (41.6%)	
T3	140 (8.35%)	43 (4.75%)	10 (13.0%)	
T4	13 (0.78%)	13 (1.43%)	2 (2.60%)	

Table 12. Baseline characteristics by groups of BMI(continued)

	Normal weight	Over weight/ Obese	Under weight	p.overall
	N=1,676	N=906	N=77	
Pathological N:				0.663
N0	1,312 (78.3%)	723 (79.8%)	61 (79.2%)	
N1-3	364 (21.7%)	183 (20.2%)	16 (20.8%)	
Adjuvant CTx:				0.414
None	984 (58.7%)	508 (56.1%)	47 (61.0%)	
Unknown	311 (18.6%)	191 (21.1%)	11 (14.3%)	
Yes	381 (22.7%)	207 (22.8%)	19 (24.7%)	
Radiation Tx:				0.102
None	1,015 (60.6%)	524 (57.8%)	47 (61.0%)	
Unknown	414 (24.7%)	256 (28.3%)	14 (18.2%)	
Yes	247 (14.7%)	126 (13.9%)	16 (20.8%)	
Recurrence:				0.106
No recur	1,374 (82.0%)	765 (84.4%)	59 (76.6%)	
Recur	302 (18.0%)	141 (15.6%)	18 (23.4%)	
Event:				<0.001
Live	1,317 (78.6%)	744 (82.1%)	46 (59.7%)	
Death	359 (21.4%)	162 (17.9%)	31 (40.3%)	

1.4. 단순 및 다중 로지스틱 회귀 분석

본 연구의 종속변수인 사망여부와 재발여부는 이항으로 표현되는 범주형 종속 변수로, 중요 변수를 식별하기 위해 단순 및 다중 로지스틱 회귀 분석을 수행하였다. 분석에는 후진제거법(Backward Elimination)을 시행하여 모든 변수를 포함한 상태에서 시작하여 각 단계마다 가장 중요도가 낮은 변수를 제거하였다. 변수들 간의 높은 상관관계가 있는 경우, 개별 변수의 영향력을 정확히 파악하기 어려워지고, 오차 추정치가 커져 모델의 해석을 어렵게 만들기 때문에 다중 공선성을 확인하였으며 분산 팽창 계수(VIF : Variation Inflation Factor)가 10 이상인 독립변수는 확인되지 않았다.

Table 13에 제시된 표는 진단 시 체질량지수(BMI)와 사망 발생에 대한 단순 로지스틱 회귀 분석 결과를 보여준다. 주관심 변수인 진단 시 체질량지수로 본 단순 로지스틱 회귀분석에서 체질량지수(BMI)가 환자의 사망에 미치는 영향을 확인할 수 있었지만, 재발에 미치는 영향은 확인 할 수 없었다(Table 13). 이 표는 BMI 범주(정상, 과체중/비만, 저체중)에 따른 생존의 오즈비(Odds Ratio, OR)를 나타낸다. 과체중이나 비만인 경우 정상 체중에 비해 사망할 오즈가 0.799배 낮으며, 통계적으로 유의미한 결과 (95%CI 0.649~0.981)를 보였다. 저체중인 경우 정상 체중에 비해 사망할 오즈가 2.473배 높으며, 이 결과 역시 통계적으로 매우 유의미하였다(95%CI 1.533~3.939).

Table 13. Odds Ratio Estimates for BMI on Simple Logistic Regression for survival

	OR	95% CI	p.overall
BMI at diagnosis (Ref. Normal weight)	0.273	0.243~0.306	<0.001
Overweight/Obese	0.799	0.649~0.981	0.033
Underweight	2.473	1.533~3.939	0.001

Table 14. Odds Ratio Estimates for BMI on Simple Logistic Regression for recurrence

	OR	95% CI	p.overall
BMI at diagnosis (Ref. Normal weight)	0.220	0.194~0.249	<0.001
Overweight/Obese	0.839	0.672~1.042	0.114
Underweight	1.388	0.786~2.339	0.236

다중 로지스틱 회귀 분석을 통해 유의미한 주요 임상 변수를 확인할 수 있었다. 선행연구를 통해 확인한 주요 임상 변수들을 포함하여 시행한 다중 로지스틱 회귀분석에서 사망과 관련해서는 연령, 성별, 진단 시 체질량지수_저체중, 진단경로, 흉막침막, 재발여부, 조직학적 형태학, 보조 항암치료, 조직학적 병기 및 방사선 치료가 사망에 영향을 주는 주요 임상 변수였다. 다른 변수를 통제하였을 때, 65세 미만 연령 그룹에 비해 65세 이상 연령 그룹에서 사망할 오즈가 2.04배 높으며 이는 통계적으로 유의하였다(95%CI 1.603~2.614). 또한 여성에 비해 남성이 사망할 오즈가 1.96배 높으며 통계적으로도 유의하였다(95%CI 1.48~2.611). 다만, 주관심 변수인 진단 시 체질량지수의 경우, 다른 변수를 통제하였을 때 정상체중에 비해 과체중/비만그룹은 0.897배 낮았으나 통계적으로 유의하지 않았지만, 저체중그룹에서 사망할 오즈가 2.53배 높았으며 통계적으로도 유의하였다(95%CI 1.384~4.575)(Table 15, Fig 5).

재발과 관련해서는 흉막 침범과 보조 항암치료 및 방사선 치료여부가 재발에 영향을 주는 중요 임상변수였다. 다른 변수를 통제하였을 때, 흉막 침범이 있는 환자가 없는 환자 대비 재발할 오즈가 1.7배 높으며 이는 통계적으로 유의하였다(95%CI 1.303~2.215). 또한 방사선 치료를 받지 않은 환자 대비 치료를 받은 환자에서 재발할 오즈가 14.572배 높으며 이 또한 통계적으로 유의하였다(95%CI 10.777~19.848)(Table 16, Fig 5)

Table 15. Odds ratio of variables associated with survival

	OR	95% CI	p.overall
Age (Ref. < 65)			
>= 65	2.052	1.609~2.627	<0.001
Sex (Ref. Female)			
Male	1.99	1.501~2.653	<0.001
BMI at diagnosis (Ref. Normal weight)			
Overweight/Obese	0.914	0.711~1.172	0.478
Underweight	2.532	1.381~4.578	0.002
Pathway (Ref. Without symptoms)			
With symptoms	1.609	1.26~2.052	<0.001
Pleural invasion (Ref. Negative)			
Positive	1.365	1.023~1.823	0.035
EGFR Mutation (Ref. Positive)			
Negative	1.367	1.001~1.875	0.051
Not tested	0.979	0.682~1.406	0.907
Pathological stage (Ref. I)			
II	1.39	0.879~2.179	0.155
III	1.959	1.014~3.76	0.044
Recurrence (Ref. No recur)			
Recur	6.464	4.843~8.659	<0.001
Pathological morphology (Ref. Adenocarcinoma)			
Other	1.617	1.018~2.54	0.039
Squamous cell	1.418	1.053~1.91	0.022
Adjuvant Ctx (Ref. None)			
Unknown	1.148	0.662~1.996	0.624
Yes	0.59	0.435~0.797	0.001
Pathological T (Ref. T1)			
T2	2.046	1.516~2.763	<0.001
T3	2.191	1.282~3.76	0.004
T4	1.954	0.673~5.653	0.216
Pathological N (Ref. N0)			
N1-3	2.381	1.443~3.957	0.001
Radiation Tx (Ref. None)			
Unknown	0.442	0.277~0.692	<0.001
Yes	1.24	0.889~1.723	0.203

Table 16. Odds ratio of variables associated with recurrence

	OR	95% CI	p.overall
Pleural invasion (Ref. Negative)			
Positive	1.7	1.303~2.215	<0.001
EGFR Mutation (Ref. Positive)			
Negative	0.824	0.618~1.1	0.188
Not tested	0.432	0.306~0.607	<0.001
Tumor size(cm) (Ref. < 1)			
1-2	4.426	1.952~11.516	0.001
2-3	2.246	1.033~5.666	0.059
3-4	3.482	1.633~8.674	0.003
4-5	2.901	1.321~7.372	0.014
>=5	3.594	1.588~9.325	0.004
Tumor number (Ref. Single)			
Multiple	2.191	1.214~3.888	0.008
Pathological stage (Ref. I)			
II	1.423	0.986~2.043	0.058
III	0.931	0.63~1.367	0.716
Adjuvant Ctx (Ref. None)			
Unknown	0.274	0.152~0.482	<0.001
Yes	1.081	0.793~1.471	0.622
Radiation Tx (Ref. None)			
Unknown	1.455	0.948~2.206	0.082
Yes	14.572	10.777~19.848	<0.001

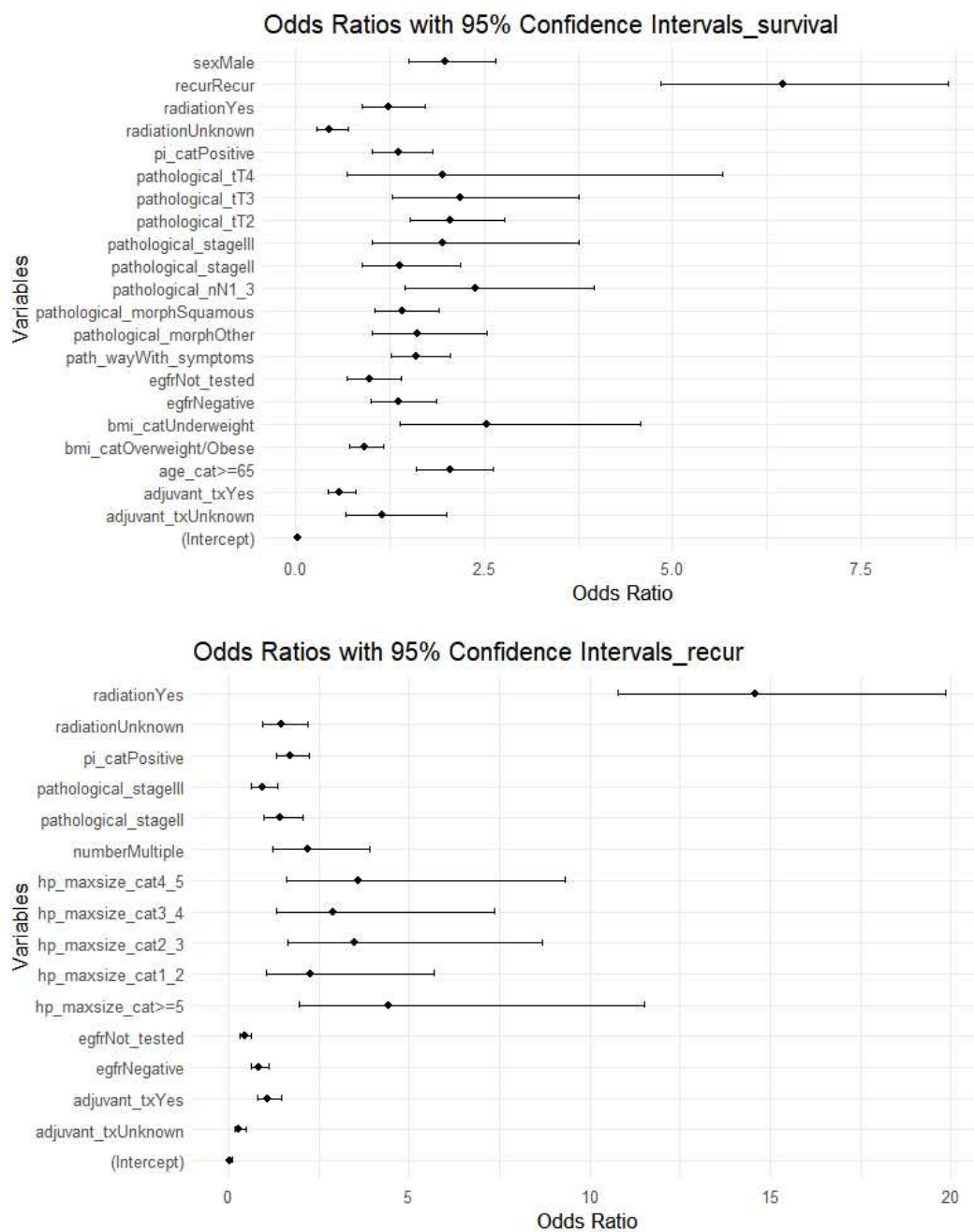


Fig. 5 Odds Ratios Plot

2. 예후 예측 모델 개발

비소세포폐암 수술 환자의 예후 예측 모델 개발을 통해 예후를 결정짓는 중요 인자를 찾고, 예측력을 알아보았다. 본 연구에서는 비소세포폐암 수술환자를 대상으로 종속변수를 사망과 재발로 설정하여 분석을 시행하였다. 모델의 과적합이나 불확실성을 방지하기 위해 먼저 변수 간의 상관관계를 조사한 후 모델을 개발하였다(Fig. 6).

일반적으로 하나의 자료를 분석할 때, 여러 가지 통계모형을 이용하여 분석하는 것이 바람직하다. 여러 가지 모형 중 최적인 모형을 선택하기 위해서는 여러 모형을 비교하고 평가해야 한다. 따라서 선형모델인 능형(Ridge) 회귀분석, Lasso 회귀분석, 앙상블 모델인 랜덤 포레스트(RF)와 XGBoost 모델을 사용하여 각 모델의 성능을 비교하였다.

모형의 성능 평가를 위해 AUC, 정확도, 민감도, 특이도, 정밀도, F1값을 사용하였으며, 선별된 환자 데이터의 70%(n=1,861)는 훈련용 셋으로 나머지 30%(n=798)는 시험용 셋으로 무작위 표본 추출을 하였고, 최적의 하이퍼파라미터를 찾아 K=5로 하는 K-fold 교차검증을 통해 모델 검증에 사용하여 예측 모델을 개발하였다.

훈련용 셋과 시험용 셋의 일반적인 특징을 살펴보았을 때, 두 그룹 간 통계적으로 유의미한 차이는 없었다(Table 17).

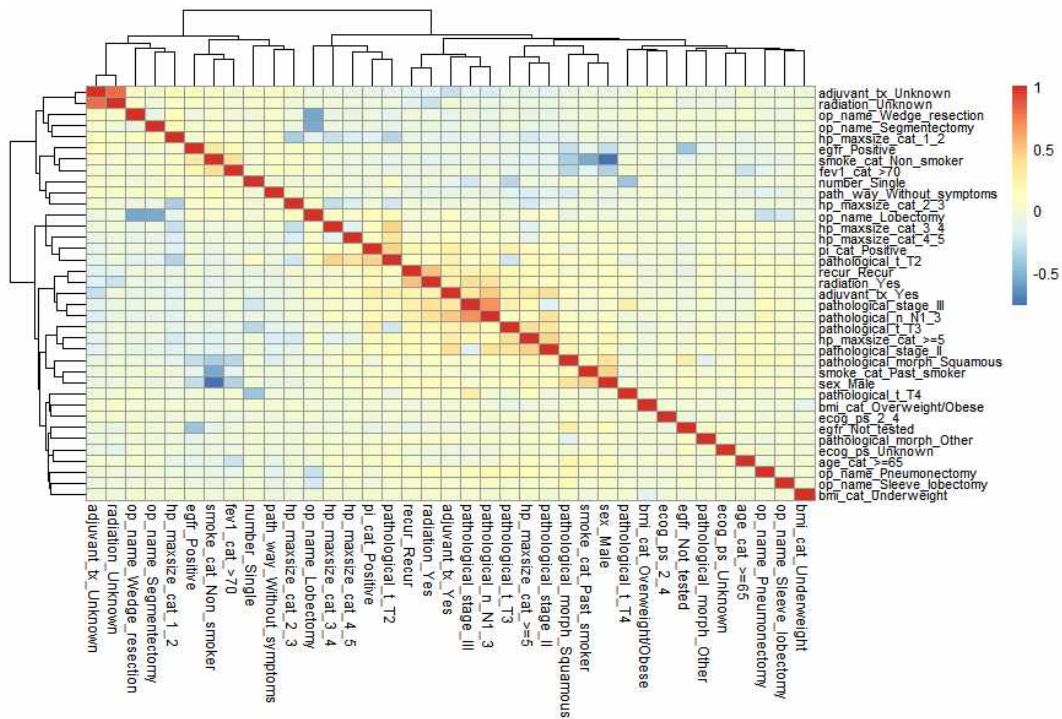


Fig. 6 Correlation Plot

Table 17. Baseline characteristics by training and testing set

	Training set N=1,861	Testing set N=798	p.overall
Event:			0.455
live	1,467 (78.8%)	640 (80.2%)	
Death	394 (21.2%)	158 (19.8%)	
Age:			0.836
< 65	931 (50.0%)	395 (49.5%)	
>= 65	930 (50.0%)	403 (50.5%)	
Sex:			0.985
Female	737 (39.6%)	315 (39.5%)	
Male	1,124 (60.4%)	483 (60.5%)	
BMI at diagnosis:			0.666
Normal weight	1,169 (62.8%)	507 (63.5%)	
Overweight/Obese	641 (34.4%)	265 (33.2%)	
Underweight	51 (2.74%)	26 (3.26%)	
Smoke:			0.408
Current smoker	536 (28.8%)	211 (26.4%)	
Non-smoker	824 (44.3%)	358 (44.9%)	
Past smoker	501 (26.9%)	229 (28.7%)	
Pathway:			0.313
With symptoms	513 (27.6%)	236 (29.6%)	
Without symptoms	1,348 (72.4%)	562 (70.4%)	
ECOG PS:			0.472
0-1	1,548 (83.2%)	657 (82.3%)	
2-4	40 (2.15%)	13 (1.63%)	
Unknown	273 (14.7%)	128 (16.0%)	
FEV1(%):			0.344
<= 70	561 (30.1%)	256 (32.1%)	
> 70	1,300 (69.9%)	542 (67.9%)	
EGFR Mutation:			0.621
Negative	746 (40.1%)	335 (42.0%)	
Not tested	584 (31.4%)	238 (29.8%)	
Positive	531 (28.5%)	225 (28.2%)	

Table 17. Baseline characteristics by training and testing set(continued)

	Training set N=1,861	Test set N=798	p.overall
Operation name:			0.51
Bilobectomy	55 (2.96%)	14 (1.75%)	
Lobectomy	1,463 (78.6%)	644 (80.7%)	
Pneumonectomy	34 (1.83%)	15 (1.88%)	
Segmentectomy	140 (7.52%)	54 (6.77%)	
Sleeve_lobectomy	21 (1.13%)	11 (1.38%)	
Wedge_resection	148 (7.95%)	60 (7.52%)	
Tumor size(cm):			0.744
<=1	129 (6.93%)	63 (7.89%)	
1-2	490 (26.3%)	207 (25.9%)	
2-3	522 (28.0%)	237 (29.7%)	
3-4	326 (17.5%)	131 (16.4%)	
4-5	188 (10.1%)	70 (8.77%)	
>=5	206 (11.1%)	90 (11.3%)	
Tumor number:			0.387
Multiple	52 (2.79%)	28 (3.51%)	
Single	1,809 (97.2%)	770 (96.5%)	
Pleural invasion:			0.583
Negative	1,424 (76.5%)	602 (75.4%)	
Positive	437 (23.5%)	196 (24.6%)	
Pathological morphology:			0.995
Adenocarcinoma	1,354 (72.8%)	580 (72.7%)	
Other	110 (5.91%)	48 (6.02%)	
Squamous cell	397 (21.3%)	170 (21.3%)	
Pathological stage:			0.975
I	1,300 (69.9%)	554 (69.4%)	
II	321 (17.2%)	140 (17.5%)	
III	240 (12.9%)	104 (13.0%)	
Pathological T:			0.865
T1	959 (51.5%)	421 (52.8%)	
T2	750 (40.3%)	308 (38.6%)	
T3	133 (7.15%)	60 (7.52%)	
T4	19 (1.02%)	9 (1.13%)	

Table 17. Baseline characteristics by training and testing set(continued)

	Training set N=1,861	Test set N=798	p.overall
Pathological N:			1
N0	1,467 (78.8%)	629 (78.8%)	
N1-3	394 (21.2%)	169 (21.2%)	
Adjuvant Ctx:			0.605
None	1,078 (57.9%)	461 (57.8%)	
Unknown	351 (18.9%)	162 (20.3%)	
Yes	432 (23.2%)	175 (21.9%)	
Radiation Tx:			0.83
None	1,117 (60.0%)	469 (58.8%)	
Unknown	475 (25.5%)	209 (26.2%)	
Yes	269 (14.5%)	120 (15.0%)	
Recurrence:			0.213
No recur	1,550 (83.3%)	648 (81.2%)	
Recur	311 (16.7%)	150 (18.8%)	

3. 머신러닝 모형 성능 비교

3.1 릿지 & 라쏘 회귀분석(Ridge & Lasso regression)

회귀계수 축소법인 Ridge와 Lasso를 모델에 적용하여 생존 분석 성능을 비교하였다. 각 모델의 성능은 Table 18과 같이 최적의 하이퍼파라미터를 위해 최적화된 인자값에 의해 분석되었으며, 3년 생존 및 5년 생존, 그리고 2년 재발 예측에 대한 성능 지표들로 나누어 비교하였다. 표에 나타난 지표들은 각 예측 기간에 따른 임계값(cutoff), 민감도(sensitivity), 특이도(specificity), 정밀도(precision), 재현율(recall), F1 점수, 정확도(accuracy), 그리고 AUC이다.

Ridge 회귀의 성능 지표는 다음과 같다. 5년 생존 예측에서 정확도는 0.798, AUC는 0.869로 3년 생존 예측에 비해 높게 나타났으며, 민감도(0.951)는 3년 생존 예측에서 더 높았다. 2년 재발 예측에서는 특이도는 0.832, 정확도는 0.821를 보였으며, AUC 또한 0.846으로 우수한 성능을 보였다.

Lasso 회귀의 경우, 5년 생존 예측에서 정확도는 0.751, AUC는 0.868이었으며, AUC는 3년 생존 예측에 비해 높게 나타났다. 3년 생존 예측에서는 특이도와 정확도가 5년 생존 예측에 비해 높게 나타났으며 AUC는 0.832이었다. 2년 재발 예측에서도 정확도 0.852, AUC 0.863으로 우수한 예측 성능을 보였다 (Table 19, Fig. 7~12).

Ridge 회귀와 Lasso 회귀 모델 모두 비교적 높은 성능을 가지고 있었으며, 두 모델 간 성능의 큰 차이는 없었으며 더 긴 기간의 장기 생존 예측에서 예측율이 높음을 확인하였다.

Table 18. Best Hyperparameters of Ridge and Lasso regression

		Ridge	Lasso
Survival	3 year	lambdas : 0.049777024	lambdas : 0.004430621
	5 year	lambdas : 0.05994843	lambdas : 0.007742637
Recur	2 year	lambdas : 0.1047616	lambdas : 0.006428073

Table 19. Performance of Ridge and Lasso regression

Model	Class	Survival		Recur
		3 year	5 year	2 year
Ridge	Sensitivity	0.821	0.782	0.737
	Specificity	0.789	0.802	0.835
	Precision	0.373	0.461	0.426
	Recall	0.821	0.782	0.426
	F1 Score	0.513	0.580	0.540
	Accuracy	0.793	0.798	0.821
	AUC	0.857	0.869	0.846
Lasso	Sensitivity	0.755	0.852	0.693
	Specificity	0.819	0.729	0.879
	Precision	0.390	0.405	0.488
	Recall	0.755	0.852	0.693
	F1 Score	0.514	0.549	0.572
	Accuracy	0.811	0.751	0.852
	AUC	0.856	0.868	0.845

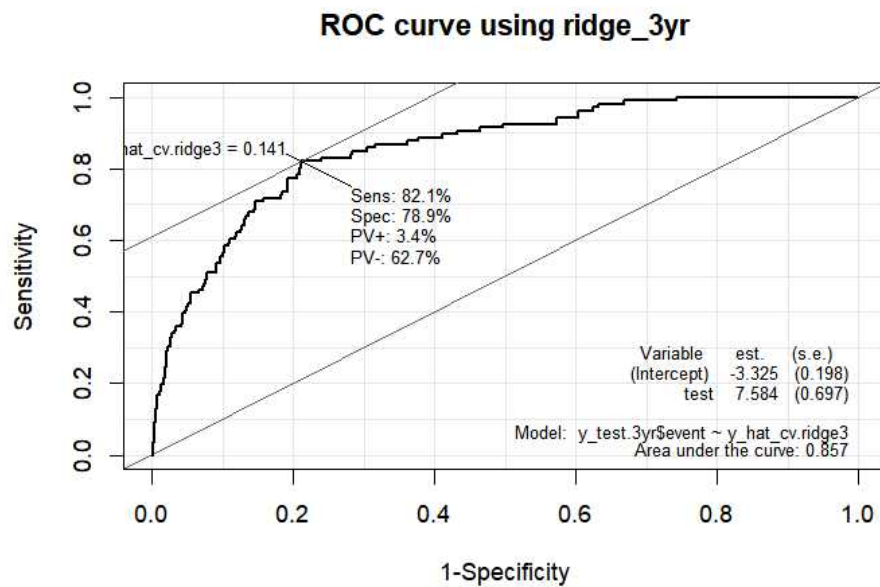


Fig. 7 3-year survival ROC curve using Ridge regression

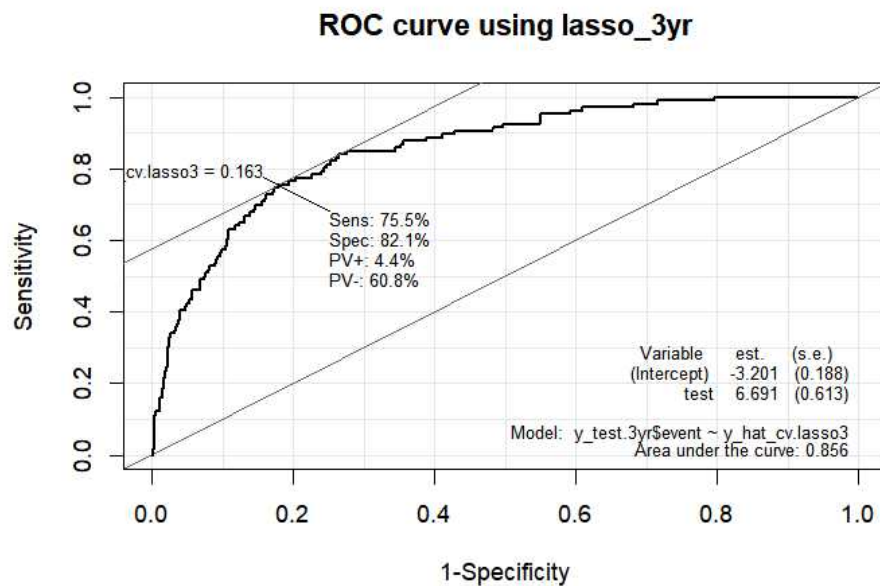


Fig. 8 3-year survival ROC curve using Lasso regression

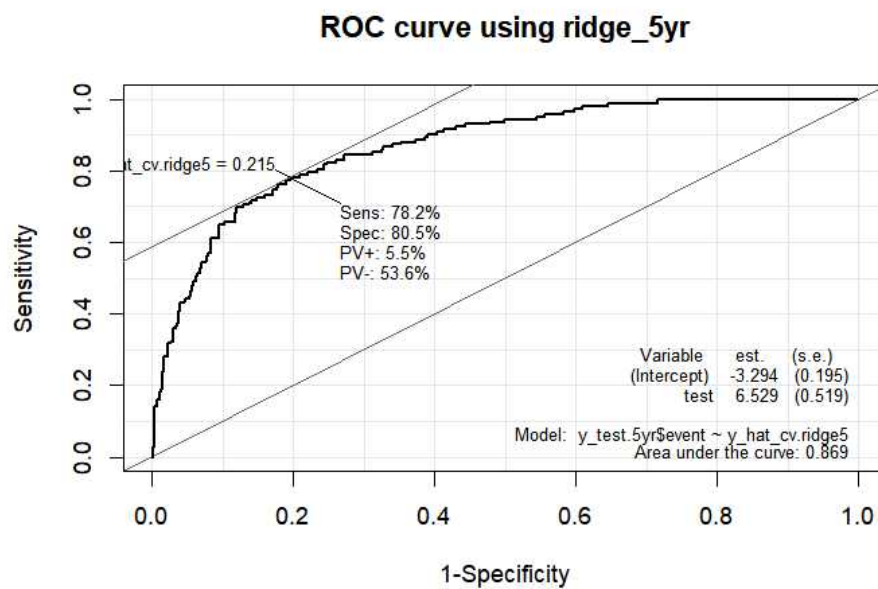


Fig. 9 5-year survival ROC curve using Ridge regression

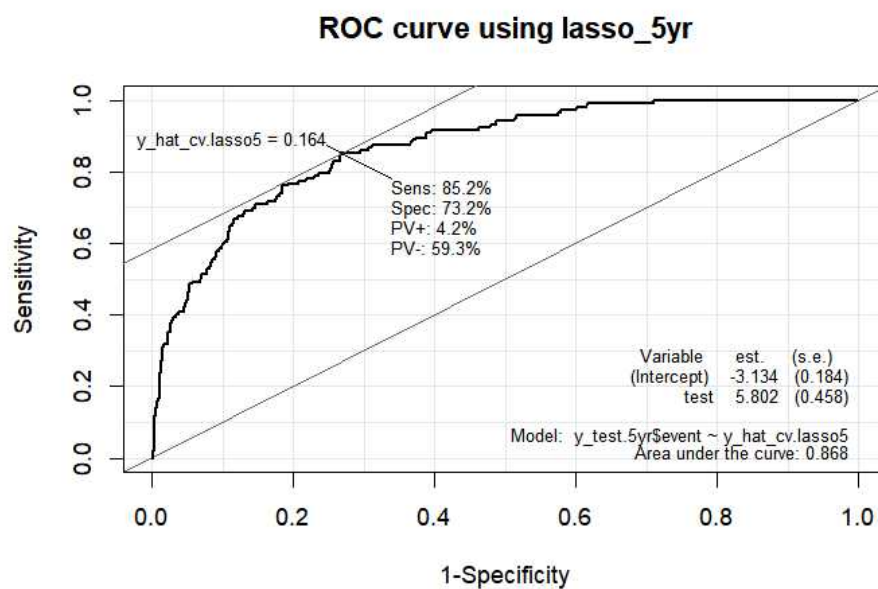


Fig. 10 5-year survival ROC curve using Lasso regression

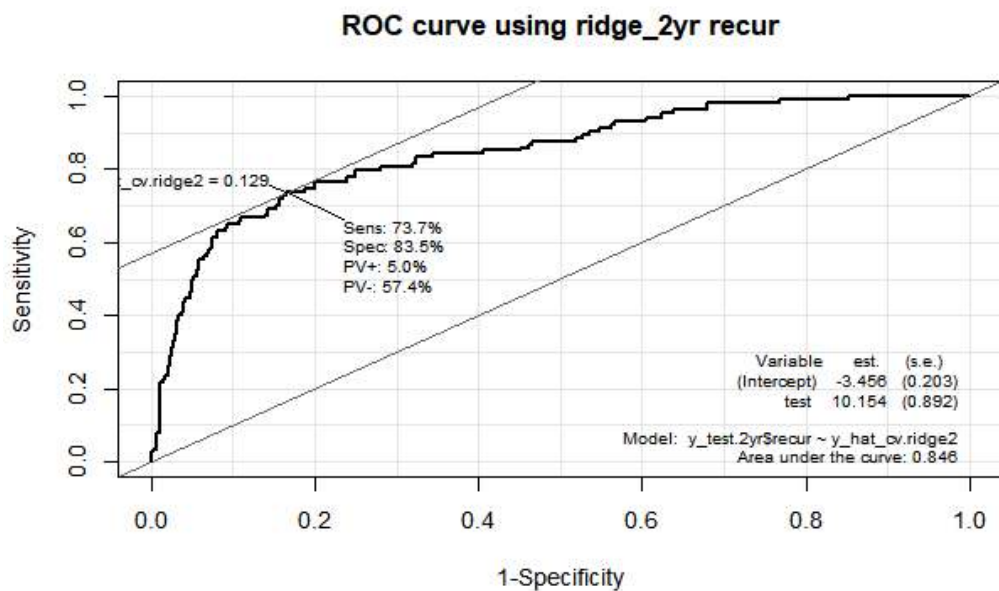


Fig. 11 2-year Recurrence ROC curve using Ridge regression

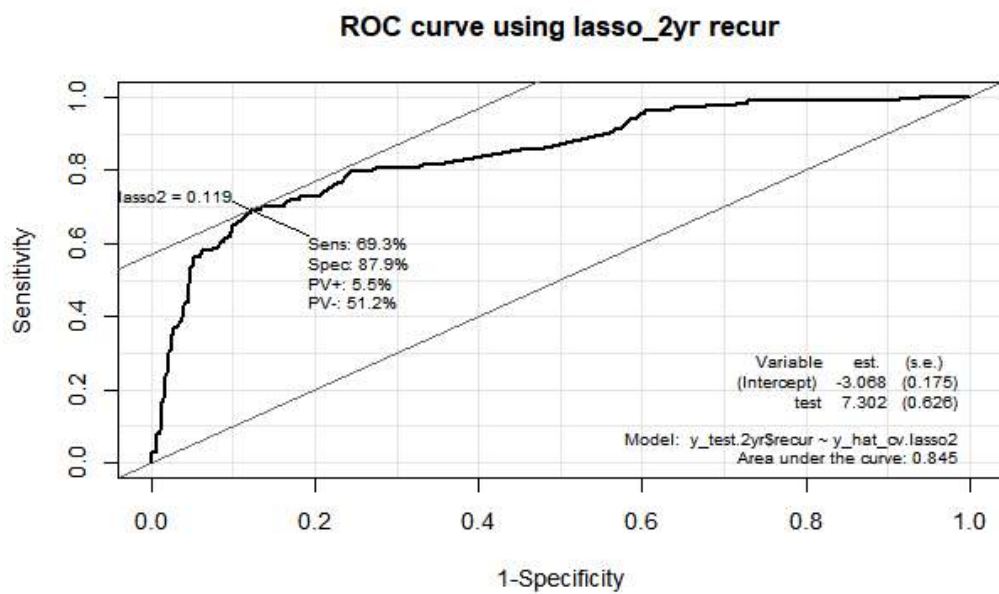


Fig. 12 2-year Recurrence ROC curve using Lasso regression

3.2 랜덤 포레스트(Random Forest)

앙상블 학습 방법의 일종인 랜덤 포레스트 모델에서 가장 적합한 인자를 결정하기 위해 mtry를 조정하여 가장 최적화된 하이퍼파라미터를 적용하였다(Table 20). 5년 생존 예측의 AUC는 0.855로 3년 생존 예측에 비해 높았으며, 생존과 재발 예측 모델 모두에서 AUC 0.8 이상의 우수한 성능을 보였다(Table 21, Fig. 13~15).

Table 20. Best Hyperparameters of Random Forest

Hyperparameters		
Survival	3 year	mtry : 2
	5 year	mtry : 6.08
Recur	2 year	mtry : 2

Table 21. Performance of Random Forest

Class	Survival		Recur
	3 year	5 year	2 year
Sensitivity	0.868	0.810	0.737
Specificity	0.717	0.771	0.808
Precision	0.319	0.433	0.391
Recall	0.868	0.810	0.737
F1 Score	0.467	0.565	0.511
Accuracy	0.737	0.778	0.798
AUC	0.853	0.855	0.826

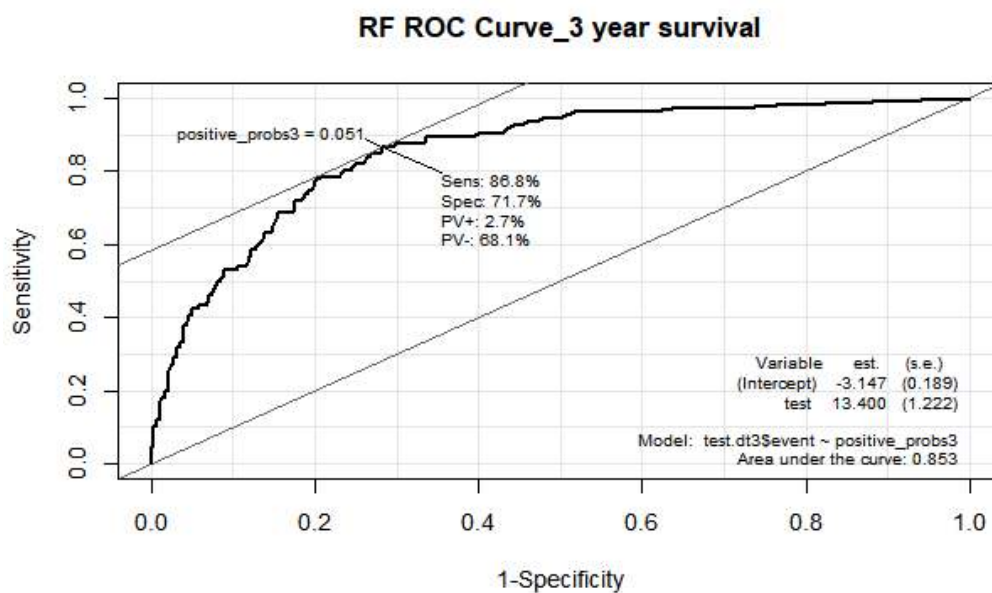


Fig. 13 3-year survival ROC curve using Random Forest

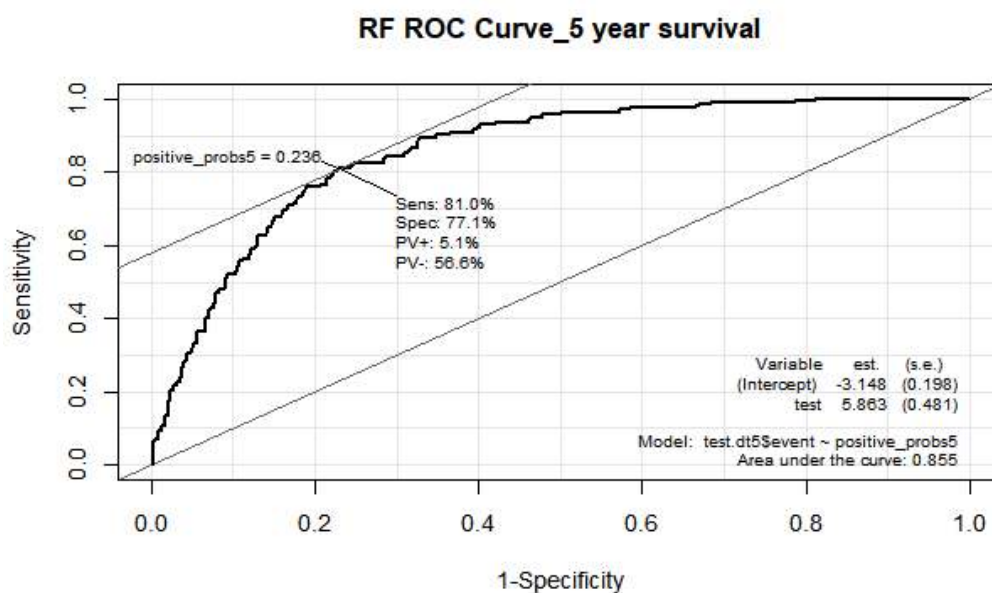


Fig. 14 5-year survival ROC curve using Random Forest

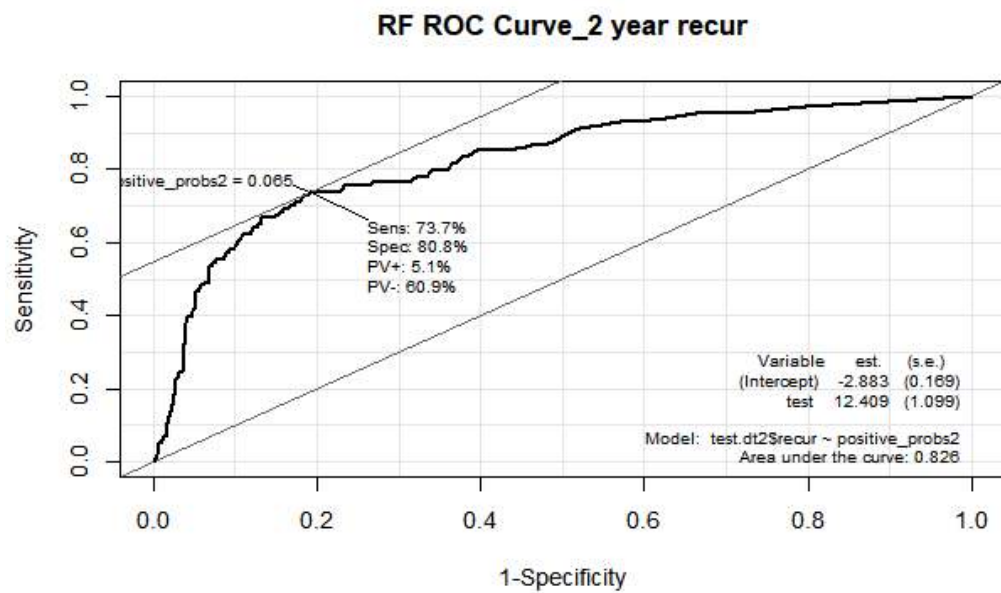


Fig. 15 2-year recur ROC curve using Random Forest

3.3 익스트림 그래디언트 부스팅(XGBoost)

익스트림 그래디언트 부스팅(XGBoost) 모델에서 가장 적합한 인자를 결정하기 위해 nrounds, eta, max_depth, colsample_bytree, min_child_weight, subsample를 조정하여 가장 최적화된 하이퍼파라미터를 적용하였다(Table 22). AUC는 폐암 5년 생존율 예측에서 0.858을 나타내며 3년 생존 예측에 비해 좋은 성능을 보였다. 수술 후 2년 이내 재발을 예측하는 모델에서는 AUC가 0.845로 우수한 성능을 확인하였다(Table 23, Fig. 16~18).

Table 22. Best Hyperparameters of XGBoost

Hyperparameters		
Survival	3 year	nrounds = 100, eta : 0.1 max_depth : 9, min_child_weight : 5 subsample : 0.5, colsample_bytree : 0.7
	5 year	nrounds = 75, eta : 0.1 max_depth : 3, min_child_weight : 5 subsample : 0.5, colsample_bytree : 0.9
	Recur 2 year	nrounds = 100, eta : 0.3 max_depth : 6, min_child_weight : 5 subsample : 0.7, colsample_bytree : 0.7

Table 23. Performance of XGBoost

Class	Survival		Recur
	3 year	5 year	2 year
Sensitivity	0.858	0.810	0.684
Specificity	0.678	0.770	0.841
Precision	0.290	0.432	0.417
Recall	0.858	0.810	0.684
F1 Score	0.433	0.564	0.518
Accuracy	0.702	0.777	0.818
AUC	0.848	0.864	0.811

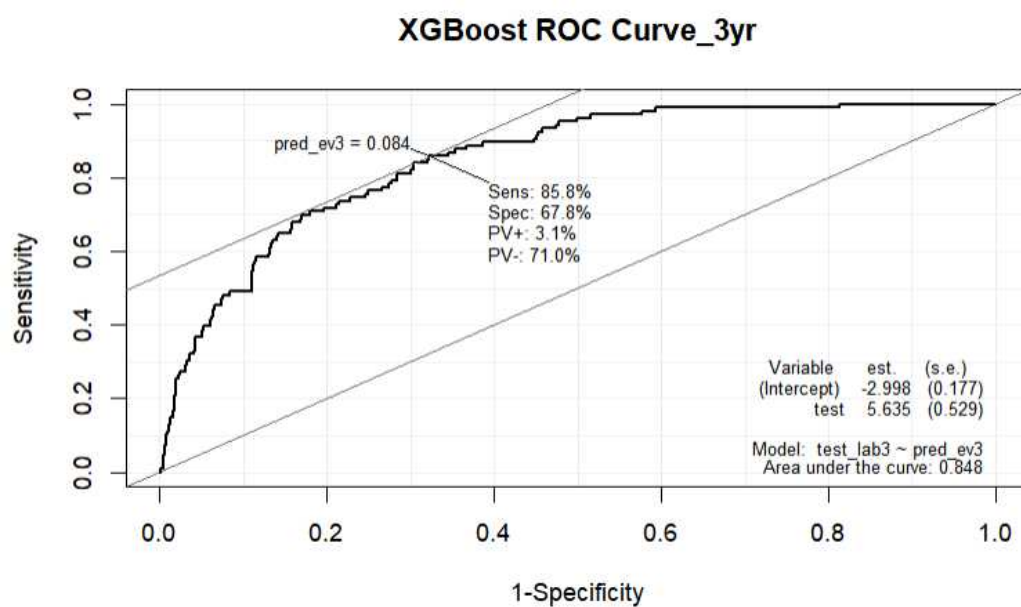


Fig. 16 3-year survival ROC curve using XGBoost

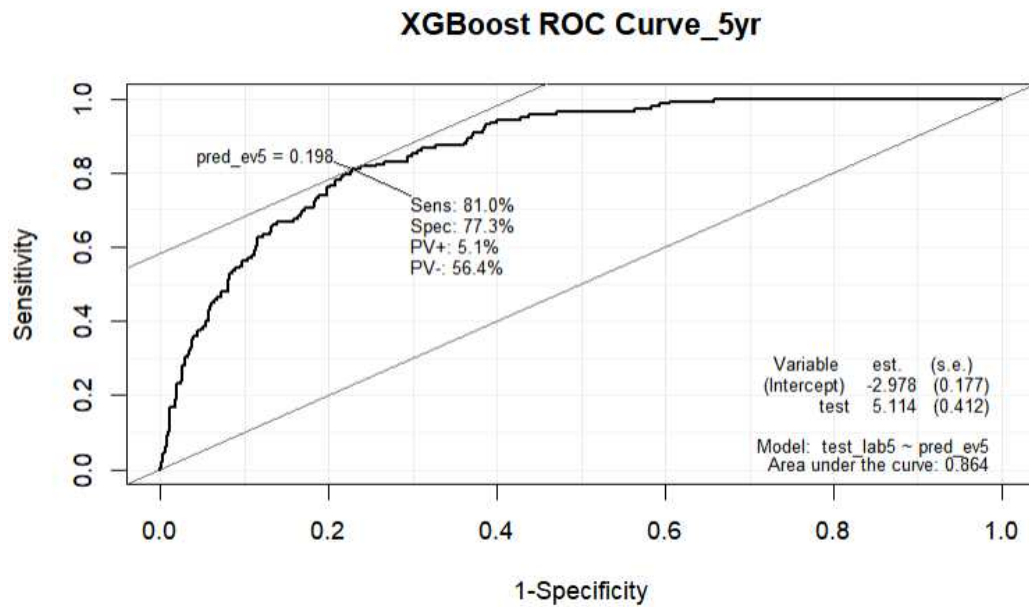


Fig. 17 5-year survival ROC curve using XGBoost

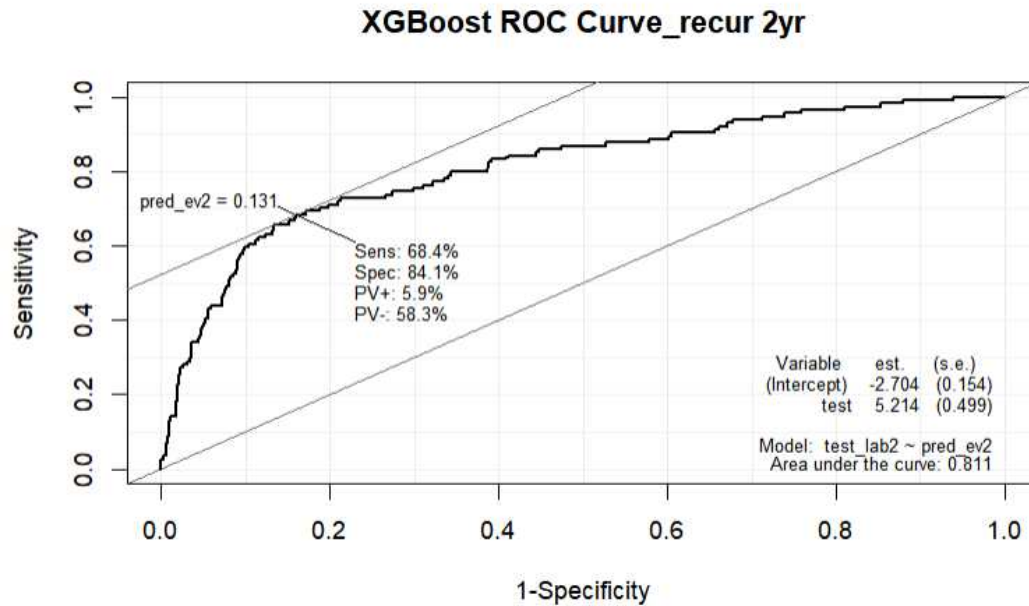


Fig. 18 2-year recur ROC curve using XGBoost

3.4 모델 결과

각 머신러닝 모델 간의 성능 차이는 Table 24와 같다. AUC로 각 모델의 성능을 평가하였을 경우, 모델 간 성능의 큰 차이는 존재하지 않았으나 생존 및 재발 예후 예측 모두에서 Ridge 회귀분석 모델의 성능이 좋게 나타났다(Table 23, Fig 19~20).

Table 24. Performance comparison between model

Set	Class	Survival		Recur
		3 year	5 year	2 year
AUC	Ridge	0.857	0.869	0.846
	Lasso	0.856	0.868	0.845
	RF	0.853	0.855	0.826
	XGBoost	0.848	0.864	0.811
Accuracy	Ridge	0.793	0.798	0.821
	Lasso	0.811	0.751	0.852
	RF	0.737	0.778	0.798
	XGBoost	0.702	0.777	0.818
Sensitivity	Ridge	0.821	0.782	0.737
	Lasso	0.755	0.852	0.693
	RF	0.868	0.810	0.737
	XGBoost	0.858	0.810	0.684
Specificity	Ridge	0.789	0.802	0.835
	Lasso	0.819	0.729	0.879
	RF	0.717	0.771	0.808
	XGBoost	0.678	0.770	0.841
F1 score	Ridge	0.513	0.580	0.540
	Lasso	0.514	0.549	0.572
	RF	0.467	0.565	0.511
	XGBoost	0.433	0.564	0.518

Model Comparison on Various Metrics of 5 year survival

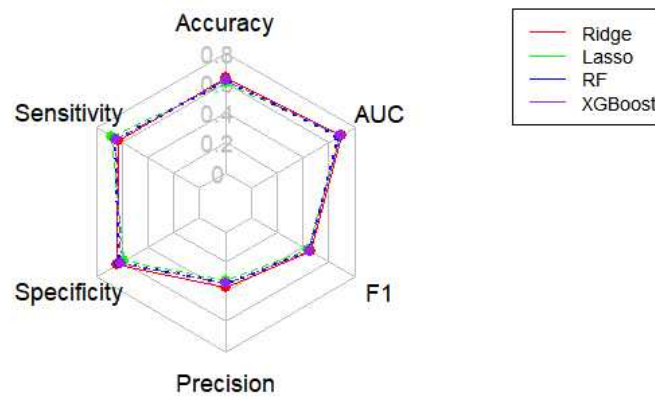


Fig. 19 Model comparison on Various metrics of 5 year survival

Model Comparison on Various Metrics of 2 year recur

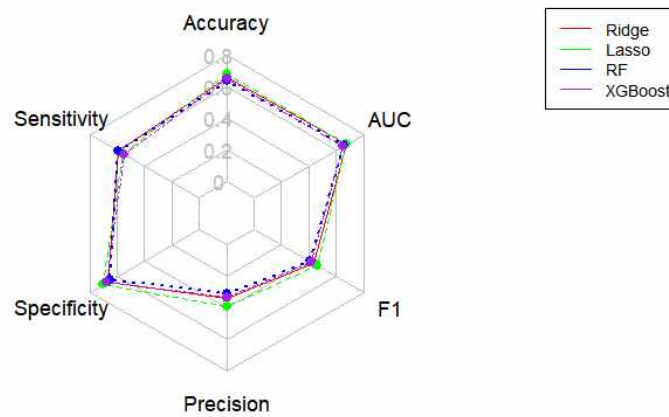


Fig. 20 Model comparison on Various metrics of 2 year recur

4. 변수 중요도(Feature importance) 비교

각 모델에서 확인된 변수 중요도는 어떤 변수가 비소세포폐암 수술환자의 사망을 예측하는데 있어 중요 변수인지를 확인하는 지표로 이를 통해 예후 예측에 가장 영향력 있는 변수와 그렇지 않은 변수들을 쉽게 식별 할 수 있다. 또한, 시각화를 통해 순위를 확인함으로써 변수 중요도를 보다 명확하게 비교할 수 있다는데 의미가 있다.

비소세포폐암 수술환자의 생존 예측에 기여한 설명변수의 중요도를 확인한 결과 모델 별 차이는 존재 하지만 공통적으로 재발여부, 조직학적 병기 및 체질량지수, 연령 및 성별, 그리고 보조적 항암치료여부 및 방사선 치료여부가 환자의 생존을 예측하는데 있어 중요한 변수인 것으로 분석되었다. 설명변수의 중요도 결과를 통해 재발여부 변수가 비소세포폐암 수술환자의 생존을 예측하는데 있어 가장 높게 영향을 미치는 것을 확인하였으며 관심변수인 체질량지수도 변수 중요도에 있어 예후 예측에 기여하는 영향도가 있음을 확인 할 수 있었다.

비소세포폐암 수술환자의 재발 예측에 기여한 설명변수의 중요도를 확인한 결과 진단 시 체질량지수는 생존 예측과는 다르게 재발에 영향을 미치는 주요 변수가 아니었으며, 모델 별 차이는 존재 하지만 공통적으로 방사선 치료여부, 보조 항암치료여부, 흉막 침범, 조직학적 병기, 나이 및 성별이 수술 후 2년 내 재발을 예측하는데 있어 중요한 변수인 것으로 분석되었다.

4.1 릿지 & 라쏘 회귀분석(Ridge & Lasso regression)

능형 회귀분석(Ridge)과 라쏘 회귀분석에서는 모델을 훈련시킨 후 ‘coef’ 함수를 사용하여 모델의 회귀 계수를 얻는다. 회귀 계수의 절대값을 기반으로 변수의 중요도를 평가할 수 있다. 릿지 회귀는 변수를 제거하지 않고 모든 변수에 소량의 가중치를 할당하며, Lasso 회귀의 경우 계수가 0인 변수는 제외된 것으로 해당 변수는 모델 예측에 덜 중요하다고 볼 수 있다.

변수 중요도(Feature importance)를 통해 능형 회귀분석(Ridge)과 라쏘(Lasso) 회귀분석 모두 재발있음(recur_Recur) 변수가 수술 후 비소세포폐암 환자의 생존 예측에 있어 가장 높은 중요도를 가지는 것을 확인 할 수 있었다. 주관심 변수인 체질량지수의 경우, 릿지 회귀분석을 이용한 5년 생존율 예측에서 저체중(bmi_cat_Underweight) 변수가 생존 예측에 중요도 3위 안에 분포하였으며, 라쏘 회귀분석에서도 변수 중요도 10위에 분포하여 진단 시 체질량지수 인자가 생존 예측에 주요 변수임을 알 수 있었다(Table 24-25, Fig. 29).

수술 후 2년 재발 예측에 있어 릿지 모델에서 방사선 치료받음(radiation_Yes) 변수와 조직학적 T4기(pathological_t_T4), 수술명이 Sleeve_lobectomy인 수술 변수가 중요한 변수였으며, 라쏘 모델에서는 방사선 치료 받음(radiation_Yes) 변수와 보조 항암치료를 알 수 없는(adjuvant_tx_Unknown) 변수가 주요 변수였다. 두 모델 모두에서 생존 예측에서는 주요 변수로 확인되지 않은, 흉막 침범 존재(pi_cat_Positive) 변수가 재발 예측에 있어 상당히 높은 중요도를 가진 변수로 확인되었다. 체질량지수 관련 변수는 생존 예측과는 다르게 수술 후 2년 재발 예측에 있어 중요도를 가지는 주요 변수로 확인되지는 않았다(Table 24-25, Fig. 30).

Table 25. Features Importance Top 10 rankings by Ridge regression

Ridge	Survival		Recur
	3year	5year	2year
recur_Recur	1	1	
op_name_Sleeve_lobectomy	4	2	3
age_cat_>=65	9	7	
bmi_cat_Underweight		3	
pathological_n_N1_3	3	6	
pathological_stage_III	2	8	9
radiation_Unknown			
ecog_ps_2_4			
radiation_Yes	8	4	1
pathological_morph_Other		5	
sex_Male	6	10	
pathological_t_T3		9	8
pi_cat_Positive			10
op_name_Pneumonectomy			
hp_maxsize_cat_>=5	5		
adjuvant_tx_Yes	7		
path_way_Without_symptoms	10		
hp_maxsize_cat_4_5			6
pathological_t_T4			2
smoke_cat_Non_smoker			
pathological_stage_II			7
number_Single			5
adjuvant_tx_Unknown			4

Table 26. Features Importance Top 10 rankings by Lasso regression

Lasso	Survival		Recur
	3year	5year	2year
recur_Recur	1	1	
radiation_Unknown		9	
age_cat_>=65	6	5	
pathological_n_N1_3	5	2	
pathological_t_T3	9	4	
sex_Male	2	6	
pathological_t_T2	10	7	
op_name_Sleeve_lobectomy		3	4
pathological_stage_III	3		
bmi_cat_Underweight		10	
adjuvant_tx_Yes	4		
path_way_Without_symptoms	7		
hp_maxsize_cat_>=5	8		8
smoke_cat_Non_smoker			
pathological_morph_Other		8	
egfr_Positive			
ecog_ps_2_4			
path_way_Without_symptoms			
radiation_Yes			1
adjuvant_tx_Unknown			2
pathological_t_T4			3
hp_maxsize_cat_4_5			5
number_Single			6
pathological_stage_II			7
pi_cat_Positive			9
egfr_Not_tested			10

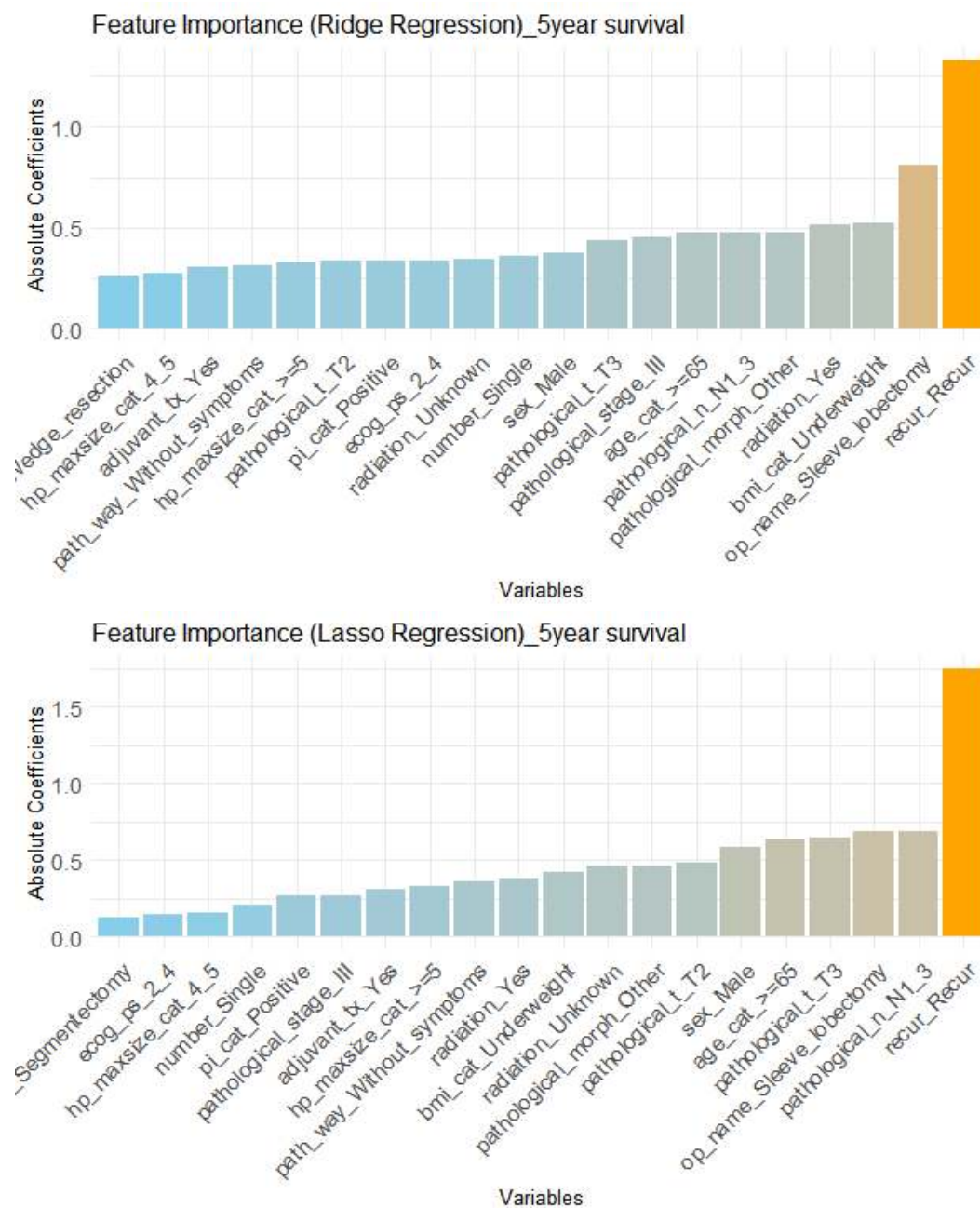


Fig. 21 Feature Importance of 5-year survival using Ridge and Lasso

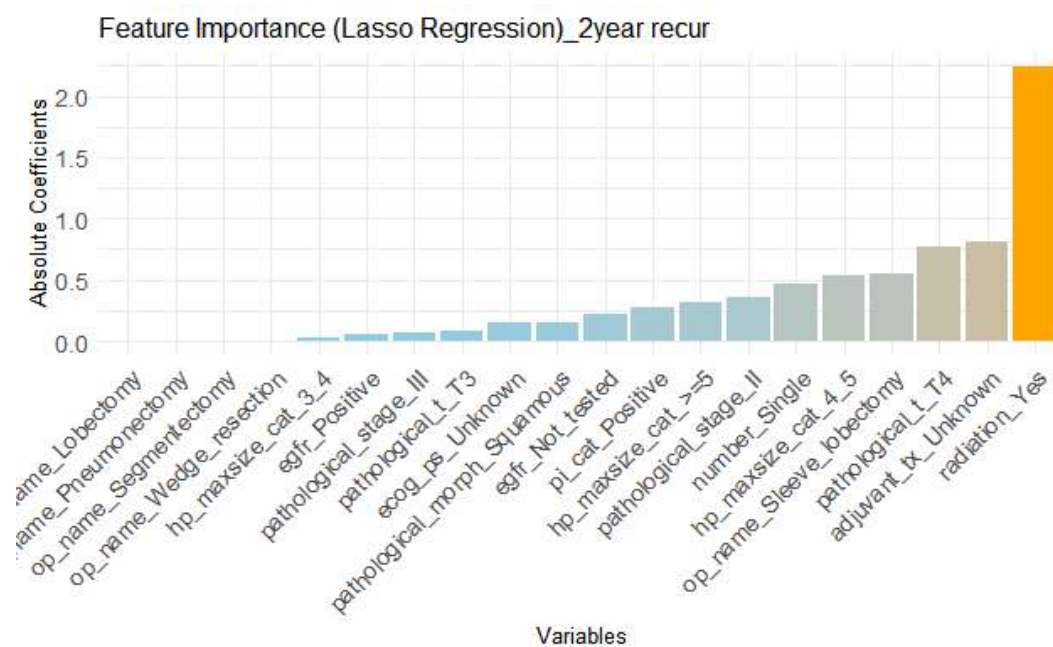
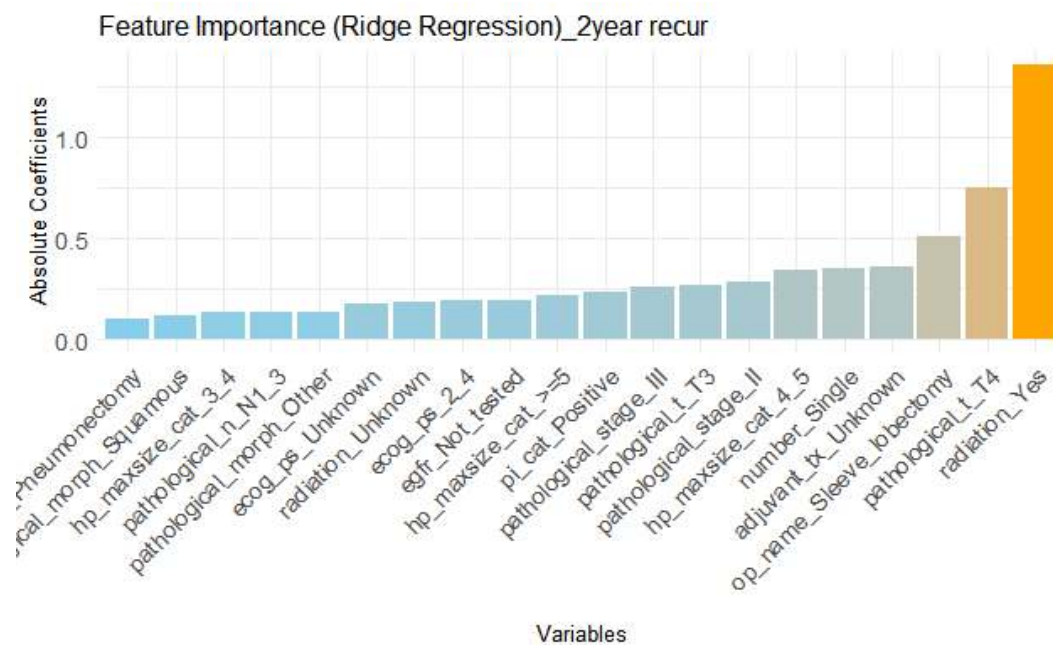


Fig. 22 Feature Importance of 2-year recur using Ridge and Lasso

4.2 랜덤 포레스트(Random Forest)

랜덤 포레스트 모델에서 변수 중요도를 평가하는데 가장 흔히 사용되는 방법 중 하나는 지니 불순도(Gini Impurity)를 기반으로 한다. 지니 불순도는 결정 트리에서 노드의 순도를 측정하는 데 사용되는 척도로 각 트리의 구성과 분할에서 변수의 중요도를 추정할 수 있다.

랜덤 포레스트로 진행한 생존 및 재발 예측 모형에서 지니 불순도를 순위화하여 변수 중요도를 나타내었다. 릿지와 라쏘 모델에서와 동일하게 재발있음(recur_Recur) 변수가 수술 후 비소세포폐암 환자의 생존 예측에 있어 가장 높은 중요도를 가지는 것을 확인 할 수 있었다. 다음으로는 방사선 치료받음(radiation_Yes) 변수와 조직학적 병기 3기(pathological_stage_III) 변수가 중요한 변수로 나타났다. 3년 생존 예측에서는 3년 생존 예측에서도 동일하였다. 랜덤 포레스트 모델에서는 진단 시 체질량 지수가 생존 예후에 영향을 미치는 주요 요인으로 5년 생존 예측에서만 과체중/비만 그룹 변수 (bmi_cat_Overweight_Obese)가 확인되었다. 재발 예측에 있어서는 릿지와 라쏘 모델에서와 동일하게 방사선 치료받음 변수의 중요도가 가장 높게 나타났으며 다음으로 조직학적 병기 3기(pathological_stage_III), 조직학적 N 병기 1-3기(pathological_n_N1_3) 순으로 나타났다(Table 27, Fig. 23).

Table 27. Features Importance Top 10 rankings by Random Forest

Ridge	Survival		Recur
	3year	5year	2year
recur_Recur	1	1	
radiation_Yes	2	2	1
pathological_n_N1_3	3	3	3
pathological_stage_III	4	8	2
pi_cat_Positive			
adjuvant_tx_Yes	8	4	4
pathological_t_T2			
hp_maxsize_cat_>=5	5		9
sex_Male	9		
age_cat_>=65		6	8
path_way_Without_symptoms	6	7	10
fev1_cat_>70		10	
egfr_Not_tested			
pi_cat_Positive	7	5	5
pathological_morph_Squamous			
smoke_cat_Past_smoker			
pathological_t_T3	10		6
bmi_cat_Overweight_Obese		9	
pathological_stage_II			

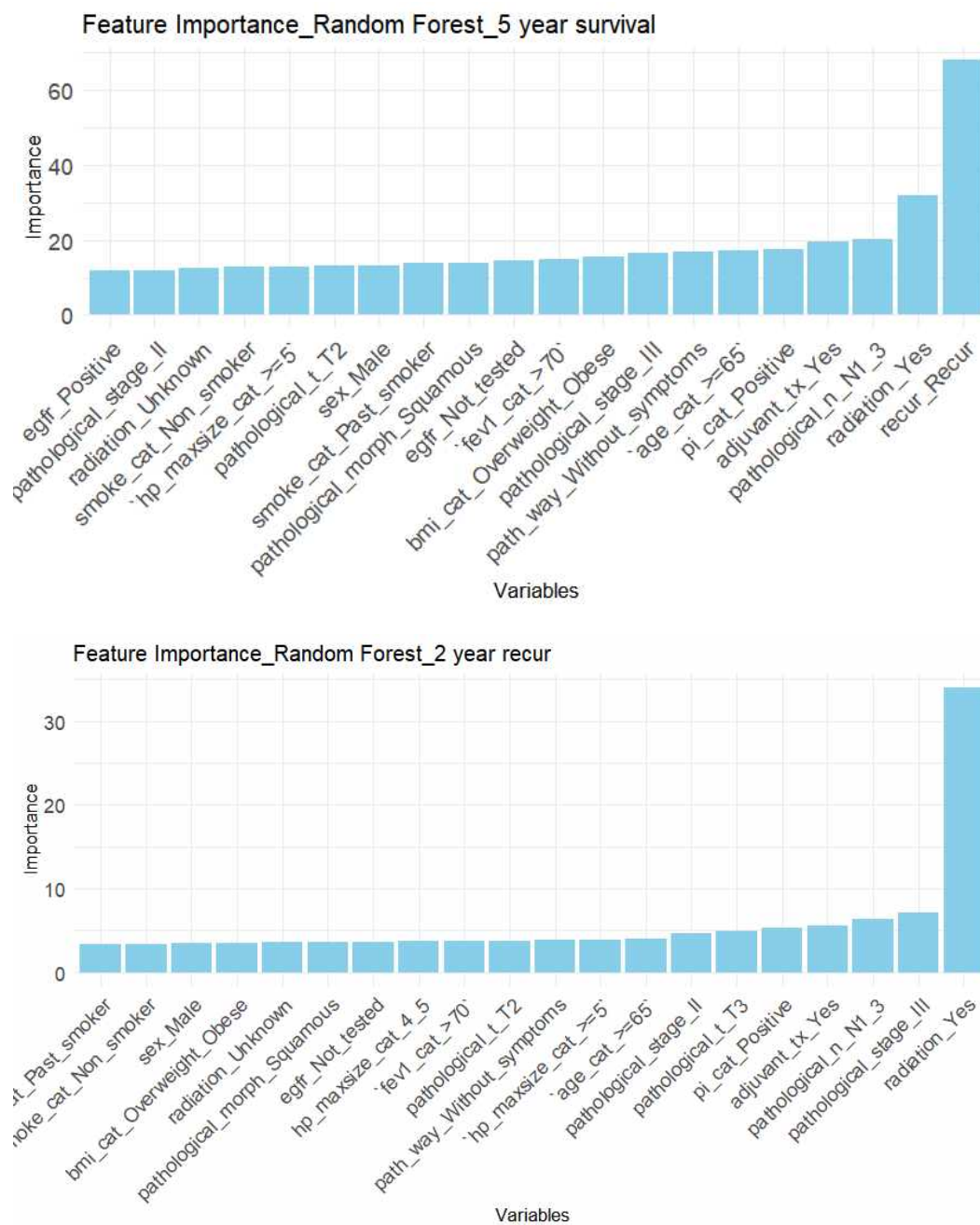


Fig. 23 Feature Importance using Random Forest

4.3 익스트림 그라디언트 부스팅(XGBoost)

XGBoost 모델을 이용하여 변수중요도를 확인하여 예후 예측에 영향력이 큰 상위 10개의 주요 변수들을 확인하였으며, 생존 예후 예측에 높은 중요도를 보이는 주요 인자로는 생존 년도 간의 차이는 존재하였으나 공통적으로 재발있음(Recur_recur), 방사선 치료받음(Radiation_Yes), 나이 65세 이상(Age_cat >= 65) 등이 상위권에 분포하였다. 5년 생존 예측에서 진단 시 체질량지수가 과체중/비만(bmi_cat_Overweight_Obese)인 경우 5년 생존 예측에 영향을 주는 주요 요인임을 확인할 수 있었다. 수술 후 2년 재발 예측과 관련한 주요 변수로는 방사선 치료받음(Radiation_Yes), 보조 항암치료 받음(adjuvant_tx_Yes) 변수처럼 치료 관련 변수였으며, 조직학적 병기 및 나이, 진단 시 체질량지수 또한 높은 중요도를 나타냈다. (Table 28, Fig. 24, Fig 26).

변수 중요도(Features importance)는 해당 변수가 예후 예측에 있어 양의 영향력을 갖는지, 음의 영향력을 갖는지는 알 수 없는 한계점이 있다. 이에 머신러닝 모델에서 성능이 좋았던 XGBoost 모델에, SHAP(SHApley Additive exPlanations) value 값을 구하여 SHAP summary plot을 추가로 나타냈다. XGBoost를 기반으로 한 SHAP summary plot에서 5년 생존 예측에 있어 가장 중요한 요인인 재발 있음(Recur_recur) 변수가 수술 후 비소세포폐암 5년 생존 예측에 있어 음의 방향보다는 양의 방향으로 즉, 생존보다 사망에 큰 영향을 미치는 변수임을 확인할 수 있었다. 또한, 체질량지수_과체중/비만 군의 경우, 5년 생존 예측에 미치는 영향력이 음의 방향으로 큰 것으로 확인되었다. 즉, 체질량지수가 과체중/비만인 경우에는 사망이 아닌 생존에 더 영향을 주는 것으로 나타났다(Fig. 25). 수술 후 2년 재발에 있어 가장 중요한 주요 요인은 방사선 치료받음(Radiation_Yes) 변수이며, 수술 후 비소세포폐암 2년 재발 예측에 있어 음의 방향보다는 양의 방향으로 즉, 재발에 큰 영향을 미치는 변수임을 확인할 수 있었다. 또한 주관심변수 체질량지수_과체중/비만 군의 경우, 2년 재발 예측에 미치는 영향력이 양의 방향으로 큰 것으로 확인됨으로써 재발에 영향을 주는 것으로 나타났다(Fig. 27)

Table 28. Features Importance Top 10 rankings by XGBoost

XGBoost	survival		recur
	3year	5year	2year
recur_Recur	1	1	
radiation_Yes	2	2	1
pathological_n_N1_3	6	3	
pi_cat_Positive		9	6
pathological_t_T2	10	6	3
age_cat_>=65	5	4	4
adjuvant_tx_Yes	4	10	2
radiation_Unknown			
path_way_Without_symptoms	3		10
pathological_stage_III	9	5	5
smoke_cat_Non_smoker			
hp_maxsize_cat_>=5	7		
bmi_cat_Overweight_Obese			7
sex_Male	8	7	
fev1_cat_>70			9
pathological_morph_Squamous		8	8

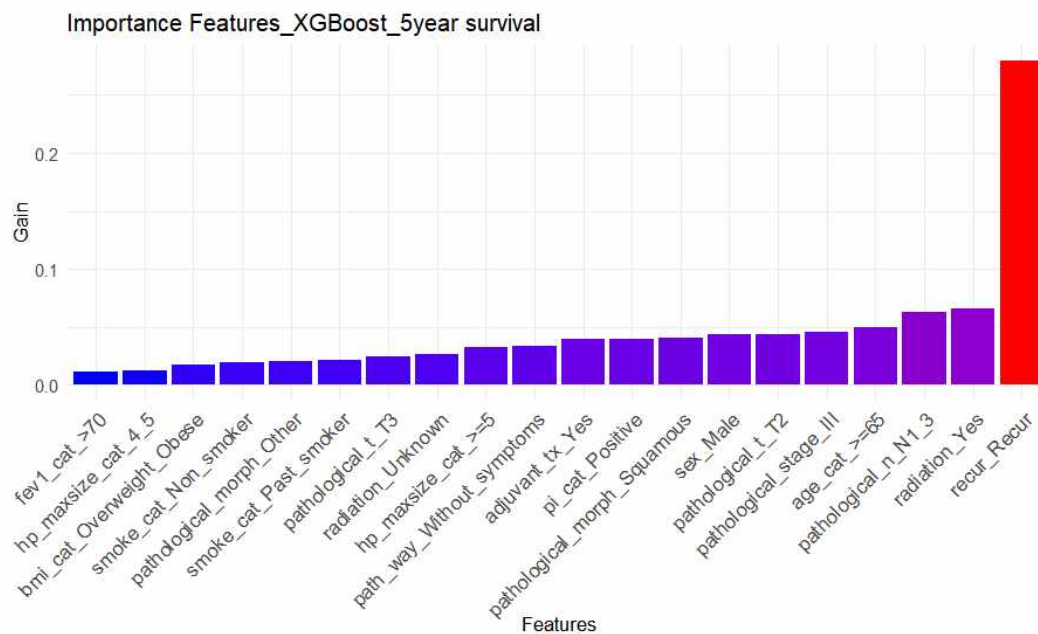


Fig. 24 Feature importance of 5-year survival using XGBoost

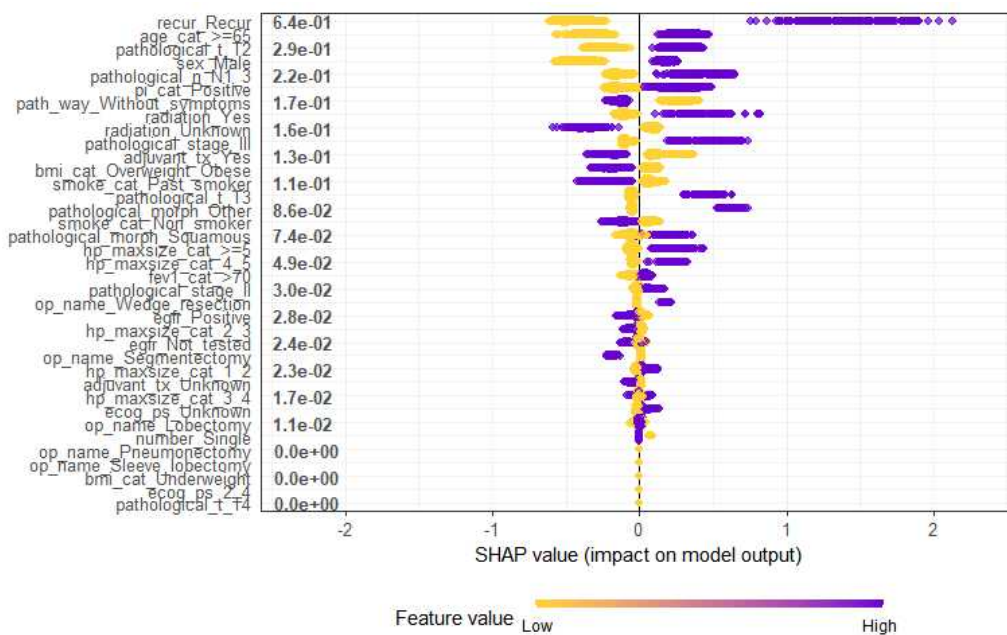


Fig. 25 SHAP summary plot of 5-year survival using XGBoost

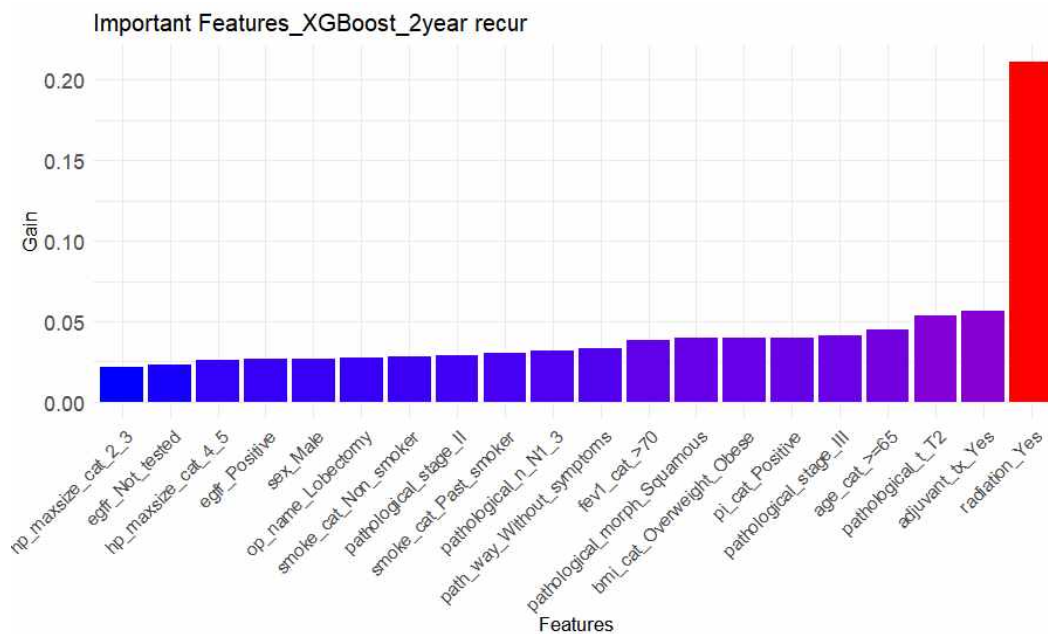


Fig. 26 Feature importance of 2-year recur using XGBoost

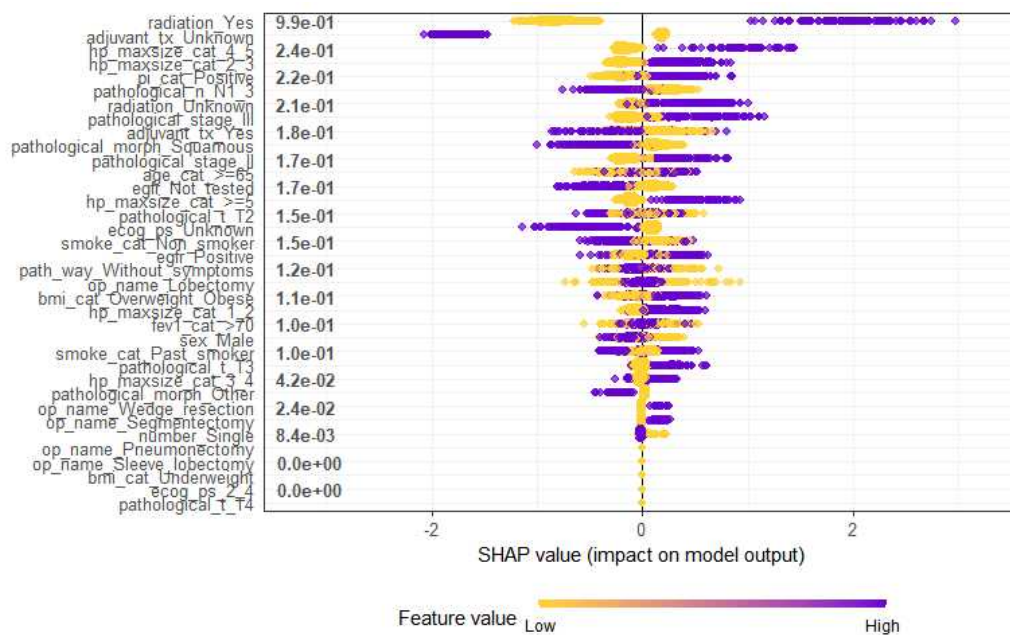


Fig. 27 SHAP summary plot of 2-year recur using XGBoost

5. 주요 변수에 따른 예후 예측

다양한 머신러닝 모델을 이용하여 생존 및 재발 예측 모델을 구축하고, 변수 중요도를 확인함으로써 어떤 변수들이 생존과 재발에 영향을 미치는지 확인하였다.

생존 예측에 있어 가장 중요 변수였던 재발 여부와 주관심 변수였던 진단 시 체질량지수의 그룹에 따라 예측 생존율이 어떻게 다를 수 있는지를 박스 플랏을 통해 시각적으로 제시하였다. 체질량지수 그룹별 5년 예측 생존율을 살펴보면 정상체중에 비해 과체중/비만 그룹이 높은 예측 생존율을 보였으며, 저체중인 그룹이 낮은 예측 생존율을 보이는 것을 확인 할 수 있었다(Fig. 28).

재발여부 그룹별 5년 예측 생존율에서는 재발이 있는 군에서 낮은 예측 생존율을 보였다(Fig. 29). 수술 후 2년 재발 예측에 있어 체질량지수 그룹 간 예측 재발률을 박스 플랏으로 확인 한 결과, 정상체중과 과체중/비만에 비해 저체중 그룹의 예측 재발율이 높게 나타났으나 큰 차이는 없었다. 가장 중요 요인이었던 방사선 치료여부에 따른 예측 재발율은 방사선치료를 받지 않은 그룹에 비해 예측 재발율이 높음을 확인하였다.

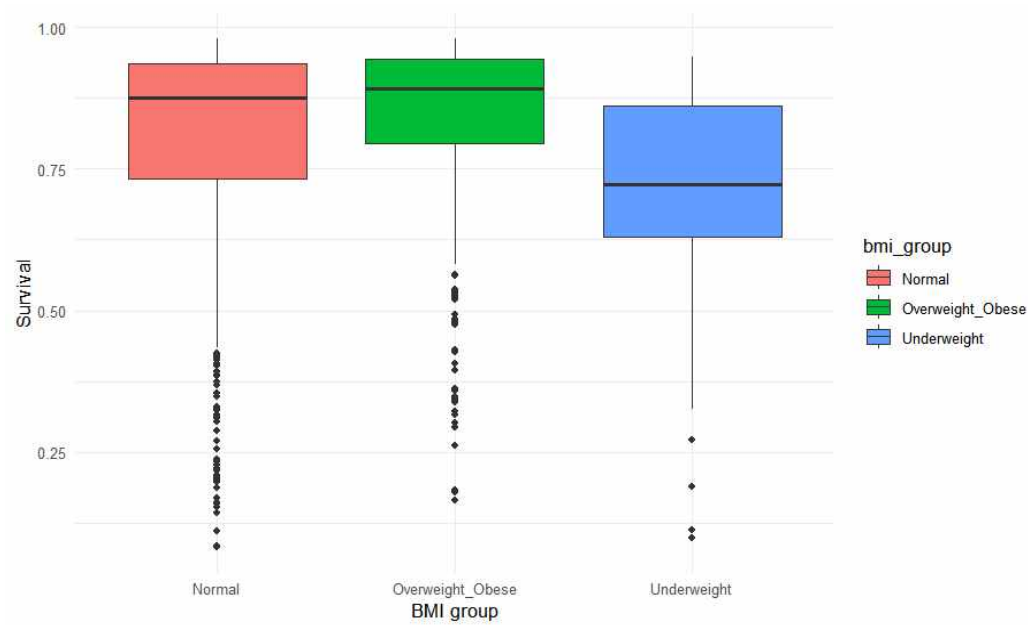


Fig. 28 Box plot of predict survival by BMI group

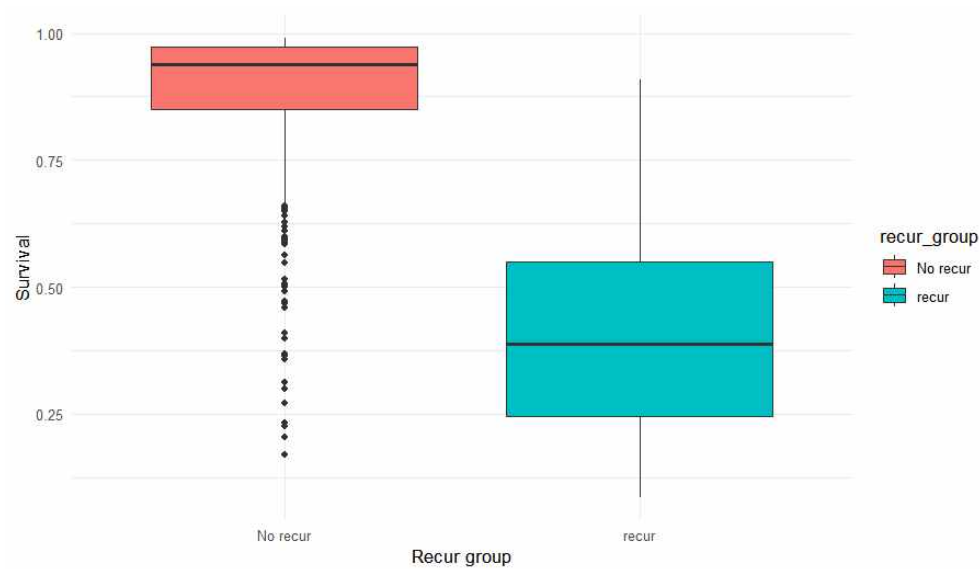


Fig. 29 Box plot of predict survival by recur group

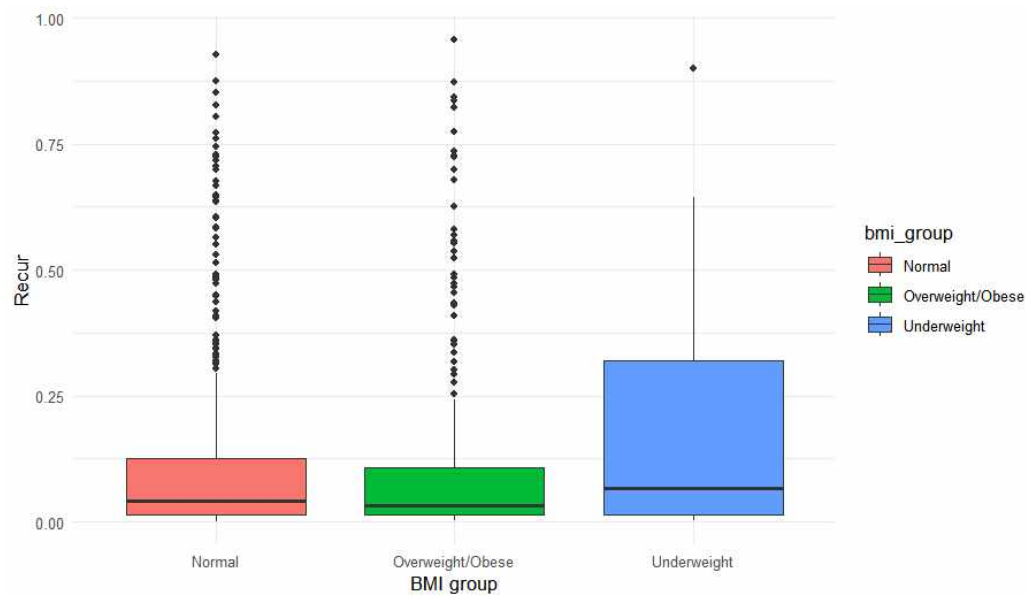


Fig. 30 Box plot of predict recurrence by BMI group

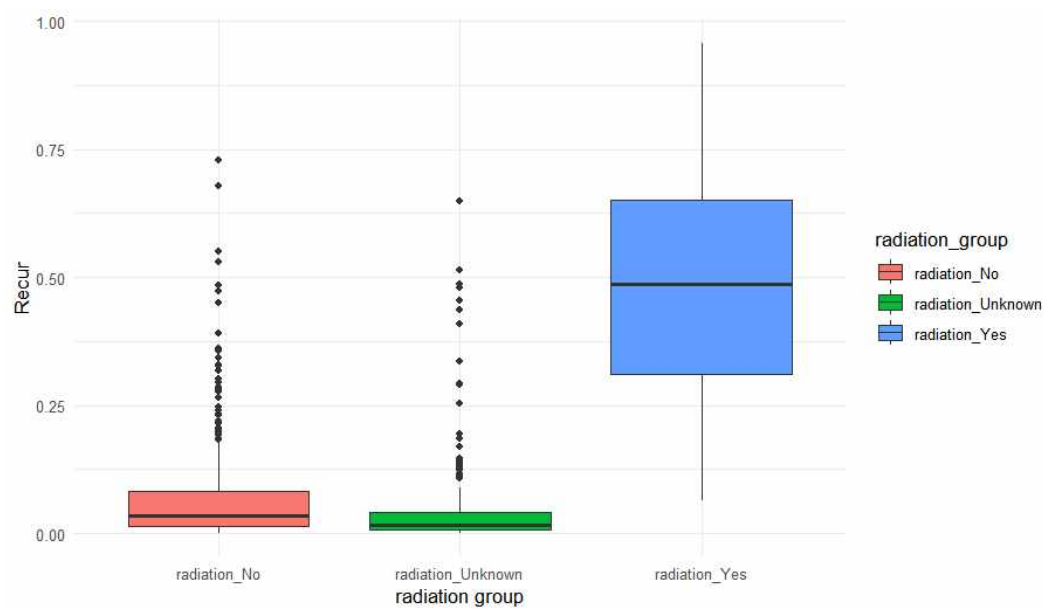


Fig. 31 Box plot of predict recurrence by radiation group

IV. 고 찰

비소세포폐암(Non-Small Cell Lung Cancer, NSCLC)은 전 세계적으로 가장 흔한 폐암 유형 중 하나로, 폐암 사망률의 주요 원인 중 하나이다. 초기 단계에서 발견되는 경우 수술적 절제가 가능하지만, 대부분의 환자들은 진단 시 이미 질병이 진행된 상태이다. 이에 따라, 환자의 생존 및 재발 예측은 임상적 의사 결정 과정에서 중요한 역할을 하며, 치료 계획 수립 및 환자 관리에 있어 핵심적인 정보를 제공한다. 머신러닝은 의료분야, 특히 질병 예측과 관련하여 점점 더 큰 주목을 받고 있다. 질병을 예측하기 위해 기계학습은 방대한 양의 임상 데이터를 평가할 수 있으며, 이를 통해 의료 전문가와 환자는 질병 관리 치료에 대해 현명한 결정을 내릴 수 있다. 폐암 생존 및 재발 예측은 임상 데이터 분석에서 가장 중요한 주제 중 하나라 할 수 있다.

폐암 환자의 생존 및 재발에 체질량지수(BMI)가 미치는 영향을 조사하기 위해, 본 연구는 설명 가능한 머신러닝 기술을 적용하여 폐암의 비만 역설을 검토하고, 전통적인 통계 방법과 현대 알고리즘을 사용하여 예측 모델을 개발하였다. 이 모델들은 다중공선성을 관리하고 정규화를 통해 특성 선택을 수행하는 능력으로 알려진 Ridge와 Lasso 로지스틱 회귀분석과, 높은 예측 정확도와 과적합에 강한 성능으로 주목받는 앙상블 방법론인 랜덤 포레스트와 XGBoost를 포함한다.

본 연구의 중요성은 전통적인 통계 방법과 현대 머신러닝 모델의 성능을 비소세포폐암(NSCLC) 환자의 생존 및 재발 결과 예측에 있어 비교하는 포괄적 접근에 있다. 본 연구에서는 모델 간 성능에 큰 차이를 보이지 않았으나 전통적 통계 방법에서는 Ridge 모델이 머신러닝 모델에서는 XGBoost의 성능이 우수했으며 모든 모델을 포함하여 Ridge 회귀분석이 가장 우수한 성능을 나타냈다. 현대 알고리즘의 진보된 기능에도 불구하고, Ridge와 Lasso 회귀분석과 같은 전통적인 통계 방법이 더 복잡한 모델들과 경쟁하며 다소 우수한 성능을 보였다. 일부 선행연구에서도 종종 이러한 결과를 보였으며, 이는 더 단순한 모델들이 과적합을 피하고 변수 간의 관계가 잘 정의된 더 작고 해석 가능한 데이터셋을 다루는데 효과적이기 때문일 수 있다(Sievering AW, 2022; Wang H, 2015). 각 모델의 장단점이 존재하기에 목적 및 상황에 맞게 전통적 통계기법과 다

양한 머신러닝 모델 모두 활용할 필요가 있을 것이다.

또한, 본 연구에서는 각 모델에서 예후 예측에 영향을 주는 변수 중요도를 확인하고 최고 성능의 머신러닝 모델의 예측을 설명하기 위해 변수중요도와 SHAP값을 적용함으로써 예측 결과에 대한 근거와 해석을 제공하며, 설명 가능한 머신러닝 기반 비소세포폐암 수술환자의 생존 및 재발 예측 모델을 구현하였다. 머신러닝 알고리즘을 사용할 때 블랙박스 특성과 설명 가능성 부족으로 인해 환자의 개별 특징이 예후 예측의 정확도 향상에 관련 있는지는 판단하기 어려울 수 있다. 본 연구에서는 변수 중요도(Feature importance) 및 SHAP 방법을 사용하여 이러한 한계점을 극복하였으며, 변수의 중요성을 명확하고 직관적으로 보여주는 SHAP plot을 통해 의료진은 어떤 요인이 환자의 생존과 재발 위험을 높이거나 낮추는데 기여하는지를 이해하는데 도움이 되고, 맞춤형 예방 또는 집중 치료 전략을 통해 환자를 도울 수 있을 것이다.

본 연구를 통해 진단 시 체질량지수가 비소세포폐암 수술환자의 생존 예후에 영향을 주는 중요 인자임을 확인할 수 있었으며, 비만 역설 현상도 확인하였다. 머신러닝을 통해 각 모델의 변수 중요도를 확인한 결과 모든 머신러닝 모델에서 진단 시 체질량 지수가 생존을 예측하는데 있어 영향력이 있는 주요 요인이었으며, 체질량지수 그룹 별 예측 생존율을 표시한 상자 도표(box plot)에서도 저체중은 정상 체중에 비해 낮은 생존율을 보이는 반면, 과체중과 비만은 높은 생존율을 보였다. 이는 주요 선행 연구에서 확인된 바와 같다. Alifano M(2021) 등은 비소세포폐암으로 폐 절제술을 받은 환자의 생존에 체질량지수가 강력하고 독립적인 예측 변수임을 확인하였으며, 저체중은 정상체중에 비해 낮은 생존율과 관련이 있었고, 과체중과 비만은 높은 생존율과 관련이 있음을 보고하였다. 또한, Matsunaga T(2015) 등도 악성 폐종양을 절제한 1,518명의 환자를 대상으로 로지스틱 회귀분석을 사용하여 예측 변수를 식별한 결과, 낮은 체질량지수가 사망에 대한 독립적인 위험인자로 확인하였다. 모델에 따라 수술 후 2년 재발 예측에 있어 진단 시 체질량지수가 중요 인자인지에 대한 부분은 차이가 있었다. 대부분의 모델에서는 체질량지수가 재발 예측에 있어 주요 요인은 아니었으나 XGBoost에서는 주요 인자로 나타났으며, XGBoost모델에 적용한 SHAP 방법을 통해서 저체중 그룹이 2년 재발에 있어 재발안함보다 재발함에 더 영향이 있음을 확인할 수 있었다. 또한 수술 후 2년 재발 예측에서는 수술 외 치료 여부 즉, 방사선 치료와 보조적 항암치료 변수, 그리고 조직학적 병기가 환자의

재발에 가장 큰 영향을 주는 인자로 확인되었다. 선행연구에서도 암 병기나 TNM 병기가 높은 환자는 예후가 좋지 않고 재발 위험도 높으며, 보조 방사선 치료를 받은 환자의 재발률이 상당히 높다. 이는 재발률이 높기 때문에 수술 전이나 수술 후 진행된 환자나 말기 환자에게 방사선치료를 주로 실시한다는 사실로 설명할 수 있다(Yang Y, 2022). 또한, 선행연구를 통해 재발의 주요 독립 인자로 보고된 흉막 침범 여부 또한 재발 예측에 영향을 주는 요인임을 확인할 수 있었다(Dziedzic DA, 2016).

본 연구는 단일기관의 데이터가 아닌 국내 폐암환자를 대표할 수 있는 중앙암 등록본부의 폐암병기조사자료를 활용하여 예측모델을 개발하고 성능을 검증하였으며, 그 성능 또한 모든 적용 모델의 AUC가 0.8 이상으로 우수한 성능을 보였다. 또한 본 연구는 폐암의 대부분을 차지하는 비소세포폐암의 수술환자만을 대상으로 했다는데 의의가 있다. 현재 비소세포폐암 수술환자의 예후 예측에 대한 선행논문은 많지 않은 실정이며, 생존과 재발 예측 모두 다뤘다는데 큰 의의가 있다.

본 연구의 제한점으로는 체질량지수의 저체중 그룹과 비만 그룹의 환자수가 적어 데이터 누락으로 인해 생략되어 변수 중요도 추출에 있어 아쉬움이 있다. 향후 연구에서는 보다 일반화된 결과를 얻고 보다 신뢰할 수 있는 모델을 구축하기 위해 더 큰 데이터 세트를 고려할 수 있을 것이다. 또한, 국내 자료를 사용하여 예측모델을 개발하였으며 외부 검증을 진행하지는 못하였기 때문에, 다른 인종과 국가를 대상으로 적용하였을 때 예측 편향이 생길 수 있다. 따라서 추후 다양한 대상자가 포함된 데이터에서의 외부 검증 작업이 필요하다. 마지막으로 본 연구에서는 EGFR mutation과 같은 유전자 데이터에 대한 변수의 종류가 다양하지 않아 다종 바이오마커를 이용해 머신러닝 알고리즘에 대입하여 다양한 변수에 대한 비소세포폐암 수술 후 예후에 대한 위험도 정보제공이 필요하며, 향후 연구에서는 정확하고 예측의 효율성을 높이기 위해 점점 더 새롭고 통합된 머신러닝 알고리즘을 사용해야 할 것이다.

본 연구를 통해 발견한 생존 및 재발에 영향을 주는 주요 인자들과 비소세포폐암 수술환자의 생존 예측에 있어 수술 전 체질량지수의 역할의 중요성은 임상에서 환자의 체질량지수 및 주요 인자들을 고려하여 맞춤형 치료 계획을 수립하는데 도움이 될 것이며, 이로써 환자의 치료 결과를 개선하고 생존율을 높일 수 있을 것이라 기대한다.

V. 결 론

본 연구는 중앙암등록본부의 폐암병기조사데이터를 활용하였으며, 다양한 머신러닝 모델로 비소세포폐암 수술환자의 3년 생존, 5년 생존과 2년 재발 예측 모델을 개발하여 각 모델링 방법에 따른 생존 및 재발 예측 성능을 비교하였다. 예측 모델의 주 성능평가 지표로 민감도, 특이도, 정확도, 정밀도, F1 점수 및 정확도, AUC를 사용하였고, 모델 간 성능이 차이는 미비하였으며, 적용한 모델 모두 AUC 0.8 이상의 우수한 성능을 나타냈고, 5년 생존 예측이 3년 생존 예측 보다 좋은 성능을 보였다.

변수 중요도를 파악하고 SHAP 분석을 적용한 머신러닝 예측 모형 결과, 비소세포폐암 수술환자에서 재발 여부, 연령, 조직학적 병기, 방사선 및 보조항암 치료여부는 본 연구에서 확인된 생존 예측에 가장 영향력을 준 주요 변수였다. 주 관심 변수인 진단 시 체질량지수(BMI) 또한 생존 예측에 독립된 인자로서 영향을 주는 주요 예측인자였으며, 특히 수술 전 진단 시 저체중인 군에서 낮은 생존율을 보이고, 과체중/비만 군에서 높은 생존율을 확인하며 폐암 비만 역설 가능성을 시사한다. 진단 시 체질량지수는 모든 모델에서 재발 예측에 영향을 주는 주요 예측인자로는 확인되지는 않았지만, 저체중 그룹의 경우 재발에 영향을 줄 수 있는 가능성을 발견하였다. 재발 예측에서의 주요 변수로는 수술 외 치료 여부 즉, 방사선 치료와 보조적 항암치료 변수, 조직학적 병기 및 흉막 침범여부로 나타났다.

비소세포폐암 환자의 수술 후 생존과 재발 가능성을 우수한 성능으로 예측하는 머신러닝 모델을 구축하였으며, SHAP 방법을 결합하여 설명 가능한 머신러닝 모델을 제시하고, 예후 예측에 영향력이 큰 중요 변수들을 결정함으로써 임상 실습에서 의료진이 더 나은 의학적 결정을 내리는 데 도움이 되는 도구로 사용되며, 귀중한 통찰력을 제공할 것이라 기대한다.

참고문헌

통계청. 사망원인통계연보, 2023

국가암정보센터. 통계로 보는 암
<https://cancer.go.kr/lay1/S1T645C646/contents.do>

중앙암등록본부 자료제공서비스
<https://kccrsurvey.cancer.go.kr/index.do>

환자 보고 성과 지표를 활용한 한국인 폐암 무병 생존자 생존 예측 모형 개발 : Cox 비례 위험 모델 및 다양한 머신 러닝 알고리즘을 적용하여[박사학위논문]. 서울: 서울대학교 대학원; 2018.

변상일. 머신러닝을 이용한 간암 환자 예후 예측 모델 개발[석사학위논문]. 서울: 연세대학교 보건대학원; 2022.

Major Cancer Mortality, National Cancer Information Center, 2023

Hosmer DW Jr, Lemeshow S, Sturdivant RX. In: Applied logistic regression. Hoboken: John Wiley & Sons; 2013.

Cunha MT, de Souza Borges AP, Carvalho Jardim V, Fujita A, de Castro G. Predicting survival in metastatic non-small cell lung cancer patients with poor ECOG-PS: A single-arm prospective study. Cancer Med. 2023; 12: 5099-109.

Wang ZH, Deng L. Establishment and Validation of a Predictive Nomogram for Postoperative Survival of Stage I Non-Small Cell Lung Cancer. Int J Gen Med. 2022 14;15:7287-98.

Yang Z, Wei S, Xia L, Liu L. [Prognostic Factors of N2 Stage Non-small Cell Lung Cancer after Surgery]. Zhongguo Fei Ai Za Zhi. 2020 20;23(1):15-20. Chinese.

Lei H, Li X, Ma W, Hong N, Liu C, Zhou W, Zhou H, Gong M, Wang Y, Wang G, Wu Y. Comparison of nomogram and machine-learning methods for predicting the survival of non-small cell lung cancer patients. Cancer Innov. 2022 30;1(2):135-145.

Jin L, Zhao Q, Fu S, Cao F, Hou B, Ma J. Development and validation of machine learning models to predict survival of patients with resected stage-III NSCLC. Front Oncol. 2023;13:1092478.

Ingle, Kshitija, Uttam Chaskar, and Shashikant Rathod. "Lung Cancer Types Prediction Using Machine Learning Approach." 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2021.

Zhang J, Fan J, Yin R, Geng L, Zhu M, Shen W, Wang Y, Cheng Y, Li Z, Dai J, Jin G, Hu Z, Ma H, Xu L, Shen H. A nomogram to predict overall survival of patients with early stage non-small cell lung cancer. J Thorac Dis. 2019;11(12):5407-16.

Fukumoto K, Mori S, Shintani Y, Okami J, Ito H, Ohtsuka T, Toyooka S, Mori T, Watanabe SI, Asamura H, Chida M, Date H, Endo S, Nagayasu T, Nakanishi R, Miyaoka E, Okumura M, Yoshino I; Japanese Joint Committee of Lung Cancer Registry. Impact of the preoperative body mass index on the postoperative outcomes in patients with completely resected non-small cell lung cancer: A retrospective analysis of 16,503 cases in a Japanese Lung Cancer Registry Study. Lung Cancer. 2020;149:120-9.

Alexander M, Wolfe R, Ball D, Conron M, Stirling RG, Solomon B, MacManus M,

Officer A, Karnam S, Burbury K, Evans SM. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. *Br J Cancer*. 2017;117(5):744-51.

Kar İ, Kocaman G, İbrahimov F, Enön S, Coşgun E, Elhan AH. Comparison of deep learning-based recurrence-free survival with random survival forest and Cox proportional hazard models in Stage-I NSCLC patients. *Cancer Med*. 2023;12(18):19272-78.

Lee HJ, Jo J, Son DS, Lee J, Choi YS, Kim K, Shim YM, Kim J. Predicting recurrence using the clinical factors of patients with non-small cell lung cancer after curative resection. *J Korean Med Sci*. 2009;24(5):824-30.

Dziedzic DA, Rudzinski P, Langfort R, Orlowski T; Polish Lung Cancer Study Group (PLCSG). Risk Factors for Local and Distant Recurrence After Surgical Treatment in Patients With Non-Small-Cell Lung Cancer. *Clin Lung Cancer*. 2016;17(5):e157-67.

Lee S, Jung J, Hong H, Kim BS. Prediction of Two-Year Recurrence-Free Survival in Operable NSCLC Patients Using Radiomic Features from Intra- and Size-Variant Peri-Tumoral Regions on Chest CT Images. *Diagnostics (Basel)*. 2022;12(6):1313.

Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, Schuchter LM, Shulman LN, Navathe AS, Patel MS, O'Connor NR. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Netw Open*. 2019;2(10):e1915997.

Zhai Y, Lin X, Wei Q, Pu Y, Pang Y. Interpretable prediction of cardiopulmonary complications after non-small cell lung cancer surgery based on machine learning and SHapley additive exPlanations. *Heliyon*. 2023;9(7):e17772.

Yi E, Bae MK, Cho S, Chung JH, Jheon S, Kim K. Pathological prognostic factors of recurrence in early stage lung adenocarcinoma. *ANZ J Surg.* 2018;88(4):327-31.

Uramoto H, Tanaka F. Prediction of recurrence after complete resection in patients with NSCLC. *Anticancer Res.* 2012;32(9):3953-60.

Maeda R, Yoshida J, Ishii G, Aokage K, Hishida T, Nishimura M, Nishiwaki Y, Nagai K. Long-term outcome and late recurrence in patients with completely resected stage IA non-small cell lung cancer. *J Thorac Oncol.* 2010;5(8):1246-50.

Zhang S, Yang L, Xu W, Wang Y, Han L, Zhao G, Cai T. Predicting the risk of lung cancer using machine learning: A large study based on UK Biobank. *Medicine (Baltimore).* 2024;103(16):e37879.

Dehdar S, Salimifard K, Mohammadi R, Marzban M, Saadatmand S, Fararouei M, Dianati-Nasab M. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. *Front Oncol.* 2023;13:1103369.

Janik A, Torrente M, Costabello L, Calvo V, Walsh B, Camps C, Mohamed SK, Ortega AL, Nováček V, Massutí B, Minervini P, Campelo MRG, Del Barco E, Bosch-Barrera J, Menasalvas E, Timilsina M, Provencio M. Machine Learning-Assisted Recurrence Prediction for Patients With Early-Stage Non-Small-Cell Lung Cancer. *JCO Clin Cancer Inform.* 2023;7:e2200062.

Sepesi B, Gold KA, Correa AM, Heymach JV, Vaporciyan AA, Roszik J, Dmitrovsky E, Liu X. The Influence of Body Mass Index on Overall Survival Following Surgical Resection of Non-Small Cell Lung Cancer. *J Thorac Oncol.* 2017;12(8):1280-7.

Yang Y, Xu L, Sun L, Zhang P, Farid SS. Machine learning application in personalised lung cancer recurrence and survivability prediction. *Comput Struct*

Biotechnol J. 2022;20:1811-20.

Sun J, Sun CK, Tang YX, Liu TC, Lu CJ. Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification. *Healthcare (Basel)*. 2023;11(14):2000.

Su L, Hounye AH, Pan Q, Miao K, Wang J, Hou M, Xiong L. Explainable cancer factors discovery: Shapley additive explanation for machine learning models demonstrates the best practices in the case of pancreatic cancer. *Pancreatology*. 2024;24(3):404-23.

Alabi RO, Almangush A, Elmusrati M, Leivo I, Mäkitie AA. An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer. *Int J Med Inform*. 2022;168:104896.

Choi JH, Choi Y, Lee KS, Ahn KH, Jang WY. Explainable Model Using Shapley Additive Explanations Approach on Wound Infection after Wide Soft Tissue Sarcoma Resection: "Big Data" Analysis Based on Health Insurance Review and Assessment Service Hub. *Medicina (Kaunas)*. 2024;60(2):327.

Zhang T, Li S, Chang J, Qin Y, Li C. Impact of BMI on the survival outcomes of non-small cell lung cancer patients treated with immune checkpoint inhibitors: a meta-analysis. *BMC Cancer*. 2023;23(1):1023.

Alifano M, Daffré E, Iannelli A, Bouchet L, Falcoz PE, Le Pimpec Barthes F, Bernard A, Pages PB, Thomas PA, Dahan M, Porcher R. The Reality of Lung Cancer Paradox: The Impact of Body Mass Index on Long-Term Survival of Resected Lung Cancer. A French Nationwide Analysis from the Epithor Database. *Cancers (Basel)*. 2021;13(18):4574.

Matsunaga T, Suzuki K, Imashimizu K, Banno T, Takamochi K, Oh S. Body Mass Index as a Prognostic Factor in Resected Lung Cancer: Obesity or Underweight, Which Is the Risk Factor? *Thorac Cardiovasc Surg.* 2015;63(7):551-7.

Wang H, Xu Q, Zhou L. Large unbalanced credit scoring using Lasso-logistic regression ensemble. *PLoS One.* 2015;10(2):e0117844.

Sievering AW, Wohlmuth P, Geßler N, Gunawardene MA, Herrlinger K, Bein B, Arnold D, Bergmann M, Nowak L, Gloeckner C, Koch I, Bachmann M, Herborn CU, Stang A. Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission. *BMC Med Inform Decis Mak.* 2022;22(1):309.

ABSTRACT

Applying Explainable Machine Learning Models for Prognostic Prediction in Non-Small Cell Lung Cancer Surgery Patients(Analyzing the Impact of Body Mass Index at Diagnosis)

Jang, Boyun
Graduate School of
Public Health Yonsei University

(Directed by professor Park, Sohee, Ph.D.)

Background:

According to the World Health Organization (WHO), lung cancer is the leading cause of cancer death globally, with approximately 2.2 million new cases annually and about 1.8 million deaths. In South Korea, lung cancer was the primary cause of cancer deaths in 2022, accounting for 22.3% (18,548 deaths) of all cancer fatalities from a total of 372,939 deaths. Lung cancer is classified histologically as either small cell lung cancer or non-small cell lung cancer (NSCLC), with the latter comprising over 85% of cases. NSCLC is characterized by its slower growth rate compared to small cell lung cancer and tends to metastasize to other parts of the body after invading surrounding tissues. Early detection and surgical intervention at a stage where surgery is viable are crucial due to the high rate of recurrence within two years post-surgery, necessitating ongoing management.

The purpose of this study is to evaluate the performance of models using traditional statistical techniques and various machine learning methodologies, and to identify key

variables, particularly Body Mass Index (BMI), that predict survival and recurrence in non-small cell lung cancer (NSCLC) patients who have undergone complete resection surgery. This research aims to understand the role of BMI in the predictive models for survival and recurrence of NSCLC patients, and to provide deep insights for clinical decision-making and patient management strategies based on these findings.

Methods:

We conducted a retrospective analysis of patients with Non-Small Cell Lung Cancer (NSCLC) who underwent complete resection (R0) from 2014 to 2017, utilizing data from the Korea lung cancer registry. To develop machine learning models, we applied regularized linear models such as Ridge and LASSO regression, along with ensemble models including Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Model performance was evaluated using sensitivity, specificity, accuracy, precision, F1 score, and AUC metrics, assessed through 5-fold cross-validation. The importance of variables was determined using the SHAP (Shapley Additive Explanation) method, and hyperparameters were optimized using grid search, followed by decision-making using explainable approaches.

Results:

All models demonstrated excellent performance with AUCs above 0.8, and among the survival predictions compared over three and five years, the five-year survival prediction performance was the highest. While there was no significant difference in performance between models, traditional statistical methods showed Ridge regression as most effective, whereas XGBoost performed best among machine learning models. The most influential variable identified for survival prediction was recurrence status, with age, histologic T stage, and treatment, as well as BMI at diagnosis, being identified as major prognostic factors influencing post-surgical survival. Notably, at the time of diagnosis, the underweight group showed a lower five-year survival rate, while the overweight/obese group exhibited a higher five-year survival rate. Although

BMI's impact on recurrence prediction was not significant, the underweight group showed a higher recurrence rate two years post-surgery compared to other groups. The main factors for recurrence prediction were identified as radiation therapy and adjuvant chemotherapy, histologic T stage, and pleural invasion.

Conclusion:

This study developed machine learning models that proficiently predict post-surgical survival and recurrence probabilities for non-small cell lung cancer (NSCLC) patients. There was no significant difference in performance between traditional statistical methods such as Ridge and Lasso logistic regression and more complex machine learning models like Random Forest (RF) and XGBoost. Additionally, the study applied explainable machine learning models that incorporate SHAP methods to assess the importance of variables in predicting survival and recurrence, and explored the potential for an obesity paradox in lung cancer.

Key words: Lung cancer, NSCLC, Machine Learning, Survival Prediction, Recurrence Prediction, BMI