MIMIC 기반 기존 연구 리뷰 및 재현 보고서

2024404060 강민혁

1 서론

1.1 선택한 논문의 기본 정보

본 보고서에서 다루는 논문의 제목은 **"A New Evaluation Model for Traumatic Severe Pneumothorax Based on Interpretable Machine Learning"**이다 [1]. 저자는 Jing Li, Yinzhen Lv, Jiayi Weng, Wei Chen, He Huang, Yuzhuo Zhao로 구성되어 있으며, 본 논문은 2024년 International Journal of Computers Communications & Control에 게재되었다. 위 논문은 MIMIC-IV 데이터셋을 기반으로 머신러닝 기법들을 활용하여 외상성 중증 기흉을 신속하고 해석 가능하게 평가하는 모델을 제안한 연구이다.

1.2 선택 이유 및 연구 주제의 중요성

여기에 연구를 선택한 이유와 주제의 임상적/데이터 과학적 중요성을 기술한다.

2 기존 논문 리뷰

2.1 연구 목적

해당 연구는 중환자실 임상 데이터베이스인 MIMIC-IV를 활용하여 외상성 중증 기흉을 신속하게 진단하고 평가할 수 있는, 해석 가능한 머신러닝 기반의 새로운 평가 모델을 제안하는 것을 주된 목적으로 한다. 레퍼런스 논문의 저자들은 기존의 영상 의학적 진단 방식이 응급 상황, 대량 사상자가 발생하는 재난 상황, 혹은 의료 자원이 부족한 환경에서 신속한 대응이 어렵다는 한계점에 주목하였다. 이에 대한 대안으로 영상 검사 결과가 도출되기 전, 환자의 활력 징후와 혈액 가스 분석 등에서 추출한 12가지 주요 임상 변수를 활용하여 기흉의 중증도를 조기에 판별할 수 있는 데이터 기반 모델을 구축하였다.

해당 연구는 XGBoost, ANN, SVM, KNN 등 네 가지 머신러닝 알고리즘의 예측 성능을 비교 분석하여 외상성 기흉 평가에 최적화된 알고리즘을 도출하고자 하였으며, SHAP을 통해 헤모글로빈, 산소화 지수, pH 등 핵심 인자가 모델의 예측 결과에 미치는 영향을 정량적으로 규명하고 임상적 타당성을 확보하는 데 중점을 두었다. 나아가 eICU 데이터셋을 활용한 외부 검증을 수행함으로써 제안된 모델의 강건성과 일반화 가능성을 입증하고자 하였다.

2.2 데이터 및 코호트 정의

해당 연구는 MIMIC-IV를 기반으로 하였다. 연구 대상자는 성인 환자 중 인구통계학적 정보와 병원 진료 기록이 온전한 환자로 한정하였으며, 분석의 신뢰도를 높이기 위해 폐질환 및 심장질환 등 기흉 발생이나 호흡기 지표에 영향을 줄 수 있는 기저 질환을 보유한 환자를 제외하였다. 또한, 데이터의 결측률이 50% 이상이거나 이상치가 포함된 샘플 역시 분석에서 배제하였다.

코호트는 외상성 중증 기흉 환자로 구성된 실험군과 비외상성 비기흉 환자로 구성된 대조군으로 분류하였다. 실험군은 외상으로 입원한 환자 중 CXR에서 대형 기흉(흉벽에서 폐 외연까지 2cm 이상 등)이 확인되었거나 흉강천자, 흉관 삽입술 등 기흉 치료를 받은 환자 174명으로 정의하였다. 대조군은 외상 및 기흉 소견이 없고 폐 관련 수술 이력이 없으며 생존 퇴원한 환자 3,697명이 포함되었다. 이때, 데이터 불균형 문제를 해소하기 위해 대조군에 대해 1:5 비율의 다운샘플링을 적용하였다.

예측 모델 구축을 위한 변수로는 활력 징후와 혈액 가스 분석 지표 33개를 1차적으로 추출하였다. 이 중 임상적 중요도와 결측률(50% 미만 기준)을 고려하여 최종적으로 Table 1와 같은 12개의 핵심 변수를 선별하였다. 선별된 데이터의 결측치는 다중 대체법을 적용하여 보정하였다.

Table 1: 예측 모델 구축을 위해 선별된 12개 핵심 변수 정의

변수명	단위	설명
pH	-	혈액의 산-염기 상태를 나타내는 지표
Hemoglobin	g/dL	혈액 내 산소 운반 능력을 반영하는 단백질 수치
PaO_2	mmHg	동맥혈 산소 분압; 동맥혈에 용해된 산소의 압력
Lactate	mmol/L	조직의 저산소증 및 혐기성 대사 상태를 반영하는 지표
Oxygenation Index	-	호흡 부전의 중증도를 평가하는 계산된 지표
SpO_2	%	경피적 산소 포화도; 말초 혈액의 산소 결합 비율
Base Excess	mEq/L	호흡성 요인을 제외한 대사성 산-염기 불균형 지표
Heart Rate	bpm	분당 심장 박동 수
$PaCO_2$	mmHg	동맥혈 이산화탄소 분압; 폐포의 환기 상태 지표
Systolic BP	mmHg	수축기 혈압; 심장이 수축할 때 혈관에 가해지는 압력
Diastolic BP	mmHg	이완기 혈압; 심장이 이완할 때 혈관에 가해지는 압력
Respiratory Rate	breaths/min	분당 호흡 횟수

Note: PaO_2, Partial pressure of oxygen; SpO_2, Peripheral oxygen saturation; $PaCO_2$, Partial pressure of carbon dioxide; BP, Blood pressure.

2.3 기존 연구 결과

기존 연구 결과, 4가지 기계학습 알고리즘 중 XGBoost가 가장 우수한 진단 성능을 보였다(Figure 1 좌측 표). MIMIC-IV 데이터셋을 기반으로 한 내부 검증에서 XGBoost는 AUROC 0.979 (95% CI: 0.966-0.989), 정확도 0.947, 재현율 0.730을 기록하며 ANN, SVM, KNN 등 타 모델을 전반적으로 상회하는 성능을 입증하였다. 또한, EICU 데이터셋을 활용한 외부 검증에서도 타 모델들과 다르게 XGBoost는 AUROC 0.806을 기록하며 데이터셋 변화에 대해서도 상대적으로 강건한 일반화 성능을 유지하였다.

SHAP 결과는 Figure 1의 우측 그래프와 같다. 분석 결과, 외상성 중증 기흉 판별에 가장 큰 영향을 미치는 상위 3가지 주요 인자는 pH, Hemoglobin, PaO_2인 것으로 확인되었다. 이 세 가지 인자의 수치가 낮을수록 모델이 기흉 위험도를 높게 평가하는 경향을 보였다. 이는 외상성 기흉 환자에게서 나타나는 병태생리학적 특징인 다량의 출혈, 호흡 부전으로 인한 저산소 및 산증과 임상적으로 일치하는 결과이다. 반면 Lactate 수치는 높을수록 기흉의 가능성을 높게 예측하는 주요 인자로 확인되었다.

Model	Dataset	Accuracy	F1 Score	Precision	Recall	AUROC (95% CI)	AUPRC	Sensitivity	NPV	PPV
XGBoost	MIMIC-IV	0.947	0.822	0.941	0.730	0.979 (0.966, 0.989)	0.926	0.991	0.730	0.941
ANN	MIMIC-IV	0.882	0.529	0.793	0.397	0.915 (0.893, 0.936)	0.717	0.979	0.397	0.793
SVM	MIMIC-IV	0.934	0.761	0.957	0.632	0.967 (0.946, 0.985)	0.911	0.994	0.632	0.957
KNN	MIMIC-IV	0.943	0.816	0.891	0.753	0.948 (0.923, 0.969)	0.900	0.982	0.753	0.891
XGBoost	80%MIMIC-IV+20%EICU	0.718	0.512	0.360	0.886	0.806 (0.740, 0.864)	0.359	0.886	0.684	0.967
ANN	80%MIMIC-IV+20%EICU	0.627	0.426	0.287	0.829	0.703 (0.628, 0.780)	0.238	0.829	0.586	0.944
SVM	80%MIMIC-IV+20%EICU	0.636	0.415	0.284	0.771	0.705 (0.621, 0.780)	0.247	0.771	0.609	0.930
KNN	80%MIMIC-IV+20%EICU	0.679	0.247	0.204	0.314	0.655 (0.565, 0.740)	0.215	0.314	0.753	0.845

XGBoost: eXtreme Gradient Boosting; ANN: Artificial Neural Network; SVM: Support Vector Machine; KNN: k-Nearest Neighbors; AUROC: Area Under Receiver Operating Characteristic Curve; AUPRC: Area Under Precision-Recall Curve; NPV: Negative Predictive Value; PPV: Positive Predictive Value.

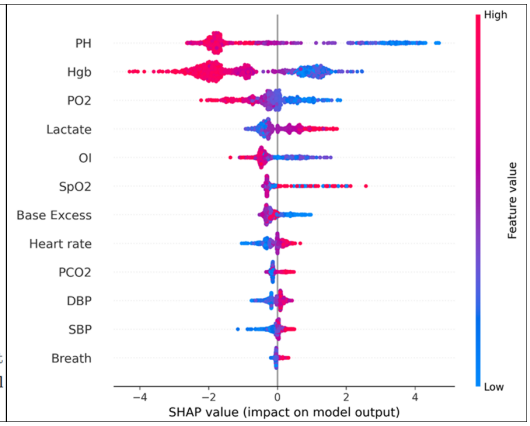

Figure 1: Original Research Performance Evaluation and SHAP Analysis Results. (Left) Model Performance Comparison Table for MIMIC-IV Internal Validation and EICU External Validation. (Right) SHAP Summary Plot for XGBoost Model, showing pH, Hgb, PO2 as the most influential factors in prediction.

3 재현 방법

3.1 코호트 구성 및 데이터 파이프라인

본 재현 연구에서는 원 논문에서 최종 선별된 12개 변수의 유효성을 검증하고, 데이터 추출 과정의 완결성을 높이기 위해 초기 후보 변수군을 확장하였다. 원 논문의 12개 변수에 FiO_2(흡입 산소 농도와 평균 기도 압력)를 추가하여 pH, Hemoglobin, PaO_2, Lactate, SpO_2, Base Excess, Heart Rate, $PaCO_2$, DBP, SBP, Respiratory Rate, FiO_2, MAP, Oxygenation Index, 총 14개의 변수를 수집하였다.

초기 변수 목록이 확장된 주된 이유는 주요 예측 인자인 Oxygenation Index의 정확한 산출을 위해서이다. MIMIC-IV 데이터베이스 특성상 Oxygenation Index가 직접 기록되지 않은 레코드가 존재할 수 있으며, 이 경우 아래의 식 1을 통해 결측치를 보완하거나 값을 검증해야 하므로 구성 요소인 FiO_2와 MAP의 수집이 필수적이었다.

$$\text{Oxygenation Index} = \frac{FiO_2 \times \text{MAP}}{PaO_2} \times 100 \tag{1}$$

수집된 14개 변수에 대해 결측률을 산출한 후, 결측률 50% 초과 변수를 기준으로, 결측 패턴에 높은 가중치를 부여하였으며, 가중치를 기준으로 우선적으로 제거 대상 샘플을 선정하도록 하였다.

기존 연구에서는 결측치 보정을 위해 다중대체법을 사용했다고 서술하였지만, 어떠한 알고리즘을 사용했는지에 대해서는 서술하지 않았다. 이에 본 보고서는 유사한 일련의 연구에서 다중대체법으로 많이 사용되어왔던 MICE를 활용하였다. 최종 모델링 단계에서는 원 논문과의 비교 가능성을 위해 계산 목적으로 사용된 FiO_2와 MAP를 제외하고, 핵심 12개 변수만을 입력값으로 확정하였다.

또한 외상성 중증 기흉 환자(실험군)와 비기흉 환자(대조군) 간의 극심한 클래스 불균형 문제를 해결하기 위해 기존 연구처럼 1:5 비율의 다운샘플링을 적용하였다. 이때, 대조군은 무작위 샘플링으로 선정하였다.

3.2 사용한 모델 및 하이퍼파라미터

본 프로젝트의 베이스라인 모델은 기존 연구에서 가장 좋은 성능을 냈던 XGBoost로 채택하였다. 이때, 기존 연구와 동일하게 이진 로지스틱 손실 함수를 목적 함수로 설정하였다.

그리드 서치를 통해 하이퍼파라미터를 탐색하였다. 단, 기존 연구에서 하이퍼파라미터 후보값을 기재하지 않아 후보값은 임의로 선정하였으며, 선정한 후보값은 Table 2과 같다. 총 972개의 조합에 대해 5-Fold Stratified Cross Validation을 수행하여 총 4,860회의 모델 학습을 진행하였다. 각 폴드에서 AUROC를 기준으로 평가하여 가장 높은 평균 AUROC를 달성한 하이퍼파라미터 조합을 최종 모델 구성에 사용하였다.

평가 지표는 기존 연구와 동일하게 Accuracy, F1 Score, Precision, Recall, AUROC, AUPRC, Sensitivity, NPV, PPV를 사용하여 성능을 측정하였다.

Table 2: XGBoost Hyperparameter Search Range and Optimal Hyperparameter

Hyperparameter	Search Range	Optimal Hyperparameter
n_estimators	[100, 200, 300]	100
max_depth	[3, 5, 7]	5
learning_rate	[0.01, 0.05, 0.1]	0.05
subsample	[0.8, 1.0]	0.8
colsample_bytree	[0.8, 1.0]	1.0
min_child_weight	[1, 3, 5]	1
gamma	[0, 0.1, 0.2]	0.2

3.3 실행 환경

본 보고서의 모든 실험은 X86환경의 Linux 환경에서 수행되었으며, 하드웨어는 AMD Ryzen 9 5900X, 64GB RAM, 그리고 NVIDIA GeForce RTX 3080 GPU로 구성되었다. 데이터 전처리, 모델 학습 및 결과 시각화를 위해 사용된 주요 소프트웨어 라이브러리의 버전은 Table 3과 같으며, 전체 소스 코드는 깃허브[1]에서 확인할 수 있다.

4 실험 결과

4.1 성능 지표 비교

원 논문과 재현 실험 간의 결과를 비교한다.

[1] https://github.com/ga111o/mimic-pneumothorax-progression-analysis

Table 3: Environment and Key Library Versions

Software / Library	Version
Python	3.12.12
CUDA	13.0
XGBoost	3.1.2
Scikit-learn	1.7.2
SHAP	0.48.0
Pandas	2.3.3
NumPy	2.3.4
SciPy	1.16.3

Table 4: Diagnostic Performance of Machine Learning Mode in Original Stuey and Reproduction

Model	Dataset	Accuracy	F1 Score	Precision	Recall	AUROC (95% CI)	AUPRC	Sensitivity	NPV	PPV
Original Study										
XGBoost	MIMIC-IV	0.947	0.822	0.941	0.730	0.979 (0.966, 0.989)	0.926	0.991	0.730	0.941
Reproduction (Ours)										
XGBoost	MIMIC-IV	0.885 ± 0.016	0.557 ± 0.076	0.779 ± 0.083	0.442 ± 0.082	0.910 ± 0.038	0.764 ± 0.066	0.442 ± 0.082	0.897 ± 0.013	0.779 ± 0.083

4.2 시각화 결과

ROC curve, PR curve, confusion matrix 등의 시각자료를 삽입한다.

5 논의 및 결론

5.1 재현 성공/실패 요인 분석

결과가 일치하지 않았거나 차이가 발생한 원인을 데이터, 모델, 실행 환경 측면에서 분석한다.

5.2 데이터/모델/환경 차이 검토

원 논문과의 차이를 구체적으로 기술하고, 재현 결과에 미친 영향을 논의한다.

5.3 연구에 대한 비판적 시각 및 개선 방안

연구 설계 상의 한계점, 일반화 가능성, 향후 연구의 발전 방향을 제시한다.

1. 기존 연구와 cohort가 다른 것. 실험군/대조군 선정 기준의 차이 의심„ 기존 연구는 실험군/대조군 선정에 자세한 그 뭐냐 그게 없었음 2. 외부 검증 안되었던 거 3. 사이즈 너무 작았던 거

References

[1] Li, J., Lv, Y., Weng, J., Chen, W., Huang, H., & Zhao, Y. (2024). A new evaluation model for traumatic severe pneumothorax based on interpretable machine learning. International Journal of Computers Communications & Control, 19(1), Article 6830. https://doi.org/10.15837/ijccc.2025.1.6830