

Tention PTX Early Detection

강동우, 강민혁, 채상윤, 최서영
DataAnalytics Team7

Abstract

긴장성 기흉은 중환자실 환경에서 즉각적인 의료 개입이 필요한 생명을 위협하는 응급 상황이다. 본 프로젝트는 MIMIC-IV 데이터베이스를 활용하여 생체 신호 및 동맥혈 가스 분석 데이터로부터 긴장성 기흉의 조기 징후를 탐지하는 머신러닝 모델을 개발하였다. 총 4,950명의 환자 데이터로부터 256개의 시계열 특성을 추출하였으며, 도메인 지식 기반 파생 변수 및 고차 통계 특성을 포함하였다. 데이터 품질 확보를 위해 결측률 가중치 기반의 탐욕적 샘플 제거 알고리즘과 비대칭 임계값 기반 특성 선택을 제안 및 적용하였으며, MICE 기법을 통한 다중 대체로 결측치를 처리하였다. 최종적으로 선별된 76개의 특성과 900명의 데이터를 이용하여 XGBoost와 LightGBM 모델을 Optuna 프레임워크로 최적화하였으며, Focal Loss와 Cost-sensitive Learning 을 통해 클래스 불균형 문제를 해결하였다. 5-fold 교차 검증 결과, XGBoost 모델이 AUROC 0.7457, AUPRC 0.6623, F1-score 0.7011을 달성하여 가장 우수한 성능을 보였다. 특히 Recall 0.7276을 기록하여 조기 탐지라는 임상적 목적에 부합하는 결과를 나타냈다. 본 프로젝트는 실시간 생체 신호 모니터링을 통한 긴장성 기흉의 조기 경보 시스템 구축 가능성을 제시한다.

Keywords: 긴장성 기흉, 조기 탐지, 머신러닝, XGBoost, 생체 신호 분석, MIMIC-IV

1 Introduction

긴장성 기흉(Tension Pneumothorax, Tension PTX)은 흉막강 내에 공기가 축적되어 폐가 허탈되고, 종격동이 편위되며, 정맥 환류가 감소하여 심혈관계 허탈을 초래하는 치명적인 응급 질환이다. 중환자실 환경에서는 기계 환기, 중심 정맥 카테터 삽입 등의 침습적 시술 중 의인성 기흉이 발생할 위험이 높으며, 신속한 진단과 치료가 이루어지지 않을 경우 수분 내에 사망에 이를 수 있다.

전통적으로 긴장성 기흉의 진단은 임상적 징후(경정맥 확장, 기관 편위, 저혈압)를 바탕으로 한 의사의 판단과 영상 검사(흉부 X-ray, CT)에 의존해왔으나, 이러한 방법은 기존에 기흉 병력이 있었거나, 증상이 명확해진 후에야 확인 가능하다는 한계가 있다. 특히 중환자실 환자의 경우 진정 상태 또는 의식 저하로 인해 주관적 증상 호소가 어렵고, 다양한 기저 질환으로 인해 생체 신호의 변화가 비특이적으로 나타나 조기 발견이 더욱 어렵다.

최근 의료 빅데이터와 머신러닝 기술의 발전으로 생체 신호의 미세한 패턴 변화를 분석하여 임상적 악화를 예측하는 연구가 활발히 진행되고 있다([rajkumar2019](#)). 특히 MIMIC(Medical Information Mart for Intensive Care) 데이터베이스는 중환자실 환자의 대규모 의료 데이터를 제공하여 다양한 예측 모델 개발의 기반이 되고 있다([johnson2016](#)).

본 프로젝트는 MIMIC-IV 데이터베이스를 활용하여 흉관 삽입이 필요한 임상적으로 유의미한 기흉 환자를 대상으로, 생체 신호의 시계열 패턴 분석을 통해 긴장성 기흉의 조기

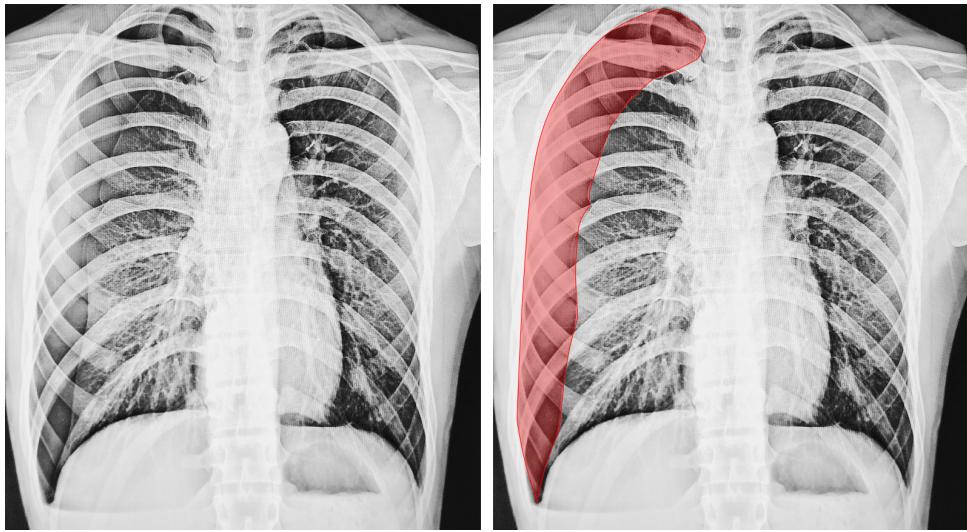


Figure 1: 긴장성 기흉 예시

탐지 가능성을 검증하고자 한다. 이를 위해 도메인 지식 기반의 특성 공학, 강건한 전처리 파이프라인, 그리고 불균형 데이터 학습에 최적화된 머신러닝 모델을 구축하였다.

2 Method

2.1 Data Collection and Cohort Definition

2.1.1 Data Sources

본 프로젝트는 교수님께서 제공해주신 SQLite 형태의 MIMIC-IV v2.2 데이터베이스와 MIMIC-CXR v2.0 영상 라벨 데이터를 활용하였다. MIMIC-IV는 2008년부터 2019년 까지 Beth Israel Deaconess Medical Center의 중환자실에 입원한 환자의 비식별화된 의료 기록을 포함한다(johnson2023mimiciv).

2.1.2 Cohort Selection Criteria

긴장성 기흉 조기 탐지 모델 학습을 위해 임상 기록 및 영상 검사 결과를 종합하여 신뢰도 수준에 따라 환자군을 세 가지 코호트로 분류하였다(Table 1). 각 코호트는 ICD 진단 코드, 시술 코드, 그리고 흉부 X-ray 영상 라벨(CheXpert, NegBio)의 조합으로 정의되었다.

GOLD 코호트는 가장 높은 신뢰도를 가진 실험군으로, 기흉 관련 ICD 진단 코드(ICD-9: 512%, ICD-10: J93%, S270%/S27.0%)와 함께 흉관 삽입 시술 코드(ICD-9 Procedure: 3491, 3404)가 모두 확인된 환자를 포함한다. 흉관 삽입술은 임상적으로 중등도 이상의 기흉에서 시행되는 침습적 치료이므로, 이 군은 긴급한 치료가 필요한 기흉 환자로 간주할 수 있다.

SILVER 코호트는 기흉 진단 코드는 있으나 흉관 삽입 시술은 시행되지 않았으며, CheXpert 또는 NegBio 라벨링 시스템 중 하나 이상에서 기흉이 양성으로 확인된 환자군이다. 이 코호트는 경증에서 중등도의 기흉으로 추정되며, 보존적 치료(관찰, 산소 요법 등)로 관리된 것으로 간주된다.

CLEAN 코호트는 음성 대조군으로, 기흉 관련 진단 코드 및 흉관 삽입 시술 기록이 전혀 없고, CheXpert와 NegBio 영상 라벨 모두에서 기흉이 음성($Pneumothorax \neq 1.0$)으로 확인된 환자를 포함한다. 이 코호트는 기흉이 없음이 임상 기록과 영상 소견 모두에서

확인되어 가장 신뢰할 수 있는 대조군이며, 표본 크기가 가장 크다. 데이터 불균형 해소를 위해 전체 대조군 중 약 5%를 무작위 샘플링하여 최종 분석에 사용하였다.

Table 1: Cohort Definition and Selection Criteria

Cohort	ICD Diagnosis Code	ICD Code	Procedure	Radiology
GOLD (Experimental)	ICD-9: 512% ICD-10: J93%, S270%	ICD-9 Proc: 3491, 3404	-	
SILVER (Diagnosis Confirmed)	ICD-9: 512% ICD-10: J93%, S270%	None	CheXpert or NegBio Positive	
CLEAN (Control)	None	None	CheXpert and NegBio Negative	

2.1.3 Reference Time Definition

조기 탐지 모델 학습을 위해 각 환자의 입원별로 데이터 추출의 기준이 되는 시점(t_{ref})을 정의하였다. 실험군의 경우 흉관 삽입술이 시행된 시점을 기준 시점으로 설정하였다. 대조군의 경우 입원 극 초기의 불안정한 상태를 배제하기 위해 입원 시점으로부터 24시간이 경과한 시점을 기준 시점으로 설정하였다.

2.2 Biological Signal and Clinical Variable Extraction

2.2.1 Target Variables

임상적 중요도와 데이터 가용성을 고려하여 호흡기 및 혈역학적 상태를 반영하는 주요 변수를 선정하였다(Table 2).

Table 2: Extracted Biological Signals and Clinical Variables

Category	Variable	Description	Main ITEMID
Vital Sign	HR	Heart Rate	220045
	RR	Respiratory Rate	220210, 224690
	SpO2	Oxygen Saturation	220277
	SBP/DBP	Systolic/Diastolic Blood Pressure	220179, 220180
	FiO2	Inhaled Oxygen Concentration	223835, 227009
Arterial Blood Gas	pH	Blood pH	50820
	PaO2	Arterial Oxygen Pressure	50821
	PaCO2	Arterial Carbon Dioxide Pressure	50818
	HCO3	Bicarbonate	50882

2.2.2 Time Window

설정된 기준 시점(t_{ref})을 기준으로 과거 3시간 동안의 데이터를 시간 윈도우로 설정하여 추출하였다.

$$t_{ref} - 3h \leq t_{chart} \leq t_{ref}, \quad \text{where } t_{ref} \geq t_{admit} + 3h \quad (1)$$

2.3 Feature Engineering

2.3.1 Domain Knowledge-based Derived Variables

단일 변수만으로는 포착하기 어려운 환자의 혈역학적 불안정성과 호흡 부전 상태를 복합적으로 반영하기 위해 임상 지표를 계산하였다(Table 3).

Table 3: Domain Knowledge-based Derived Variables

Category	Variable	Formula/Definition	Clinical Significance
Hemodynamics	MAP	$DBP + \frac{1}{3}(SBP - DBP)$	Mean arterial pressure, indicator of organ perfusion
	Pulse Pressure	$SBP - DBP$	Pulse pressure, reflects cardiovascular stiffness
	Shock Index	HR/SBP	Early warning for acute hemorrhage and septic shock
Respiratory	MSI	HR/MAP	Modified shock index
	P/F Ratio	PaO_2/FiO_2	Severity classification criterion for ARDS
	ROX Index	$(SpO_2/FiO_2)/RR$	Predictor of high-flow nasal cannula therapy success
	AA Gradient	$P_AO_2 - P_aO_2$	Alveolar-arterial oxygen pressure difference
	RDI	$RR/SpO_2 \times 100$	Respiratory distress index

2.3.2 Time Series Feature Analysis

Raw Signal의 고차원성을 해소하고, 트리 기반 모델이 학습하기 어려운 시계열의 비선형적 동특성과 단기 변동성을 명시적 특징으로 변환하여 모델의 예측 성능과 임상적 설명력을 동시에 확보하고자 하였기에, 원시 시계열 데이터를 직접 사용하기보다 통계량 기반 특성 추출 방식을 채택하였다. 이벤트 발생 전 생체 신호의 미세한 변화를 포착하기 위해 고차 통계량과 신호 복잡도를 나타내는 특성을 추출하였다. 먼저, 데이터 분포의 비대칭도와 첨도를 계산하여 급격한 값의 변화나 이상치 발생 경향을 정량화하였다. 변동성 동역학을 평가하기 위해 인접한 측정값 간 차이의 제곱평균제곱근과 중위수 절대 편차를 산출하였다. 또한 생체 신호의 규칙성과 복잡성을 측정하기 위해, 즉, 병리학적 상태에서 신호가 지나치게 규칙적이거나 무질서해지는 경향을 반영하기 위해 샘플 엔트로피를 계산하였다. 마지막으로 비모수적 추세 검정인 Mann-Kendall Tau를 적용하여 관측 윈도우 내에서 수치가 일관되게 증가하거나 감소하는 경향성을 정량화하였다.

2.3.3 Multiple Time Window

임박한 이벤트의 조기 징후를 포착하면서도 전반적인 환자 상태의 변화 추이를 반영하기 위해, 기준 시점으로부터 30분, 60분, 120분의 세 가지 시간 윈도우를 설정하였다. 각 윈도우별로 통계적 특성을 독립적으로 추출하여, 모델이 생리학적 악화의 시간적 패턴을 학습할 수 있도록 하였다. 모든 시간 윈도우와 변수를 결합하여 총 256개의 특성이 추출되었다. 추출된 특성의 상세 목록은 함께 위치한 부록(Appendix.pdf)에 제시하였다.

2.4 Data Preprocessing

2.4.1 Missing Value Weight based Sample Removal for Data Quality Assurance

데이터셋 D 를 N 개의 표본과 M 개의 피쳐로 구성된 행렬 $X \in \mathbb{R}^{N \times M}$ 이라 하자. 각 표본 i 는 그룹 $g_i \in \{G_{ctrl}, G_{exp}\}$ 에 속한다.

결측 여부를 나타내는 지시 행렬 R 을 다음과 같이 정의한다.

$$r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

각 피쳐 j 에 대한 그룹 G 에서의 결측률 $\mu_j(G)$ 는 다음과 같다.

$$\mu_j(G) = \frac{1}{|G|} \sum_{i \in G} r_{ij} \quad (3)$$

Greedy Sample Removal 이 알고리즘의 목적은 대조군 표본 집합 S_{ctrl} 의 부분집합 S'_{ctrl} 을 찾아, 모든 특성에 대해 결측률이 주어진 임계값 θ_{ctrl} 이하가 되도록 하는 것이다. 본 프로젝트에서는 대조군과 실험군에 대해 각각 다른 임계값을 설정하였으며, 대조군의 경우 $\theta_{ctrl} = 50\%$, 실험군의 경우 $\theta_{exp} = 65\%$ 로 설정하였다. 또한 대조군의 최소 표본 수를 실험군의 5배인 750개로 유지하도록 제한하였다.

이는 표본의 수를 최대한 유지하는 최적화 문제이나, 피쳐 엔지니어링으로 늘어난 피쳐의 수와 대조군의 샘플 수로 인한 계산 복잡도를 고려하여 탐욕적 알고리즘을 사용하였다.

Removal Priority Score 각 반복 t 에서, 임계값을 초과하는 특성들의 집합을 $F_{excess}^{(t)} = \{j \mid \mu_j(S_{ctrl}^{(t)}) > \theta_{ctrl}\}$ 이라 할 때, 표본 i 의 제거 점수 $Score_i^{(t)}$ 는 다음과 같이 정의된다.

$$w_j^{(t)} = \mu_j(S_{ctrl}^{(t)}) - \theta_{ctrl} \quad (\text{Excess Weight}) \quad (4)$$

$$Score_i^{(t)} = \sum_{j \in F_{excess}^{(t)}} w_j^{(t)} \cdot r_{ij} \quad (5)$$

즉, 결측률이 높게 초과된 특성을 많이 가지고 있지 않은 표본일수록 높은 점수를 받아 우선적으로 제거된다.

일반적으로 결측률이 높은 특성을 일괄 삭제하면 중요한 임상 정보를 잃을 위험이 있다. 본 프로젝트에서는 대조군의 샘플 수가 충분하다는 점에 착안하여, 특성을 보존하기 위해 질 낮은 대조군 샘플을 제거하는 결측률 가중치 기반 알고리즘을 기반한 샘플 제거 알고리즘을 제안 및 적용하였다. 이 알고리즘은 특정 특성의 결측률이 허용 임계값($\theta_{control} = 50\%$)을 초과할 경우, 해당 결측에 대한 기여도가 높은 대조군 샘플을 우선적으로 반복 제거하는

방식으로 작동한다. 이를 통해 대조군 샘플 크기를 일부 희생하더라도 희소한 실험군을 유지하면서 모델링에 사용 가능한 유효 특성의 수를 최대화하였다.

2.4.2 Asymmetric Threshold-based Feature Selection

희소한 실험군 데이터와 풍부한 대조군 데이터의 특성을 고려하여, 특성 선정 기준에 비대칭 임계값을 적용하였다. 실험군의 경우 희소한 양성 환자 데이터를 최대한 보존하기 위해 결측률이 65% 미만인 특성을 유지하였다. 반면 대조군에는 고품질 데이터만을 선별하고 학습 안정성을 확보하기 위해 50%라는 더 엄격한 임계값을 적용하였다. 이러한 비대칭적 접근법은 희귀 질환 사례로부터 최대한의 정보를 추출하면서도, 더 큰 규모의 대조군에 대해서는 데이터 품질 기준을 유지하는 균형을 이루었다.

2.4.3 Multiple Imputation for Missing Values

단순 평균 대치법이 야기하는 변수 간 상관관계 왜곡과 분산 축소 문제를 방지하기 위해, 특히, 의료 도메인에서 생체 신호 간의 다변량 상관관계를 학습하여 생리학적으로 타당한 값을 생성해낼 수 있도록 MICE(Multivariate Imputation by Chained Equations) 기법을 적용하였다 **buuren2011**. 단일 대체의 편향을 줄이기 위해 서로 다른 난수 시드를 적용하여 5개의 독립적인 대체 데이터셋을 생성하였으며, 이들의 평균값을 최종 결측 보정값으로 사용하였다.

2.4.4 Normalization

StandardScaler를 적용하여 모든 특성의 평균을 0, 표준편차를 1로 정규화하였다.

2.4.5 Final Dataset Construction

전처리 이후, SILVER 코호트를 배제하고 확진된 GOLD 코호트와 정제된 CLEAN 코호트만을 사용하여 최종 데이터셋을 구성하였다. 최종적으로 결측치가 존재하지 않는 완전한 데이터셋이 생성되었다. 각 단계별 코호트의 변화는 Table 4와 같다.

Table 4: Number of Subjects at Each Preprocessing Step

Step	GOLD	SILVER	CLEAN	Total
Initial Cohort	704	409	539,842	540,955
Feature Extraction	150	409	4,391	4,950
Sample Removal	150	-	750	900

2.5 Machine Learning Pipeline

2.5.1 Feature Selection

Permutation Importance 기반의 특성 선택을 수행하였다(반복 횟수: 5회). 최종적으로 총 256개의 특성 중 76개의 유의미한 특성이 모델 학습에 사용되었다. 사용한 특성의 상세 목록은 함께 위치한 부록(Appendix.pdf)에 제시하였다.

2.5.2 Cost-sensitive Learning for Class Imbalance

실제 임상 환경에서는 실험군보다 대조군이 훨씬 많다는 점을 고려하여, 별도의 합성 샘플 생성 기법을 적용하지 않고 Cost-sensitive Learning 방식을 채택하였다. 소수 클래스 가중치 최적화를 위해 모델의 양성 클래스 가중치 파라미터를 제어하는 승수를 Optuna TPE 샘플러로 5.0에서 25.0 범위에서 탐색하였으며, 최종 가중치는 다음과 같이 계산하였다.

$$\text{scale_pos_weight} = \frac{n_{\text{negative}}}{n_{\text{positive}}} \times \text{multiplier} \quad (6)$$

또한 불균형 데이터 학습의 효율성을 높이기 위해 Focal Loss를 도입하였다^{lin2017}. Focal Loss는 쉽게 분류되는 샘플의 손실 가중치를 줄이고, 어려운 샘플에 집중하도록 설계된 손실 함수로, 다음과 같이 정의된다.

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (7)$$

여기서 p_t 는 올바른 클래스에 대한 예측 확률이며, α_t 는 클래스별 가중치, γ 는 focusing parameter이다. p_t 와 α_t 는 다음과 같이 정의된다.

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases}, \quad \alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{if } y = 0 \end{cases} \quad (8)$$

본 연구에서는 $\alpha = 0.25$, $\gamma = 2.0$ 을 사용하였으며, XGBoost와 LightGBM 모두에서 objective function으로 구현하여 적용하였다.

2.5.3 Cross-validation

강건한 모델 학습을 위해 동일 환자의 데이터가 학습 데이터와 검증 데이터에 동시에 포함되지 않도록 GroupKFold 방식을 사용하여 데이터를 분리하여 학습한 후, 일반화 성능을 평가하였다(5-Fold).

2.5.4 Baseline Model

비교를 목적으로 해석력이 높은 로지스틱 회귀 기반의 기반 모델을 구축하였다. 기반 모델은 이벤트 시점 이전 3시간 동안의 생체 신호를 6개의 시간 구간으로 나누어 각 구간별로 평균, 최솟값, 최댓값 통계치를 추출하여 총 54개의 특성을 생성하였다. 결측치 처리는 중앙값을 사용하였으며, StandardScaler로 정규화를 수행하였다. 클래스 불균형 문제는 SMOTE를 통해 해결하였다. 로지스틱 회귀 모델은 L-BFGS 솔버를 기반으로 구축하였으며, L2 페널티를 적용해 일반화 성능을 높였다. 학습 과정의 최대 반복 횟수는 1,000회로 제한하였다. 회귀계수 분석 결과, SBP_min_t1, HR_max_t5, DBP_min_t4 등 혈압 및 심박수 관련 특성의 변동성이 기흉 예측에 주요 기여를 하는 것으로 나타났다.

2.5.5 Model Selection and Hyperparameter Optimization

본 프로젝트에서는 데이터의 수가 적고, Tabular Feature Engineering이 중요하여 LSTM, GRU, Transformer와 같은 시계열 딥러닝 모델이 아닌 트리 기반 모델인 XG-Boost와 LightGBM을 사용하였다. 하이퍼파라미터 최적화는 Optuna 프레임워크^{akiba2019}를 활용하였으며, Tree-structured Parzen Estimator Sampler를 탐색 알고리즘으로 채택하였다. 비효율적인 하이퍼파라미터 조합의 조기 종료를 위해 MedianPruner를 적용

하였으며, 과적합 방지를 위해 검증 성능이 30 라운드 동안 향상되지 않을 경우 학습을 조기 종료하도록 설정하였다. 각 모델에 대해 100회의 최적화 시도를 수행하였으며, 최적화 목표 지표로 LogLoss를 최소화하도록 설정하였다.

2.5.6 Threshold Optimization

기본 임계값(0.5) 대신, 각 Fold 및 모델별로 F1-score를 최대화하는 최적 임계값을 탐색하여 적용하였다.

2.5.7 Ensemble Strategy

단일 모델의 한계를 극복하고 예측 안정성을 높이기 위해 두 가지 앙상블 전략을 적용하였다. Soft Voting 방식에서는 XGBoost와 LightGBM의 예측 확률값을 가중 평균하였으며, 0.00에서 1.00 범위에서 가중치를 탐색하여 AUROC를 최대화하는 최적 가중치를 선정하였다. Stacking 방식에서는 XGBoost와 LightGBM을 Base model로 사용하고, 이들의 예측 확률값을 메타 특성으로 하여 Logistic Regression을 Meta-learner로 학습시켰다. Meta-learner의 최적화는 L-BFGS solver와 정규화 파라미터 C=1.0를 사용하였다.

3 Results

3.1 Optimal Hyperparameters

Optuna 최적화를 통해 도출된 각 모델의 최적 하이퍼파라미터는 Table 5와 같다. 파이프라인 전체에서 사용된 하이퍼파라미터의 상세 목록은 최적 하이퍼파라미터 값은 함께 위치한 부록(Appendix.pdf)에 제시하였다.

Table 5: Optimal Hyperparameters Results

Parameter	XGBoost	LightGBM
n_estimators	158	263
max_depth	4	7
learning_rate	0.0411	0.0122
pos_weight_multiplier	13.21	17.69

3.2 Ablation Study on Preprocessing Strategies

Table 6: Ablation Study of Preprocessing Strategies

Strategy	AUROC	AUPRC	Recall	Precision	Accuracy	F1-score
Baseline(Without Preprocessing)	0.7657	0.1140	0.2418	0.1086	0.9028	0.1499
Greedy Sample Removal	0.6470	0.4122	0.3673	0.4148	0.7855	0.3896
Greedy Sample Removal + Asymmetric Thresh (Proposed)	0.7457	0.6623	0.7276	0.6764	0.7678	0.7011

제안한 결측률 가중치 기반 샘플 제거 알고리즘과 비대칭 임계값 기반 특성 선택이 들어가지 않은 베이스라인 모델은 높은 정확도(0.9028)와 AUROC(0.7657)를 보였으나, AUPRC(0.1140)와 F1-score(0.1499)가 극도로 저조하였다. 이는 모델이 다수 클래스인 대조군에 과적합되어, 소수 클래스인 실제 긴장성 기흉을 거의 탐지하지 못하는(Recall 0.2418) 문제가 발생한 것으로, 단순 고정 임계값 적용으로 인해 실험군의 중요한 정보가 손실되었거나, 질 낮은 대조군 데이터가 노이즈로 작용한 결과로 판단된다.

여기에 제안한 결측률 가중치 기반 샘플 제거 알고리즘을 적용한 결과, AUROC와 정확도는 다소 감소하였으나, AUPRC가 0.1140에서 0.4122로, F1-score가 0.1499에서 0.3896으로 대폭 향상되었다. 이는 결측률이 높은 저품질 샘플을 선별적으로 제거함으로써 데이터의 질적 향상을 도모하고, 모델이 소수 클래스의 패턴을 학습하는 데 긍정적인 영향을 미쳤음을 보여준다. 비록 AUROC은 불안정해졌으나, 정밀도와 재현율의 균형이 개선되기 시작했다.

최종적으로 비대칭 임계값 기반 특성 선택 알고리즘을 추가 적용했을 때, 모든 성능 지표에서 비약적인 향상이 확인되었다. 특히 F1-score는 0.7011로 베이스라인 대비 4배 이상 증가하였으며, Recall 또한 0.7276으로 크게 개선되어 조기 탐지라는 임상적 목적을 달성하였다. AUROC 또한 0.7457로 회복되어, 일반화 성능과 탐지 능력의 균형을 맞추었다. 이는 희소한 실험군의 특성은 최대한 보존하고, 풍부한 대조군은 엄격하게 선별하는 비대칭 전략이 클래스 불균형 문제 해결과 정보 보존에 결정적인 기여를 했음을 입증한다. 즉, 데이터 품질 확보와 불균형 해소를 통해 모델의 실질적인 탐지 성능을 극대화할 수 있었음을 확인할 수 있다.

3.3 Model Performance Evaluation

5-fold 교차 검증을 통해 평가된 모델별 성능은 Table 7와 같다.

Table 7: Model Performance Comparison (Mean \pm SD)

Model	AUROC	AUPRC	Accuracy	Precision	Recall	F1-score
Baseline (LR)	0.4989	0.5486	0.5246	0.5417	0.4194	0.4728
XGBoost	0.7457 (± 0.0238)	0.6623 (± 0.0563)	0.7678 (± 0.0288)	0.6764 (± 0.0338)	0.7276 (± 0.1147)	0.7011 (± 0.0516)
LightGBM	0.7103 (± 0.0310)	0.6411 (± 0.0653)	0.7344 (± 0.0927)	0.6512 (± 0.0516)	0.6844 (± 0.1144)	0.6674 (± 0.0405)
Soft Voting	0.7324 (± 0.0324)	0.6607 (± 0.0511)	0.7167 (± 0.0621)	0.6410 (± 0.0625)	0.6934 (± 0.0682)	0.6662 (± 0.0465)
Stacking	0.7324 (± 0.0324)	0.6607 (± 0.0511)	0.7167 (± 0.0621)	0.6410 (± 0.0625)	0.6934 (± 0.0682)	0.6662 (± 0.0465)

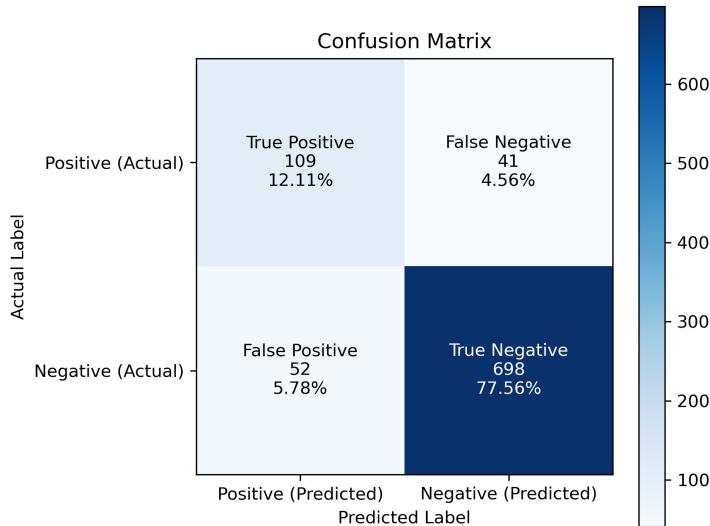


Figure 2: Confusion Matrix (XGBoost). TP: 109, FN: 41, FP: 52, TN: 698.

3.4 Statistical Analysis and Clinical Utility

3.4.1 Statistical Significance Test

XGBoost와 LightGBM 모델 간의 성능 차이에 대한 통계적 유의성을 검증하기 위해, 5-fold 교차 검증을 통해 산출된 F1-score에 대하여 Paired t-test를 수행하였다. 검정 결과 p-value는 0.1514 ($p > 0.05$)로 산출되어, 5% 유의 수준에서 두 모델 간의 성능 차이는 통계적으로 유의하지 않은 것으로 나타났다. 이는 XGBoost가 수치적으로 더 높은 평균 성능을 보였으나, 폴드 간의 변동성을 고려할 때 두 모델이 통계적으로 유사한 예측 성능을 가짐을 시사한다.

3.4.2 Calibration Analysis

모델의 예측 확률 신뢰도를 평가하기 위해 Calibration Curve를 분석하였다(Figure 3).

분석 결과, LightGBM은 대각선보다 현저히 낮은 위치에 분포하여, 실제 양성 발생 빈도에 비해 예측 확률을 과소평가하는 경향이 뚜렷하게 관찰되었다. 반면, XGBoost는 전

구간에서 대각선에 상대적으로 더 근접한 패턴을 보이며 더 나은 보정 성능을 입증하였다. 이를 정량적으로 뒷받침하는 Brier Score 또한 XGBoost가 0.142로 LightGBM의 0.213 보다 낮게 산출되어, XGBoost가 예측한 확률값이 실제 임상적 위험도를 더 신뢰성 있게 반영하고 있음을 확인하였다.

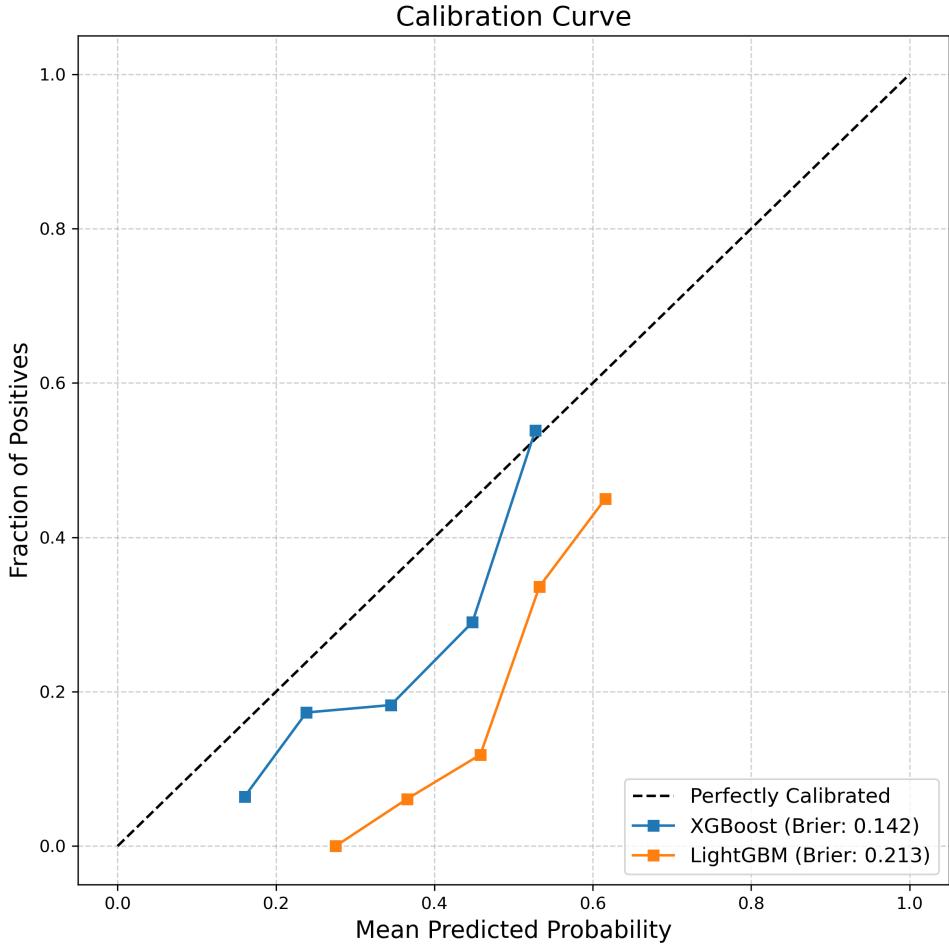


Figure 3: Calibration curves comparing XGBoost and LightGBM. The XGBoost model (Brier: 0.142) demonstrates better calibration compared to LightGBM (Brier: 0.213), which tends to underestimate the risk probabilities.

3.4.3 Decision Curve Analysis

전통적인 성능 지표를 넘어 모델의 임상적 실효성을 평가하기 위해 Decision Curve Analysis(DCA)를 수행하였다(Figure 4).

Figure 4에서 XGBoost 모델은 약 0.0에서 0.2 사이의 낮은 임계값 구간에서 Treat All 및 Treat None 전략 대비 높은 Net Benefit을 보였다. 이는 해당 모델이 긴장성 기흉을 놓치지 않아야 하는 조기 선별 도구로서 임상적 가치가 있음을 시사한다.

그러나 임계값이 0.25를 초과하는 구간에서는 Net Benefit이 0 미만으로 감소하거나 불안정한 양상을 보였다. 이는 모델이 높은 확신을 요구하는 확진 단계보다는, 낮은 확률 임계값을 적용하여 위험군을 1차적으로 선별하는 조기 경보 시스템으로 활용될 때 최대의 효용을 가짐을 의미한다. 반면, LightGBM은 대부분의 구간에서 Treat None 대비 유의미한 이득을 보이지 못하였다.

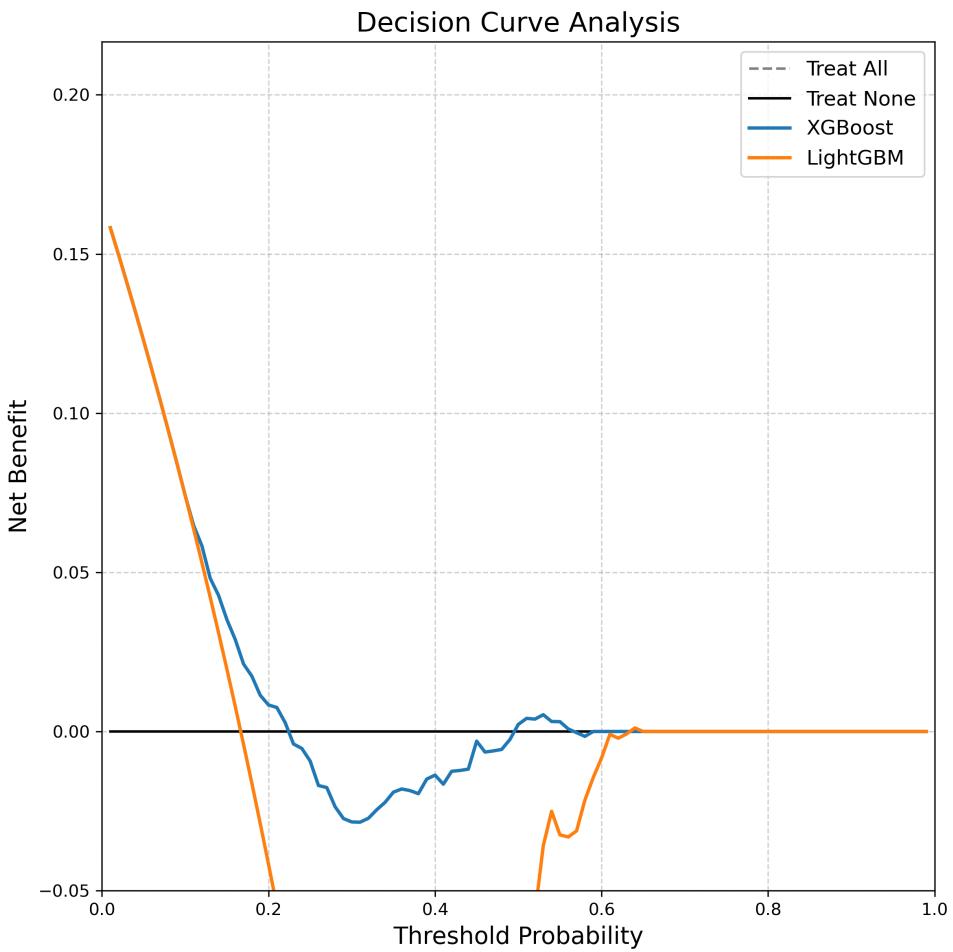


Figure 4: Decision Curve Analysis (DCA). The y-axis measures the net benefit. The XGBoost model shows a positive net benefit in the low threshold probability range (0.0 – 0.2), suggesting its utility as a screening tool. In contrast, LightGBM fails to provide clinical utility across relevant thresholds.

3.5 Key Predictive Features

Permutation Importance 기반으로 선정된 상위 5개 특성은 Table 8와 같다.

Table 8: Key Predictive Features (XGBoost)

Feature Name	Importance	Clinical Significance
w120_RR_kurt	0.0417	Kurtosis of respiratory rate over the past 2 hours. Captures abrupt changes or irregularities in breathing patterns preceding tension pneumothorax
all_SHOCK_INDEX_kurt	0.0300	Kurtosis of shock index over the entire observation period. Cardiovascular stress from pneumothorax manifests as rapid fluctuations in shock index
all_RR_kurt	0.0291	Kurtosis of respiratory rate over the entire observation period. Instability of breathing patterns throughout the 3-hour window
w120_SpO2_mad	0.0265	Median absolute deviation of SpO ₂ over the past 2 hours. Increased variability in oxygen saturation during ventilation impairment caused by pneumothorax
w120_DBP_kurt	0.0245	Kurtosis of diastolic blood pressure over the past 2 hours. Tension pneumothorax reduces venous return to the heart, causing blood pressure instability

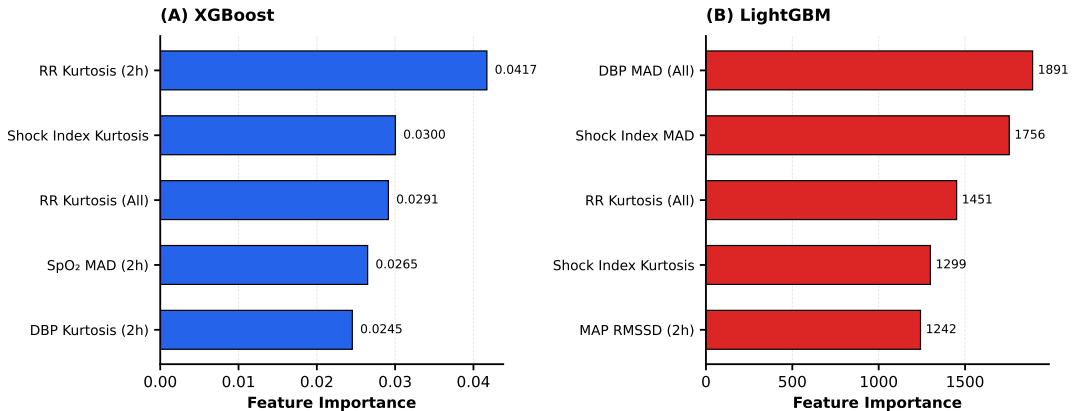


Figure 5: Top 5 features ranked by permutation importance in (A) XGBoost and (B) LightGBM. RR: respiratory rate; MAD: median absolute deviation; DBP: diastolic blood pressure; MAP: mean arterial pressure; RMSSD: root mean square of successive differences.

3.6 SHAP Analysis for Model Interpretability and Local Interpretability

3.6.1 Global Interpretability and Feature Impact

Figure 6은 XGBoost 모델의 SHAP summary plot으로 분석 결과, 전체 쇼크 인덱스 평균절대편차(all_SHOCK_INDEX_mad)와 최근 2시간 호흡수 첨도(w120_RR_kurt)는 붉은 점들이 주로 양의 SHAP 값에 분포하여, 해당 생체 신호의 급격한 변동성이 긴장성 기흉의 위험도를 높이는 주요 인자임을 시각적으로 입증하였다. 이는 앞서 Figure 5의 결과와 일치하며, 기흉 발생 시 호흡 및 혈역학적 불안정성이 급증한다는 임상적 지식과 부합한다.

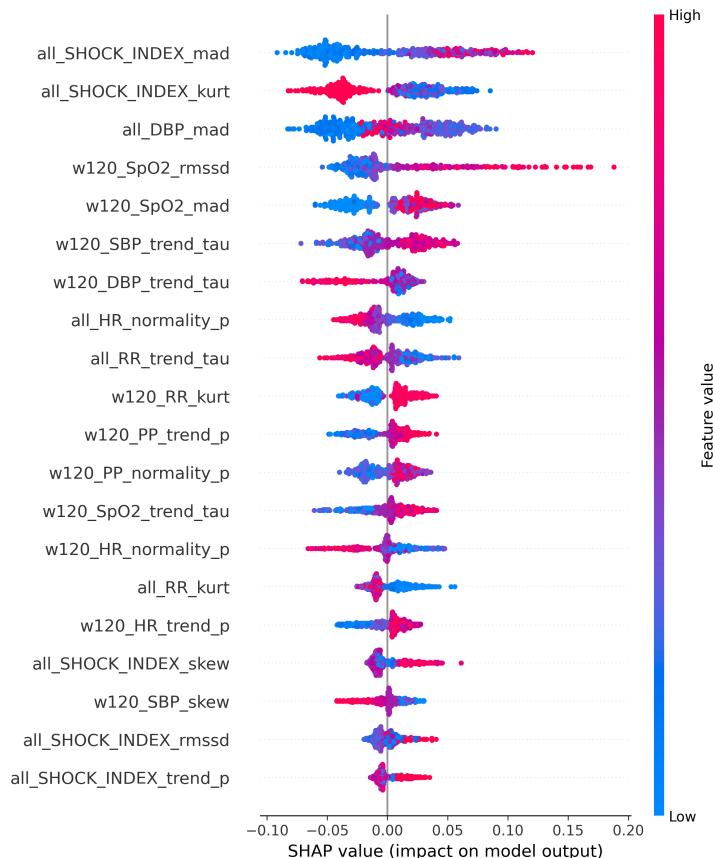


Figure 6: SHAP beeswarm plot summarizing the impact of features on model output. Red dots represent high feature values, blue dots represent low values. Positive SHAP values indicate a higher predicted risk of tension pneumothorax.

3.6.2 Feature Dependence and Non-linearity Analysis

Figure 7는 주요 상위 변수들에 대한 SHAP dependence plot을 도시화한 것으로, 이를 통해 모델이 학습한 변수와 예측값 간의 비선형적 관계 및 변수 간 상호작용 효과를 규명하였다.

본 분석에서 관찰된 주된 특징은 혈역학적 불안정성을 대변하는 변수들이 모델의 의사 결정에 지배적인 영향을 미친다는 점이다.

첫째, 좌상단의 쇼크 인덱스의 절대 표준 편차(all_SHOCK_INDEX_mad) 그래프에서는 변동성이 0에 근접한 기저 상태를 벗어나는 즉시 SHAP 값이 급격히 상승하는 양상이 관찰

된다. 이는 모델이 쇼크 인덱스의 미세한 동요조차 긴장성 기흉의 조기 징후로 식별하며, 해당 변수에 대해 매우 높은 민감도를 가지고 있음을 시사한다.

둘째, 좌하단의 $w120_SpO2_rmssd$ 는 산소포화도의 변동성이 증가함에 따라 위험도가 선형적으로 상승하는 뚜렷한 양의 상관관계를 보인다. 이는 호흡 부전 진행에 수반되는 생체 신호의 불규칙성을 모델이 주요 위험 인자로 학습하였음을 뒷받침한다.

셋째, 우측 하단의 혈압추세($w120_SBP_trend_tau$)는 명확한 시그모이드 형태의 비선형 패턴을 나타낸다. 이는 모델이 혈압 추세의 변화를 단순 선형 결합이 아닌, 특정 임계값을 기준으로 위험 구간을 구분하는 비선형적 메커니즘을 통해 탐지하고 있음을 의미한다.

아울러, 각 그래프의 색상은 상호작용 변수의 분포를 나타낸다. 특히 $all_SHOCK_INDEX_mad$ 그래프에서 관찰되는 수직적 분산은 호흡수 첨도($w120_RR_kurt$) 값에 의해 매개된다. 이는 모델의 최종 위험도 예측이 단일 변수의 변화뿐만 아니라, 호흡수 패턴 등 타 생체 신호와의 복합적인 상호작용에 기인함을 입증한다.

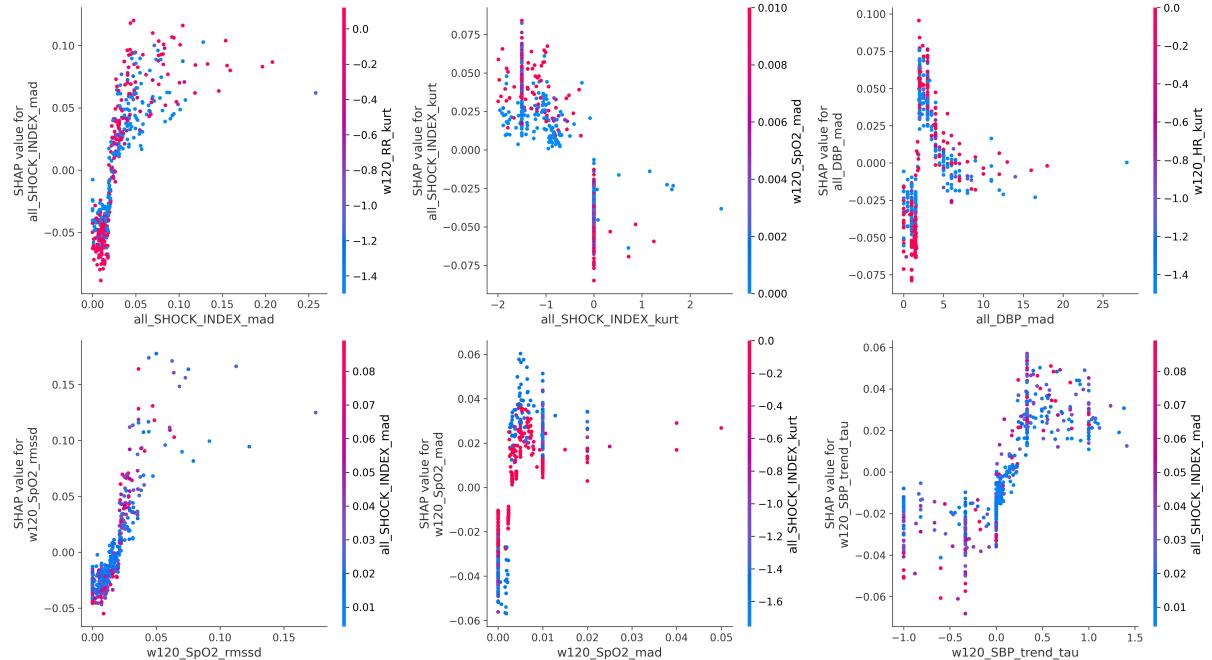


Figure 7: SHAP dependence plots for the top 6 features. These plots reveal non-linear relationships between feature values (x-axis) and their impact on the model's prediction (y-axis).

3.6.3 Local Interpretability analysis using SHAP Force Plots

Figure 8는 XGBoost 모델의 개별 예측 사례에 대한 SHAP force plot이다.

Case 1: True Positive Figure 8의 상단 그래프는 실제 긴장성 기흉을 정확하게 탐지한 사례(Prob: 0.94)이다. 여기서 모델의 판단을 주도한 핵심 변수는 $w120_SpO2_mad$ 와 all_DBP_mad 였다. 이는 산소포화도와 이완기 혈압의 급격한 변동성이 기흉 발생 시 나타나는 저산소증 및 혈역학적 불안정성을 잘 반영하고 있음을 시사한다. 또한 $all_SHOCK_INDEX_kurt$ 가 양의 방향으로 작용하여, 쇼크 인덱스의 분포적 특이성 또한 중요한 탐지 근거로 활용되었다.

Case 2: False Positive 중단 그래프는 실제로는 위험 상황이 아니라, 모델이 고위험군으로 오분류(Prob: 0.79)한 사례이다. a11_SHOCK_INDEX_kurt와 w120_RR_kurt가 붉은색 막대의 큰 비중을 차지하고 있다. 이는 호흡수와 쇼크 인덱스의 일시적인 이상 패턴이나 노이즈를 모델이 심각한 징후로 과대평가했기 때문이다. 특히 w120_SBP_trend_tau와 같은 트렌드 변수까지 결합되어, 모델이 단순한 생체 신호의 흔들림을 병리적인 패턴으로 오인했음을 확인할 수 있다.

Case 3: False Negative 하단 그래프는 실제 긴장성 기흉을 놓친 사례(Prob: 0.46)이다. 예측 확률이 0.46으로 임계값(0.5)에 매우 근접하였으나 최종적으로 음성으로 분류되었다. 주목할 점은 심박수 샘플 엔트로피(w120_HR_sampen)는 붉은색으로 작용하여 위험 신호를 감지했으나, a11_SHOCK_INDEX_mad와 a11_SHOCK_INDEX_kurt가 음의 기여으로 강하게 작용하여 이를 상쇄했다는 것이다. 즉, 심박수의 복잡도는 증가했음에도 불구하고, 쇼크 인덱스의 변동성이 낮게 유지되어 모델이 이를 혈역학적으로 안정된 상태로 오판하는 결과를 초래했다. 이는 단일 생체 신호 지표가 안정적일지라도, 엔트로피와 같은 복합적인 신호 패턴을 더 면밀히 고려해야 함을 시사한다.

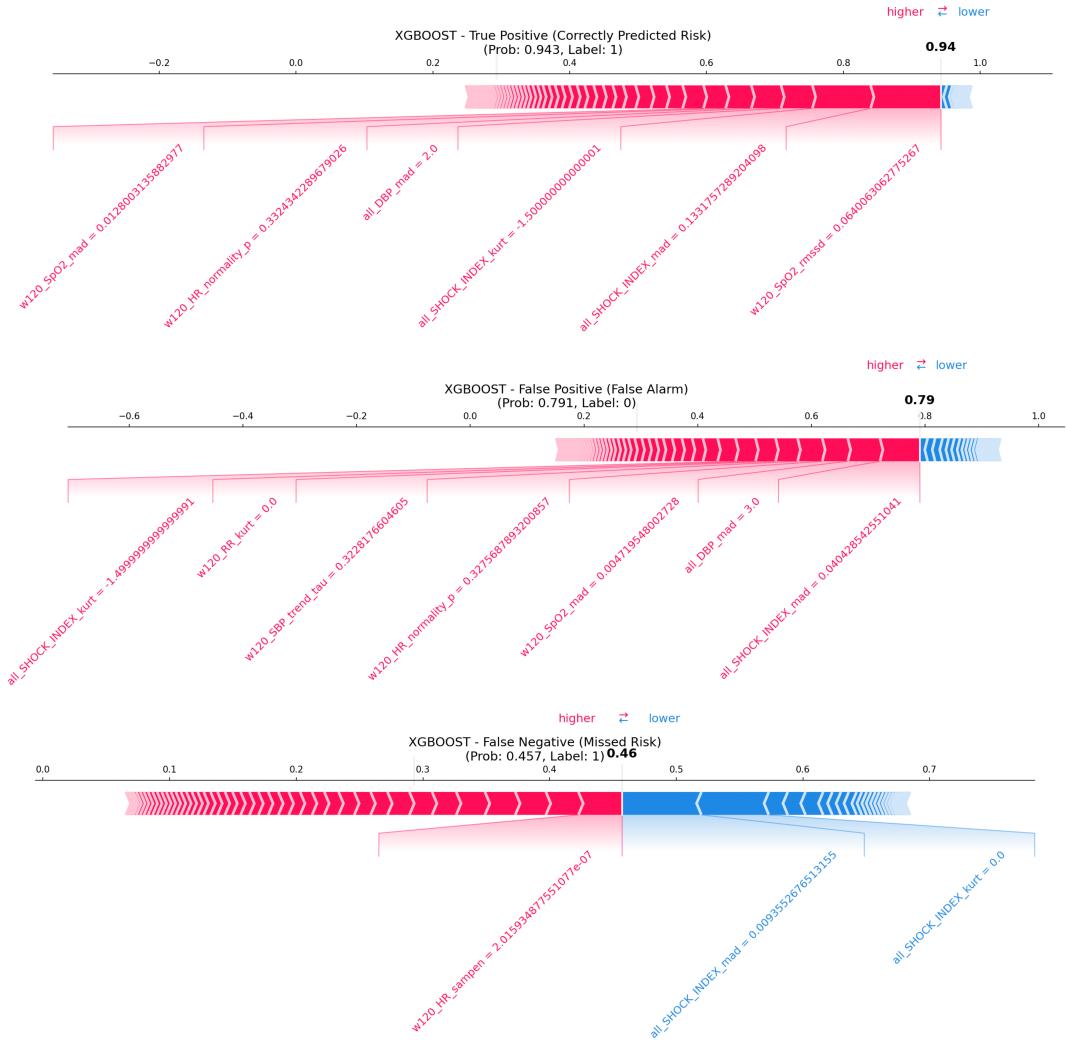


Figure 8: SHAP force plots for individual predictions. (Top) True Positive case where the model correctly identified high risk. (Bottom) False Positive case illustrating potential false alarms. Red bars indicate features pushing the prediction towards positive (risk), while blue bars push towards negative (safe).

3.7 Error Analysis

모델의 예측 신뢰성을 제고하고 오분류 원인을 규명하기 위해, XGBoost 모델의 False Negative(FN) 및 False Positive(FP) 사례에 대한 정성적 분석을 수행하였다.

3.7.1 False Negative Analysis

실제 긴장성 기흉 환자를 놓친(FN) 사례들의 주요 특징은 데이터의 '비정상적 평탄화'였다. Figure 9에서 볼 수 있듯이, 주요 생체 신호의 첨도가 -1.5로 나타나는 경우가 빈번하였다. 이는 데이터 분포가 완전히 평탄하거나, 결측치가 단일 값으로 대체되어 변동성이 소실된 상태를 의미한다. 또한 데이터 포인트의 개수가 현저히 적은 패턴도 확인하였다. 즉, 센서 접촉 불량이나 데이터 수집 오류로 인해 신호가 소실되거나, 전처리 과정에서 과도하게 평활화된 경우 모델이 이를 안정적인 정상 상태로 오인하여 위험 징후를 포착하지 못한 것으로 해석된다.

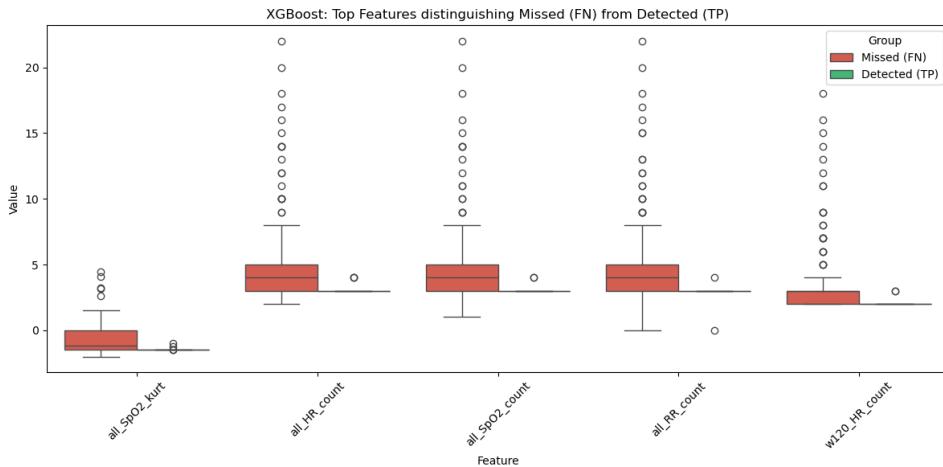


Figure 9: Distribution of top distinguishing features for False Negative cases in XGBoost. The presence of features with kurtosis values of -1.5 suggests that signal loss or artificial flatness due to imputation contributed to missed detections.

3.7.2 False Positive Analysis

반면, 거짓 양성 사례에서는 신호의 복잡도가 0.0이거나 첨도가 -1.5인 극단적인 패턴이 주요 원인으로 식별되었다(Figure 10). 일반적으로 생체 신호는 일정 수준의 불규칙성을 가지나, 장기간 신호가 결측되어 MICE로도 대체가 어려웠거나, 직선 형태의 아티팩트가 포함된 경우 엔트로피가 0으로 수렴한다. XGBoost 모델은 이러한 정적을 기흉 발생 임박 시의 전조 증상이나 심각한 이상 신호로 과잉 해석하여 거짓 경보를 발생시킨 것으로 판단된다. 이는 데이터 품질 저하가 모델의 성능을 저해하는 양면적인 원인임을 시사한다.

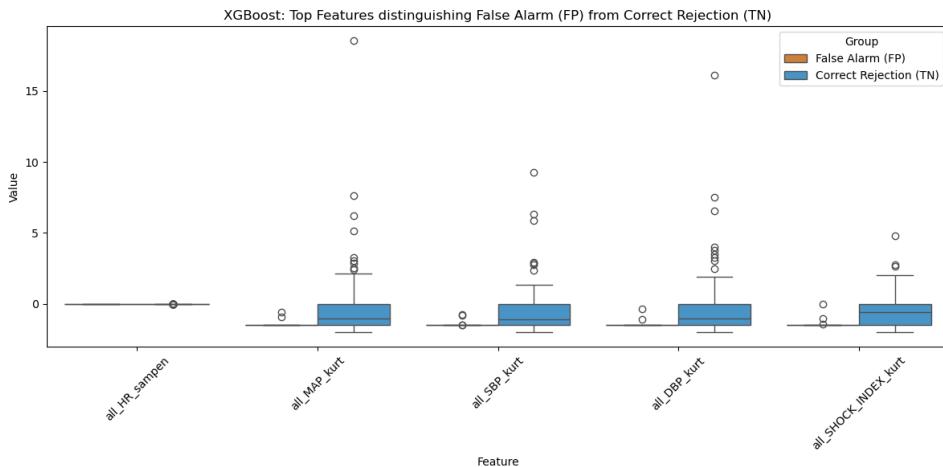


Figure 10: Distribution of top distinguishing features for False Positive cases in XGBoost. Features with zero sample entropy or extreme kurtosis indicate that the model may misinterpret signal artifacts or flatlines as pathological anomalies.

4 Discussion

4.1 Interpretation

본 프로젝트에서 개발한 XGBoost 기반 긴장성 기흉 조기 탐지 모델은 AUROC 0.7457, Recall 0.7276을 달성하였다. 특히 Recall 72.76%는 실제 긴장성 기흉 환자의 약 73%를 조기에 탐지할 수 있음을 의미하며, 이는 생명을 위협하는 응급 상황에서 거짓 음성을 최소화하는 것이 중요한 임상적 목표를 고려할 때 유의미한 수준이라 판단된다. AUROC 0.7457은 모델이 무작위 분류기(0.5)보다 현저히 우수하며, 이와 같은 불균형 데이터셋을 고려하면 특히 더 좋은 결과라고 판단된다.

제안 모델은 생체 신호의 정교한 해석을 위해 기반 모델 대비 고도화된 특성 추출 및 학습 전략을 채택하였다. 기존 모델이 시계열 데이터를 단순 통계치로 요약한 반면, 본 모델은 256개의 고차원 특성을 기반으로 고차 통계량 및 신호 복잡도를 분석하여 생체 신호 고유의 동역학적 패턴을 구체화하였다. 특히 30분에서 120분에 이르는 다중 시간 윈도우를 활용함으로써 단일 윈도우의 한계를 넘어 급성 변화와 장기 추세를 아우르는 다각적 분석을 수행하였다. 데이터 품질 및 학습 측면에서도, 단순 대체나 SMOTE에 의존하기보다 생체 신호 간 다양성 상관관계를 고려한 결측치 보정과 Cost-sensitive Learning 및 Focal Loss를 적용하여, 생리학적 타당성과 실제 임상 데이터 분포를 유지하면서 클래스 불균형 문제를 강건하게 해결하였다.

주요 예측 특성 분석 결과는 긴장성 기흉의 병태생리학적 특징과 일치하는 패턴을 보였다. 첫째, 상위 5개 특성 중 4개가 kurtosis로 나타났다. 이는 긴장성 기흉은 흉막강 내 압력이 급격히 증가하여 폐 허탈과 종격동 편위를 유발하는 급성 병리 현상이므로, 환자의 생체 신호가 정상 범위를 벗어나 급격히 변동하는 경향을 보인다. 따라서 kurtosis가 높은 값을 가지는 특성들이 주요 예측 인자로 선정된 것은 이 질환의 급성 발현 특성을 반영한 것으로 해석된다.

또한, 상위 5개 특성 중 3개가 120분(2시간) 시간 윈도우에서 추출되었다. 이는 30분 단기 윈도우나 180분(3시간) 전체 윈도우보다 중간 범위의 시간 척도가 더 높은 예측력을 보임을 의미한다. 30분 윈도우는 측정 오차나 일시적 변동에 취약할 수 있으며, 180분 윈도우는 초기의 안정적인 상태를 포함하여 이벤트 직전의 변화를 희석시킬 수 있다. 반면 120분 윈도우는 병리학적 변화가 본격적으로 진행되는 시기를 포착하면서도 충분한 데이터 포인트를 확보하여 통계적 안정성을 유지하는 최적 구간으로 판단된다.

그리고, 호흡수 관련 특성이 나머지를 차지하였다. 긴장성 기흉의 핵심 병태생리는 환기-관류 불균형과 폐 허탈로 인한 호흡 부전이며, 이는 심박수나 혈압의 변화보다 선행하는 것으로 알려져 있다. 환자는 저산소증에 대한 보상 기전으로 호흡수를 증가시키며, 흉막강 내 압력 변동으로 인해 호흡 패턴의 불규칙성이 나타난다. 따라서 호흡수의 첨도가 주요 예측 인자로 식별된 것은 긴장성 기흉의 임상적 발현 순서와 일치하며, 호흡기계 모니터링이 조기 경보 시스템에서 핵심적 역할을 할 수 있음을 시사한다.

4.2 Methodological Strengths

4.2.1 Weight-based Sample Removal Algorithm

의료 데이터에서는 특성 간 상호의존성이 높아 단순 특성 제거 시 임상적으로 중요한 정보가 손실될 수 있다. 탐욕적 샘플 제거 알고리즘 기반한 제안한 특성 추출 방식은 listwise deletion이나 완전 특성 제거와 달리, 동맥혈 가스 분석과 같이 임상적으로 핵심적인 변수들을 보존하면서도 결측률이 높은 샘플만 선택적으로 제거한다.

이를 통해 특성 기반 접근법들이 가진 정보 손실 문제를 해결하고, 의료진의 임상적 판단에 필수적인 생체 지표들을 온전히 유지할 수 있었다. 알고리즘은 각 샘플의 결측

패턴을 분석하여 전체 데이터셋의 통계적 대표성을 훼손하지 않는 범위에서 최적의 샘플을 제거하도록 설계되었다.

4.2.2 Asymmetric Threshold-based Feature Selection

중환자실 데이터는 본질적으로 클래스 불균형을 내포하며, 희소한 실험군과 풍부한 대조군 간 데이터 특성이 상이하다. 단일 임계값 기반 방법은 이러한 비대칭성을 고려하지 못해 소수 클래스의 중요 특성을 과도하게 제거할 위험이 있다.

비대칭 임계값 전략은 각 클래스별로 독립적인 결측률 기준을 설정하여, 희귀 질환군에서는 더 관대한 임계값을 적용하고 대조군에서는 엄격한 기준을 적용한다. 이를 통해 symmetric feature selection이나 class-agnostic filtering 등에서 발생하는 편향을 최대한 방지하고자 하였으며, 결과적으로 클래스별 임상적 특성을 균형있게 보존할 수 있었다.

4.2.3 MICE

의료 생체 신호는 생리학적 메커니즘에 의해 강한 다변량 상관관계를 보이며, 단순 대체 방법으로는 이러한 관계를 재현할 수 없다. Mean/median imputation이나 forward fill 같은 단변량 방법은 변수 간 의존성을 무시하여 비현실적인 값을 생성할 수 있다. 이 외는 다르게 MICE는 조건부 확률분포를 학습하여 혈압-심박수, 산소포화도-호흡수 등 의 생리학적 연관성을 반영한 결측치 보정을 수행한다. KNN imputation이나 matrix factorization과 달리, MICE는 각 변수의 예측 모델을 반복적으로 업데이트하여 복잡한 비선형 관계까지 포착하고 불확실성을 고려한 다중 데이터셋을 생성한다.

4.2.4 Cost-sensitive Learning and Focal Loss

의료 진단에서 실제 환자 분포는 불균형하며, 단순 oversampling은 소수 클래스의 합성 샘플을 생성하여 실제 임상 분포를 왜곡시킴을 확인하였다. 또한, Undersampling은 다수 클래스의 중요한 패턴을 손실시켜 일반화 성능을 저하시킨다.

따라서 Cost-sensitive Learning를 사용하여 오분류 비용을 차등 적용하여 실제 데이터 분포를 유지하면서도 모델이 희귀 클래스에 더 집중하도록 유도하였다. Focal Loss는 쉬운 샘플의 기여도를 감소시켜 분류가 어려운 경계 사례에 학습을 집중시키며, 두 기법의 결합은 class weight balancing이나 threshold moving 같은 후처리 방법보다 근본적이고 안정적인 해결책을 제공한다.

4.2.5 GroupKFold Cross-validation

의료 데이터는 동일 환자에서 시계열로 수집된 다중 관측치를 포함하며, 기존의 교차 검증은 같은 환자의 데이터가 훈련-검증 세트에 분산되어 데이터 누출을 발생시킨다. 이는 모델이 환자별 특이 패턴을 학습하여 과적합되고, 실제 성능을 과대평가하게 만든다. 의료 데이터는 이러한 과적합으로 인한 성능 과대 포장이 절대로 일어나서는 안되는 도메인으로, 최대한 강건한 모델을 만들기 위해 노력하였다.

4.3 Clinical Implications

본 프로젝트의 결과는 실시간 생체 신호 모니터링을 통한 긴장성 기흉의 조기 경보 시스템 구축 가능성을 제시한다. 특히 중환자실 환경에서 의식이 저하되거나 진정된 환자의 경우, 주관적 증상 호소가 어려워 임상적 악화의 조기 발견이 어렵다. 본 모델을 중환자실 모니터

링 시스템에 통합할 경우, 의료진이 긴장성 기흉의 발생 가능성을 사전에 인지하고 신속한 진단 및 치료를 시작할 수 있어 환자 예후 개선에 기여할 수 있을 것으로 기대된다.

본 프로젝트에서 제안한 모델은 의료 데이터의 클래스 불균형, 환자 간 이질성, 시계열적 변동성 등 실제 중환자실 환경의 제약조건을 고려하여 다층적 규제 전략을 체계적으로 도입함으로써 임상 환경에서의 강건성을 극대화하도록 설계되었다. 손실 함수 수준에서는 Cost-sensitive Learning과 Focal Loss를 결합하여 클래스 불균형에 대한 알고리즘적 규제를 적용하였으며, 학습 과정에서는 early stopping으로 과적합을 방지하였다. 또한 XGBoost 모델의 내재적 규제 파라미터인 L1/L2 정규화, 트리 깊이 제한, 그리고 낮은 학습률을 Optuna 최적화를 통해 조정함으로써 모델 복잡도를 제어하였다. GroupKFold 교차 검증은 환자 단위 데이터 누출을 염격히 방지하여 일반화 성능 평가의 신뢰성을 확보하는 구조적 규제로 작용하였다. 이러한 다단계 규제 전략을 통해 실제 중환자실에서 관찰되는 불안정한 데이터 분포 하에서도 안정적인 예측 성능을 유지하도록 설계하였다.

또한, 주요 예측 특성으로 식별된 호흡수의 첨도, 쇼크 지수의 첨도, 산소포화도의 변동성 등은 임상적으로 관찰 가능한 지표들이므로, 모델의 예측 결과에 대한 해석 가능성을 높여 의료진의 신뢰도를 확보할 수 있다. 이는 단순히 예측 정확도가 높은 모델을 제시하는 데 그치지 않고, 실제 임상 의사결정 과정에서 활용 가능한 의사결정 보조 도구로서의 잠재력을 강화한다. Permutation Importance 기반의 특성 선택 역시 모델의 복잡도를 256개에서 76개 특성으로 감소시켜 과적합을 방지하고 해석 가능성을 향상시키는 특성 수준의 규제로 작용하였다.

4.4 Limitations

본 프로젝트는 몇 가지 제한점을 가진다. 첫째, 흉관 삽입이라는 침습적 치료를 받은 환자를 긴장성 기흉의 대리 지표로 사용하였으나, 생리학적 정의에 따른 완벽한 긴장성 기흉(호기-흡기 전 구간의 지속적 양압)을 직접 검증하지는 않았다. 따라서 본 모델은 엄밀히 말해 임상적 개입이 필요한 기흉을 예측하는 것으로 해석되어야 한다. 둘째, 단일 기관의 데이터만을, 즉, 외부 검증 없이 교수님께서 제공해주신 MIMIC-IV 데이터셋만을 사용하여 외부 타당도 검증이 이루어지지 않았으므로, 다른 의료 기관에서의 일반화 가능성은 추가 검증이 필요하다. 셋째, 후향적 분석 설계로 인해 실제 임상 환경에서의 실시간 성능 평가가 이루어지지 않았다. 넷째, SILVER 코호트를 최종 학습에서 배제함으로써 경증 기흉에 대한 모델의 민감도가 제한될 수 있으며, 이는 조기 탐지 시스템의 적용 범위를 제한할 가능성이 있다.

5 Conclusion

본 프로젝트는 MIMIC-IV 데이터베이스를 활용하여 생체 신호 및 동맥혈 가스 분석 데이터로부터 긴장성 기흉의 조기 탐지 가능성을 검증하였다. 도메인 지식 기반의 특성 공학, 강건한 전처리 파이프라인, 그리고 불균형 데이터 학습에 최적화된 XGBoost 모델을 통해 AUROC 0.7457, Recall 0.7276의 성능을 달성하였다. 특히 급격한 생체 신호 변화를 나타내는 첨도 특성과 최근 2시간 동안의 호흡수 및 산소포화도 변동성이 주요 예측 인자로 식별되었다.

본 프로젝트의 결과는 실시간 생체 신호 모니터링을 통한 긴장성 기흉의 조기 경보 시스템 구축 가능성을 제시하며, 중환자실 환경에서 생명을 위협하는 응급 상황에 대한 의료진의 신속한 대응을 지원할 수 있을 것으로 기대된다. 향후 다기관 외부 검증 및 전향적 임상 시험을 통해 실제 임상 환경에서의 유용성을 평가하고, 해석 가능성을 강화한 모델 개발이 필요하다.