의학석사 학위논문

MIMIC-IV로 학습한 중환자실 사망 예측 모델의 성능 외부 검증

2023년 8월

서울대학교 대학원

의학과 마취통증의학 전공

이 호 종

MIMIC-IV로 학습한 중환자실 사망 예측 모델의 성능 외부 검증

지도교수 이 형 철

이 논문을 의학석사 학위논문으로 제출함
2023년 4월

서울대학교 대학원
의학과 마취통증의학 전공
이 호 종

이호종의 석사 학위논문을 인준함
2023년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

초 록

연구배경: 의료에서 빅데이터(big data) 구축 및 인공지능 모델 개발을 넘어, 의료 현장에서 인공지능을 어떻게 활용할 것인지에 대한 논의가 활발하다. 특히 Medical Information Mart for Intensive Care (MIMIC), eICU Collaborative Research Database 등 공개 데이터베이스를 활용한 연구는 많으나 다른 나라 다른 의료기관 중환자실에서 좋은 외부 검증(external validation) 결과가 나온 연구는 거의 없다. 본 연구에서는, 외국의 공개 데이터베이스를 활용하여 실제 중환자실 환경에서 자동으로 수집 가능한 데이터를 이용하여 원내 사망을 실시간으로 예측하는 기계학습 모델을 개발하고, 이를 지리적 시간적으로 구분된 국내 의료기관에 적용하였을 때 유의미한 성능이 나오는지 검증해 보았다.

대상 및 방법: MIMIC-IV의 환자들 중 18세 이상이며 24시간 이상 체류한 중환자실 환자 23,152명을 개발용 코호트(development cohort)로 구축하였다. 이전 약 24시간 동안의 활력 징후 데이터(vital data)를 1시간 간격으로 추출하여, 이후 24시간 이내 사망 여부(mortality)를 예측하는 모델을 개발하였다. 일반적으로 사용되는 gradient boosting machine (GBM), 장단기 메모리(long short-term memory, LSTM), 트랜스포머(transformer)의 기계학습 방법을 검토하였으며, 사후 확률 보정(post-hoc probability calibration) 기법을 적용하였다. 평가 기준으로 area under the receiver operating

characteristic (AUROC) curve, area under the precision–recall curve (AUPRC), F1 점수, F2 점수, 확률 보정 곡선(calibration curve) 등을 사용하여, 내부 검증(internal validation)용 시험용 데이터 세트(testing dataset)에서 모델의 성능을 평가하였다. 개발용 코호트와 지리적, 시간적으로 구분된 국내 의료기관인 서울대학교병원에서 같은 조건의 환자 5,745명으로 검증용 코호트(validation cohort)를 구축한 후, 이를 대상으로 하여 외부 검증하였다.

결과: 내부 검증에서 GBM 모델이 가장 우수하여 AUROC 0.903, AUPRC 0.346 (기준값 0.021), F1 점수 0.383, F2 점수 0.378를 보였으며, 확률 보정을 통해 과다 추정(overestimation) 양상이 교정되었다. 외부 검증에서 AUROC는 0.933으로 잘 유지되었으나, AUPRC 0.181 (기준값 0.009), F1 점수 0.202, F2 점수 0.341로 감소하였고, 확률 보정 곡선상 과다 추정 양상이 교정되지 않는 등, 종합적인 성능이 감소하였다.

결론: 공개 데이터베이스를 활용하여 만든 기계학습 모델이 지리적 시간적으로 구분된 타 국가의 의료기관을 대상으로 한 외부 검증에서 성능이 감소하였다. 각 의료기관의 데이터에 특화된 모델 생성 혹은 기존 모델의 재학습을 포함하여 임상에서 인공지능 모델을 활용할 수 있는 방안에 대한 연구가 필요하다.

주요어 : 외부검증, 기계학습, 실시간, 사망예측, 중환자실
학 번 : 2020-23022

목 차

표 목차

그림 목차

제 1 장 서 론

전 세계적으로 의료 빅데이터(big data) 구축 및 활용에 대한 논의가 활발하다. 빅데이터로부터 패턴을 감지하여 예후 예측이나 치료 방향 설정 등 진료에 활용하는 데에 있어서 주로 인공지능이 활용되고 있으며, 관련 논문의 숫자도 폭발적으로 증가하고 있다 [1]. 그러나 각 국가별로 상이한 의료 시스템과 인종 차이, 의료보험 제도 차이 등으로 인해 이러한 외국 데이터를 기반으로 학습된 인공지능 모델들을 타 국가의 의료기관에도 적용 가능한지 검증할 필요가 있다. 만약 이러한 모델이 지리적 시간적으로 구분된 외부 검증(temporal and geographical external validation)에서도 좋은 성능을 보인다면 해당 모델의 범용성과 실용성을 입증한 것이 될 것이고, 성능이 좋지 않다면 각 의료기관의 데이터에 특화된 모델을 생성하여 사용해야 할 것이다. 아쉽게도, 사전 문헌 조사 상 외부 검증(external validation)이 이루어지지 않았거나 내부 검증(internal validation)과 성능 차이가 나는 경우가 많았고, 특히 다른 나라의 의료기관에서도 좋은 성능을 보인 것은 거의 없었다 [2].

환자의 원내 사망 여부(in-hospital mortality)는, 그 자체로 중요한 임상적 예후이면서 판정의 정확도에 있어 지리적 시간적 차이가 거의 없기 때문에, 사망 예측 모델은 국가 간 외부 검증에 적합하다. 또한 사망 예측은 의료 자원의 공정한 배분 문제도 중요하게 고려해야 하는 중환자실(intensive care unit, ICU)에서 특히 중요하다. 그러나 현재 임상에서 널리 사용 중인 중환자 사망 예측 지표들인 Acute Physiology

And Chronic Health Evaluation (APACHE) score [3], Simplified Acute Physiology Score (SAPS) [4], Mortality Probability Model (MPM) score [5] 는 사람의 입력을 필요로 하는 모델이기에 모든 중환자실 환자에서 반복적으로 실시간 적용하기는 어렵다. 만약 자동으로 수집 가능한 데이터로 이를 실시간 예측하는 기계학습 모델을 개발한다면 임상에서의 활용도가 높을 것으로 기대된다.

중환자실 사망 예측 모델 개발에 사용 가능한 외국 공개 데이터베이스로는 Medical Information Mart for Intensive Care (MIMIC) [6], eICU Collaborative Research Database (eICU-CRD) [7] 등이 있으며, 그 중 MIMIC은 미국의 3차 의료기관인 Beth Israel Deaconess Medical Center의 중환자실 환자 데이터로 만들어진 대표적인 공개 데이터베이스로, 2000년 3월 발표된 이후 꾸준한 업데이트가 이루어지고 있고, 인용 논문 수 역시 꾸준히 증가해오고 있다는 점에서 높은 데이터 신뢰성을 가지고 있다.

이에 본 연구에서는, MIMIC의 최신 버전인 MIMIC-IV를 사용하여 실시간 중환자실 사망 예측 모델을 개발 및 내부 검증한 후, 국내 3차 의료기관인 서울대학교병원의 중환자실 재원 환자에서의 사망 예측에 적용하여 서로 다른 국가 간의 외부 검증 시 모델 성능의 차이가 있는지 검증해보고자 하였다. 본 연구의 가설은 임상적으로 중요한 의미를 갖는 적은 수의 입력 변수로 개발된 인공지능 모델의 성능은 외부 검증에서도 잘 유지될 것이라는 것이었다.

제 2 장 연구 방법

제 1 절 데이터 출처 및 연구 윤리

본 후향적 연구는 서울대학교병원 의학연구윤리심의위원회 (승인 번호 2304-027-1420)로부터 승인을 받았으며, 후향적 연구라는 연구 디자인상의 특성으로 인해 환자의 서면동의는 면제되었다.

제 2 절 데이터 수집

모델 개발을 위하여, 해외의 공개 데이터베이스인 MIMIC-IV의 데이터를 후향적으로 분석하였다. 2008년부터 2019년까지 미국 Beth Israel Deaconess Medical Center의 내과계 중환자실(medical ICU, MICU) 및 외과계 중환자실(surgical ICU, SICU)에 24시간 이상 체류한 18세 이상의 환자 23,152명을 개발용 코호트(development cohort)로 정의하였다. 제외 기준은 1) 만 18세 미만 혹은 89세 초과, 2) 24시간 미만 체류, 3) 입원 중 유의미한 활력 징후 데이터(vital data)가 기록되지 않은 경우였다 (그림 1).

특정 증례의 과평가를 피하기 위해, 각 환자의 첫번째 중환자실 입원만을 다루었다. 여기서 첫번째 입원이란, 해당 환자의 제외 기준에 해당하지 않는 입원 중 첫번째를 의미하는 것으로, 가령 17세에 처음 중환자실에 입원하였던 환자라도 18세에 다시 입원하였다면 이때의 기록을 코호트에 포함시켰다. 입원 정보로는 사망 여부, 입원 횟수, 입실 시각, 퇴실 시각, 사망 시각을 수집하였다.

인구통계학적(demographics) 데이터로 나이, 성별, 키, 체중을 수집하였다. 나이의 경우 개인정보보호를 위한 비식별화로 인하여, 출생일을 기준으로 한 나이 계산이 불가능하여 입실 시각의 나이를 사용하였다.

활력 징후 데이터로 수축기 혈압(systolic blood pressure, SBP), 이완기 혈압(diastolic blood pressure, DBP), 평균 혈압(mean blood pressure, MBP), 심박수(heart rate, HR), 호흡수(respiratory rate, RR), 체온(body temperature, BT), 산소 포화도(saturation of peripheral oxygen, SpO_2), 흡기 산소 분율(fraction of inspired oxygen, FiO_2), 글래스고 혼수 척도(Glasgow coma scale, GCS)를 수집하였다. 수축기, 이완기 및 평균 혈압의 경우, 동맥압(arterial blood pressure, ABP)과 비침습적 혈압(non-invasive blood pressure, NIBP)을 모두 사용했으며, 평균 혈압이 없는 경우 $MBP = (SBP + DBP \times 2) \div 3$ 으로 계산하여 사용하였다. 체온의 경우 섭씨와 화씨를 모두 사용하되, 섭씨는 화씨로 변환하여 사용하였다. FiO_2의 경우 0~1 사이의 값은 100을 곱하기 했으며, 21 미만인 경우 산소 공급량(oxygen flow rate, L/min)을 적은 것으로 보고 배제하였다. GCS의 경우, 언어 반응(verbal response, GCS_v) 값을 사용할 수 없는 경우, 안구 반응(eye response, GCS_e)과 운동 반응(motor response, GCS_m) 값을 이용하여 $GCS_v = 2.3976 + GCS_m \times (-0.9253) + GCS_e \times (-0.9214) + GCS_m^2 \times 0.2208 + GCS_e^2 \times 0.2318$ 으로 계산하여 사용하였다 [8].

검사실 소견(laboratory findings)으로 각 장기 기능을 대표하는

지표인 알부민(albumin), 혈당(glucose), 프로트롬빈 시간(prothrombin time/international normalized ratio, PT/INR), 동맥혈 산소 분압(partial pressure of oxygen in arterial blood, PaO_2), 동맥혈 이산화탄소 분압(partial pressure of carbon dioxide in arterial blood, $PaCO_2$), 혈중 요소 질소(blood urea nitrogen, BUN), 크레아티닌(creatinine) 값을 수집하였다.

수집된 데이터가 다음과 같은 경우 없는 값으로 취급하였다: 1) 각 입력 변수의 값이 명백하게 생리학적 정상 범위를 벗어난 경우, 2) 사망 이후 기록된 경우.

최종 모델을 외부 검증하기 위하여, 개발용 코호트와 지리적 시간적으로 구분된 국내 3차 의료기관인 서울대학교병원의 중환자실 데이터를 사용하여 검증용 코호트(validation cohort)를 구축하였다. 서울대학교병원의 2006년도부터의 전자 의무 기록(electronic medical record, EMR) 데이터를 분석한 결과 다음 이유로 2017년도부터의 데이터를 사용하기로 하였다: 1) 데이터 정제(data cleansing) 완성도 문제, 2) 검사실 소견 데이터의 빈도 부족. 2017년 1월부터 2021년 2월 사이에 입실과 퇴실이 모두 이루어진 환자들을 대상으로 하였으며, 나머지 규칙들은 모두 개발용 코호트와 동일하였다. 그 결과 환자 5,745명이 검증용 코호트에 포함되었다 (그림 1). MIMIC-IV과 다르게 이미 검토된 데이터베이스가 아니기에 각각의 증례에 대해 전자 의무 기록과 병원 정보 시스템(hospital information system, HIS)을 일일이 대조하여 수작업으로 확인하였다. 여기에 고식적인 방법과의 비교를

위하여, 입실 시 APACHE II 점수를 추가로 수집하였다.

하위집단 분석(subgroup analysis)을 위해, 환자들을 나이, 성별, 중환자실 종류별로 구분하였다. 나이는 청년(young, < 45), 중년(middle age, 45~65), 노년(old, > 65)의 세 그룹으로 구분하였다 [9].

제 3 절 입력 변수 후보 및 결과 지표 선정

본 연구는 중환자실 입실 중 실시간 예측이라는 목표와, 적은 수의 입력 변수로 개발된 모델의 성능은 외부 검증에서도 잘 유지될 것이라는 가설 하에 진행되었다. 이에, 모든 의료기관에서 공통적으로 수합되는 인구통계학적 데이터와, 의료기관 간 차이가 적고 측정빈도가 높은 활력징후 데이터 만을 입력 변수 후보로 선정하였다.

검사실 소견은 기저 특성(baseline characteristics) 비교에만 사용하고 입력 변수 후보에서는 제외하였는데, 1) 일반적으로 검사실 소견 측정은 활력 징후 데이터 측정에 비해 주기가 일정치 않고 측정빈도가 훨씬 낮아 시간에 따른 변화를 더 민감하게 반영하지 못하여 실시간 예측에 부적합하고, 2) 검사실 소견에 대한 의료기관 별 치료 방침의 차이가 크며, 3) 환자의 상태가 악화되면 결국에는 활력 징후 데이터에 실시간으로 반영될 것이기 때문이다.

결과 지표로는, 1차 결과 지표(primary outcome)로 향후 24시간 이내의 원내 사망 여부를 평가하였다. 좀더 정확하게 사망 여부 및 사망 시각을 판정하기 위하여, 데이터베이스에 기록된 사망 여부 값 외에도, 입원 횟수, 입실 시각, 퇴실 시각, 사망 시각을 추가로 활용하였다: 1)

입원 횟수가 여러 번인 환자의 첫 입원은 사망 안함 처리, 2) 입실 시각과 퇴실 시각의 차이가 충분치 않은 경우 배제, 3) 퇴실 시각과 사망 시각이 다른 경우 좀더 빠른 쪽을 사망 시각으로 사용.

제 4 절 표본 추출

중환자실 입실 1시간 후부터 사망 혹은 퇴실 전까지 1시간 간격으로 기준 시각(reference time point)을 이동(sliding)하면서 표본(sample)을 생성하였다 (그림 2). 이는 환자 한 명당 첫번째 중환자실 체류시간에서 소수점 첫째자리 이하를 버림한 수만큼의 표본을 추출하였음을 의미한다. 중환자실 입실 후의 데이터를 포함시키기 위해 입실 1시간 후부터 추출하였다.

각 표본은, 활력 징후 데이터에서는 기준 시각 23시간 전부터 기준 시각 0시간 전까지 1시간 간격으로 총 24개 시점들(time points)에서의 입력값, 즉 24개의 시점값(time point value)을 가지며, 이들로부터 대푯값(representative value) 5개(처음값, 마지막값, 최소값, 최대값, 중앙값)를 추가로 계산하였다. 인구통계학적 데이터에서는 기준 시각 0시간 전 1개 시점에서의 시점값 1개 만을 가진다. 따라서 표본 하나당 9개 활력 징후 데이터에 대한 216(=9×24)개 시점값과 45(=9×5)개 대푯값, 4개 인구통계학적 데이터에 대한 4(=4×1)개 시점값을 가진다.

활력 징후 데이터의 시점값으로는 각 시점에서 최대 24시간 전까지의 범위에서 마지막 값을 취하였으며, 값이 없으면 결측값으로 남겨두었다. 인구통계학적 데이터의 시점값으로는 시간 제한 없이 각

시점까지 알려진 가장 마지막 값을 사용하였다.

각 표본은, 기준 시각 0시간 후부터 24시간 후 사이의 원내사망 여부를 결과 변수로 가진다. 따라서 생존하여 퇴실한 환자로부터는 음성 표본만 추출되지만, 사망한 환자로부터는 양성 표본과 음성 표본이 모두 추출된다. 가령, 입실 30시간 후 사망한 환자라면, 입실 1시간 후의 표본에서는 환자가 입실 1시간 후부터 25시간 후 사이에 생존 중이었으므로 결과값이 0(음성)이며, 입실 10시간 후의 표본에서는 환자가 입실 10시간 후부터 34시간 후 사이에 사망하였으므로 결과값이 1(양성)이다. 사망하지 않은 환자에 대한 표본은 재원 기간에 비례하여 계속 생기므로 매우 심한 불균형 데이터가 되겠으나, 실제 중환자실의 실시간 환경에 부합하므로 추출 과정에서는 별도의 표본 선별 작업은 하지 않고, 모델 개발 및 평가 시에 불균형 문제를 다루기 위한 방법들을 고려하였다.

제 5 절 입력 변수 선정

입력 변수를 선택하기 위한 사전 학습으로, 트리(tree) 기반 알고리즘(algorithm)인 gradient boosting machine (GBM) [10]을 기준으로 Boruta SHAP (Shapley additive explanation)을 통해 입력 변수들의 예측 기여도를 확인하였다 (그림 3). Boruta SHAP이란, 특정 변수가 이를 복원 추출(sampling with replacement)하여 만든 변수(shadow)보다 예측에의 기여도가 떨어지면 중요하지 않은 변수로 보고 배제하는 Boruta 알고리즘 [11] 에서, 기여도 계산에 게임

이론(game theory) [12] 에서 사용하는 섀플리값(Shapley value) [13]을 사용하는 방법으로, 트리 기반 알고리즘에서 입력 변수를 선별하기 위해 자주 사용된다. 9개 활력 징후 데이터에 대한 45개 대푯값과 4개 인구통계학적 데이터의 시점값을 합쳐 총 49개 값 중, DBP의 처음값과 MBP의 처음값의 2개 값만이 음의 기여도를 보였다. 다만 DBP, MBP도 처음값을 제외한 나머지 대푯값은 양의 기여도를 보였다.

이에, 이후 학습에서 GBM에 대해서는 DBP의 처음값과 MBP의 처음값을 제외한 43개 대푯값과 4개 인구통계학적 데이터의 시점값을 입력 변수로 사용하였으며, 시계열 데이터 학습을 하는 모델들에 대해서는 활력 징후 데이터와 인구통계학적 데이터의 220개 시점값 전체를 입력 변수로 사용하였다.

제 6 절 기계학습 모델 개발

개발용 코호트의 환자들을 80:20으로 무작위로 나눈 후 각자의 표본을 각각 학습용 데이터 세트(training dataset)와 시험용 데이터 세트(testing dataset)로 정의하였다. 따라서 같은 환자의 표본이 서로 다른 데이터 세트에 들어가는 일은 없었다. 이후의 모델 개발은 오로지 학습용 데이터 세트만을 사용하여 이루어졌으며, 개발된 모델에 대한 성능 평가는 시험용 데이터 세트를 대상으로 이루어졌다.

모델 종류 3가지, 표집(sampling) 방법 3가지, 학습 중 평가 지표(evaluation metrics) 2가지의 조합에 따라 총 18개 기계학습 모델을 개발하여 성능을 비교하였다.

기계학습 모델 종류로 GBM, 장단기 메모리(long short-term memory, LSTM) [14], 트랜스포머(transformer) [15]의 3가지를 고려하였다. GBM은 대표적인 트리 기반 앙상블 기계학습 기법으로, 원리상 입력 변수가 많을 때도 잘 동작하는 것으로 알려져 있어 의료 인공지능 분야에서 많이 쓰이고 있기에 고려하였으며, LSTM은 표본이 시계열 데이터라는 특성상 순환 신경망(recurrent neural network, RNN)의 대표적 모델로 고려하였고, 트랜스포머는 자기 주의(self-attention) 기법을 기반으로 복잡한 시계열 데이터에 대해 잘 동작한다고 알려져 있어 여러 분야에서 각광받고 있기에 고려하였다.

양성 표본의 비율이 매우 낮은 심한 불균형 데이터 문제(imbalanced data problem)이므로, 표집 방법으로 1) 과소표집(under-sampling), 2) 과대표집(over-sampling), 3) 원본 비율대로 사용 후 부류 가중치 적용(class weight)의 3가지를 고려하였다.

학습 중 평가 지표로는 개발된 모델 검증 시 사용할 지표 중 임계값(threshold value)의 영향을 받지 않는 area under the receiver operating characteristic (AUROC) curve, area under the precision-recall curve (AUPRC)의 2가지를 고려하였다.

모델 개발은 모델 종류, 표집 방법, 학습 중 평가 지표의 조합 각각에 대해 1) 초모수(hyperparameters) 최적화, 2) 확률 보정 기법 최적화, 3) 임계값 조정 의 순서로 이루어졌다 (그림 4).

먼저 GBM 모델은, 초모수로 number of trees = [1~600], max

depth = [3, 4, 5, 6], minimum child weight = [1, 2, 4], gamma = [0, 1, 10, 100], subsampling rate = [0.5, 0.8], column sampling rate = [0.5, 0.8]를 격자 탐색(grid search)하였으며, 5겹 교차 검증(5-fold cross validation)으로 얻은 평가 지표의 평균값을 모델의 성능으로 삼아 최적의 초모수를 선별하였다. 조기 종료(early stopping) 기법은 격자 탐색 중에는 사용하지 않았으나, 격자 탐색 전에 딥러닝(deep learning)에서의 에포크(epoch)에 해당하는 트리 수(number of trees)의 범위를 산정하는데에 활용하였으며, 모든 초모수 조합에서 해당 범위 내에 과적합(overfitting)이 발생함을 확인하였다.

이후 확률 보정 기법 중 1) Platt 비례법(Platt's scaling) [16], 2) 등장 회귀법(isotonic regression) [17], 3) 스플라인 기반 확률 보정법(spline-based probability calibration) [18] 의 3가지에 대하여 각각 5겹 교차 검증을 통해 확률 보정을 수행한 후, expected calibration error (ECE) [19], Brier 점수(Brier score) [20] 및 보정 곡선(calibration curve)을 그려 보정 성능을 확인하고, ECE가 가장 낮은 기법을 최종 선택하였다.

마지막으로 F1 점수를 최대화하는 임계값을 구하여 적용하였으며, 추후 논의를 위하여 재현율(recall)과 특이도(specificity)의 합을 최대화하는 Youden 지수(Youden index) [21] 나 F2 점수를 최대화하는 임계값도 미리 계산해두었다.

모든 교차 검증용 분할에서 층화 표집(stratified sampling) [22] 기법을 사용하였으며, 같은 환자의 표본이 서로 다른 분할에 나뉘어

들어가지 않도록 특별히 주의하였다.

같은 절차로 LSTM 및 트랜스포머 모델을 개발하였으나, GBM과 달리 모델 학습 전에 입력 변수별로 최소-최대 정규화(min-max normalization)하였고 [23], 초기값 편향(initial bias) 기법을 추가로 적용하였다. 초모수는 LSTM에서 epochs = [1~200], number of hidden layers = [1, 2], number of hidden nodes = [16, 32, 64], number of dense nodes = [16, 32, 64, 128], dropout rate = [0.2, 0.5]를 격차 탐색하였으며, 트랜스포머에서는 epochs = [1~50], number of filters = [32, 64], number of heads = [2, 3], embedded dimension = [32, 64], number of transformer layers = [1, 2, 3], dropout rate = [0.1, 0.2]를 격자 탐색하였다.

제 7 절 기계학습 모델 검증

내부 검증으로, 개발용 코호트 중 시험용 데이터 세트를 대상으로 앞서 개발된 18개 모델들의 성능을 평가하였다. 성능 평가를 위하여 확인한 지표는 AUROC, AUPRC, F1 점수, F2 점수, 정확도(accuracy), 균형 정확도(balanced accuracy), ECE이며, 모델 종류별로 F1 점수가 가장 높은 모델을 선정하여 GBM 대표 모델, LSTM 대표 모델, 트랜스포머 대표 모델로 명명한 후 이후의 분석에 사용하였고, 그 중에서도 F1 점수가 가장 높은 모델을 본 연구의 최종 모델로 삼았다.

외부 검증은 검증용 코호트를 대상으로 이루어졌다. 내부 검증을 할 때와 같은 지표들로 GBM 대표 모델, LSTM 대표 모델, 트랜스포머

대표 모델의 성능을 확인한 후, 최종 모델을 기준으로 내부 검증과 외부 검증의 성능을 비교 판단하였다. 하위집단 분석 역시 최종 모델을 기준으로 이루어졌다.

모델 개발 및 검증의 전 과정은 Python (version 3.8.10, Python Software Foundation, Wilmington, DE, USA) [24] 언어로 TensorFlow (version 2.10.1) [25], XGBoost (version 1.4.2) [26], scikit-learn (version 1.2.2) [27], BorutaShap (version 1.0.16) [28], ML Insights (version 1.0.2) [18] 라이브러리를 이용하여 자체 개발한 프로그램을 이용하여 수행되었다.

제 8 절 통계 분석

숫자형 변수(numerical variables)는 평균과 표준편차로 표기하였고, 범주형 변수(categorical variables)는 빈도/백분율로 표기하였다. 기저 특성 비교 시 환자 별 측정항목별 대푯값은 중앙값(median)으로 하였으며, 범주형 변수는 Chi-squared 방법으로 비교하였고, 연속 변수(continuous variables)는 Kruskal-Wallis 방법으로 비교하였다. 모든 비교에서 결측값은 제외하였다.

모델 성능 평가에는 불균형 데이터라는 점을 고려하여 AUROC 뿐만 아니라, AUPRC, F1 점수를 확인하였으며, 양성 표본의 의학적 중대성을 고려하여 F2 점수를 추가로 확인하였고, F1 및 F2 점수 계산에 사용한 임계값을 함께 표기하였다. AUROC는 값이 0.8 이상일 때 우수함(excellent)으로 간주하였는데, 이는 의료 분야에서 많이

쓰이는 기준이다 [29, 30]. 모델 간의 AUROC를 비교할 때는 DeLong's test [31] 를 사용하였고, 95% 신뢰구간을 함께 표기하였다.

각 모델의 예측값과 실제값의 일치 정도를 평가하기 위하여 보정 곡선을 그려서 확인하였고, ECE, Brier 점수를 계산하였다.

통계 분석을 위하여 Python (version 3.8.10, Python software Foundation)이 사용되었다. $P < 0.05$ 일 때 통계적으로 유의미하다고 간주하였다.

제 3 장 연구 결과

MIMIC-IV에는 총 53,150명의 중환자실 입원 환자가 있었으며, 그 중 기준에 맞는 23,152명의 첫 중환자실 입원 데이터로 모델을 개발하였다. 서울대학교병원에는 총 28,103명의 중환자실 입원 환자가 있었으며, 그 중 기준에 맞는 5,745명의 첫 중환자실 입원 데이터로 모델을 외부 검증하였다 (그림 1). 기저 특성을 표에 정리하였다 (표 1). 수집한 모든 값에서 개발용 코호트와 검증용 코호트 간에 유의미한 차이가 있었다. 개발용 코호트 내에서는, 학습용 데이터 세트와 시험용 데이터 세트 간에 성별(sex, $P = 0.036$)을 제외하면 유의미한 차이가 없었다 ($P \geq 0.05$).

입원 중 사망 비율은 개발용 코호트에서 23,152명 중 2,734명 (11.8%), 검증용 코호트에서 5,745명 중 508명 (8.8%)으로 개발용 코호트에서 유의미하게 높았다 ($P < 0.001$). 개발용 코호트와 검증용 코호트 모두에서 연령이 더 높은 하위집단 일수록 사망률이 높았는데, 개발용 코호트에서는 청년층 3,384명 중 177명 (5.2%), 중년층 8,864명 중 892명 (10.0%), 노년층 10,904명 중 1,665명 (15.3%)이 사망하였으며, 검증용 코호트에서는 청년층 823명 중 51명 (6.2%), 중년층 2,390명 중 190명 (8.0%), 노년층 2,532명 중 267명 (10.5%)이 사망하였다. 개발용 코호트에서 획득한 표본 수는 2,373,760개였으며, 그 중에서 양성 표본은 2.1%인 50,529개였다. 검증용 코호트에서 획득한 표본 수는 735,723개였으며, 그 중에서 양성 표본은 0.9%인 6,617개로, 양성 표본의 비율이 개발용 코호트에서

유의미하게 높았다 ($P < 0.001$). 중환자실 체류 기간(length of stay, LOS)은 개발용 코호트에서 103.025±133.142시간, 검증용 코호트에서 128.550±312.735시간으로 유의미한 차이가 있었으며 ($P < 0.001$), 사망 환자 체류 기간 역시 개발용 코호트에서 150.786±171.851시간, 검증용 코호트에서 286.260±499.382시간으로 유의미한 차이가 있었다 ($P < 0.001$).

개발된 18개 모델의 내부 검증 결과를 표로 정리하였다 (표 2). 시험용 데이터 세트에 대한 결과이며, 임계값으로 학습용 데이터 세트에 대한 F1 점수를 최대화하는 값을 사용하였다. 모델 종류별로 F1 점수가 가장 높은 모델을 해당 모델 종류의 대표 모델로 보았다 (표 2, 3).

AUROC는 0.892–0.904의 값을 보여 모델 종류 무관하게 우수한 성능을 보여주었으며 (AUROC ≥ 0.8), 대표 모델을 기준으로 GBM이 0.903 (95% 신뢰구간 0.900–0.906)을 보여 LSTM의 0.894 (95% 신뢰구간 0.891–0.898, $P < 0.001$)이나 트랜스포머의 0.898 (95% 신뢰구간 0.895–0.901, $P < 0.001$)보다 유의미하게 높았다. AUPRC는 0.297–0.349의 값을 보여 기준값(baseline)인 양성 표본의 빈도 0.021에 대비하여 모델 종류와 무관하게 높은 값을 보여주었으며, 대표 모델을 기준으로 GBM이 0.346을 보여 LSTM의 0.327이나 트랜스포머의 0.340 대비 우수하였다. 그 외 불균형 데이터의 주요 평가 지표인 F1 점수는 0.353–0.383, F2 점수는 0.342–0.378, 균형 정확도는 0.661–0.681의 범위를 보였으며, 세 가지 모두에서 GBM이 가장 우수하였다. 정확도는 0.973–0.976으로 모든 모델에서 높았으며,

ECE는 모든 모델에서 0.001–0.004로 낮게 나와 확률 보정이 이루어졌음을 확인할 수 있었다. 최종적으로는, AUROC, AUPRC, F1 점수, F2 점수, 균형 정확도 모두에서 가장 높은 값을 보인 GBM 대표 모델을 가장 우수한 모델로 보고 최종 모델로 선정하였다.

GBM 대표 모델은 AUROC를 학습 중 평가 지표로 사용하였으며, 부류 가중치와 스플라인 기반 확률 보정법이 적용된 모델로, number of trees = 337, max depth = 4, minimum child weight = 1, gamma = 10, subsampling rate = 0.5, column sampling rate = 0.5 을 초모수로 사용하였다. F1 점수가 최대인 임계값은 0.171이었다.

LSTM 대표 모델은 AUPRC를 학습 중 평가 지표로 사용하였으며, 과대표집과 스플라인 기반 확률 보정법이 적용된 모델로, epochs = 59, number of hidden layers = 1, number of hidden nodes = 16, number of dense nodes = 128, dropout rate = 0.5 를 초모수로 사용하였다. F1 점수가 최대인 임계값은 0.174이었다.

트랜스포머 대표 모델은 AUPRC를 학습 중 평가 지표로 사용하였으며, 부류 가중치와 등장 회귀법을 통한 확률 보정이 적용된 모델로, epochs = 17, number of filters = 64, number of heads = 3, embedded dimension = 64, number of transformer layers = 1, dropout rate = 0.2 를 초모수로 사용하였다. F1 점수가 최대인 임계값은 0.180이었다.

채택된 대표 모델들에 대한 외부 검증 결과를 내부 검증 결과와 함께 표로 정리하고 (표 3), receiver operating characteristic (ROC)

곡선과 정밀도-재현율(precision-recall, PR) 곡선을 그려 확인하였다 (그림 5, 6). 외부 검증은 검증용 코호트를 대상으로 하였고, 내부 검증에서와 동일하게 학습용 데이터 세트에 대한 F1 점수를 최대화하는 값을 임계값으로 사용하였다.

외부 검증에서 AUROC는 0.914-0.933의 값을 보여 모델 종류 무관하게 우수한 성능을 보여주었으며, GBM이 0.933 (95% 신뢰구간 0.931-0.936)을 보여 LSTM (0.914, 95% 신뢰구간 0.911-0.918, $P < 0.001$)이나 트랜스포머 (0.930, 95% 신뢰구간 0.927-0.933, $P < 0.001$)보다 유의미하게 높았다. 이는 입실 24시간 내의 데이터를 기준으로 재원 중 사망을 예측한 APACHE II (0.861, 95% 신뢰구간 0.844-0.878, $P < 0.001$)보다 유의미하게 높은 수치이다. AUPRC는 0.153-0.181의 값을 보여 기준값인 양성 표본 빈도 0.009에 대비하여 모델 종류와 무관하게 높은 값을 보여주었으며, GBM이 0.181을 보여 LSTM 0.153이나 트랜스포머 0.174 대비 우수하였다. 그 외 F1 점수 0.202-0.220, F2 점수 0.341-0.354, 균형 정확도 0.769-0.793 였으며, F1 점수와 F2 점수는 트랜스포머가, 균형 정확도는 GBM이 가장 높았다. 정확도는 0.955-0.963으로 모든 모델에서 높았으며, ECE는 0.020-0.026이었다. AUROC, AUPRC, 균형 정확도를 기준으로 할 경우 여전히 GBM이 가장 우수하였으나, F1 점수, F2 점수를 기준으로 할 경우 트랜스포머가 좀더 우수하였다. 종합적으로 볼 때 GBM과 트랜스포머는 유사한 성능을 보였고, 두 모델 모두 LSTM보다 우수하였다.

보정 곡선의 경우, 세 대표 모델 모두 내부 검증에서는 1:1 직선에 수렴하여 확률 보정이 되지만, 외부 검증에서는 확률 보정이 잘 되지 않고 곡선이 크게 기울어 있는 것을 확인하였다 (그림 7). 이는 개발된 모델들이 외부 검증에서 과다 추정(overestimation) 양상을 가지며, 확률 보정 기법을 사용하였음에도 교정되지 않았음을 의미한다.

다른 임계값을 사용했을 때의 결과를 표로 정리하였다 (표 4). 모든 임계값 계산의 기준은 학습용 데이터 세트였으며, AUROC, AUPRC, ECE는 임계값 변경으로는 바뀌지 않으므로 표에서 생략하였다. 임계값으로 Youden 지수를 취할 경우 내부 검증과 외부 검증 모두에서 균형 정확도는 증가하나 F1 점수, F2 점수와 정확도는 크게 감소하였다. F2 점수를 최대화하는 값을 취할 경우 내부 검증에서 F2 점수와 균형 정확도는 증가하나 F1 점수와 정확도는 감소하였으며, 외부 검증에서는 F2 점수도 감소하였다. 이러한 경향은 모델의 종류와 무관하였다.

최종 모델인 GBM 대표 모델을 기준으로 내부 검증과 외부 검증의 성능을 비교하면 (표 3, 그림 7), AUROC는 0.903에서 0.933으로, 균형 정확도는 0.681에서 0.793으로 증가하였으나, AUPRC는 0.346에서 0.181로, F1은 0.383에서 0.202로, F2는 0.378에서 0.341로 감소하였으며, 낮을수록 좋은 ECE가 0.002에서 0.026으로 증가하였고, 확률 보정 곡선상 과다 추정 양상이 교정되지 않는 등, 종합적인 성능이 감소하였다. 이러한 경향은 임계값 설정 방법과 무관하였다 (표 4).

하위집단 분석은 검증용 코호트를 대상으로 최종 모델인 GBM

모델을 사용하여 연령별, 성별, 중환자실 종류별로 이루어졌다 (표 5). 성별로는 남성에서, 연령별로는 중년에서, 중환자실별로는 SICU에서 상대적으로 좋은 지표를 보였다.

제 4 장 고 찰

본 연구에서는, 국내와는 환경이 다른 외국 의료기관 데이터베이스 중, 상대적으로 사망률이 높고 의료 자원의 공정한 배분 문제가 중요하게 고려되어야 하는 중환자실 환자를 대상으로, 모든 의료기관에서 필수적으로 계측되거나 자동으로 수집되어 해석의 여지가 적은 인구통계학적 데이터와 활력 징후 데이터들 만을 입력 변수로 삼아, 지리적 시간적 차이 없이 논쟁에서 자유로운 결과 변수인 사망 여부를 실시간으로 예측하는 모델을 만든 후, 이를 국내 의료기관에 적용하였을 때 성능이 잘 유지되는지 확인하고자 하였다. 이를 위해 실제 임상 환경에서의 실시간 예측을 전제로 한 입력 변수와 결과 변수를 담은 표본 형태를 설계하였으며, 주요 기계학습 모델을 기반으로, 초기값 편향, 데이터 정규화, 학습시 과대 및 과소 표집, 부류 가중치, 그룹화된 k-겹 층화 표집 교차 검증, 초모수 최적화, 사후 확률 보정, 임계값 조정 등 여러 기법들을 적용한 후, AUROC, AUPRC, F1, F2, 균형 정확도, ECE, 보정 곡선 등 다양한 지표로 확인하였다. 그 결과, 내부 검증에서 좋은 결과를 얻었으나, 지리적 시간적 외부 검증에서는 그 성능이 유지되지 않음을 확인하였다.

모델 개발에는 MIMIC-IV를 사용하였는데, 이는 1) 중환자실 환자를 대상으로 하고, 2) 20년 이상 꾸준히 업데이트되고 오랫동안 검증되어 높은 신뢰성을 가지면서, 3) 단일 기관 데이터베이스라는 명확한 한계점도 동시에 가지고 있기에, 지리적 시간적 외부 검증이라는 본 연구의 설계에 가장 적합하다고 보았기 때문이다. 외부 검증에

사용된 서울대학교병원의 데이터 역시 1) 서울대학교병원이 MIMIC의 Beth Israel Deaconess Medical Center와 같은 3차 의료기관이라는 점, 2) 데이터 양이 중요한 기계학습에서 서울대학교병원이 국내 최대 규모의 의료기관 중 하나라는 점, 그리고 3) 전자 의무 기록을 국내 최초로 도입하고 오랫동안 유지보수해온 대형 의료기관이라는 점에서 본 연구의 설계에 가장 적합하다고 보았다.

모델 개발에 사용한 중환자실을 내과계 중환자실과 외과계 중환자실 위주로 제한하였는데, 이는 다른 국가 다른 의료기관 간 모델 검증에서 좋은 성능을 유지하기 위해, 임상 현장에서의 차이를 고려하여 의료 보험 정보, 진료과 정보 등을 입력 변수에서 배제하다 보니 이질적인 환자군을 다루는 것에 한계가 있을 것으로 보았기 때문이다. 심혈관 중환자실(cardiac vascular ICU, CVICU)은 기저 심장 질환의 종류나 체외막 산소 공급(extracorporeal membrane oxygenation, ECMO) 사용 여부 등 특수 장비의 사용으로 인해 별도의 모델 개발이 더 적절할 것으로 보고 배제하였으며, 외상 외과 중환자실 (trauma SICU, TSICU)은 해당 환자군을 서울대학교병원에서 거의 보지 않기 때문에 배제하였다. 그 외에 두 의료기관에 공통되지 않은 중환자실들, 가령 COVID-19 환자 위주인 Disaster ICU (DICU)나 그 외 여러 부속 중환자실들은 배제하였다.

본 연구에서는 실시간 예측이라는 목적과 입력 변수의 가짓수를 줄임으로써 외부 검증에서도 성능이 유지될 것이라는 가설에 맞춰 입력 변수의 종류를 인구통계학적 데이터(나이, 성별, 키, 체중)와 활력 징후

데이터(SBP, DBP, MBP, HR, RR, BT, SpO_2, FiO_2, GCS)로 제한하고 시작하였는데, 내부 검증에서 높은 AUROC, 기저값 대비 우수한 AUPRC, 1:1에 수렴하는 보정 곡선 등 좋은 결과를 얻을 수 있었다는 점에서 유의미하였다. 관련하여 소아 중환자실(pediatric ICU, PICU)을 대상으로 한 사망 예측 연구에서 7개 활력 징후 데이터(SBP, DBP, MBP, HR, RR, BT, SpO_2)와 2개 인구통계학적 데이터(나이, 체중)를 입력 변수로 사용한 바 있다 [32]. Boruta SHAP으로 GBM 기준 예측 기여도를 확인하였을 때도 일부 대푯값을 제외한 모든 입력 변수가 양의 기여도를 보였는데, 이는 본 연구에서 사용한 인구통계학적 데이터와 활력 징후 데이터 전체가 성능에 보탬이 되므로 모두 입력 변수로 사용함이 옳음을 의미한다. 설계상 검사실 소견을 배제한 것과 관련하여 본 연구자가 2006년도부터의 서울대학교병원의 데이터를 조사한 결과, 2016년까지는 $PaCO_2$ 등의 검사실 소견을 잘 측정하지 않았고, 2017년 및 그 이후에도 한국의 주요 의료기관들 중 그 측정 빈도가 낮은 편으로 알려져 있다. GCS와 생체 신호들(vital signs)은 입원한 환자들에서 기본적으로 측정하거나 쉽게 측정할 수 있으며, 검사실 소견들에 비해 측정 빈도나 치료 방침 등에서 의료기관 간 차이가 적을 것으로 예상되는 지표들이기 때문에 국경을 넘어 타 의료기관에도 쉽게 적용할 수 있고 의료기관 간 모델의 성능 차이도 상대적으로 적을 것으로 보인다. 또한 이들 입력 변수를 사람이 실시간으로 빠르게 인지하고 통합하여 분석하기는 어려우나 인공지능의 예측 결과에 대한 사후 분석에서는 쉽게 납득 가능하다는 점에서 임상의가 데이터를

모니터링하는 피로도를 낮춰 의료 현장에서의 인공지능 활용 가능성을 높인다고 볼 수 있겠다.

본 연구에서는 결과 지표로 24시간 이내 사망 여부를 실시간 예측하였는데, 개별 환자의 예후에서 사망 여부보다 중대한 것이 없음은 부인할 수 없는 사실로, 많은 중환자실에서 사망률을 낮추기 위해 노력하고 있으며 그 결과 국내외에서 중환자실 사망률이 지속적으로 감소해오고 있음이 보고되었다 [33, 34]. 최근 연구들에서는 성인 중환자실 기준으로도 9~10% 정도를 보고하는 경우가 많아지고 있으며 [35], 본 연구의 MIMIC-IV 11.8%, 서울대학교병원 8.8%도 이에 부합하는 수치로 보인다. 그러나 중환자실 입실 환자 수와 평균 비용은 지속적으로 증가하는 경향을 보이고 있으며 [36], 관련 인력도 부족한 실정이기에 [37], 사망 확률은 의료 자원 분배 측면에서도 고려되어야 한다. 본 연구자는 자동화된 사망 예측 모델이 실제 중환자실 환경에서 자동으로 수집 가능한 데이터로부터 사망 확률을 예측하여 임상의의 의사결정에 도움이 된다면 이 문제의 해결에 기여할 수 있을 것으로 보았으며, 특히 입원시에는 모두 중환자라는 중환자실의 특성상 입원 후의 경과가 중요하기에 실시간성을 확보하고자 하였다.

본 연구에서는 검증이라는 연구 목적상, 극한의 성능 향상을 목표로 동작이 무겁고 복잡도가 높게 모델을 설계하기 보다는, 의료 분야에서 일반적으로 많이 쓰이고 있고 본 연구에서 풀고자 하는 문제의 속성에 적합해 보이는 모델들을 선별하여 비교 테스트하였다. 임상의의 관점에서 볼 때 다수의 입력 변수를 종합해가며 판단하는 임상의들의

의사결정 방식과 동작 원리가 비슷하다는 점에서 GBM을, 데이터의 관점에서 볼 때 실시간 예측을 위해 시계열 데이터를 다룬다는 점에서 순환 신경망(recurrent neural network, RNN)의 대표격인 LSTM을, 컴퓨팅(computing)의 관점에서 볼 때 전체 시계열 데이터를 한꺼번에 병렬로 처리하면서도 중요한 정보에 대한 가중치를 잃지 않는 트랜스포머를 채택하여 모델을 만들고 학습시킨 후 성능을 비교하였다. 그 결과 세 모델 모두에서 APACHE II보다 우월한 성능을 보였는데, 이는 시계열 데이터를 이용하여 예측하기 때문에 시간에 따른 입력 변수들의 변화 양상을 포함하여 판단한다는 점과, 실시간으로 예측할 수 있다는 점 때문에 좋은 성능을 가진 것으로 보인다. 시계열 데이터라는 점 때문에 LSTM이나 트랜스포머의 성능이 더 좋을 것이라고 예상할 수 있으나, 최종적으로 가장 성능이 좋은 것은 GBM이었다. 관련하여 중환자실 재원 중 사망 확률을 예측하는 Risk of Inpatient Death (RIPD) 모델이 GBM을 사용하여 좋은 결과를 보여준 바 있으며 [35], 소아 중환자실 대상 연구에서도 GBM이 LSTM보다 더 우수한 결과를 보여준 바 있다 [32]. 이는 자동화 및 표준화 되어있는 활력 징후 데이터들을 입력 변수로 사용하였기 때문에, 비교적 전체 경우의 수가 제한되어, 경우의 수를 잘 암기하는 GBM의 결과가 좋았던 것으로 판단된다. LSTM과 트랜스포머 중에서는 트랜스포머의 성능이 대체로 더 우수하였으며, 이는 최근의 인공지능 분야에서의 두 모델에 대한 일반적인 인식과 일치한다.

본 연구에서는 1시간 간격인 이전 24개 시점의 데이터를 입력 변수,

이후 24시간 이내의 사망 여부를 출력 변수가 되게끔 표본을 생성하였는데, 설계상 중환자실 재원 중 어느 시점에서도 입력 변수 산출이 가능하므로 실시간 예측이 가능해진다. 계측 빈도와 각 입력 변수의 변동성을 고려하여, 인구통계학적 데이터는 각 시점까지 알려진 가장 마지막 값을, 활력 징후 데이터는 각 시점에서 24시간 이내의 마지막 값을 사용하게 하였는데, 실제 임상 환경에서 인구통계학적 데이터 값은 항상 있으며, 활력 징후 데이터는 자주 측정되기에 외부 의료기관에의 모델 적용에 무리가 없을 것으로 생각된다.

이러한 표본 생성 방식은 사망률 이상의 불균형 데이터 문제를 야기할 수 있으나, 실제 중환자실에 도입하여 실시간으로 데이터를 수합한다면 피할 수 없는 문제이기에, 본 연구에서는 불균형 데이터를 다루는 다양한 기법들을 고려하여 이를 극복하고자 노력하였다. 대표적인 기법으로 과소표집, 과대표집, 부류 가중치가 있으며, 일반적으로 서로 비슷한 성능을 가지는 것으로 알려져 있지만, 부류 가중치의 경우 양성 표본에 대한 가중치 배율 적용으로 인해 높고 낮은 2가지 학습률(learning rate)을 동시에 가지는 효과가 있어서 일반적으로 모델 최적화에서 불리하다고 알려져 있고, 과소표집은 양성 표본의 비율이 낮을수록 정보 손실의 우려가 커지며, 과대표집 기법으로 많이 고려되는 synthetic minority oversampling technique (SMOTE)의 경우 본 연구에서처럼 한 환자의 표본이 서로 다른 그룹으로 섞여 들어가면 안 되어 표본의 그룹화가 필요한 경우엔 적용 관련 방법론이 명확하게 정리되어 있지 않아 적용이 어렵다. 이에 본

연구는 학습용 데이터 세트를 다룰 때 임의 과소표집(random under-sampling), 임의 과대표집(random over-sampling), 부류 가중치의 결과를 각각 확인하였으며, 모델 종류 및 학습 중 평가 지표와의 조합에 따라 최선의 기법이 달라짐을 확인하였다. 결과적으로 GBM과 트랜스포머 대표 모델에서는 부류 가중치, LSTM 대표 모델에서는 임의 과대표집이 가장 좋은 결과를 보였다. 그 외에 교차 검증 과정에서 표본의 치우침이 생기는 것을 억제하기 위해 항상 층화 표집 기법을 적용하였고, LSTM과 트랜스포머 학습시에는 학습 수렴 속도 및 결과 개선을 도모하기 위하여 초기값 편향을 부여하였다.

본 연구는 심한 불균형 문제를 다루기 때문에 모델 개발 및 평가 시에 다양한 지표들을 고려하였다. 일반적으로 분류 문제에서 모델 성능 지표로 AUROC를 흔히 사용하는데, 이는 기준값이 0.5이고 값의 범위가 0과 1 사이이며, 임계값의 영향을 받지 않기에 서로 다른 모델 간의 비교가 용이하기 때문이다. 하지만 이는 재현율와 특이도를 동시에 다루는 지표로, 양성 표본 중 양성 예측의 비율, 음성 표본 중 음성 예측의 비율을 의미하기 때문에, 양성과 음성 표본의 빈도가 비슷하거나 양성과 음성 여부의 중요도가 비슷한 모델들을 평가하는 데에는 유용하나, 양성 표본의 빈도가 낮을수록 고평가하기 쉽기에 [38], 양성과 음성의 빈도 및 중요도가 불균형할 때는 다른 지표를 함께 고려해야만 한다. 정밀도(precision)와 재현율은 둘다 예측과 실제 모두 음성인 표본(true negative)을 다루지 않기 때문에 불균형 문제에서 더 큰 의의를 가지며, 임계값을 조정하여 각각을 1로 만들 수 있기에, 둘을

동시에 다루면서 임계값에 영향을 받지 않는 AUPRC는 매우 중요한 지표이다 [39, 40]. 하지만 AUPRC는 기준값이 양성 표본의 빈도이며, 데이터 세트가 얼마나 불균형한가에 따라 달성 불가능한 영역이 있어 값의 범위가 0에서 1이 될 수 없다 [41]. 따라서 같은 데이터 세트를 대상으로 같은 데이터 전처리(data preprocess)를 하지 않았다면 모델 간 비교에 사용하기 어렵다. 이에 본 연구에서는 최종 평가 목적이 아닌 학습 중 평가 지표로 AUPRC를 고려하였으며, AUROC 최적화와 AUPRC 최적화에서 모델의 성능 차이가 날 수 있다고 알려져 있기에 [40], 교차 검증 시 두 가지 방법을 모두 확인하였다. 비슷한 지표로는 정밀도와 재현율의 조화 평균인 F1 점수와, F1 점수에서 재현율의 가중치를 높인 F2 점수가 있으나, 임계값에 따라 값이 크게 달라지기에, 양성과 음성 표본의 비율이 다른 부류 가중치 학습에서 적당한 값을 미리 선택할 수 없어서 학습 중 평가 지표에서는 배제하였다. 다만 실제 임상에서 사용하려면 임계값을 정해야 하고, 내부 검증된 모델을 그대로 외부 검증하려 했기 때문에, 학습용 데이터 세트에서 F1 점수를 최대화하는 임계값을 설정한 후 F1 점수를 기준으로 모델을 선별하였다. 양성 표본인 환자 사망이 임상적으로 중대한 위험임을 고려하여 성능 평가시 F2 점수도 확인하였다.

본 연구는 모델의 출력값 그 자체를 환자의 사망 확률로 해석 가능하게끔 함으로써 의료 현장에서의 활용도를 높이기 위하여 확률 보정 기법을 적용하였다. 확률 보정 성능이 좋지 않은 모델은 그 출력값을 독립적인 확률로 해석할 수 없기 때문에 통계적인 활용도가

떨어지며, 또한 의료 인공지능의 경우 후향적 연구를 기반으로 전향적 연구를 설계하여 데이터 세트가 계속 추가되는 경우가 많아 지속적인 학습과 피드백을 통해 모델을 꾸준히 개선하게 되는데 확률 보정이 되지 않았던 경우 업데이트된 모델의 성능에 대한 일관성 있는 판단이 어렵다 [42]. 아쉽게도 기계학습 모델에서 과다 추정하는 방향으로 확률 보정 성능이 좋지 않은 것은 흔히 관찰되는 현상이며 특히 불균형 데이터 문제에서는 더욱 그러하다고 알려져 있다 [43, 44]. 본 연구의 기계학습 모델들 역시 확률 보정 전에는 실제 확률을 과다 추정하였다. 이에 본 연구는 Platt 비례법, 등장 회귀법, 스플라인 기반 확률 보정의 세 가지 방식으로 사후 보정을 시도해 보았으며, 세 방법 다 확률 보정 성능에 개선이 있었으나, Platt 비례법보다는 등장 회귀법이나 스플라인 기반 방식이 좀더 우수한 결과를 보였다. 이는 Platt 비례법의 경우 모델의 예측과 실제 확률이 S자 모양의 로지스틱(logistic) 관계에 있다는 가정 하에 동작하는 방식이고 [16], 등장 회귀법이나 스플라인 기반 방식의 경우 비모수적(non-parametric)이기에 다양한 모양의 신뢰도 그림에 적용 가능하며, 데이터가 충분하여 과적합을 피할 수 있었기 때문으로 보인다. 등장 회귀법과 스플라인 기반 방식의 차이는 구간별로 상수 근사(piecewise-constant approximations)를 하였는지 와 매끄러운 입방 다항식(smooth cubic polynomial)을 사용하는지 였으며 [18], 어느 쪽이 우수한지는 모델 별로 달랐으나 대체로 비슷한 ECE 점수를 보여주었다. 보정 성능 확인을 위하여 ECE, Brier 점수, 보정 곡선 육안 확인 세 가지 방법 모두 검토하였으며, 보정 방법 간 성능이 비슷할

경우 육안 비교는 어려움이 있고, Brier 점수는 보정 손실(calibration loss)와 정제 손실(refinement loss)의 조합이기 때문에 보정 성능 만을 대변하지는 않아서, 최종적으로는 ECE를 기준으로 비교하였다.

본 연구는 모델의 예측에 사용하는 임계값으로 학습용 데이터 세트에서 F1 점수를 최대화하는 값 외에 재현율과 특이도를 최대화하는 값(Youden 지수), F2 점수를 최대화하는 값도 추가로 계산하여 확인하였는데, 내부 검증을 기준으로 Youden 지수를 사용할 경우 균형 정확도가 증가하고, F2 점수를 최대화하는 값을 사용할 경우 균형 정확도와 F2 점수가 상승하였다. 이는 임상적 상황이나 목표에 맞춰서 임계값을 사후 조정하는 것을 고려해볼 수 있음을 의미한다. 우리의 결과 지표인 원내 사망의 경우 일반적으로 재현율이 정밀도보다 더 중요하겠으나, 의료 자원의 분배 측면에서는 정밀도 역시 중요하므로, 이를 도입하는 의료기관의 판단이 필요한 부분이라고 할 수 있겠다.

본 연구는 모델 개발에 사용된 데이터를 획득된 의료기관과 시간적, 지리적, 문화적으로 완전히 독립적인 타 국가의 의료기관에서도 해당 모델을 그대로 사용 가능할지를 확인하고자 하였는데, 이러한 예측 모델의 외부 유효성은 환자 집단 전반에 걸쳐 명확한 예측을 보장하고 임상 도구로서 모델의 유용성을 확립하는데 중요하다 [45]. 아쉽게도 불균형 데이터 문제에서는 한 가지 지표만으로 성능을 파악할 수 없는 데다가, 내부 검증과 외부 검증에서 불균형도의 차이가 클수록 더욱 해석하기 어렵다. 이는 불균형도가 심할수록 AUROC는 오르고, AUPRC, F1 점수는 내려가기 때문이다. 본 연구의 외부 검증에 사용된 검증용

코호트의 양성 표본 비율은 0.9%에 불과하여 절대값 자체가 낮은 데다가 상대적 관점에서도 내부 검증에서의 2.1%의 절반 이하였기에, 외부 검증에서 AUPRC, F1, F2 점수가 낮아진 것만으로 성능이 감소했다고 결론짓기는 어렵다. 또한 AUROC는 0에서 1이라는 공통된 범위와 0.5라는 공통된 기준값을 가지기에, 아무리 불균형 데이터 문제일지라도 외부 검증에서 0.9 이상의 값을 유지했다는 점에서는 성능이 유지되었다고도 볼 수 있다. 그럼에도 불구하고 본 연구자는 확률 보정 정도를 기준으로 하여 외부 검증에서 성능이 감소했다고 최종 판단하였다. 이는 본 연구의 목적이 개발된 모델의 외부 기관 적용 가능성 확인인데, 확률 보정이 안되면 모델의 출력값을 확률로 해석할 수 없어서 임상 현장에의 적용이 어렵기 때문이다. 내부 검증에서는 잘 동작하였음을 감안하면 각 의료기관의 데이터에 특화된 모델의 필요성을 의미한다고 볼 수 있으며, 외부 검증에서도 AUROC 등이 유지되었음을 감안하면 적용할 기관의 데이터로 기존 모델을 재학습시키는 방안을 고려할 수 있겠다. 해당 의료기관의 방침이나 상황에 따라 임계값 역시 새로이 조정 가능하겠다. 결과적으로 개발된 모델의 성능이 뛰어나다고 하더라도, 기관별로 성능이 달라질 수 있으며, 적용을 위한 기관별 추가 조치가 필요하기에, 이와 같은 인공지능 의료기기의 식약처 인허가 과정에서 해당 기관에서의 성능을 확인하고 보장할 수 있는 방법의 마련이 필요하다.

인공지능을 활용하여 사망 확률을 예측하려 한 다양한 선행 연구들이 존재하며 (표 6), 대체로 AUROC ≥ 0.85로 APACHE,

Pediatric Index of Mortality (PIM) 등 기존의 전통적인 예측 수단에 비해 좋은 결과를 보여주었다.

입력 변수 측면에서는 본 연구처럼 입력 변수의 부류나 개수를 제한하고 변수 간 속성 중복을 배제하고자 한 연구와, 가능한 많은 입력 변수를 활용하려 한 여러 연구들이 있었다. Landon Brand 등은 4개 생체 신호 만을 사용하였으며 [46], Jacob Deasy 등은 전자 의무 기록으로부터 얻을 수 있는 모든 값을 사용하였다 [47]. Hans-Christian Thorsen-Meyer 등은 개별 입력 변수의 예측에의 기여도를 확인하는데 주력하였으며 [48], Huizhen Jiang 등은 입력 변수의 부류별로 예측력에 차이가 발생하는지 확인하려 하였다 [49].

결과 지표 측면에서는 예측 시점이나 사망 원인에 대한 차이를 보이고 있다. 장기 예후 예측을 시도한 연구들이 대부분이나, 본 연구처럼 단기 예후 예측을 시도한 연구도 있었다. Stephanie Baker 등은 직전 24시간 데이터를 통해 3, 7, 14일 이내 사망여부를 예측하는 연구를 수행하였는데 [50], 이는 의료 자원 분배를 위한 실시간성을 목적으로 시간 단위인 24시간 이내 사망여부 예측을 하는 본 연구와 차이를 보인다. 또한 본 연구처럼 모든 원인에 의한 사망(all-cause mortality) 확률 예측을 시도한 연구들이 대부분이나, 특정 원인에 의한 사망 예측으로 좁혀 시도한 연구도 있었다. Joon-Myoung Kwon 등은 병실환자의 심정지(cardiac arrest)를 예측하는 연구를 수행하였다 [51].

환자군 측면에서는 중환자실 대상 연구가 대부분인데, 이에는 데이터가 집중적으로 수집되는 중환자실의 특성이 큰 역할을 한 것으로

보인다. 중환자실 전체 환자를 대상으로 한 연구들이 많으나, 소아 중환자실이나, 외과계 중환자실과 같이 범위를 좁힌 연구들도 있었다. Soo Yeon Kim 등은 소아 중환자실 환자를 대상으로 한 연구를 수행하였으며 [32], Kyongsik Yun 등은 수술 후 외과계 중환자실을 입원했던 환자들을 대상으로 한 연구를 수행하였다 [52]. 본 연구는 MIMIC의 가장 최신 버전인 MIMIC-IV를 개발용 코호트로 삼았으며, 내과계 중환자실 및 외과계 중환자실을 대상으로 하여 너무 범위를 좁히지 않으면서도 환자군의 차이에 영향을 덜 받고자 하였다.

연구에 사용한 인공지능의 모델 종류도 계속 달라져왔다. 초기 연구들에서는 다층 퍼셉트론(multi-layer perceptron, MLP)이나 합성곱 신경망(convolutional neural network, CNN) 위주로 성과를 보여주었다면, 최근의 연구들은 RNN이나 GBM 위주의 연구들이 성과를 보여주고 있다. Ryan J Delahanty 등은 입원 후 24시간 이내 데이터를 사용하여 입원 중 사망을 예측하는 연구에서 GBM을 사용하여 APACHE보다 더 높은 AUROC를 얻을 수 있음을 주장하였다 [35]. 본 연구는 LSTM, GBM 외에 트랜스포머라는 좀더 최근에 등장한 모델도 시도하고 결과를 비교하였다.

인공지능 모델에 대한 외부 검증의 필요성이 대두되고, 국가 및 의료기관별 의료 빅데이터 구축이 진행되어 사용 가능한 데이터베이스가 늘어나면서, 외부 검증을 포함하는 경우가 점차 많아지고 있는데, 대체로 내부 검증과 외부 검증 간에 상당한 성능 차이를 보이고 있다. Yixi Xu 등은 COVID-19 환자의 입원 중 사망 및 장기 부전(organ

failure)에 대한 위험도를 예측하는 기계학습 모델을 미국의 University of Washington에서 개발한 후 중국의 Tongji Hospital에 외부 검증하는 연구를 진행하였으며 내부 검증에서 AUROC 0.72, 외부 검증에서 AUROC 0.85를 보고하였다 [53]. Chiang Dung-Hung 등은 공개 데이터베이스인 eICU-CRD 데이터로 개발된 기계학습 모델인 Hemodynamic Stability Index (HSI)를 Taipei Veteran General Hospital (TPEVGH)에 검증하는 연구를 하였으며 HSI 개발자들이 보고한 AUROC 0.82보다 낮은 AUROC 0.76 정도가 나옴을 보고하였다 [54]. Yanni Kang 등은 공개 데이터베이스인 eICU-CRD 데이터를 이용하여 중환자실 입실 직후부터 48시간동안의 모든 전자 의무 기록 데이터로부터 입원 중 사망을 예측하는 모델을 개발한 후 역시 공개 데이터베이스인 MIMIC-III 데이터를 대상으로 외부 검증하였으며 내부 검증에서 AUROC 0.899, 외부 검증에서 AUROC 0.855를 보고하였다 [55]. Ryoung-Eun Ko 등은 Yonsei Cancer Center (YCC) 데이터를 이용하여 중환자실에 입원한 성인 암 환자를 대상으로 생체 및 검사실 소견 데이터로부터 28일 내 사망 여부를 예측하는 기계학습 모델을 개발한 후 Samsung Medical Center (SMC)와 MIMIC-III 데이터를 대상으로 외부 검증하였으며 내부 검증에서 AUROC 0.939, 외부 검증에서 AUROC 각각 0.775, 0.753을 보고하였다 [56]. 본 연구는 불균형 데이터임을 고려하더라도 내부 검증과 외부 검증 모두에서 높은 AUROC를 보였는데, 실시간 예측이며, 의료기관 간의 차이가 적을 것으로 예상되는 입력 변수만을 사용하였고, 데이터 전처리 과정을

충분히 검토하였으며, 불균형 데이터 및 결측값에 대한 여러 기법을 적용한 것이 도움이 되었을 가능성이 있다. 다만 AUPRC나 F1 점수, F2 점수가 모델 종류와 무관하게 감소하였고 외부 검증 시 확률 보정 성능이 좋지 않아 종합적으로 볼 때는 본 연구에서도 외부 검증 성능이 감소하였다고 생각된다.

본 연구에 몇 가지 한계점들이 있다. 첫째, 전체 중환자실이 아닌 내과계 중환자실과 외과계 중환자실 만을 다루었다는 점이다. 이는 국가 간 의료기관 간 차이를 고려하여 입력 변수를 단순화하다 보니 너무 이질적인 환자군을 다루는 것에 한계가 있을 것으로 보았기 때문이다. 향후 배제된 중환자실 종류들을 포함한 후속 연구가 필요하다. 둘째, 서울대학교병원의 경우 전자 의무 기록에 기록된 전체 기간이 아닌 2017년 1월부터 2021년 2월까지의 데이터만을 사용하면서 증례 숫자가 감소했다는 점이다. 이는 2017년 이전 기록에서의 검사실 소견 데이터의 부족과 데이터 정제 완성도 문제를 우려하였기 때문이다. 서울대학교병원이 국내 최대의 의료기관 중 하나이기에 기간 제한에도 불구하고 외부 검증 가능한 숫자를 확보할 수는 있었으나, 향후 2006년까지의 데이터를 더 정제하고, 2021년 3월부터의 데이터를 추가하여, 충분한 양성 표본의 수를 확보하고 특정 하위집단의 사망 환자의 수가 너무 적지 않게 한 후 재확인해볼 필요가 있다. 셋째, 심한 불균형 데이터 문제를 충분히 잘 다루었는지에 대해 추가적인 시도가 필요할 수 있다. 두 기관의 입원 중 사망률 자체가 낮은 데다가, 향후 24시간 이내 사망한 경우에만 양성 취급하였기에 양성 표본의 비율은

더욱 낮았다. 이를 극복하기 위한 방법으로 상술한 다양한 기법을 적용 후 비교하였으나, 충분하지 않았을 수 있다. 향후 좀더 다양하고 고도화된 기법을 시도해볼 필요가 있으며, 근본적으로는 실시간 예측에 활용 가능하면서도 불균형 문제를 줄이는 방향으로 표본을 설계할 필요가 있다. 넷째, 보정 전 모델의 확률 보정 성능이 좋지 않았다는 점이다. 인공지능 모델에서 흔한 현상이고 사후보정 자체를 두려워할 것이 없다는 주장도 있지만 [43], 인공지능 모델이라 할지라도 학습 데이터 양이 충분할수록 확률 보정 성능이 좋은 경향이 있음을 고려하면 [42], 후속 연구에서는 모델 개발에 있어서 좀더 많은 수의 증례를 확보할 필요가 있겠다. 다섯째, AUPRC가 낮아 거짓 양성(false positive) 예측으로 인한 경보 피로(alarm fatigue)가 있을 수 있다는 점이다. 하지만 예측하려는 결과 지표의 양성 표본 비율 자체가 낮다는 점에서 AUPRC가 낮은 것은 어쩔 수 없는 측면이 있으며, 환자 사망의 중대성과, 본 모델이 자동 수집 데이터로부터의 실시간 예측이라는 점에서 의료 자원의 부담을 줄여주는 이점이 더 클 것이라고 조심스럽게 주장해본다. 여섯째, 외부 검증에서 성능 감소 여부를 충분히 잘 다루었는지에 대한 추가적인 고찰이 필요할 수 있다. 본 연구는 성능 비교를 위해 기존에 알려진 여러 지표들을 제시하였으나 사후 확률 보정 정도를 제외한 나머지 지표들에 대해서는 성능 증감 여부에 대해 명확히 결론 내리지 못했다. 관련하여 좀더 많은 기관에 대해 외부 검증하면서 각 지표의 유용성에 대한 추가적인 고찰과 이를 통한 결론이 필요할 수 있다. 일곱째, 의료기관별 특화 모델 생성 혹은 기존 모델의 재학습을

포함한 외부 기관에의 모델 적용 및 이를 감안한 인공지능 의료기기 식약처 인허가 방안의 필요성을 제언하였으나 관련하여 구체적인 논의는 이루어지지 못했다. 마지막으로, 본 연구는 후향적 연구로, 실제 중환자실 환경에서 기계학습 모델 예측이 임상의에게 어느 정도 도움이 되는지에 대한 평가는 이루어지지 않았다. 기계학습 모델의 현장 투입을 위한 후속 연구가 필요하다.

결론적으로, 본 연구에서 외국의 공개 데이터베이스를 활용하여 실제 중환자실에서 자동으로 수집 가능한 활력 징후 데이터로부터 환자 사망을 실시간으로 예측하는 고성능의 인공지능 예측 모델을 만들 수 있었다. 단, 이렇게 생성된 모델을 지리적, 시간적, 문화적으로 완전히 독립적인 의료기관에 그대로 적용할 경우 성능이 감소하기 때문에, 기초 모델의 재학습 등을 통하여 각 의료기관의 데이터에 적합한 활용도 높은 인공지능 모델을 개발하는 후속 연구가 필요할 것으로 생각된다.

참고 문헌

1. Pastorino, R., et al., *Benefits and challenges of Big Data in healthcare: an overview of the European initiatives.* Eur J Public Health, 2019. **29**(Supplement_3): p. 23-27.
2. Siontis, G.C., et al., *External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination.* J Clin Epidemiol, 2015. **68**(1): p. 25-34.
3. Knaus, W.A., et al., *APACHE II: a severity of disease classification system.* Crit Care Med, 1985. **13**(10): p. 818-29.
4. Le Gall, J.R., S. Lemeshow, and F. Saulnier, *A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study.* JAMA, 1993. **270**(24): p. 2957-63.
5. Lemeshow, S., et al., *Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients.* JAMA, 1993. **270**(20): p. 2478-86.
6. Johnson, A., et al., *MIMIC-IV (version 1.0).* PhysioNet, 2021.
7. Pollard, T.J., et al., *The eICU Collaborative Research Database, a freely available multi-center database for critical care research.* Sci Data, 2018. **5**: p. 180178.
8. Rutledge, R., et al., *Appropriate use of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the*

Glasgow verbal score from the Glasgow eye and motor scores. J Trauma, 1996. **41**(3): p. 514-22.

9. Livingston, G., et al., *Dementia prevention, intervention, and care: 2020 report of the Lancet Commission.* Lancet, 2020. **396**(10248): p. 413-446.

10. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.

11. Kursa, M.B. and W.R. Rudnicki, *Feature selection with the Boruta package.* Journal of statistical software, 2010. **36**: p. 1-13.

12. v. Neumann, J., *Zur theorie der gesellschaftsspiele.* Mathematische annalen, 1928. **100**(1): p. 295-320.

13. Shapley, L.S., *Notes on the n-person game—ii: The value of an n-person game.(1951).* Lloyd S Shapley, 1951.

14. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

15. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

16. Platt, J., *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in large margin classifiers, 1999. **10**(3): p. 61-74.

17. Niculescu-Mizil, A. and R. Caruana. *Predicting good*

probabilities with supervised learning. in *Proceedings of the 22nd international conference on Machine learning*. 2005.

18. Lucena, B., *Spline-based probability calibration.* arXiv preprint arXiv:1809.07751, 2018.
19. Guo, C., et al. *On calibration of modern neural networks*. in *International conference on machine learning*. 2017. PMLR.
20. Brier, G.W., *Verification of forecasts expressed in terms of probability.* Monthly weather review, 1950. **78**(1): p. 1-3.
21. Youden, W.J., *Index for rating diagnostic tests.* Cancer, 1950. **3**(1): p. 32-5.
22. Parsons, V.L., *Stratified sampling.* Wiley StatsRef: Statistics Reference Online, 2014: p. 1-11.
23. Ahsan, M.M., et al., *Effect of data scaling methods on machine learning algorithms and model performance.* Technologies, 2021. **9**(3): p. 52.
24. Van Rossum, G. and F.L. Drake Jr, *Python tutorial*. Vol. 620. 1995: Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.
25. Abadi, M. *TensorFlow: learning functions at scale*. in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*. 2016.
26. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting*

system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

27. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.

28. Keany, E., *BorutaShap: A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values.* Zenodo, 2020.

29. D'Agostino, R.B., Sr., et al., *Cardiovascular Disease Risk Assessment: Insights from Framingham.* Glob Heart, 2013. **8**(1): p. 11-23.

30. Muller, M.P., et al., *Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia?* Clin Infect Dis, 2005. **40**(8): p. 1079-86.

31. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988. **44**(3): p. 837-45.

32. Kim, S.Y., et al., *A deep learning model for real-time mortality prediction in critically ill children.* Crit Care, 2019. **23**(1): p. 279.

33. Zimmerman, J.E., A.A. Kramer, and W.A. Knaus, *Changes in*

hospital mortality for United States intensive care unit admissions from 1988 to 2012. Crit Care, 2013. **17**(2): p. R81.

34. Park, J., et al., *A nationwide analysis of intensive care unit admissions, 2009-2014 - The Korean ICU National Data (KIND) study.* J Crit Care, 2018. **44**: p. 24-30.

35. Delahanty, R.J., D. Kaufman, and S.S. Jones, *Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients.* Crit Care Med, 2018. **46**(6): p. e481-e488.

36. Halpern, N.A., et al., *Trends in Critical Care Beds and Use Among Population Groups and Medicare and Medicaid Beneficiaries in the United States: 2000-2010.* Crit Care Med, 2016. **44**(8): p. 1490-9.

37. Halpern, N.A., et al., *Critical care medicine in the United States: addressing the intensivist shortage and image of the specialty.* Crit Care Med, 2013. **41**(12): p. 2754-61.

38. Movahedi, F., R. Padman, and J.F. Antaki, *Limitations of ROC on imbalanced data: Evaluation of LVAD mortality risk scores.* arXiv preprint arXiv:2010.16253, 2020.

39. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PloS one, 2015. **10**(3): p. e0118432.

40. Davis, J. and M. Goadrich. *The relationship between Precision-Recall and ROC curves*. in *Proceedings of the 23rd international conference on Machine learning*. 2006.

41. Boyd, K., et al., *Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation.* Proc Int Conf Mach Learn, 2012. **2012**: p. 349.

42. Van Calster, B., et al., *Calibration: the Achilles heel of predictive analytics.* BMC Med, 2019. **17**(1): p. 230.

43. Wang, D.-B., L. Feng, and M.-L. Zhang, *Rethinking calibration of deep neural networks: Do not be afraid of overconfidence.* Advances in Neural Information Processing Systems, 2021. **34**: p. 11809-11820.

44. Wei, H., et al. *Mitigating neural network overconfidence with logit normalization*. in *International Conference on Machine Learning*. 2022. PMLR.

45. Girardat-Rotar, L., et al., *Temporal and geographical external validation study and extension of the Mayo Clinic prediction model to predict eGFR in the younger population of Swiss ADPKD patients.* BMC Nephrol, 2017. **18**(1): p. 241.

46. Brand, L., et al. *Real Time Mortality Risk Prediction: A Convolutional Neural Network Approach*. in *HEALTHINF*. 2018.

47. Deasy, J., P. Lio, and A. Ercole, *Dynamic survival prediction in*

intensive care units from heterogeneous time series without the need for variable selection or curation. Sci Rep, 2020. **10**(1): p. 22129.

48. Thorsen-Meyer, H.C., et al., *Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records.* Lancet Digit Health, 2020. **2**(4): p. e179-e191.

49. Jiang, H., et al., *Noninvasive Real-Time Mortality Prediction in Intensive Care Units Based on Gradient Boosting Method: Model Development and Validation Study.* JMIR Med Inform, 2021. **9**(3): p. e23888.

50. Baker, S., W. Xiang, and I. Atkinson, *Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach.* Sci Rep, 2020. **10**(1): p. 21282.

51. Kwon, J.M., et al., *An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest.* J Am Heart Assoc, 2018. **7**(13).

52. Yun, K., et al., *Prediction of Mortality in Surgical Intensive Care Unit Patients Using Machine Learning Algorithms.* Front Med (Lausanne), 2021. **8**: p. 621861.

53. Xu, Y., et al., *Machine learning-based derivation and external*

validation of a tool to predict death and development of organ failure in hospitalized patients with COVID-19. Sci Rep, 2022. **12**(1): p. 16913.

54. Dung-Hung, C., et al., *External validation of a machine learning model to predict hemodynamic instability in intensive care unit.* Crit Care, 2022. **26**(1): p. 215.

55. Kang, Y., et al., *A Clinically Practical and Interpretable Deep Model for ICU Mortality Prediction with External Validation.* AMIA Annu Symp Proc, 2020. **2020**: p. 629-637.

56. Ko, R.E., et al., *Machine Learning-Based Mortality Prediction Model for Critically Ill Cancer Patients Admitted to the Intensive Care Unit (CanICU).* Cancers (Basel), 2023. **15**(3).

Abstract

External Validation of the performance of an Intensive Care Unit Mortality Prediction Model Trained with MIMIC-IV

Ho-Jong Lee
Department of Anesthesiology and Pain Medicine
The Graduate School
Seoul National University

Background: While there has been significant progress in the build of big data and development of artificial intelligence (AI) models for healthcare, the practical application of these models remains a topic of active debate. Many AI studies utilize public intensive care unit (ICU) databases, such as the Medical Information Mart for Intensive Care (MIMIC) and eICU Collaborative Research Database, but few have demonstrated meaningful external validation results in diverse healthcare settings. This study aimed to develop a machine learning model to predict patient mortality in real-time from automatically collectable data in a real ICU environment by utilizing foreign public databases, and to validate its performance when applied to a geographically and temporally distinct Korean medical institution.

Methods: A development cohort was established from MIMIC-IV, consisting of 23,152 ICU patients aged 18 or above with a minimum stay of 24 hours. Vital data from the preceding 24 hours or so, collected in 1-hour intervals, was used to develop a model predicting mortality within the next 24 hours. Three commonly used machine learning methods, gradient boosting machine (GBM), long short-term memory (LSTM), and transformer, were examined, and post-hoc probability calibration techniques were applied. The performance of the model was evaluated on a testing dataset for internal validation using metrics including the area under the receiver operating characteristic (AUROC) curve, the area under the precision-recall curve (AUPRC), F1 and F2 scores, and probability calibration curves. Subsequent to the formation of a validation cohort of 5,745 patients at Seoul National University Hospital, a facility distinct in geography and time from the cohort used for development, an external validation process was carried out.

Results: In the internal validation, the GBM model performed best, exhibiting an AUROC of 0.903, an AUPRC of 0.346 (baseline: 0.021), an F1 score of 0.383, and an F2 score of 0.378. Overestimation was rectified through probability calibration. In the external validation, the AUROC was well-maintained at 0.933, but the AUPRC decreased to

0.181 (baseline: 0.009), F1 score to 0.202, and F2 score to 0.341. Additionally, overestimation trends were not corrected on the probability calibration curve, leading to an overall decrease in performance.

Conclusion: The performance of the machine learning model, developed using a public database, decreased during external validation targeting a medical institution in another country that is geographically and temporally distinct. Research is needed on ways to utilize artificial intelligence models in clinical settings, including the creation of models specialized for each medical institution's data or retraining of existing models.

Keywords : external validation, machine learning, real-time, mortality prediction, ICU
Student Number : 2020-23022

Table 1. Baseline characteristics in the datasets.

		Development cohort			Validation cohort	P-value
		Training dataset	Testing dataset	Total		
Patients		18,521	4,631	23,152	5,745	
	Survival	16,323(88.1%)	4,095(88.4%)	20,418(88.2%)	5,237(91.2%)	<0.001
	Death	2,198(11.9%)	536(11.6%)	2,734(11.8%)	508(8.8%)	
Samples		1,910,403	463,357	2,373,760	735,723	
	Negative	1,869,820(97.9%)	453,411(97.9%)	2,323,231(97.9%)	729,106(99.1%)	<0.001
	Positive	40,583(2.1%)	9,946(2.1%)	50,529(2.1%)	6,617(0.9%)	
LOS		103.643±135.959	100.554±121.207	103.025±133.142	128.550±312.735	<0.001
	Survival	97.230±128.726	94.239±112.833	96.630±125.703	113.252±283.720	<0.001
	Death	151.269±173.617	148.802±164.553	150.786±171.851	286.260±499.382	<0.001
Age		62.531±16.434	62.613±16.595	62.547±16.466	61.296±15.144	<0.001
	Young (<45)	2,714(14.7%)	670(14.5%)	3,384(14.6%)	823(14.3%)	
	Middle age (45~65)	7,094(38.3%)	1,770(38.2%)	8,864(38.3%)	2,390(41.6%)	
	Old (>65)	8,713(47.0%)	2,191(47.3%)	10,904(47.1%)	2,532(44.1%)	
Sex	Male	9,825(53.0%)	2,537(54.8%)	12,362(53.4%)	3,347(58.3%)	<0.001
	Female	8,696(47.0%)	2,094(45.2%)	10,790(46.6%)	2,398(41.7%)	
ICU	MICU	7,032(38.0%)	1,757(37.9%)	8,789(38.0%)	1,290(22.5%)	<0.001
	SICU	5,507(29.7%)	1,418(30.6%)	6,925(29.9%)	2,012(35.0%)	
	Neuro SICU	563(3.0%)	138(3.0%)	701(3.0%)	2,443(42.5%)	
	MICU/SICU	5,419(29.3%)	1,318(28.5%)	6,737(29.1%)	0(0.0%)	
Weight (kg)		82.287±24.639	82.345±24.298	82.298±24.570	62.898±12.743	<0.001
Height (cm)		168.234±14.362	168.770±13.921	168.342±14.276	161.069±15.967	<0.001
Vital signs	SBP	121.291±16.594	121.476±16.426	121.328±16.561	128.789±17.626	<0.001
	DBP	64.142±10.997	64.210±11.026	64.155±11.002	65.563±9.779	<0.001
	MBP	79.363±11.059	79.397±11.047	79.369±11.056	86.880±10.345	<0.001
	HR	85.242±14.982	85.059±14.945	85.205±14.974	84.967±16.089	0.007
	RR	19.311±3.824	19.279±3.906	19.305±3.841	19.072±3.990	<0.001
	BT (°F)	98.370±1.203	98.386±0.791	98.373±1.133	98.391±0.848	<0.001
SpO_2		96.825±2.013	96.792±2.013	96.818±2.013	98.038±2.504	<0.001
FiO_2		47.733±15.995	48.394±16.496	47.866±16.098	44.095±15.363	<0.001
GCS		13.440±2.806	13.432±2.810	13.439±2.807	13.224±2.971	<0.001
Laboratory	Albumin	3.187±0.703	3.200±0.694	3.189±0.701	3.137±0.507	<0.001
findings	Glucose	135.213±46.944	135.124±45.444	135.195±46.647	153.854±57.910	<0.001
	PT/INR	1.401±0.638	1.395±0.665	1.399±0.643	1.239±0.535	<0.001
	PaO_2	104.833±71.641	105.082±70.534	104.882±71.421	116.923±36.356	<0.001
	$PaCO_2$	41.399±10.255	41.330±10.246	41.385±10.253	37.940±6.114	<0.001
	BUN	26.673±22.035	26.976±22.570	26.734±22.143	22.748±16.263	<0.001
	Creatinine	1.443±1.615	1.426±1.640	1.439±1.620	1.204±1.389	<0.001

Data are represented as number(%) and median±standard deviation. LOS, length of stay; ICU, intensive care unit; MICU, medical ICU; SICU, surgical ICU; SBP, systolic blood pressure; DBP, diastolic blood pressure; MBP,

mean blood pressure; HR, heart rate; RR, respiratory rate; BT, body temperature; SpO_2, saturation of peripheral oxygen; FiO_2, fraction of inspired oxygen; GCS, Glasgow coma scale; PT/INR, prothrombin time/international normalized ratio; PaO_2, partial pressure of oxygen in arterial blood; $PaCO_2$, partial pressure of carbon dioxide in arterial blood; BUN, blood urea nitrogen.

Table 2. Summary of the performance of the models during internal validation.

Type	Sampling	Metric	AUROC	AUPRC	F1	F2	ACC	BA	ECE	PC
GBM	RUS	AUROC	0.903	0.349	0.376	0.374	0.974	0.680	0.001	Spline
GBM	RUS	AUPRC	0.904	0.349	0.380	0.370	0.975	0.676	0.002	Spline
GBM[1, 4)]	CW	AUROC	0.903	0.346	0.383	0.378	0.974	0.681	0.002	Spline
GBM	CW	AUPRC	0.900	0.345	0.379	0.375	0.974	0.680	0.002	Spline
GBM	ROS	AUROC	0.901	0.343	0.382	0.370	0.975	0.676	0.002	Spline
GBM	ROS	AUPRC	0.901	0.341	0.374	0.362	0.974	0.671	0.001	Spline
LSTM	RUS	AUROC	0.894	0.297	0.354	0.348	0.973	0.665	0.004	Isotonic
LSTM	RUS	AUPRC	0.894	0.297	0.354	0.348	0.973	0.665	0.004	Isotonic
LSTM	CW	AUROC	0.895	0.309	0.353	0.342	0.974	0.661	0.002	Isotonic
LSTM	CW	AUPRC	0.892	0.318	0.359	0.344	0.974	0.661	0.001	Spline
LSTM	ROS	AUROC	0.898	0.323	0.372	0.361	0.974	0.671	0.002	Isotonic
LSTM[2)]	ROS	AUPRC	0.894	0.327	0.375	0.373	0.973	0.679	0.002	Spline
TF	RUS	AUROC	0.895	0.332	0.363	0.350	0.974	0.665	0.001	Isotonic
TF	RUS	AUPRC	0.896	0.341	0.376	0.354	0.976	0.665	0.001	Spline
TF	CW	AUROC	0.895	0.335	0.372	0.354	0.975	0.666	0.003	Isotonic
TF[3)]	CW	AUPRC	0.898	0.340	0.377	0.360	0.975	0.669	0.002	Isotonic
TF	ROS	AUROC	0.897	0.340	0.375	0.354	0.976	0.665	0.002	Isotonic
TF	ROS	AUPRC	0.896	0.332	0.368	0.366	0.973	0.675	0.002	Isotonic

Thresholds that maximize F1 scores are used. 1) best of GBM models, 2) best of LSTM models, 3) best of Transformer models, 4) best of all models. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; F1, F1 score; F2, F2 score; ACC, accuracy; BA, balanced accuracy; ECE, expected calibration error; PC, probability calibration; GBM, gradient boosting machine; LSTM, long short-term memory; TF, transformer; RUS, random under-sampling; CW, class weight; ROS, random over-sampling; Spline, spline-based probability calibration; Isotonic, isotonic regression.

Table 3. Summary of internal versus external validation results.

Type	Threshold	Development cohort (testing dataset)								Validation cohort							
		AUROC	AUPRC	F1	F2	ACC	BA	ECE	Brier	AUROC	AUPRC	F1	F2	ACC	BA	ECE	Brier
GBM	0.171	0.903	0.346	0.383	0.378	0.974	0.681	0.002	0.026	0.933	0.181	0.202	0.341	0.955	0.793	0.026	0.045
LSTM	0.174	0.894	0.327	0.375	0.373	0.973	0.679	0.002	0.027	0.914	0.153	0.218	0.346	0.963	0.769	0.020	0.041
Transformer	0.180	0.898	0.340	0.377	0.360	0.975	0.669	0.002	0.025	0.930	0.174	0.220	0.354	0.962	0.780	0.021	0.038

Thresholds that maximize F1 scores are used. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; F1, F1 score; F2, F2 score; ACC, accuracy; BA, balanced accuracy; ECE, expected calibration error; Brier, Brier score; GBM, gradient boosting machine; LSTM, long short–term memory.

Table 4. Summary of threshold adjustment results.

Type	Method	Threshold	Development cohort (testing dataset)				Validation cohort			
			F1	F2	ACC	BA	F1	F2	ACC	BA
GBM	F1	0.171	0.383	0.378	0.974	0.681	0.202	0.341	0.955	0.793
	Youden	0.026	0.183	0.341	0.846	0.825	0.069	0.155	0.776	0.847
	F2	0.082	0.327	0.437	0.950	0.761	0.127	0.256	0.903	0.845
LSTM	F1	0.174	0.375	0.373	0.973	0.679	0.218	0.346	0.963	0.769
	Youden	0.021	0.176	0.331	0.840	0.819	0.069	0.154	0.783	0.836
	F2	0.075	0.316	0.428	0.948	0.758	0.125	0.250	0.908	0.822
TF	F1	0.180	0.377	0.360	0.975	0.669	0.220	0.354	0.962	0.780
	Youden	0.018	0.169	0.322	0.830	0.819	0.070	0.157	0.783	0.847
	F2	0.080	0.309	0.418	0.948	0.751	0.139	0.272	0.916	0.835

F1, F1 score; F2, F2 score; ACC, accuracy; BA, balanced accuracy; GBM, gradient boosting machine; LSTM, long short-term memory; TF, transformer.

Table 5. Subgroup analysis. GBM model predictions for the validation cohort.

		N	Death	AUROC	AUPRC	F1	F2	ACC	BA	ECE	Brier
Total		5,745	508 (8.8%)	0.933	0.181	0.202	0.341	0.955	0.793	0.026	0.045
Age	Young (<45)	823 (14.3%)	51 (6.2%)	0.942	0.111	0.199	0.305	0.976	0.724	0.011	0.024
	Middle age (45~65)	2,390 (41.6%)	190 (8.0%)	0.950	0.227	0.257	0.403	0.967	0.811	0.019	0.033
	Old (>65)	2,532 (44.1%)	267 (10.5%)	0.920	0.172	0.174	0.308	0.939	0.791	0.037	0.061
Sex	Male	3,347 (58.3%)	318 (9.5%)	0.934	0.190	0.210	0.353	0.956	0.801	0.026	0.044
	Female	2,398 (41.7%)	190 (7.9%)	0.932	0.168	0.189	0.322	0.955	0.781	0.027	0.045
ICU	MICU	1,290 (22.5%)	384 (30.0%)	0.885	0.189	0.208	0.351	0.894	0.773	0.048	0.106
	SICU	2,012 (35.0%)	57 (2.8%)	0.976	0.253	0.268	0.405	0.986	0.800	0.010	0.014
	Neuro SICU	2,443 (42.5%)	67 (2.7%)	0.907	0.054	0.101	0.190	0.987	0.725	0.018	0.013

The threshold that maximizes the F1 score is used (0.171). AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; F1, F1 score; F2, F2 score; ACC, accuracy; BA, balanced accuracy; ECE, expected calibration error; Brier, Brier score; GBM, gradient boosting machine; ICU, intensive care unit; MICU, medical ICU; SICU, surgical ICU.

Table 6. Comparing ICU mortality prediction studies.

Author	Input type	Input count	Input window	Outcome window	Cause of death	ICU type	Development cohort	Validation cohort	Model type	AUROC of IV	AUROC of EV
Brand [46]	VS	4	First 47h	In-hospital mortality	All	All	MIMIC-III	-	CNN	0.87	-
Deasy [47]	DG, VS, lab, chart	∞	First 48h	In-hospital mortality	All	All	MIMIC-III	-	LSTM	0.85	-
Thorsen-Meyer [48]	DG, VS, lab, chart	44	Up to the reference time point	90-day mortality	All	All	4 (Denmark)	1 (Denmark)	LSTM	0.88	0.83
Jiang [49]	VS, lab, SS	72	Up to the reference time point	After 2h	All	All	1 (China)	-	GBM	0.848	-
Baker [50]	DG, VS	9	24h	Within 3, 7, 14d	All	All	MIMIC-III	-	CNN + LSTM	0.858-0.884	-
Kim [32]	DG, VS	9	24h	Within 6h-60h	All	PICU	1 (S. Korea)	1 (S. Korea)	CNN	0.887-0.965	0.881-0.922
Yun [52]	DG, VS, lab, chart	43	Not specified	30-day mortality	All	SICU	1 (S. Korea)	-	Decision tree	0.96	-
Delahanty [35]	DG, VS, lab, chart	17	First 24h	In-hospital mortality	All	All	36 (USA)	17 (USA)	GBM	0.951	0.943
Xu [53]	DG, VS, lab	20	First 24h	In-hospital mortality	COVID-19	All	1 (USA)	1 (China)	Elastic net LR	0.72	0.85
Kang [55]	DG, VS, lab, chart, SS	67	First 48h	In-hospital mortality	All	All	eICU-CRD	MIMIC-III	LSTM + Attention	0.899	0.855
Ko [56]	VS, lab, chart	9	24h prior to admission	28-day mortality	Cancer	All	1 (S. Korea)	1 (S. Korea) + MIMIC-III	RF	0.939	0.753-0.775
This study	DG, VS	13	24h	Within 24h	All	MICU + SICU	MIMIC-IV	1 (S. Korea)	GBM	0.903	0.933

DG, demographics; VS, vital signs; lab, laboratory findings; SS, scoring systems; ICU, intensive care unit; PICU, pediatric ICU; SICU, surgical ICU; MICU, medical ICU; MIMIC, Medical Information Mart for Intensive Care; S. Korea, South Korea; USA, United States of America; eICU-CRD, eICU Collaborative Research Database; CNN, convolutional neural network; LSTM, long short-term memory; LR, logistic regression; RF, random forest; GBM, gradient boosting machine; AUROC, area under the receiver operating characteristic curve; IV, internal validation; EV, external validation.

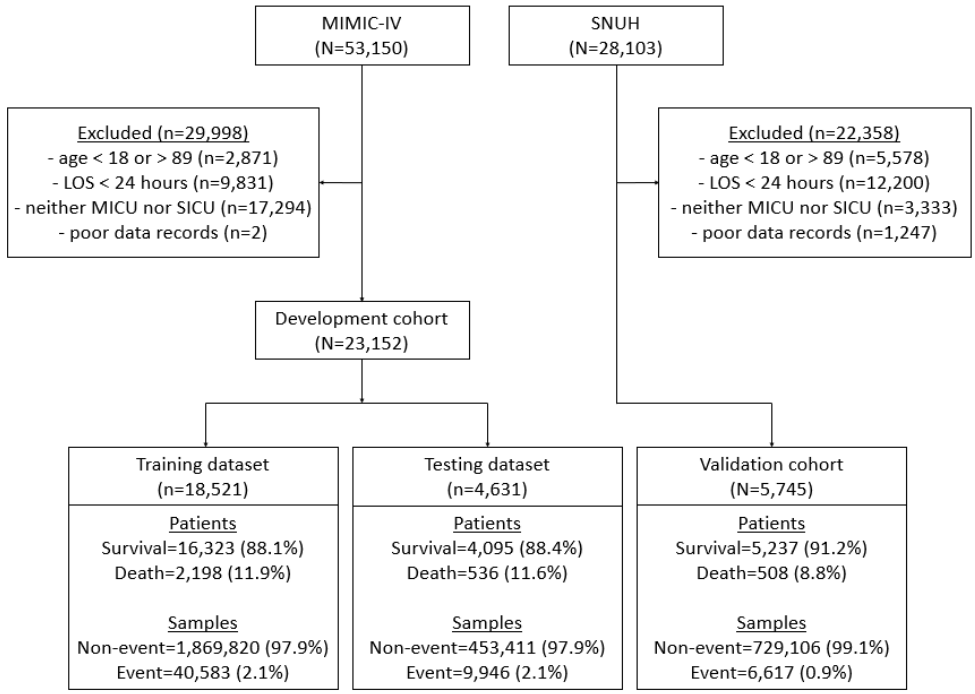
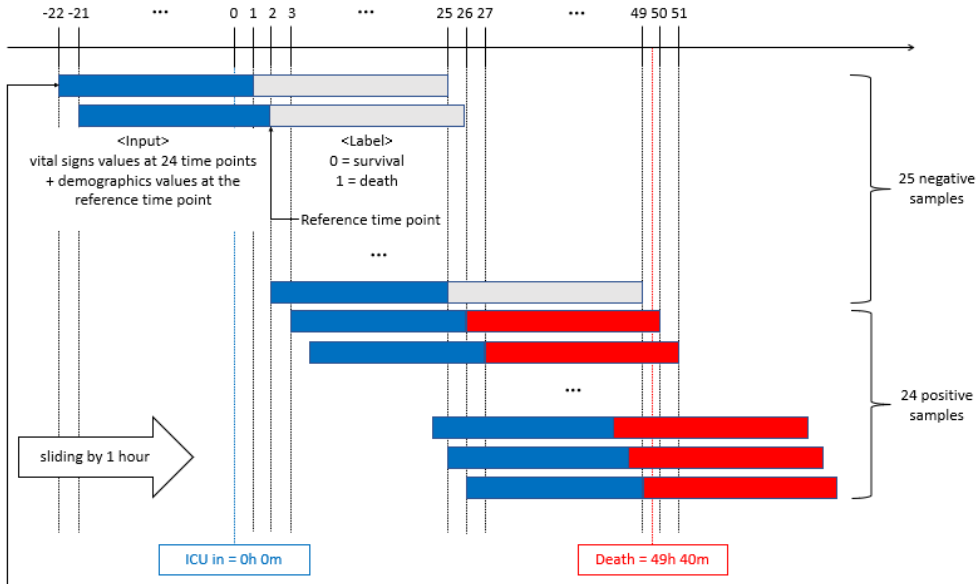


Fig 1. Flow chart presenting patient selection. MIMIC, Medical Information Mart for Intensive Care; LOS, length of stay; MICU, medical intensive care unit; SICU, surgical intensive care unit; SNUH, Seoul National University Hospital.



	Vital signs data			Demographics data	
Time point (hours)	SBP	GCS	GCS	Weight	Height
-72	128	N/A	15	58	**167 (V)**
...	...	N/A	N/A	N/A	N/A
-25	145	N/A	15	N/A	N/A
-24	140	N/A	N/A	N/A	N/A
-23	137	15	N/A	**56 (V)**	N/A
...	...	N/A	N/A	N/A	N/A
-2	131	**14 (V)**	N/A	N/A	N/A
-1	123	N/A	N/A	N/A	N/A
0	**125 (V)**	N/A	N/A	N/A	N/A
Value at a time point	**125**	**14**	**N/A**	**56**	**167**

Fig 2. Explain sliding time window sampling and how to get the value at each time point. ICU, intensive care unit; SBP, systolic blood pressure; GCS, Glasgow coma scale; N/A, not available.

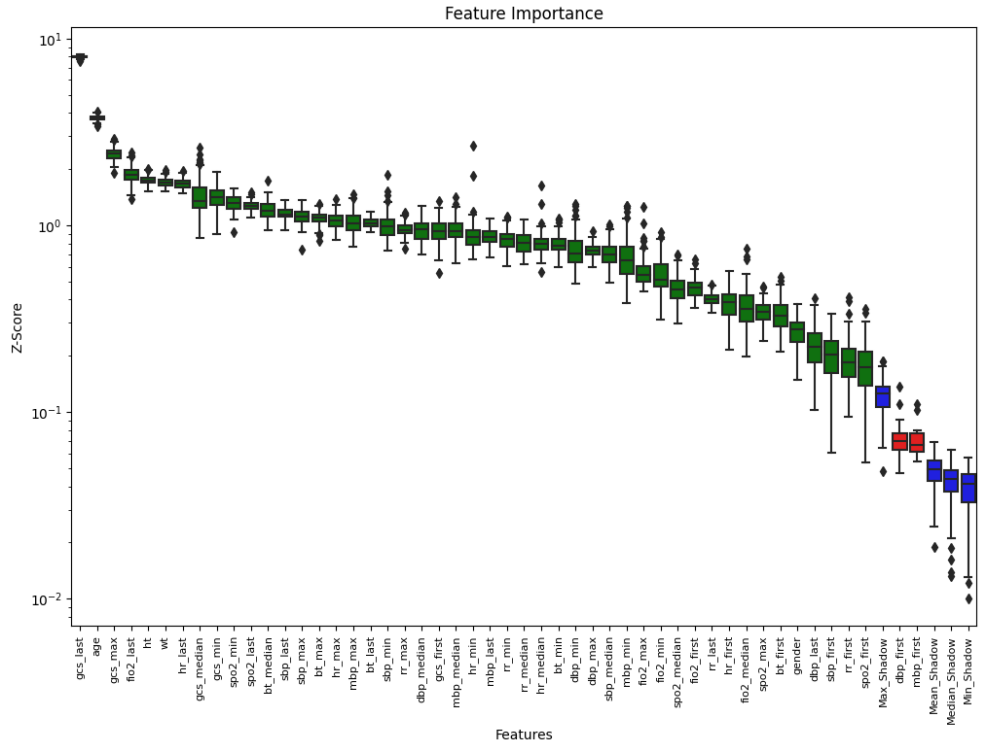


Fig 3. Feature Selection using Boruta SHAP. SHAP, Shapley additive explanation; SBP, systolic blood pressure; DBP, diastolic blood pressure; MBP, mean blood pressure; HR, heart rate; RR, respiratory rate; BT, body temperature; SpO_2, saturation of peripheral oxygen; FiO_2, fraction of inspired oxygen; GCS, Glasgow coma scale.

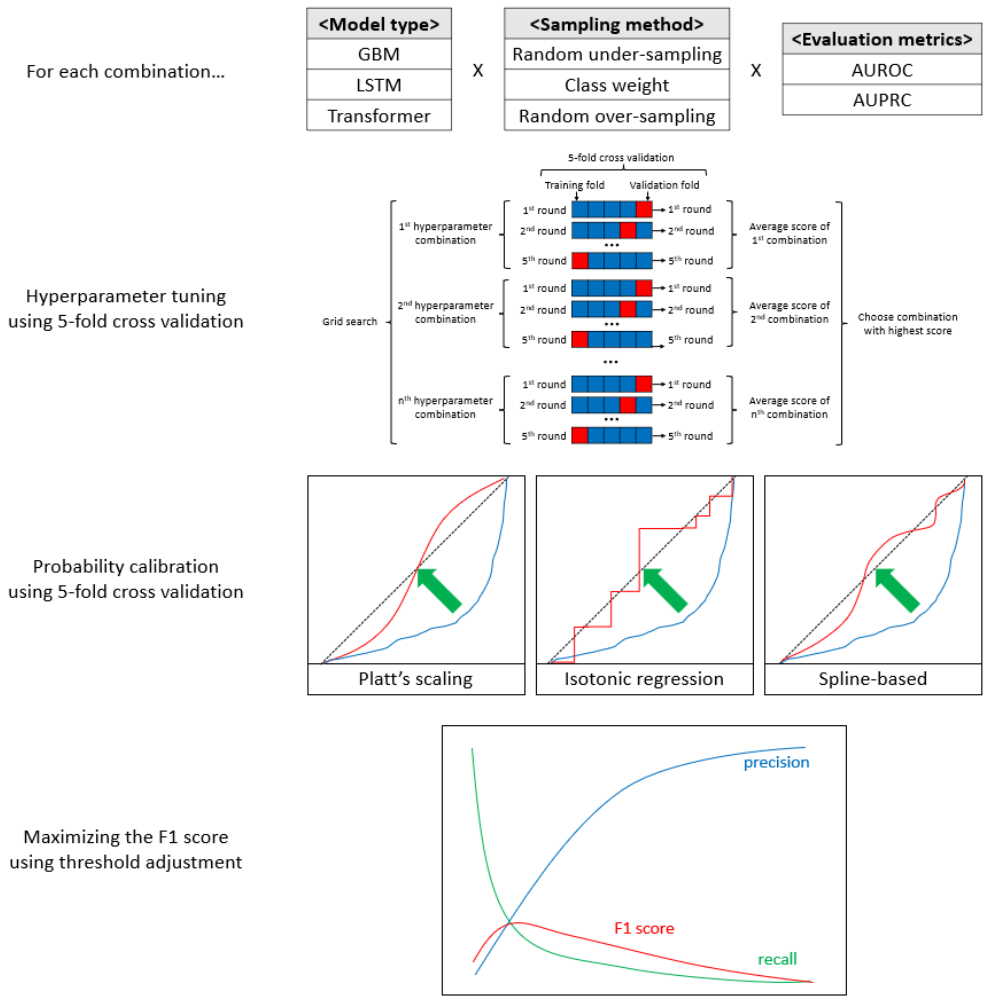


Fig 4. Development sequence for machine learning models. GBM, gradient boosting machine; LSTM, long short-term memory; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve.

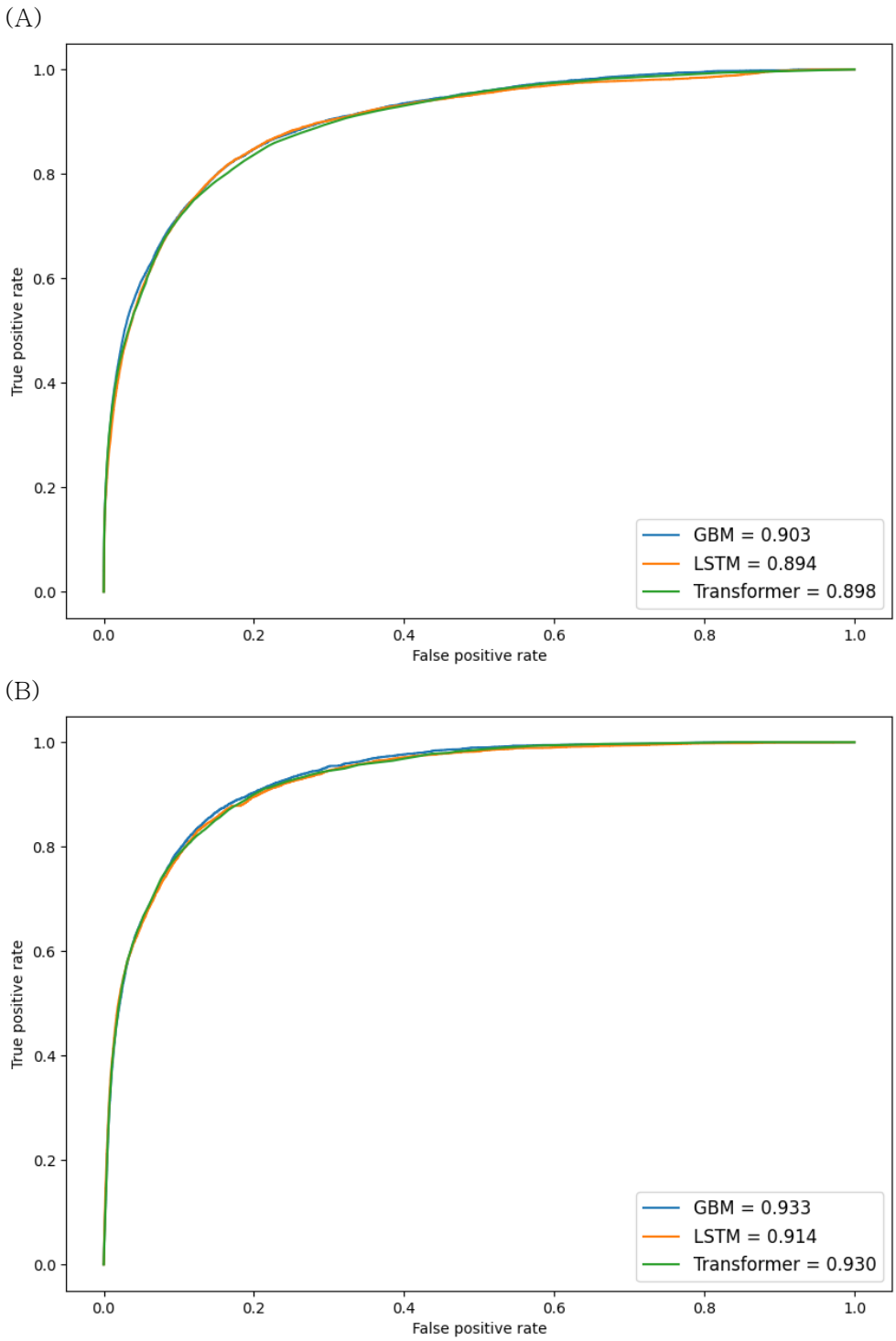


Fig 5. Comparison of area under the receiver operating characteristic curves of the prediction in (A) development cohort (B) validation cohort. GBM, gradient-boosting mode; LSTM, long short-term memory.

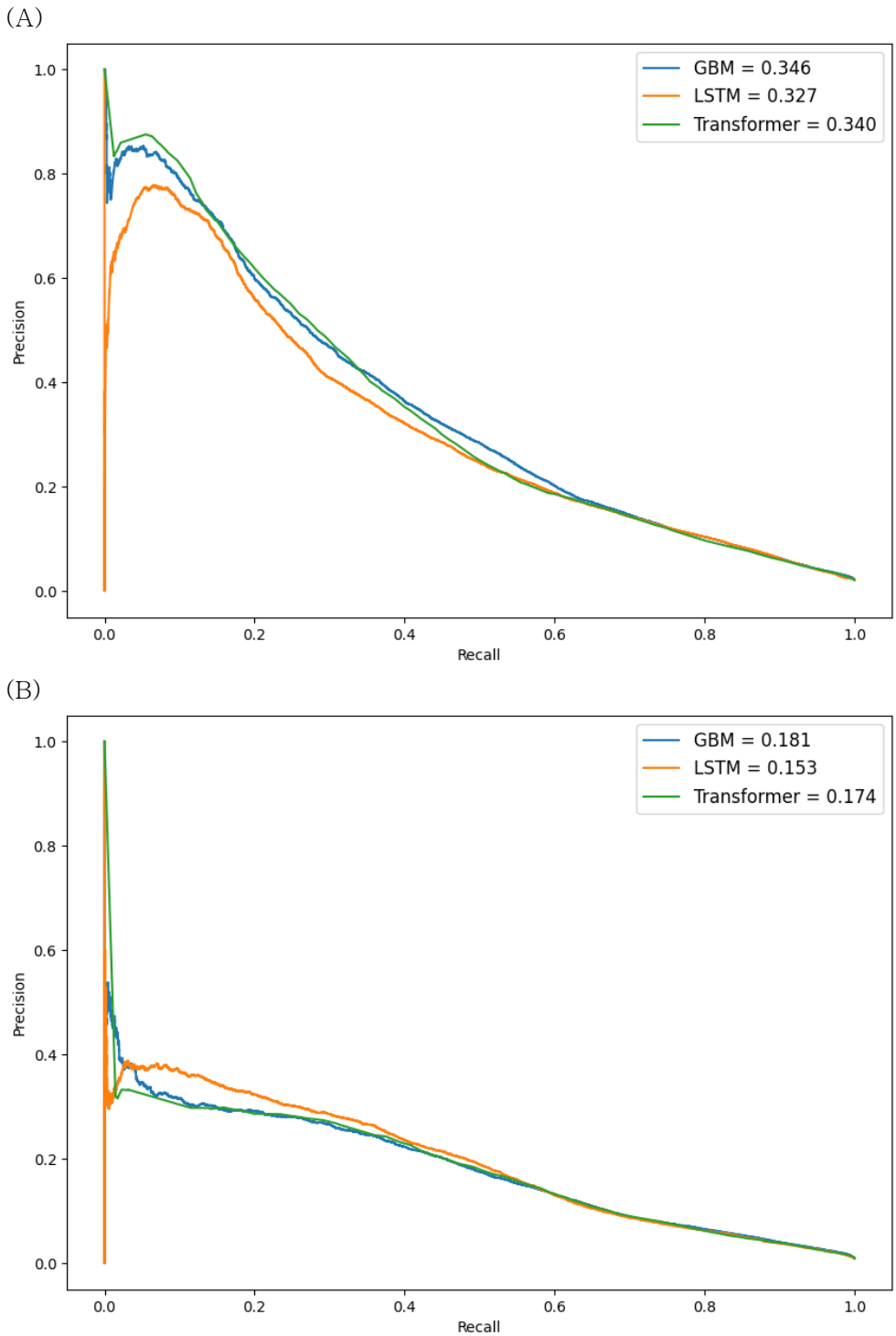
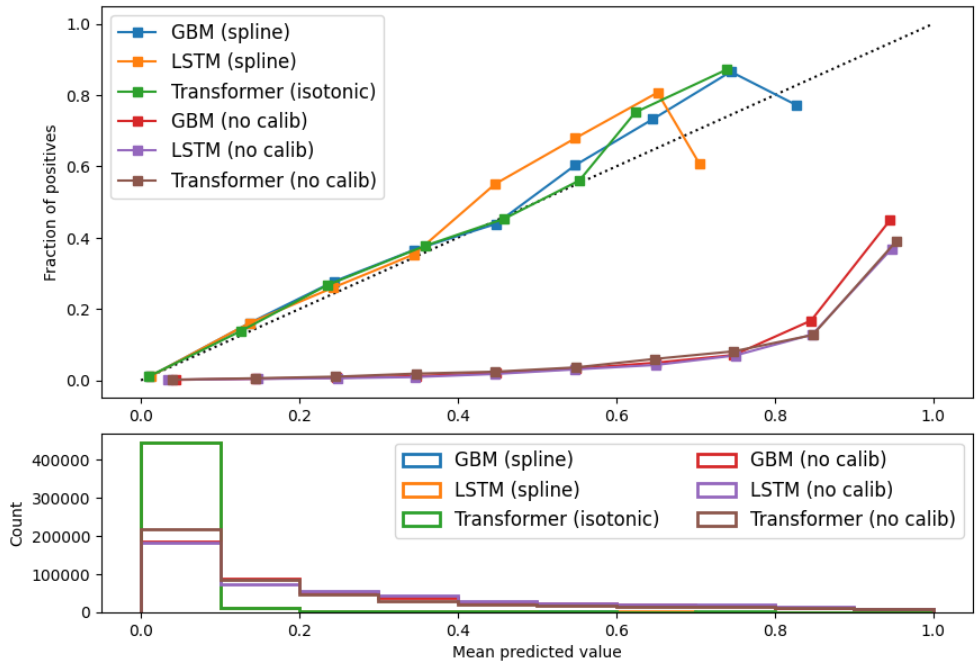


Fig 6. Comparison of area under the precision-recall curves of the prediction in (A) development cohort (B) validation cohort. GBM, gradient-boosting mode; LSTM, long short-term memory.

(A)



(B)

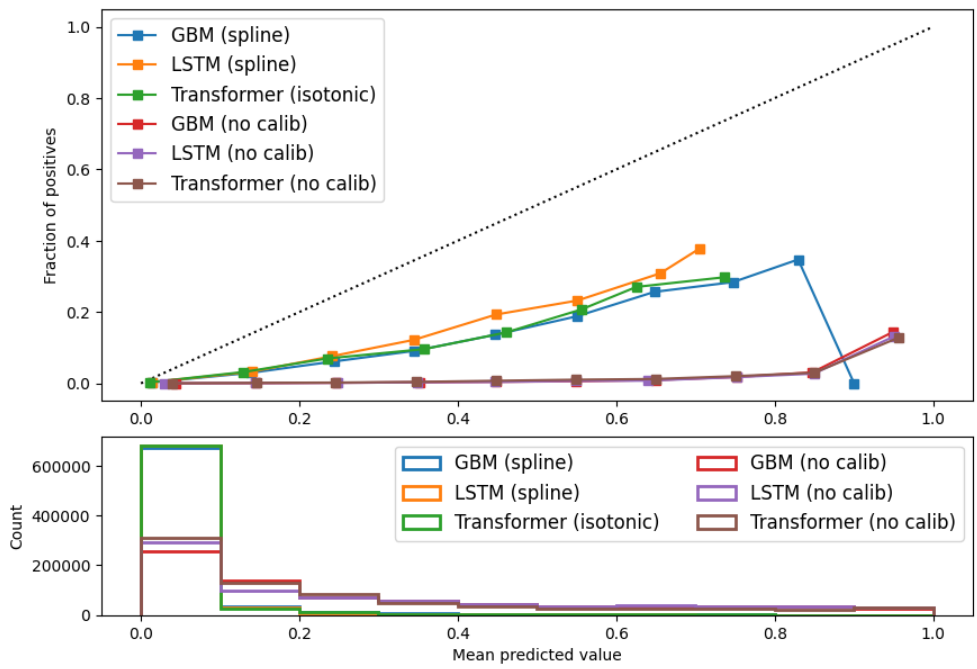


Fig 7. Comparison of calibration curve of the prediction in (A) development cohort (B) validation cohort. GBM, gradient-boosting mode; LSTM, long short-term memory.