

탐색적 데이터 분석과 기계학습을 통한 상부 요로감염 환자 Ciprofloxacin 항생제 내성 예측 연구

(Prediction of Antibiotic Resistance to Ciprofloxacin in
Patients with Upper Urinary Tract Infection through
Exploratory Data Analysis and Machine Learning)

이 종 범 ^{*} 이 현 규 ^{**}
 (Jongbub Lee) (Hyungyu Lee)

요약 응급의학과 의사는 요로감염 환자에 대해 임상적으로 감염 병원균 또는 항생제 내성 프로파일 이 확인되기 전에 항생제를 선택해야 하는 경험적 치료전략을 사용한다. 지역사회 의 요로 병원균의 항생제 내성 증가를 우려하였을 때 경험적 치료는 도전적인 과업이다. 본 연구는 단일 기관 후향적 연구로써, 응급실에서 상부 요로 감염 진단을 받은 환자를 대상으로 기계학습 알고리즘을 통한 항생제 내성 예측 방법을 제안한다. 먼저, 통계적 검정 방법과 게임 이론적 방식 기반의 SHAP(SHapley Additive exPlanation) 을 수행하여 유의미한 예측 변수를 선택한다. 그리고 4개의 분류기의 성능을 비교하고 예측 확률 임계치 조절을 통해 의사의 경험적 치료를 보조할 수 있는 알고리즘을 제안한다. 결과적으로, SHAP를 통해 선별된 전체 예측 변수의 65%만을 사용한 SVM 분류기는 실험에 사용된 모든 분류기 중에 AUROC 0.775로 가장 높은 성능을 보였고, 모든 예측 변수를 사용했을 때보다 AUROC 0.015 증가했다. 그리고 예측 확률 임계치를 조절 과정을 통해 의사의 경험적 치료의 민감도를 98% 수준으로 유지하면서 특이도가 3.9배 향상된 분류 정확도를 달성했다.

키워드: 기계학습, 특징 선택, 항생제 내성, 설명 가능한 인공지능

Abstract Emergency medicine physicians use an empirical treatment strategy to select antibiotics before clinically confirming an antibiotic resistance profile for a patient with a urinary tract infection. Empirical treatment is a challenging task in the context of concern for increased antibiotic resistance of urinary tract pathogens in the community. As a single-institution retrospective study, this study proposed a method for predicting antibiotic resistance using a machine learning algorithm for patients diagnosed with upper urinary tract infection in the emergency department. First, we selected significant predictors using statistical test methods and a game theory based SHAP (SHapley Additive exPlanation), respectively. Next, we compared four classifier performances and proposed an algorithm to assist decision-making in empirical treatment by adjusting the prediction probability threshold. As

· 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1F1A1071574)과 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2022-00155915, 인공지능융합 혁신인재양성(인하대학교)과 인하대학교의 지원을 받아 수행된 연구임.

^{*} 비 회 원 : 인하대학교 컴퓨터정보공학과 학생
 22231241@inha.edu

^{**} 정 회 원 : 인하대학교 의예과 교수(Inha Univ.)
 hglee@inha.ac.kr
 (Corresponding author임)

논문접수 : 2022년 11월 2일
 (Received 2 November 2022)
 논문수정 : 2022년 12월 22일
 (Revised 22 December 2022)
 심사완료 : 2023년 1월 3일
 (Accepted 3 January 2023)

Copyright©2023 한국정보과학회: 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
 정보과학회논문지 제50권 제3호(2023. 3)

a result, the SVM classifier using predictors selected through SHAP (65% of the total) showed the highest AUROC (0.775) among all conditions used in the experiment. By adjusting the predictive probability threshold in the SVM, we achieved classification accuracy with a specificity that was 3.9 times higher than empirical treatment while preserving the sensitivity of the doctor's empirical treatment at 98%.

Keywords: machine learning, feature selection, antibiotic resistance, explainable artificial intelligence

1. 서론

항생제 내성 증가는 최근 세계보건기구가 코로나 이후 최대 보건 위기가 될 것이라 경고한바 있을 만큼 중요한 문제다. 박테리아 감염을 치료하는 데 사용하는 약물인 항생제의 사용 빈도 증가는 해당 항생제에 대한 박테리아의 내성균 증가로 이어질 수 있으며, 내성 증가는 장기 입원 환자 및 의료비 증가뿐만 아니라 나아가 환자의 사망을 초래할 수 있다[1]. 또한 내성균 보균자는 균을 전파할 수도 있으며, 특히 약제 내성균의 경우 의료환경에서 심각한 문제를 발생시킬 수 있다[2].

항생제를 사용하는 많은 감염 중 특히 상부 요로 감염은 요로계에 세균이 침입하여 염증성 반응을 유발하는 것으로, 응급실에서 흔히 볼 수 있는 질병이다. 감수성은 한가지 이상의 항균제 존재 시에 미생물이 성장할 수 없는 상태를 서술하는 용어로 박테리아가 항생제에 민감도를 측정하는 감수성 검사는 3~5일이 소요되어 긴급한 상황에서는 사용할 수 없다. 따라서, 응급실에서는 내성 프로파일이 확인되기 전에 의사들이 경험적으로 감수성이 높을 것으로 예상되는 항생제를 선택해야 한다[3]. 하지만, 최근 국내외 항생제 오남용과 그로 인한 항생제의 내성 증가로 인해 경험적 치료를 위한 항생제 선택은 응급 의사에게 임상적 도전과제이다[4]. 특히 항생제 ciprofloxacin에 대한 내성률은 개발도상국에서 55.5~85.5%, 선진국에서 5.1~32.0%를 보인다[5].

이를 극복하기 위해 전자의무기록 데이터로부터 환자의 항생제 종류별 내성 및 감수성을 예측하기 위한 다양한 연구들이 시도되고 있다. 기계학습을 통해 캄보디아 어린이 병원 243명 환자의 항생제 내성 예측연구가 시도되었다[6]. 혈액 배양검사 데이터, 인구 통계학적 자료와 일상생활 환경 정보를 예측 변수로 선정하였고, 랜덤 포레스트, 에이다 부스트, 서포트 벡터 머신, 로지스틱 회귀 등의 분류기를 사용하였으며, 항생제 내성 예측을 위한 주요 특징 등을 보고했다. 하지만, 전자의무기록 외의 가정환경에 관한 설문이 포함되어 캄보디아 외의 국가 및 응급실 환경에 적용하기에 한계가 있다. 또 다른 연구로, 임상 미생물학 연구실 정보 시스템(LIS)의 간단한 인구 통계학적 자료, 그람 염색 및 44종의 항균 물질 검사 자료를 Microsoft Azure AutoML을 이용해

항생제 내성을 예측하였으며, 기계학습에 많은 도메인 지식 없이도 의사결정을 돕는 모형을 제안했다[7]. 하지만, 같은 환자의 중복된 데이터들을 훈련과 평가 데이터에 포함하여 정확한 성능 파악이 어렵고 전자의무기록 자료 외의 긴 시간 소요가 필요한 검사 결과를 포함하여 응급실에서 긴급하게 사용하기 어렵다.

그 외에도 의료데이터의 디지털화로 인해, 임상데이터에 관한 기계학습 연구가 활발하게 시도되고 있다. 특히 전자의무기록 분석은 데이터 수집의 용이성으로 인해 질병 분류, 사망 여부 예측, 당뇨병 저혈압 예측 등과 같이 환자의 의무기록으로부터 위험 질병군을 분류하거나 치료 경과를 예측하는 연구에 다양하게 이용되고 있다[8]. 그 사례로, 중환자의 상태 확인 및 사망을 예측하기 위해 활력 징후를 시각화하고 기계학습을 활용한 방법이 연구되었고, 저혈당 발생을 예측하기 위해 혈당 자료와 랜덤 포레스트 분류기를 이용하는 예측 방법이 연구되었다[9,10].

따라서, 본 논문에서는 상부 요로감염 환자를 대상으로 국내 응급실에서 적용할 수 있는 ciprofloxacin 항생제에 대한 내성 예측 기계학습 알고리즘을 제안한다. 또한, 긴급하게 결정을 내려야 하는 응급실에서는 데이터 수집 과정이 효율적이어야 하지만, 특정 의료 데이터의 수집은 큰 비경제적 비용을 요구한다. 예를 들어, 사생활 보호를 위한 행정 처리와 검사 방법에 따른 시간 및 노력을 요구한다. 그러므로 특징 선택을 통해 감수성과 관련한 특징 선택을 통해 성능을 높이고 실효성을 높이는 방법을 제안한다[11]. 통계적 방법과 게임 이론적 방식 기반의 특징 선택 방법을 각각 적용하여 성능을 비교하였으며, 상부 요로감염에 대한 항생제 내성 예측에 필요한 예측 변수를 제안한다. 또한, 다양한 분류기를 비교하여 특징 선택으로 감소한 차원에 적합한 분류기를 비교 분석하고, 기존의 경험적 치료의 효과를 높일 수 있도록 예측 임계치를 조정한다.

본 연구는 다음과 같은 장점이 있다. 첫째, 긴급하게 결정을 내려야 하는 응급실에서 효용성을 높일 수 있는 특징 선택 방법들을 비교 분석하여 유효한 예측변수를 제안한다. 둘째, 환자의 전자의무기록 데이터에서 선택된 특징들과 여러 기계학습 알고리즘 조합의 성능을 비교하여 응급실 환경의 희소 데이터에 적합한 분류기를

제안한다. 셋째, 상부 요로감염 환자 항생제 치료를 위한 진단 보조 연구로, 임계치 조절 과정을 통해 기존의 경험적 항생제 선택의 효과를 높일 가능성을 제시한다.

2. 제안 방법

본 논문에서 제안하는 방법은 그림 1로, (i) 탐색적 데이터 분석을 통한 특징 추출, (ii) 분류기 학습, (iii) 임계치 조절의 세 단계로 구성된다. 먼저 탐색적 데이터 분석 방법에서는 크게 통계적 분석 방법과 게임 이론적 방식 기반의 분석 방법으로 구성되며, 이 과정을 통해 유의한 예측변수를 추출하게 된다. 두 번째로 분류기 학습에서는 최소한 의료 데이터에 적합한 기계학습 모델을 찾기 위해, 서로 다른 특성을 가진 로지스틱 회귀분석, 그래디언트 부스트, 다층 퍼셉트론과 선형 SVM을 선택된 변수들과 학습시키고 성능을 비교 분석한다. 마지막으로, 임계치 조절 과정에서 각 모델의 결과로 나온 항생제 내성 예측 확률의 임계치를 조절한다. 본 논문에서 감수성을 예측하는 ciprofloxacin 항생제는 높은 내성률을 보이기에, 치료 실패를 방지하기 위해 더 높은 수준의 민감도가 필요하다. 일반적으로 이진 분류에 사용하는 확률 임계치 0.5가 아닌 조절된 확률 임계치를 사용하여, 민감도를 경험적 치료 방법과 같은 수준으로 유지하여 결과를 평가한다.

의료 데이터는 특성상 같은 환자에서 추출된 많은 중복된 데이터와 결측값을 포함하며, 특히 병과 관련 없이 측정된 특징들을 가지고 있다[12]. 이로 인해, 적은 샘플 수에 비해 불필요한 특징 차원이 커져 기계학습 모형에

과적합을 발생시킬 수 있다. 또한, 차원의 저주(Curse of dimensionality) 개념에 의하면 입력 특징 수가 커짐에 따라 필요한 데이터 수는 기하급수적으로 증가하게 된다[13]. 따라서, 기계학습을 활용한 의학적 예측에서는 더욱 연관성이 높은 특징을 선택하여 불필요한 예측변수로부터 모형의 예측 성능 및 일반화 성능을 높일 수 있다[11,12]. 본 논문에서는 통계적 방법과 기계학습의 출력을 설명하기 위한 게임 이론적 접근방법을 통해 특징을 선택한다. 먼저, 통계적 방법으로는 Kolmogorov-Smirnov 검정[14]을 통해 수치형 자료들이 감수성 여부와 무관하게 분포가 동일하지, 카이제곱 독립성 검정[15]을 통해 범주형 자료들이 감수성 여부와 독립인지 분석한다. 검정 결과, 동일성과 독립성이 기각된다면 특징으로 채택하고 p-value를 변수 중요도로 활용한다. 게임 이론적 방식을 활용한 탐색적 데이터 분석은 SHAP(Shapley additive explanation)[16]로부터 모형의 예측값에 대한 각 특징의 기여도를 활용한다.

2.1 이표본 KS(Two Sample Kolmogorov-Smirnov) 검정을 통한 수치형 특징 선택

수치형 자료에서 특징을 선택하기 위해, 예측 변수의 자료가 ciprofloxacin 항생제에 감수성일 때의 분포와 내성일 때의 분포가 다른지 유의한지 검정한다. 하지만, 해당 변수의 분포를 가정할 수 없기에 분포무관 검정법을 사용한다. 이표본 KS 검정은 Smirnov(1939)에 의해 제안된 검정법으로 두 표본 간 분포의 동일성을 검정하는 데에 가장 널리 사용되는 분포 무관 검정이다. 이표본 KS 검정에서 통계량은 두 집단의 경험적 누적

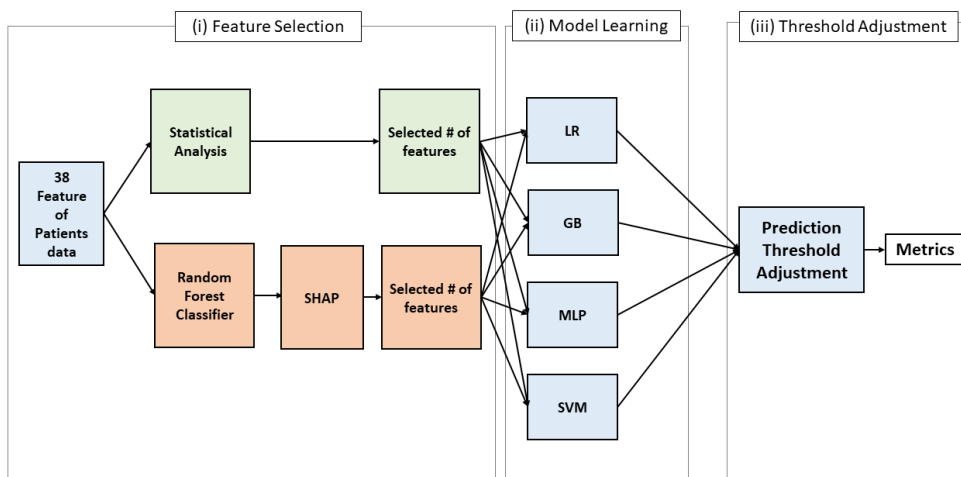


그림 1 특징 선택과 알고리즘 평가 과정

Fig. 1 Procedure of feature selection and assessment of algorithms. LR: Logistic Regression, GB: Gradient Boost, MLP: Multi-Layer Perceptron, SVM: Linear Support Vector Machine, Prediction Threshold Adjustment

분포함수의 차이를 통해 계산된다[17]. 본 연구에 활용하기 위해서는, 수치형 예측 변수 X 를 환자의 감수성 여부에 따라 확률 변수 X_{Sus}, X_{Imm} 로 구분하고 이에 대한 경험적 누적분포함수를 계산한다. Y 는 종속변수로 0은 감수성을, 1은 내성을 의미할 때 X_{Sus} 와 X_{Imm} 의 경험적누적분포함수 $F_n(x)$ 와 $G_m(x)$ 는 다음 식 (1)과 같다. n 과 m 은 각각 내성, 감수성 환자 표본의 수이다.

$$X_i = \begin{cases} X_{Imm_i}, & Y_i = 1, \\ X_{Sus_i}, & Y_i = 0, \end{cases}$$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_{Imm_i} \leq x), \quad i = 1, \dots, n \quad (1)$$

$$G_m(x) = \frac{1}{m} \sum_{j=1}^m I(X_{Sus_j} \leq x), \quad j = 1, \dots, m$$

두 경험적누적분포함수 $F(x)$ 와 $G(x)$ 서로 동일한지 확인하는 가설 검정을 위한 검정 통계량은 다음 식 (2)와 같다.

$$D_{m,n} = \sup_x (|F_n(x) - G_m(x)|), \quad (2)$$

검정 통계량 $D_{m,n}$ 은 경험적누적분포 F_n 과 G_m 의 차의 절댓값 중 최댓값을 의미한다[18]. 검정 통계량이 계산되면, 유의 확률 p-value를 계산한다. p-value가 유의 수준 0.05 이하이면 귀무가설을 기각하고 해당 수치형 변수를 선택한다. 또한 p-value는 오름차순으로 수치형 변수 간에 변수 중요도로 활용한다.

2.2 카이제곱 독립성 검정을 통한 범주형 특징 선택

두 범주형 변수가 독립인지 검정하는 카이제곱 독립성 검정은 최초의 독립성 검정이자 가장 많이 사용되는 검정이다. 카이제곱 독립성 검정은 귀무가설을 만족하는 기대빈도를 계산하고, 관찰빈도와 기대빈도의 차이가 클 경우 기각하는 원리를 가진다[19]. 본 연구에서는 성별, 침상 환자 여부 등의 범주형 변수와 감수성 여부의 독립성을 검정한다. KS 검정과 마찬가지로 p-value가 유의 수준 0.05 이하이면, 두 변수는 독립이라는 귀무가설을 기각하고 해당 범주형 변수를 선택한다. 또한 p-value는 오름차순으로 범주형 특징 간에 특징 중요도로 활용한다.

2.3 SHAP를 이용한 특징 선택

게임이론적 방식의 탐색적 데이터 분석은 입력 특징의 중요도를 활용해 분석을 수행한다. 중요도를 산정하는 기준 중 SHAP(Shapley Additive exPlanation) 값이 가장 신뢰도가 높다고 알려져 있다[13]. SHAP 값은 표본마다 계산될 수 있는 값으로, 입력 특징 중요도를 계산하기 위해 여러 특성의 조합을 구성하여 해당 특성의 유무에 따른 평균적인 변화를 통해 계산된다. 특징 중요도를 계산하기 위해서는 식 (3)과 같이 특징별 표본들의 SHAP 값을 절댓값 연산 후 평균을 계산한다.

$$I = \frac{1}{n} \sum_{i=1}^n |\phi_{(i)}|, \quad (3)$$

$\phi_{(i)}$ 은 각 표본 별 SHAP 값이고 n 은 표본의 수이다 [20].

하지만, 비교 모형마다 SHAP를 활용하여 입력 특징을 선정한다면, 모형에 따라 입력 특징의 중요도는 달라져 해석이 어려워지고 동등한 평가를 진행할 수 없다. 분류기들 중 랜덤 포레스트는 배깅(Bagging)을 활용하여 예측 모델의 분산이 감소하고, 학습 데이터가 충분하지 않아도 학습 효과를 얻을 수 있다. 랜덤 포레스트를 활용해 특징 중요도를 산정하면 안정적으로 특징을 선택하고, 과적합에 의한 결과를 방지한다[21]. 따라서, 본 연구에서는 랜덤 포레스트 SHAP 결과로 특징 중요도를 계산하고, 중요도를 내림차순으로 정렬하여 특징을 선택 후 모형 학습에 사용한다.

3. 실험

3.1 실험 환경

3.1.1 데이터 셋

해당 데이터 셋은 인하대학교 병원 응급의학과에서 2020년부터 2021년 6월까지 수집된 상부 요로감염 환자의 ciprofloxacin 항생제 감수성 자료이다. 입력 특징은 총 38개로, 18개의 범주형 자료와 20개의 수치형 자료로 구성되며, 결측치 82건을 제외한 총 468 데이터 중 감수성 환자 자료는 233건, 내성 환자 자료는 235건이다. 입력 특징들은 임상적 인구 통계학적 데이터로 연령, 성별, 동반 질환, 수술 및 입원 여부, 요로 감염력 등이 있다. 그 중 침상 환자 여부, 최근 수술 및 입원 여부, 1년 내 요로 감염력, 간호 여부 등이 주요한 특징으로 알려져 있다[2,3]. 결측치 82건 중 73건은 동맥혈 가스 측정에서 나온 자료로, 타 특징들과의 상관계수가 모두 0.2 이하로 보간을 수행하기 어려워 결측치들을 제외하였으며, 나머지 9건의 결측치들도 제외했다.

3.1.2 분류기 구성 및 학습 환경

사용된 알고리즘은 로지스틱 회귀분석, 그래디언트 부스트, 다중 퍼셉트론과 선형 SVM 총 4가지다. SHAP를 이용한 특징 추출 시 랜덤 포레스트에 적합한 변수가 추출이 될 수 있기에, 통계적 방법과 게임이론적 방법 모두 모형에서 제외한다. 훈련, 테스트 데이터 분할 비율은 8:2이며, 공통으로 수치형 데이터들은 표준화를 통해서 각 예측 변수의 스케일이 조절된다. 이때는 표본 평균이고는 표준편차를 의미한다. 각 모형은 5겹 교차검증으로 하이퍼 파라미터 튜닝을 수행한다. 다중 퍼셉트론은 은닉 층 수, 노드 개수를, SVM은 규제항을 탐색한다. 로지스틱 회귀분석은 별도의 정규화 없이 최

대가능도 추정으로 학습을 수행한다. 각 모형 별 종합적인 성능을 평가하기 전, 모형의 특징 수를 달리하여 최적화하는 실험을 수행한다. 식 (4)는 실험을 수행한 모형과 변수의 가능한 조합들을 나타낸다.

$$M_s = \{LR, GB, MLP, SVM\}, s = \{STAT, SHAP\},$$

$$n = \begin{cases} 10, 15, & \dots & \text{if } s = STAT \\ 10, 15, 17, 20, 25, 30, & \text{if } s = SHAP \end{cases} \quad (4)$$

M 은 모형의 종류, 첨자 s 는 변수 선택을 위한 탐색적 데이터 분석의 종류, n 은 사용한 변수의 수를 의미한다. $s = STAT$ 에서 n 의 최댓값은 통계적으로 선택된 특징의 수이다.

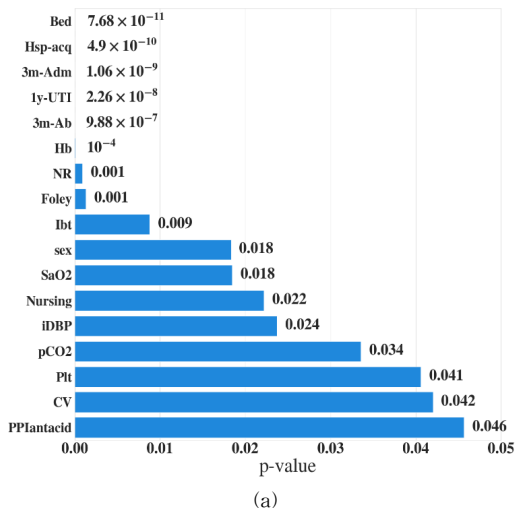
3.1.3 모형 성능 평가 지표

모형들의 종합적인 성능을 비교하기 위해, ROC(Receiver Operating Characteristic) 곡선의 영역으로 계산되는 AUROC(Area Under ROC Curve) 값을 사용했다. 그 외에도 민감도(sensitivity), 정밀도(precision), F1 Score, 특이도(specificity)를 사용해 모형을 비교한다. 수식은 (5)와 같다.

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

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$F1 \text{ score} = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$$



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

3.2 선별된 예측 특징

3.2.1 통계적 방법에 의해 선택된 특징

그림 2의 (a)은 통계적 검정 결과 유의 확률 0.05로 기각된 특징들이며, p-value를 특징 중요도로 오름차순한 결과이다. 총 38개 특징 중 17개 특징이 선택되었으며, 범주형 자료는 11개 선택되어 더 주요한 특징임을 알 수 있다. 알려진 주요 변수인 침상 환자 여부(Bed), 최근 수술 입원 여부(HSP-acq), 3개월 내 입원 여부(3m-Adm), 요로 감염력(1y-UTI) 순으로 높은 중요도를 보였으며, 추가로 헤모글로빈 농도(Hb), 신경퇴행성 질환 여부(NR), 침습적 측정 체온(iBT), 침습적 측정 이완기 혈압(iDBP) 등이 선택된 것을 관찰할 수 있다. 그림 2의 (b)은 선별된 예측 변수들의 값이 항생제 감수성에 미치는 영향을 확인하기 위해, 다중상응분석(MCA)을 활용해 범주형 자료 간에 관계를 요약한 도표이다. 변수들의 수준이 가까이 위치할수록 유사한 패턴을 보인다. 주요한 3가지 변수 Bed, Hsp-acq, 3m-Adm를 확인한 결과, X축 기준 좌측의 범주들은 침상 환자가 아니고, 최근 1달 및 3달 내 입원 이력이 있고, 1년 내 요로 감염력과 유치 도뇨관 사용경험이 있는 환자의 수준이 있다. 우측은 그에 반대되는 환자들의 범주 수준이 있다. 즉, 범주들의 수준이 좌측은 감수성에, 우측은

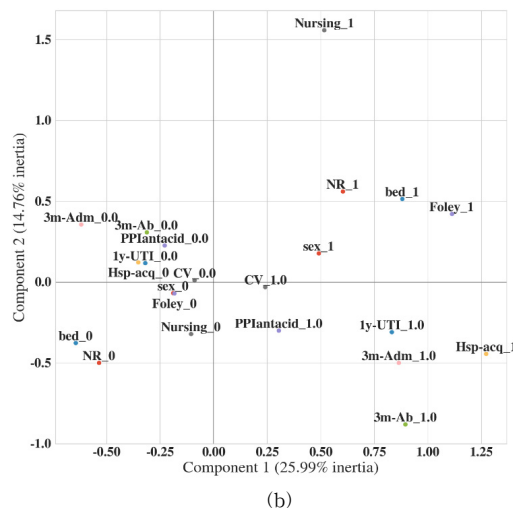


그림 2 통계적 방법에 의한 탐색적 데이터 분석. (a) 통계적 검정에 의해 선택된 입력 특징들, (b) 범주형 자료 MCA(다중 상응 분석), 각 점의 명칭은 앞부분이 변수명, 뒷부분이 수준을 뜻한다. MCA 도표의 X축과 Y축은 제1,2 주성분을 의미한다. 괄호 안의 관성(inertia)은 주성분이 기존 차원을 설명하는 정도의 비율

Fig. 2 Exploratory data analysis by statistical methods. (a) Input features selected by statistical tests, (b) MCA (Multiple Correspondence Analysis) of categorical data, with the first part of the name of each dot meaning variable name and the second part meaning the level

내성에 군집화 되었다. 이를 통해, 예측 변수들의 값이 감수성에 미치는 영향을 확인하였다.

3.2.2 게임이론적 방법에 의해 선택된 특징

그림 3는 모든 특징을 사용하여 학습한 모형에서

SHAP를 활용하여 식 3의 특징 중요도로 정렬한 자료이다. 통계적 검정과 달리 기각 역이 없기에, 20개 특징을 선택하여 나타낸다. 그 중, 그림 3의 (a)는 랜덤 포레스트를 활용하여 특징 중요도를 확인한 자료로, 17개

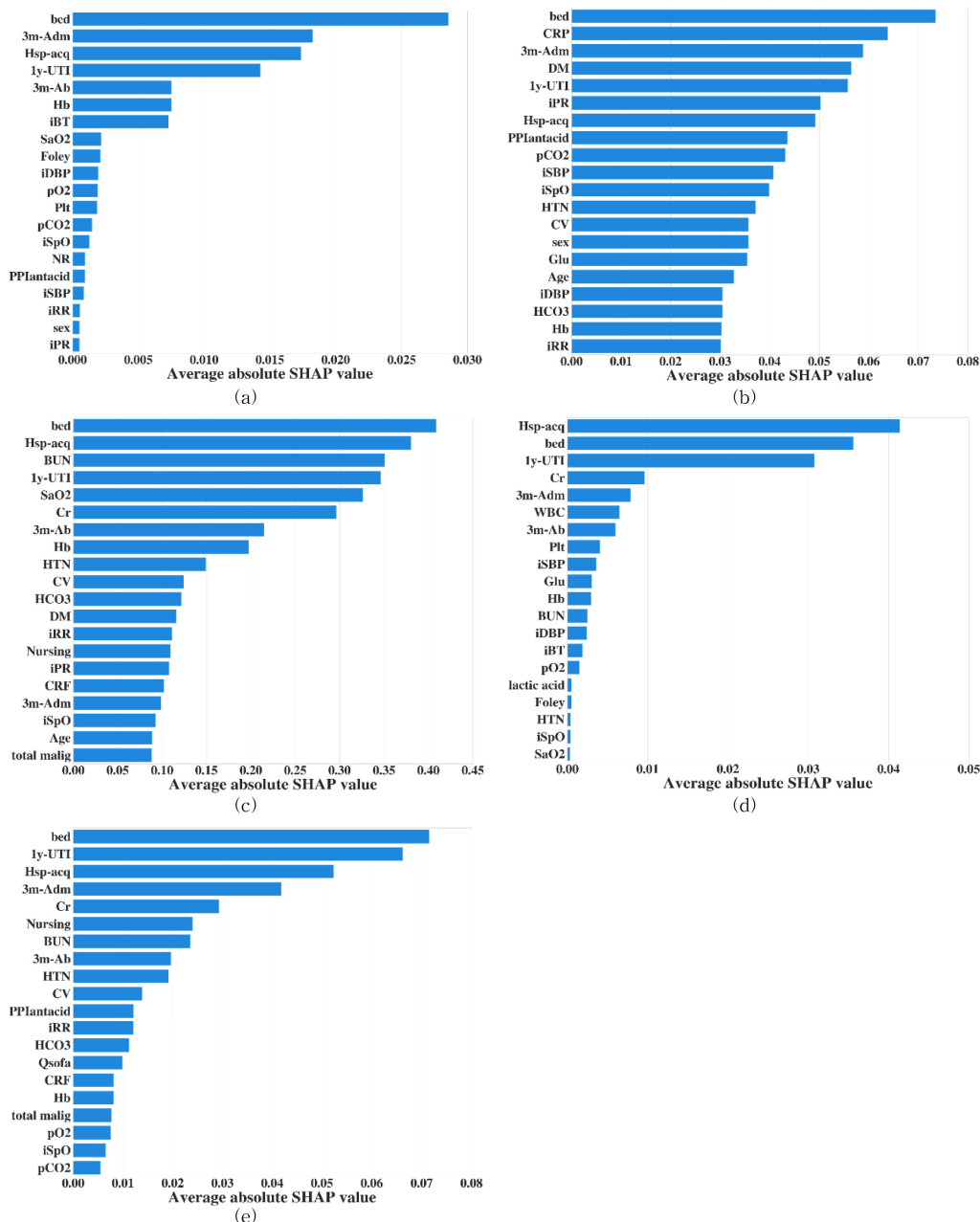


그림 3 모형 별 SHAP 특징 중요도. (a) 랜덤 포레스트, (b) 다층 퍼셉트론, (c) 로지스틱 회귀분석, (d) 그래디언트 부스트, (e) Linear Support Vector Machine

Fig. 3 SHAP feature importance for each model. (a) Random Forest, (b) MLP, (c) Logistic regression, (d) Gradient Boost, (e) Linear Support Vector Machine

의 특징만을 통계적 방법 결과와 비교하면, 알려진 주요 특징을 포함하여 상위 6개의 특징은 공통으로 선택되었다. 하지만, 요양원 및 가정간호 여부(Nursing), 혈압제외 심장 질환(CV) 등의 범주형 자료가 제외되고 산소분압(pO₂), 침습적 측정 산소포화도(iSpO₂), 침습적 측정 혈압(iSBP)이 추가되어 수치형 특징이 총 11개로 통계적 방법에 비해 더 많이 선택되었다. 이는 통계적 검정과 달리 랜덤 포레스트는 특징들의 교호작용을 반영하고, 부트스트래핑 방법으로 여러 분포를 학습하기 때문이다.

3.3 실험 결과

그림 4는 식 (4)에 따라 모형의 특징 수를 달리하여 최적화하는 실험을 수행한 결과이다. 통계적 검정에 의해 선택된 특징의 수는 총 17개로 p-value를 기준으로 특징 수 (10,15,17)개의 순서로 실험하였다. SHAP에 의한 선택 방법은 (10,15,17,20,25,30)개의 순서로 실험하였다.

SHAP로 선택한 특징의 SVM이 AUROC 0.775로 가장 높은 성능을 보였으며, 특징을 10개로 선택 시에는 통계적 방법이 AUROC 0.768로 가장 높았다. SVM은 차원 대비 작은 표본 수에서도 원활히 작동하기에 가장 우수한 결과가 나온 것으로 볼 수 있다. 하지만, SHAP 방법에 의해서 20개 이하의 특징 선택 시, 모든 변수를 사용한 모형에 비해 AUROC 0.014가 감소했으며, 통계적 검정 방법에서 15개 특징 선택 시 AUROC 0.0195 감소했다. 두 경우 모두, 수치형 자료들이 추가되어 AUROC가 감소했으며, 이는 수치형 자료가 많이 포함되면 선형으로 분리하기 어려워졌기 때문이다. 로지스틱 회귀분석은 통계적, SHAP 방법 모두 특징 선택 시 AUROC가 상승한 것을 관찰할 수 있으며, 특히 통계적

방법에서 모든 변수를 사용한 모형에 비해 AUROC 0.06이 상승했다. 이는 일부 특징을 선택함으로써 다중 공선성이 해결된 것을 의미한다. 그라디언트 부스팅은 통계적, SHAP 방법 모두 특징 수를 줄일수록 AUROC가 증가했다. 부스팅(Boosting)은 배경에 비해 과적합에 취약해 특징 선택 결과가 유의했음을 알 수 있다. 하지만, 10개의 특징만 사용 시 AUROC가 감소하여, 부스팅이 범주형 자료들이 많은 경우 사용이 어려운 것을 의미하며, 적절한 특징의 비율을 찾는 것이 중요함을 알 수 있다. 다층 퍼셉트론은 10개의 특징 선택 시 모든 변수를 사용한 모형에 비해 통계적 방법 시 AUROC 0.1072, SHAP 방법 시 AUROC 0.0869로 모형 중 가장 크게 향상했다. 이는 MLP가 특징 선택이 필요한 모형임을 의미한다.

그림 3의 (b)에서 MLP 특징 중요도 범위는 (0.03, 0.073)이고 최대, 최소 비율이 2.3:1로 모든 특징들의 변수 중요도가 유사하게 측정되었다. 또한, 검정 결과 유의하지 않은 것으로 관찰된 특징인 만성 신부전 여부(CRP)와 침습적 측정 맥박(iPR)이 상위 2번째, 5번째의 변수 중요도를 가진다. 즉, MLP는 유의하지 못한 특징들이 출력 결과에 큰 영향을 줄 수 있으며, 특징 선택이 중요하다는 것을 알 수 있다. 마찬가지로, 그림 3의 (c)에서 로지스틱 회귀분석의 특징 중요도 범위의 최대, 최소 비율이 4.6:1로 모든 특징들의 변수 중요도가 높게 측정된다. 이와 달리, 비교적 정확도가 높은 그림 3의 (d)와 (e)에서 그라디언트 부스팅과 SVM의 특징 중요도 범위의 최대, 최소 비율은 각각 13:1, 127:1로 유의한 특징들만 출력 결과에 강한 영향을 준다. 이를 통해, 기계학습 모형의 낮은 정확도는 고른 특징 중요도에 기인하는 경향을 알 수 있다.

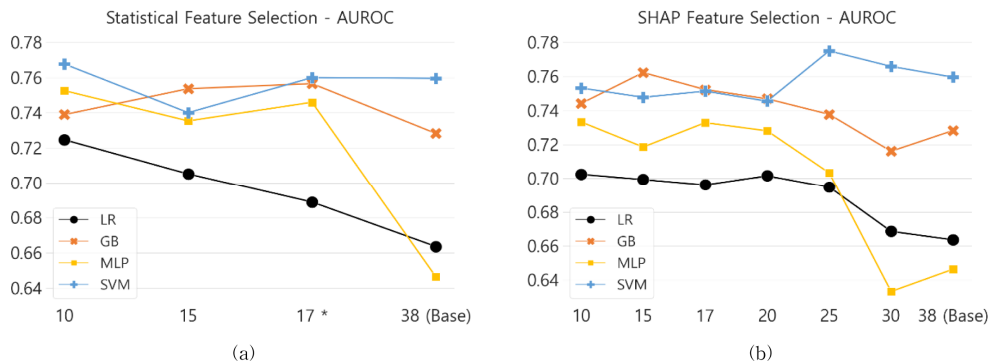


그림 4 특징 선택에 따른 성능 비교. (a) 통계적 방법에 의한 특징 선택 성능 결과, (b) SHAP 방법에 의한 특징 선택 성능 결과, *: 통계적으로 기각된 특징의 수, X축은 특징의 수를 의미한다.

Fig. 4 Performance comparison of according to feature selection. (a) Performance of feature selection by statistical method, (b) Performance of feature selection by SHAP, *: Number of features rejected by statistical methods. X-axis means the number of features

표 1 모형 별 우수한 특징들의 예측 확률 임계치 조절결과. 괄호안의 값은 테스트 데이터와 검증 데이터의 예측 결과 차이

Table 1 Results of prediction threshold adjustment for each model with best features (value in parentheses means difference between predicted results of test data and validation data. ET: Empirical Treatment, LR: Logistic Regression, GB: Gradient Boost, MLP: Multi-Layer Perceptron, SVM: Linear Support Vector Machine)

Classifier	AUROC	Accuracy	Sensitivity	Precision	Specificity	F1 score
ET	-	0.489	0.872	0.493	0.106	0.631
LR _{STAT} ¹⁰	0.725	0.614 (0.022)	0.827 (0.071)	0.584 (0.030)	0.402 (0.027)	0.682 (0.044)
GB _{STAT} ¹⁷	0.757	0.638 (0.028)	0.810 (0.110)	0.609 (0.021)	0.466 (0.047)	0.693 (0.016)
MLP _{STAT} ¹⁰	0.753	0.668 (0.007)	0.791 (0.158)	0.641 (0.087)	0.545 (0.003)	0.705 (0.012)
SVM _{STAT} ¹⁰	0.762	0.643 (0.086)	0.838 (0.061)	0.610 (0.002)	0.449 (0.139)	0.703 (0.020)
SVM _{SHAP} ²⁵	0.775	0.641 (0.034)	0.861 (0.036)	0.606 (0.010)	0.421 (0.042)	0.707 (0.003)

표 1은 식 (4)를 통해 통계적 검정과 게임이론적 방법에서 모형별로 AUROC가 가장 우수한 특징을 선별하고, 학습 데이터에서 10겹 교차검증을 통해 경험적 치료의 민감도와 가장 유사한 수치를 예측 확률 임계치로 사용하여 테스트 셋에 적용한 결과이다. 기존의 경험적 치료에서는 정확도가 0.489으로 무작위 선택과 같은 열악한 수치를 보였으나, 민감도는 0.872로 가능한 ciprofloxacin 내성인 환자들에게 효용성이 있는 항생제를 선택했음을 나타낸다. 반면, 특이도가 0.106으로 ciprofloxacin 감수성인 환자들에게 ciprofloxacin이 아닌 다른 항생제 치료를 실시했음을 의미한다. ciprofloxacin 내성인 환자에게 해당 항생제를 처방한다면 치료 실패가 발생할 수 있기에, 경험적 치료에서는 특이도에 비해 민감도가 상대적으로 높다. 따라서, 본 연구에서는 치료 실패를 방지하기 위해, 민감도가 가장 중요한 평가 요소이다. 다층 퍼셉트론은 정밀도와 특이도가 가장 높은 결과가 있지만, 민감도가 모든 방법 중 가장 낮으며 검증 셋과 테스트 셋의 민감도 간에 차이로 재현력 또한 가장 낮았다. 이에 비해, SVM은 테스트 셋에서 민감도가 가장 높았으며, 검증 셋과 테스트 셋의 민감도 차이는 가장 낮은 0.036을 보여 재현력도 가장 우수했다. SHAP 방법으로 특징 선택된 SVM은 민감도를 경험적 치료와 비교하여 98% 수준을 유지하고 특이도는 0.315 향상해 0.421로 가장 좋은 성능을 보였다. 로지스틱 회귀는 다른 모형들에 비교하여 전반적으로 가장 낮은 성능을 보였다. 그래디언트 부스트는 특이도가 SVM에 비해 약간의 개선이 있지만 민감도는 떨어졌다.

4. 결론

본 논문에서는 상부 요로감염 환자 항생제 내성 예측을 위한 특징 선택 방법과 기계학습 알고리즘들의 성능을 비교하여, 유의미한 특징을 선택하고 희소 데이터에 적합한 알고리즘을 제안하였다. 특징 선택 방법으로는

통계적 검정 방법과 SHAP를 활용한 게임 이론적 특징 추출 방법을 이용했고, 추출된 특징에 따라 로지스틱 회귀분석, 선형 SVM, 그래디언트 부스트, 다층 퍼셉트론 등 기계학습 모형의 성능을 비교했다. 또한 예측 확률 임계치 조절을 수행하여 경험적 치료 수준의 민감도를 가능한 유지하면서, 특이도를 높이는 방법을 제안했다. 결과적으로, 특징 선택방법에서 통계적 선택방법은 적은 수의 데이터에서도 좋은 성능을 보이고 SHAP 특징 추출 방법은 가장 높은 성능을 보였다. 모형들 중에서는 선형 SVM이 통계적 검정 방법과 SHAP를 활용한 특징 추출 방법에서 모두 가장 높은 AUROC를 보였다. 그뿐만 아니라 SHAP 특징 추출 방법의 SVM은 예측 임계치 조절을 통해 민감도를 조절하여 경험적 치료의 민감도를 98% 유지하면서 특이도를 3.9배 향상했다. 임상에서 ciprofloxacin 의사결정에 제안하는 방법의 도움을 받는다면, 향상된 특이도로 인해 항생제 오남용을 일부 개선할 수 있을 것으로 기대된다. 이를 통해 장기적으로 지역사회 항생제 내성 증가를 완화하는데 기여할 수 있다. 향후에는 ciprofloxacin 외에도 다른 항생제로 연구의 대상을 확대할 계획이다.

References

- [1] S.-W. Oh, H. Lee, J.-Y. Shin, and J.-H. Lee, "Antibiotics-Resistant Bacteria Infection Prediction Based on Deep Learning", *The Journal of Society for e-Business Studies*, Vol. 24, No. 1, pp. 105-120, 2019. (in Korean)
- [2] Alexander J McAdam, David C Hooper, Alfred DeMaria, Brandi M Limbago, Thomas F O'Brien, and Betsy McCaughey, "Antibiotic Resistance: How Serious Is the Problem, and What Can Be Done?", *Clinical Chemistry*, Vol. 58, No. 8, pp. 1182-1186, 1 August 2012.
- [3] Hyunjoon Pai, "Treatment of Community-Acquired

- Uncomplicated Urinary Tract Infection", *The Korean Journal of Medicine*, Vol. 81, No. 6, pp. 685-689, 2011. (in Korean)
- [4] Tae Uk Ha, Hwang Yong, Seung Chol Park, and Lee Jea Whan, "Changes in Causative Organisms and Antimicrobial Susceptibility of the Urinary Tract Infection", *Journal of Korea Academia-Industrial cooperation Society*, Vol. 18, No. 3, pp. 85-93, 2017. (in Korean)
- [5] Barbara Kot, "Antibiotic Resistance Among Uropathogenic Escherichia coli", *Microbiol*, Vol. 68, pp. 403-15, 2019.
- [6] Oonsivilai, M., Mo, Y., Luangasanatip, N., Lubell, Y., Miliya, T., Tan, P., Loeluk, L., Turner, P., and Cooper, B. S, "Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia", *Wellcome open research*, Vol. 3, 2018.
- [7] Feretzakis, G., Sakagianni, A., Loupelis, E., Kalles, D., Skarmoutsou, N., Martsoukou, M., Christopoulos, C., Lada, S., Petropoulou, S., Velentza, A., Michelidou, S., Chatzikyriakou, R., and Dimitrellos, E, "Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric Antimicrobial Therapy", *Healthcare informatics research*, Vol. 27, No. 3, pp. 214-221, 2021.
- [8] Doohyun Pak, Mingyu Hwang, Minji Lee, Sung-Il Woo, Sang-Woo Hahn, Yeon Jung Lee, and Jaek Hwang, "Application of Text-Classification Based Machine Learning in Predicting Psychiatric Diagnosis", *Journal of Korean Society of Biological Psychiatry*, Vol. 27, No. 1, pp. 18-26, 2020. (in Korean)
- [9] Su Min Kim and Jiyoung Yoon, "Data Processing and Visualization Method for Retrospective Data Analysis and Research Using Patient Vital Signs", *Journal of biomedical Engineering Research*, Vol. 42, No. 4, pp. 175-185, 2021. (in Korean)
- [10] Zale, A. D., Abusamaan, M. S., McGready, J., and Mathioudakis, N, "Development and validation of a machine learning model for classification of next glucose measurement in hospitalized patients", *EClinicalMedicine*, Vol. 44, 2022.
- [11] Teisseyre P. and Klonecki, T, "Controlling Costs in Feature Selection: Information Theoretic Approach", *Computational Science - ICCS 2021, Lecture Notes in Computer Science*, pp. 1-14, 2021.
- [12] R. Tang and X. Zhang, "CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification", *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pp. 80-84, 2020.
- [13] Yong-Geon Lee, Jae-Young Oh, and Gibak Kim, "Improvement of solar power forecasting using SHAP value", *The Korean Institute of Electrical Engineers*, pp. 1886-1887, 2020. (in Korean)
- [14] Frank J. and Massey, Jr, "The Kolmogorov-Smirnov Test for Goodness of Fit", *Journal of the American Statistical Association*, Vol. 46, No. 253, pp. 68-78, 1951.
- [15] McHugh M. L, "The chi-square test of independence", *Biochemia medica*, Vol. 23, No. 2, pp. 143-149, 2013.
- [16] Scott M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 4768-4777, 2017.
- [17] S. B. Kim and J. W. Lee, "Power comparison of distribution-free two sample goodness-of-fit tests", *The Korean Journal of applied Statistics*, Vol. 30, No. 4, pp. 513-528, 2017. (in Korean)
- [18] Myung HoonBaek and Jun Geol, "Root Cause Analysis Methods in Semiconductor Manufacturing System Using Modified K-S(Kolmogorov-Smirnov) Test", *Journal of the Korean Institute of Industrial Engineers*, Vol. 44, No. 2, pp. 132-140, 2018. (in Korean)
- [19] Youngsuk Suh, "Historical Review on Independence Tests: From Chi-square Test to Bootstrapping Method", *Journal of Educational Evaluation*, Vol. 33, No. 2, 323-351, 2020. (in Korean)
- [20] Y. A. Yoon, S. H. Lee, and Y. S. Kim, "A Study on the Remaining Useful Life Prediction Performance Variation based on Identification and Selection by using SHAP", *Journal of Korean Society of Industrial and Systems Engineering*, Vol. 44, No. 4, pp. 1-11, 2021. (in Korean)
- [21] Won Hee Chung, Da Woon Jeong, Kang Ah Young, Yeong Hyeon Gu, and Yoo Seong joon, "Optimal Machine Learning-Based Demand Prediction Model Performance Comparison by Demand Pattern", *The Journal of Korean Institute of Next Generation Computing*, Vol. 16, No. 6, pp. 76-89, 2020. (in Korean)

이 종 법

2023년 인하대학교 컴퓨터공학과, 통계학과 졸업(학사). 2023년~현재 인하대학교 전기전자컴퓨터공학과 인공지능전공 석사과정. 관심분야는 의료인공지능, 딥러닝



이 현 규

2010년 인하대학교 컴퓨터공학과 졸업(학사). 2017년 인하대학교 컴퓨터공학과 졸업(박사). 2018년 LG전자 소재/생산기술원 선임연구원. 2019년~2021년 두에이아이㈜ 연구소장. 2021년~현재 인하대학교 의과대학 조교수. 관심분야는 기계학습, 컴퓨터비전 등



부 록

(표 1) 본 연구에 사용된 예측변수 목록

	Feature	Type of feature	Mean / Std
General	Age	Numeric	72.13 / 15.86
	Gender	Category (0,1)	0.26 / 0.44
	Bed-ridden	Category (0,1)	0.38 / 0.49
	Mental change	Category (0,1)	0.21 / 0.41
	Nursing home residence	Category (0,1)	0.15 / 0.35
	Use of ppi antacid	Category (0,1)	0.38 / 0.49
Medical device	Urinary catheter/Cystostomy	Category (0,1)	0.13 / 0.34
Comorbid conditions	Diabetes mellitus	Category (0,1)	0.42 / 0.49
	Hypertension	Category (0,1)	0.59 / 0.49
	Cardiovascular disease	Category (0,1)	0.23 / 0.42
	Chronic renal failure	Category (0,1)	0.13 / 0.34
	Neurodegenerative disease	Category (0,1)	0.43 / 0.50
	Malignancy	Category (0,1)	0.21 / 0.41
Past history	Hospital-acquired	Category (0,1)	0.19 / 0.39
	Hospitalization within 3 months	Category (0,1)	0.37 / 0.48
	Antibiotics use within 3 months	Category (0,1)	0.23 / 0.42
	UTI within a year	Category (0,1)	0.24 / 0.43
Disease	Leukopenia	Category (0,1)	0.06 / 0.24
Assessment models	qSOFA	Category (0,1,2,3)	0.65 / 0.83
Vital signs on presentation	Systolic blood pressure	Numeric	127.55 / 30.48
	Diastolic blood pressure	Numeric	72.98 / 17.83
	Pulse rate	Numeric	97.94 / 22.04
	Respiratory rate	Numeric	20.25 / 3.81
	Body temperature	Numeric	38.55 / 15.00
	Peripheral oxygen saturation	Numeric	78.63 / 30.29
Complete blood cell counts	Leukocyte count	Numeric	12000 / 6495
	Hemoglobin	Numeric	11.78 / 9.0
	Platelet	Numeric	210.0 / 101.0
Other laboratory findings	C reactive protein	Numeric	12.13 / 9.0
	Glucose	Numeric	165.02 / 94.30
	Blood urea nitrogen	Numeric	28.35 / 44.08
	Creatinine	Numeric	1.34 / 1.32
Arterial blood gas analysis	Ph	Numeric	38.69 / 482
	Pco2	Numeric	33.50 / 7.03
	PO2	Numeric	78.63 / 30.29
	HCO3	Numeric	23.53 / 6.93
	Arterial oxygen saturation	Numeric	96.19 / 38.45
	Lactic acid	Numeric	2.65 / 6.34