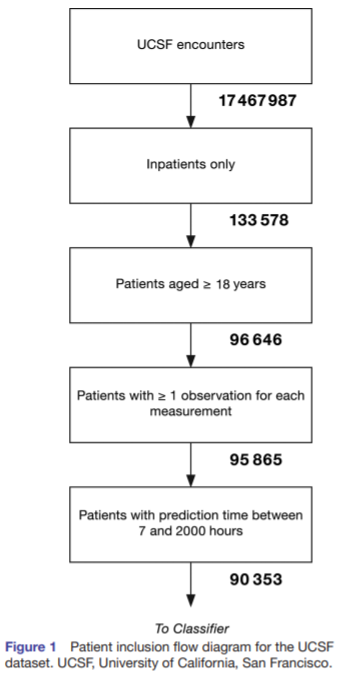
Multicentre validation of a sepsis prediction algorithm  
using only vital sign data in the emergency department,  
general ward and ICU

Qingping Mao, et al.

Sepsis 예측을 위한 머신러닝 알고리즘인 InSight에 6개 vital sign만 사용하여, SIRS, SOFA, MEWS 대비 3 gold standard (Sepsis, Severe Sepsis, Septic Shock)들의 detection, prediction 능력을 비교 검증한다.

ML algorithm: Gradient Tree Boosting

Dataset: UCSF, MIMIC-3 for transfer-learning

Evaluation tool: AUROC, accuracy, sensitivity, specificity, LR+, LR-

데이터의 추출은 Figure 1과 같이 이루어졌으며  
Data Imputation을 위해서 carry-forward imputation 방식을 취하였다.

Feature Engineering: (V*onset*, V*t-1*, V*t-2*, dVt01, dVt12 for each of the six measurement channels)  
그러므로 6개 feature마다 5개씩 총 30개의 feature가 만들어진다.

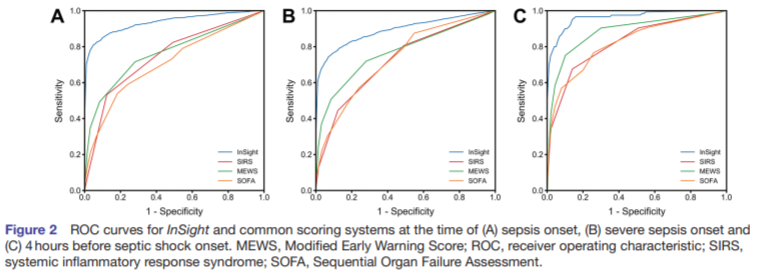
텍스트이(가) 표시된 사진

자동 생성된 설명

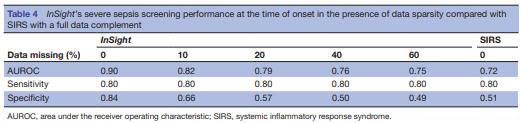
Gradient Boosting은 약한 분류기인 leaf node들을 여러 개 결합하여 강한 분류기를 만들어내는 것이다. 이 원리는 loss function *L*을 주목하면 이해하기 편하다. L은 오차값을 갖고있는 2차 방정식(squared error)으로, 기하학적으로는 아래로 볼록한 곡선이 2차함수 곡선이 그려진다. 이 함수에서 오차를 최소화시키는 것이 목적이다. 그러므로 이를 미분하면, 오차의 변화율이 나오고, 이 값이 0이 되는 지점, 즉 저점을 찾는 것이 목적이다. 함수를 미분하면 negative gradient 방정식이 나오게 된다. 이는 아래로 내려가고자 하는 방향성을 가지고 있기 때문에 Gradient descent + boosting이라고도 불린다.

여기에 10-fold cross-validation을 진행한다. Training set은 80%이고, Test set은 20%이다.   
이외에도 여러가지 검증을 수행하였다.  
Missing data가 존재하는 data에 대해서 robustness 등을 측정하기도 하였다.

그후 UCSF와 동일한 목적을 갖고 있지만, 환자 통계, 특징 등이 다른 MIMIC-3의 데이터를 이용해서 Transfer Learning을 시행하였다. MIMIC-3로 훈련시키고, UCSF의 target을 찾게 한 것이다.



결과적으로 모든 측면에서 3Gold Standard 대비 InSight가 뛰어난 민감도와 특이도를 보여주었다. C의 결과가 특히 좋은 이유는 Septic Shock의 4 시간 전 예측을 위한 측정 기준이 훨씬 엄격하고, 상대적으로 덜 까다로웠기 때문이다. 그리고 이 모델이 일반화가 잘 이루어졌는지 확인하기 위해 4개 병원 (Stanford, Oroville, BHH, CRMC)에서도 test를 진행하였다. 또한 InSight의 성능에 dropout을 단계적으로 적용하여 성능을 검증해보았다.



Dropout은 overfitting을 막기위한 regularization 기법이다. <https://algopoolja.tistory.com/50>   
주로 신경망 알고리즘에서 사용되는데, 임의의 뉴런 노드들을 epoch마다 사용하지 못하게 하여서 학습이 한쪽으로 지나치게 강화되지 않게 한다.

1. 이 연구만으로 일반화 능력이 있다고 주장할 수는 없다.
2. ICD-9 code만을 이용하였기에, 진단되지 않았거나, 올바르게 기록되지 않은 환자들의 데이터를 사용하지는 못하였다.
3. Imputation 등으로 인해서 사용한 dataset의 분산을 과소평가하였다고 볼 수 있다.
4. 후향적 연구로 인해서 해석이 들어갔다.

추가적으로 주목할 것은 Supplementary Table 3-4를 통해 확인할 수 있다.  
테이블이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

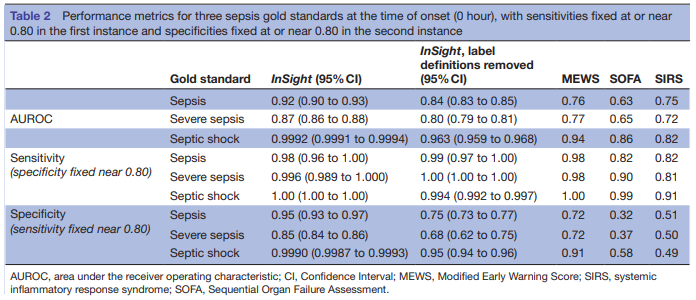
자동 생성된 설명재밌는 것은  
Actual Positive 즉, 실험군의 환자수가  
Actual Negative 대조군의 환자수보다  
압도적으로 많다는 것이다.

오히려 반대인 것이 자연스럽다.  
양성인 환자가 정상인 환자보다  
비율이 적을 테니 말이다.

그러므로 이 논문의 실험결과는  
모델이 양성에 대하여 과적합이  
일어날 수 있는 가능성이 농후한,  
**과적합 성능에 우호적인 실험**이라고  
볼 수 있다.

만약 대조군이 더 많아지게 된다면,  
위 결과는 전혀 다른 결과로  
나올 수 있다.  
즉, 본 결과는 신뢰하기 어렵다.

내 생각



InSight는 위의 지나치게 실험군 환자수가 많은 데이터를 사용했는데,  
어떻게 민감도와 특이도가 둘 다 높게 나온 것일까? 이는 해당 모델이 UCSF에 지나치게 과적합 되었다고 의심해볼 수 있다. 대조군 데이터가 실험군 대비 현저히 부족하기 때문에 이런 불균형한 데이터로 좋은 성능이 나왔다면 과적합을 의심할 수 있다.

MIMIC-3로 Transfer Learning을 시켰다고 하였는데, 그 결과 Table이 없는 부분은 아쉽다.

테이블이(가) 표시된 사진

자동 생성된 설명

고작 4개 병원의 결과로 해당 모델이 일반화되었다고 주장할 수는 없다.  
이 4개 병원이 UCSF dataset의 환자 demography와 유사하여 좋은 결과가 나왔을지도 모른다.  
**과적합 모델이라는 의심이 생길 수 있는 여지를 주었다는 시점**에서부터 이 논문의 사용에 대한 주저를 낳았다고 볼 수 있다.