Sử dụng mô hình hồi quy Cox để dự đoán sớm nhiễm trùng huyết

Lê Quang Hưng

Giới thiệu chung

Dự đoán sớm nhiễm trùng huyết

Nhiễm trùng huyết

- Nhiễm khuẩn huyết hay nhiễm trùng máu (nhiễm trùng huyết) là tình trạng nhiễm trùng rất nghiêm trọng. Vi sinh vật gây bệnh không cư trú tại một cơ quan bị tổn thương ban đầu, mà theo đường máu lan đi khắp cơ thể.

- Nếu không được phát hiện sớm và điều trị kịp thời dẫn đến các biến chứng nặng về tuần hoàn, rối loạn đông máu, hô hấp, suy gan thận và các tạng khác.

- Việc phát hiện sớm nhiễm trùng huyết là cần thiết để có những can thiệp kịp thời và giảm tỷ lệ tử vong.

Dữ liệu: ở 2 bệnh viện A, B

A

B

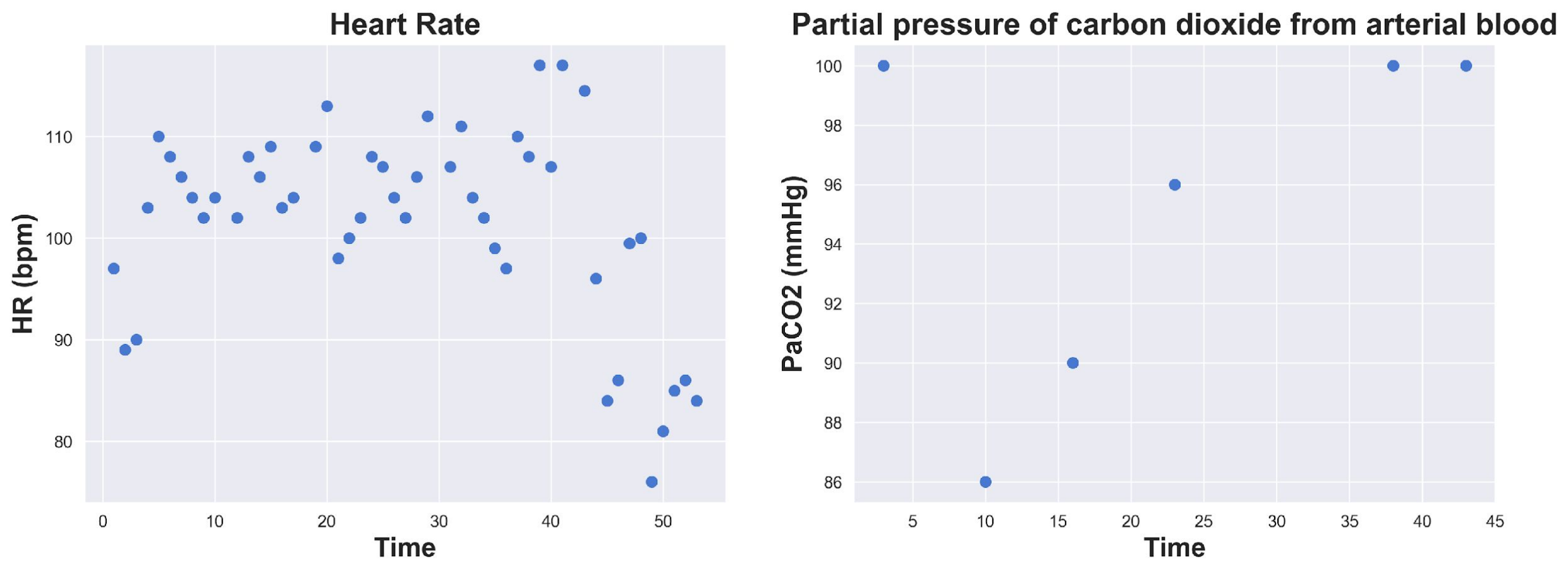
Mô tả chung dữ liệu

Bệnh viện	A	B
Số lượng bệnh nhân	20,336	20,000
Số lượng bệnh nhân nhiễm trùng huyết	1790	1142
Tỷ lệ nhiễm trùng huyết	8.80%	5.70%
Số dòng	739,663	648,508
Số dữ liệu đầu vào	5,536,849	4,950,064
Mật độ dữ liệu đầu vào	20.60%	19.10%

Đặc trưng của dữ liệu

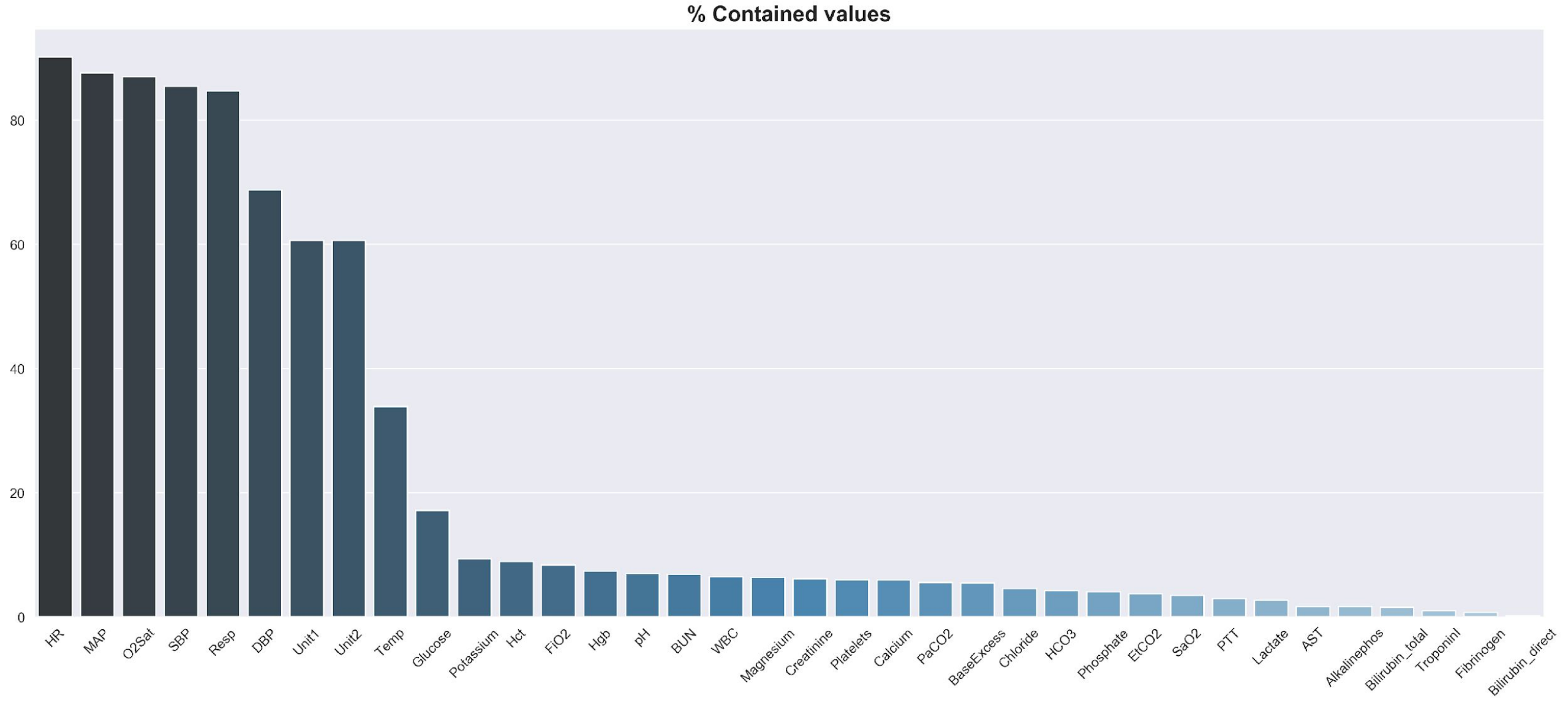
- Tổng cộng 40 đặc trưng bao gồm:
  - **8 dấu hiệu sinh tồn** – Heart Rate, Temperature, MAP, …
  - **28 chỉ số xét nghiệm** – FiO2, Lactate, Bilirubin, …
  - **6 đặc trưng nhân khẩu học** – Age, Gender, Hospital Unit, …

Dữ liệu



- Các dữ liệu phòng thí nghiệm bị thiếu nhiều

Tỉ lệ dữ liệu bị thiếu



Xử lý dữ liệu bị thiếu

- Dữ liệu sẽ được điền vào theo từng bệnh nhân với phương pháp KNN Imputation (cho rằng các chỉ số của cơ thể người có liên quan đến nhau, ta có thể suy ra dữ liệu bị thiếu từ các dòng dữ liệu gần giống)

- Một số features ở một vài bệnh nhân bị thiếu hoàn toàn, khi đó dữ liệu sẽ được điền vào bằng trung bình của features đó trên tập huấn luyện.

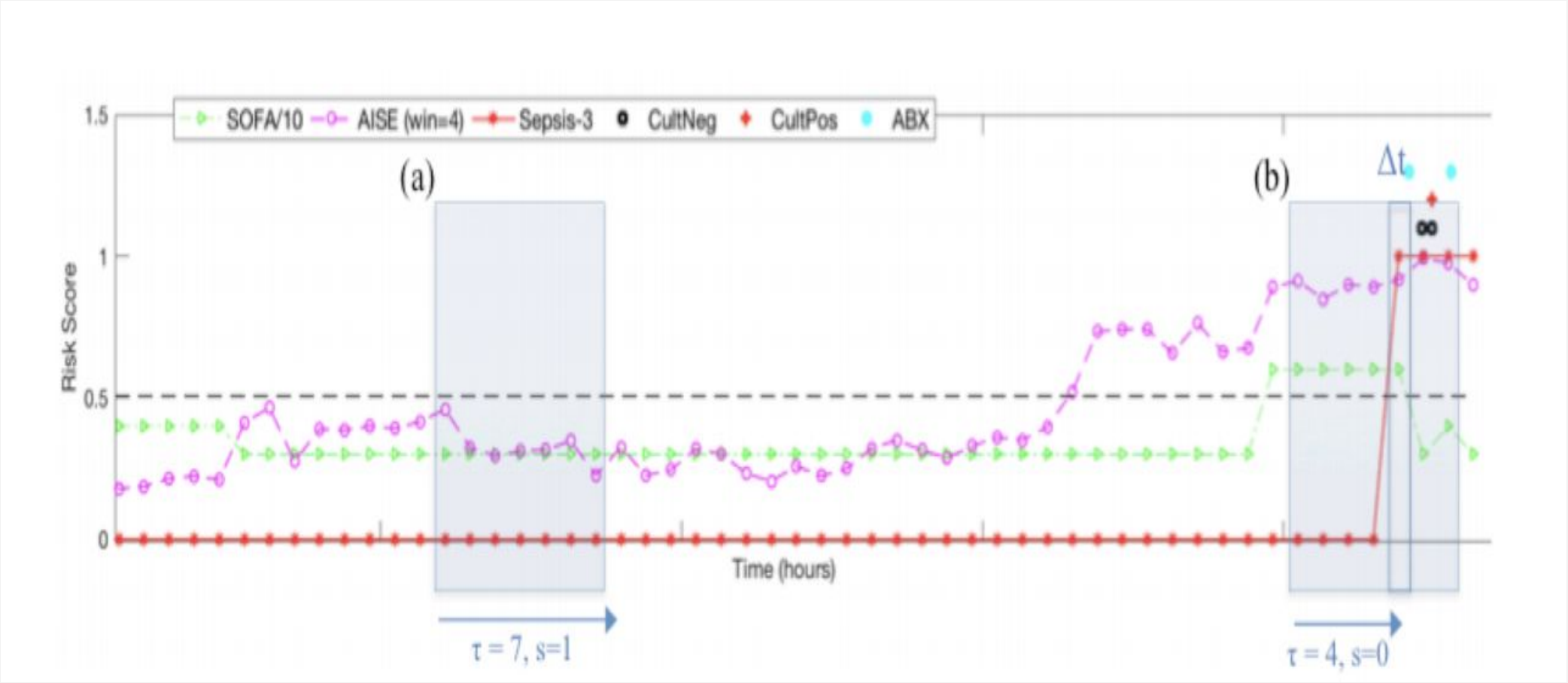
Một số dữ liệu bổ sung

Feature Name	Description
ShockIndex	Ratio of heart rate to systolic blood pressure (HR/SBP)
BUN/CR	Bilirubin / creatinine ratio
SOFA Deterioration	Binary marker of deterioration
MEWS	Modified Early Warning Score
SIRS	Systemic Inflammatory response syndrome
qSOFA	Quick SOFA score

Trích xuất thông tin chuỗi thời gian

- Một cửa sổ trượt sẽ được sử dụng để trích xuất thông tin từ dữ liệu chuỗi thời gian. (cửa sổ được sử dụng là 6 giờ)
- Nếu không có sự kiện (ở đây là nhiễm trùng huyết) xảy ra thì đánh si=1 đánh dấu right censoring, >6; nếu có sự kiện xảy ra thì đánh si=0, thời gian xảy ra sẽ được ghi lại (giả sử =4).

Trích xuất thông tin chuỗi thời gian



Mô hình hồi quy Cox

- Là một mô hình được sử dụng trong phân tích sống còn (Survival Analysis)

- Cox Regression xây dựng một mô hình dự đoán “thời gian đến sự kiện”. Mô hình tạo ra một hàm sống sót (survival function) dự đoán xác suất sự kiện quan tâm xảy ra tại một thời điểm nhất định t.

- Mô hình hồi quy Cox cho phép ước lượng tỷ lệ nguy hiểm dựa trên nhiều yếu tố.

Mô hình hồi quy Cox

Mô hình Cox được biểu thị bằng hàm nguy hiểm ký hiệu là $h(t)$. Một cách ngắn gọn, hàm nguy hiểm có thể được hiểu là nguy cơ tử vong tại thời điểm t. Nó có thể được ước tính như sau:

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + ... + b_px_p)$$

Trong đó:

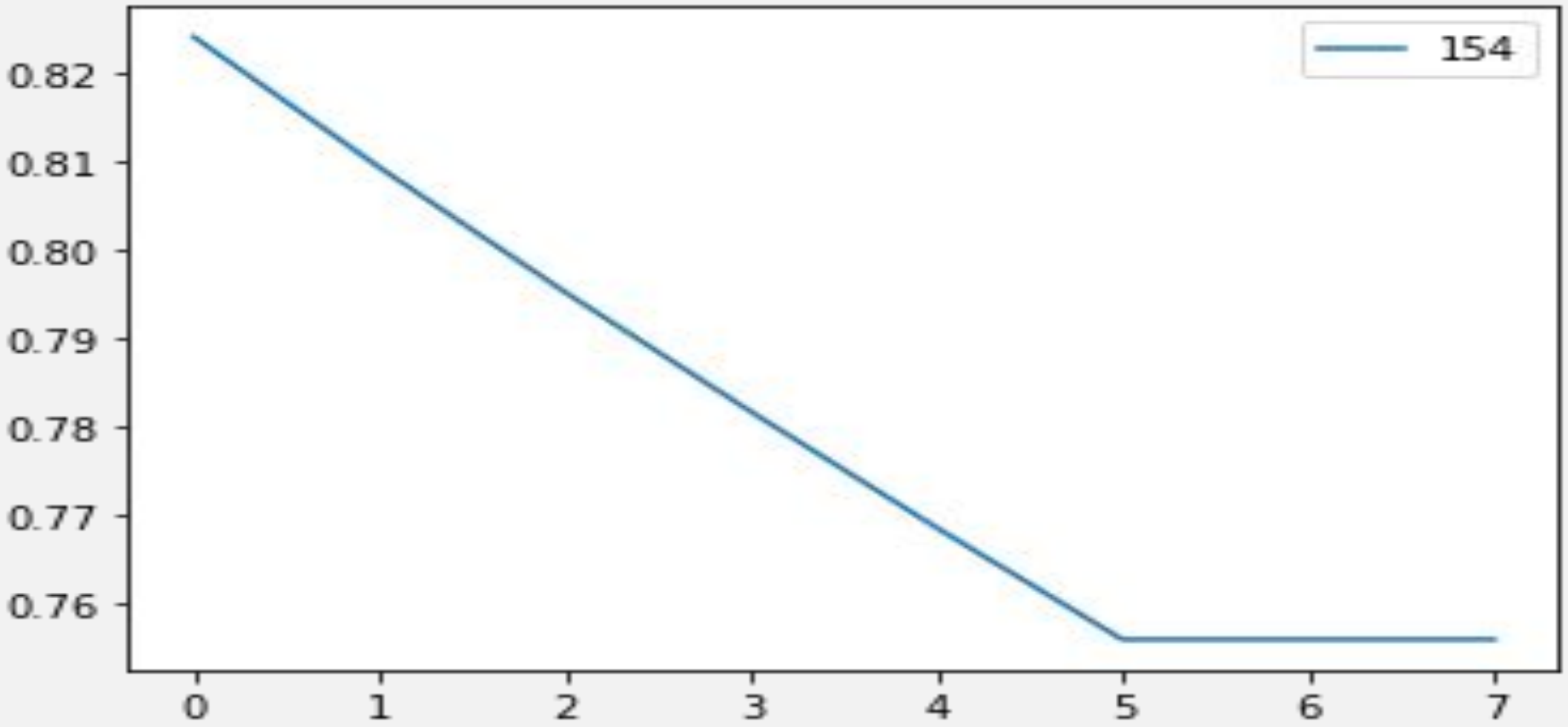
t là thời gian sống sót

$h(t)$ là hàm nguy hiểm được xác định bởi một tập hiệp biến p

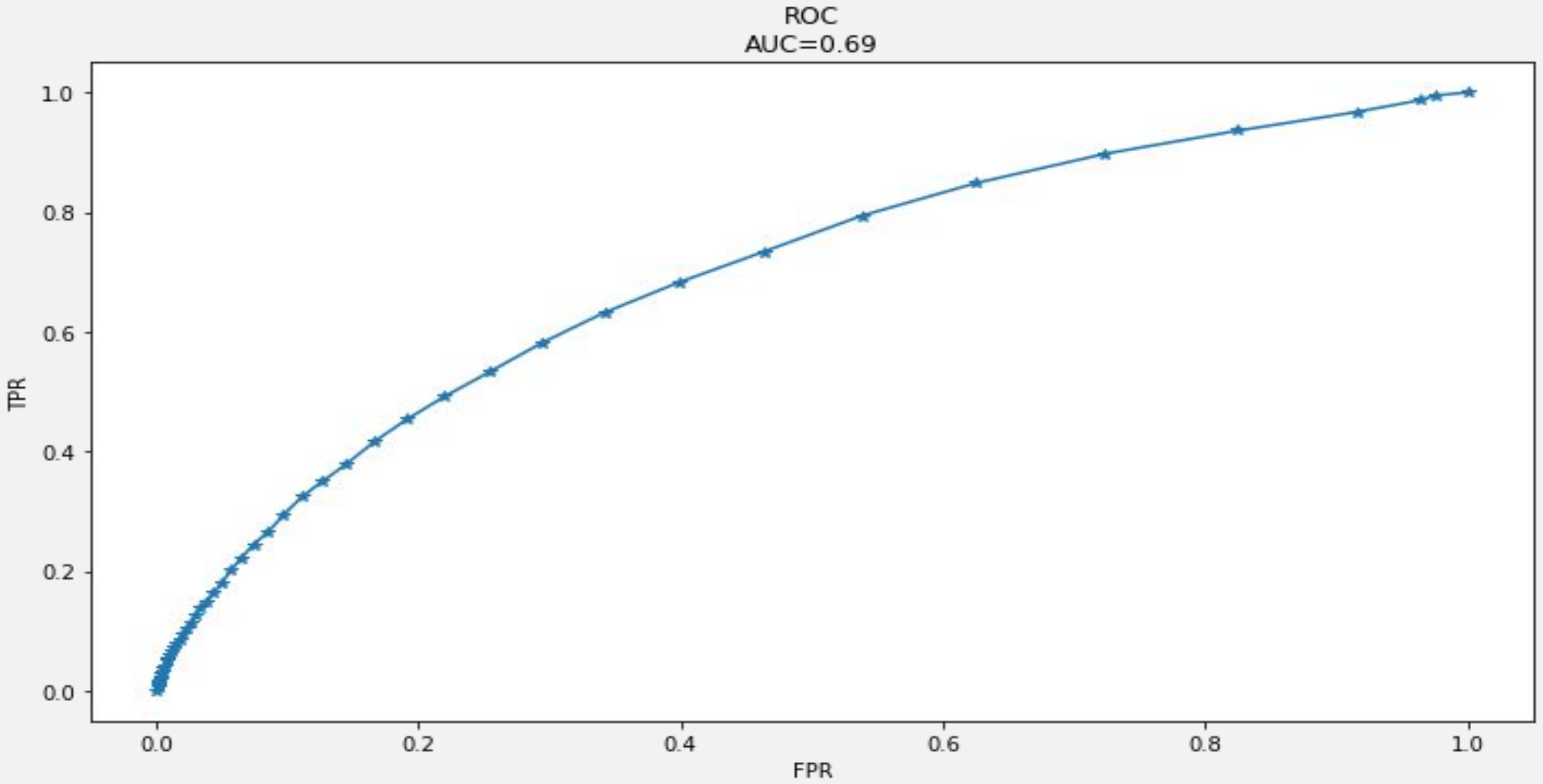
Các trọng số b thể hiện sự ảnh hưởng của các yếu tố x lên tỷ lệ.

$h_0(t)$ được gọi là baseline hazard tương ứng với giá trị của mối nguy hiểm nếu tất cả các x_i đều bằng 0 ($exp(0)$ bằng 1). Đại lượng 't' trong $h(t)$ thể hiện rằng mối nguy hiểm có thể thay đổi theo thời gian.

Mô hình hồi quy Cox



Kết quả



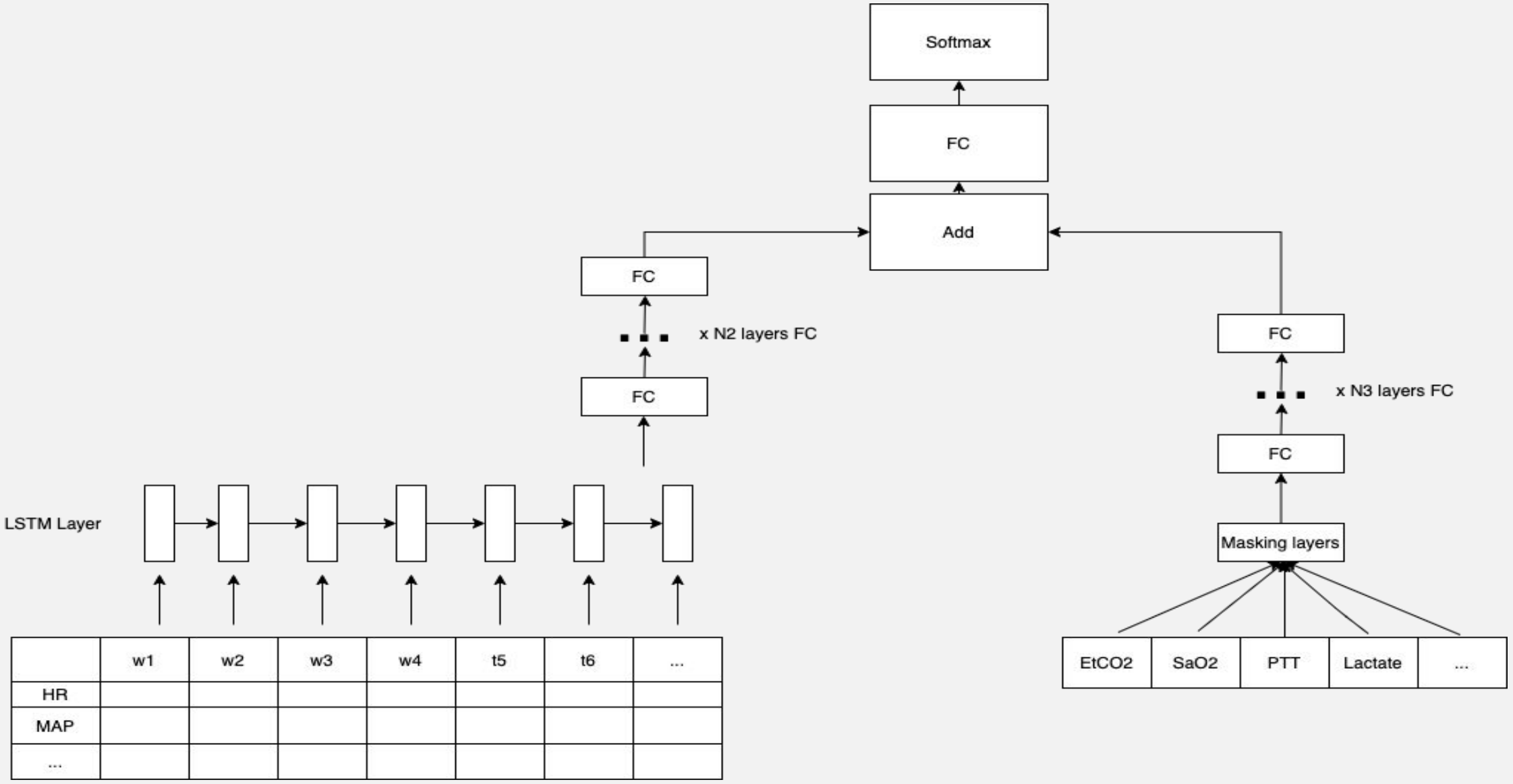
Kết quả

- Mô hình đạt điểm AUROC 0.69
- Tuy nhiên điểm utility không cao

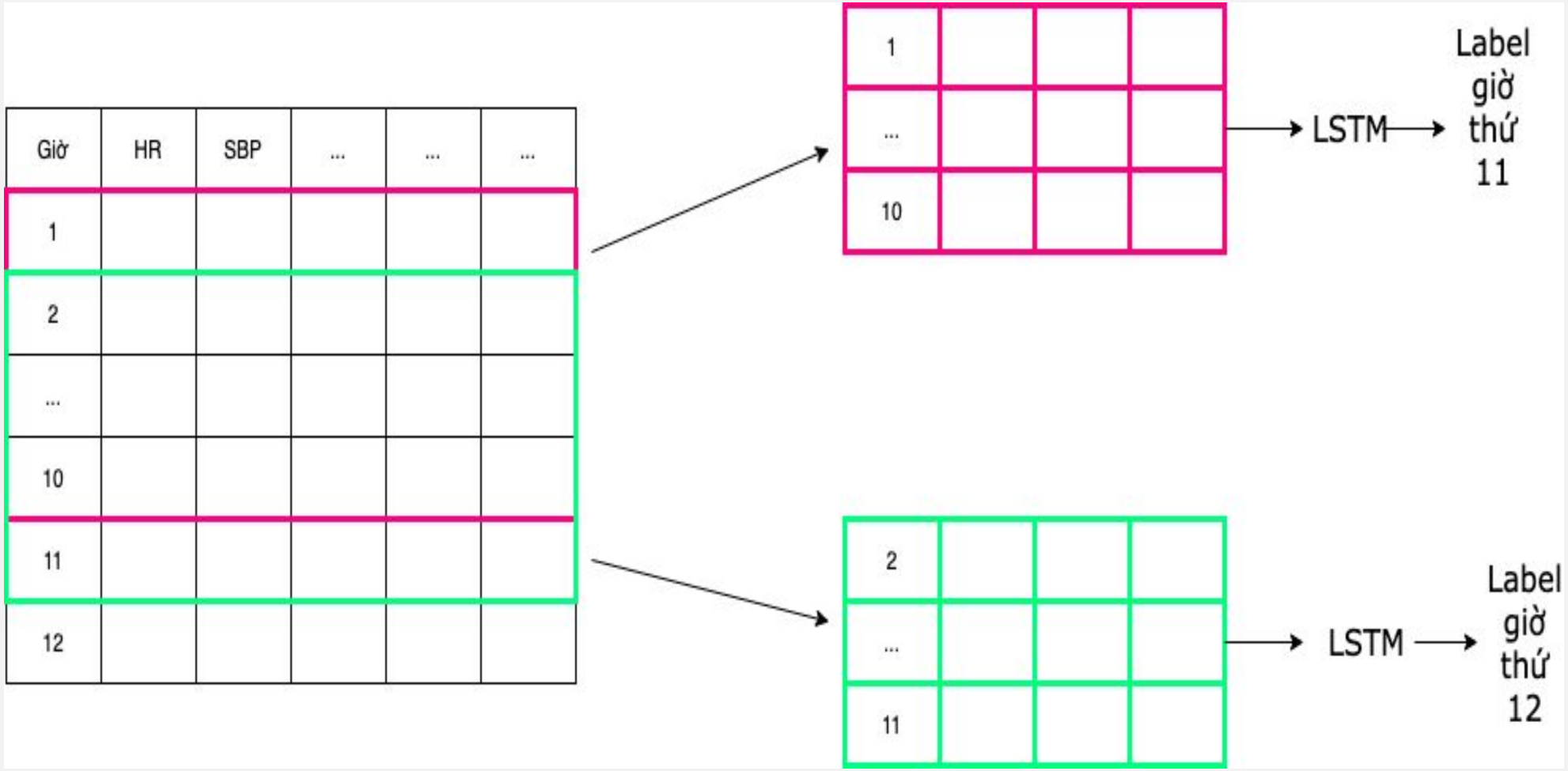
Future work

- Trong tương lai, sử dụng Weibull baseline hazard có thể mang lại hiệu quả cao hơn.
- Thử nghiệm các mô hình tập hợp tree-base như random forest, random survival forest, gradient boosting,...

Mô hình LSTM


Softmax
FC
Add
FC
x N2 layers FC
FC
FC
x N3 layers FC
FC
Masking layers
LSTM Layer
w1
w2
w3
w4
t5
t6
...
HR
MAP
...
EtCO2
SaO2
PTT
Lactate
...

Mô hình LSTM


Giờ
HR
SBP
...
...
...
1
2
...
10
11
12
1
...
10
LSTM
Label giờ thứ 11
2
...
11
LSTM
Label giờ thứ 12

Mô hình LSTM