



مرکز آموزش عالی  
علوم پزشکی  
وارسنگان

# Prediction of sepsis mortality in ICU patients using machine learning methods

پیش‌بینی مرگ و میر ناشی از سپسیس در بیماران بخش مراقبت‌های ویژه با  
استفاده از روش‌های یادگیری ماشین

- عارف عقیلی
- ۱۴۰۴/۰۹/۱۷
- ژونال کلاب

# نویسندگان اصلی مقاله



**USC** University of  
Southern California



**Negin Ashrafi**

**H-Index : 5-7**  
**Total CITE : +50**



**Maryam Pishgar**

**H-Index : +10**  
**Total CITE : +566**

## مشخصات ژورنال

|                  |                                                            |
|------------------|------------------------------------------------------------|
| <b>Name</b>      | BMC Medical Informatics and Decision Making                |
| <b>INDEXING</b>  | PubMed , Scopus , Web of Science , MEDLINE , EMBASE , DOAJ |
| <b>IF</b>        | ۳/۸                                                        |
| <b>Publisher</b> | Springer Nature                                            |
| <b>Category</b>  | Q2                                                         |

# جدول اختصارات

|            |                                                        |                                                                                                |
|------------|--------------------------------------------------------|------------------------------------------------------------------------------------------------|
| MIMIC-IV   | Medical Information Mart for Intensive Care IV         | اطلاعات پزشکی مارت برای مراقبت‌های ویژه                                                        |
| SMOTE      | Synthetic Minority Over-sampling Technique             | تکنیک بیش‌نمونه‌گیری مصنوعی اقلیت                                                              |
| SHAC       | Sequential Halving and Classification                  | الگوریتمی برای تنظیم ابرپارامترهای مدل‌های ML به‌جای Grid Search کامل                          |
| AUROC      | Area under the Receiver Operating Characteristic Curve | مساحت زیر منحنی ROC به منظور بررسی عملکرد الگوریتم‌های یادگیری ماشین                           |
| SOFA score | Sequential Organ Failure Assessment                    | یک سیستم امتیازبندی در پزشکی برای ارزیابی وضعیت بیماران بدحال و پیش‌بینی احتمال مرگ‌ومیر آن‌ها |
| SHAP       | SHapley Additive exPlanations                          | روشی برای توضیح‌پذیر کردن خروجی مدل یادگیری ماشین                                              |

# فهرست مطالب

|    |                                                                       |    |                                |
|----|-----------------------------------------------------------------------|----|--------------------------------|
| 01 | مقدمه                                                                 | 04 | یافته ها                       |
|    | اهمیت موضوع و بیان مسئله ، ضرورت انجام تحقیق                          |    | نمودار ROC و مقایسه نتایج حاصل |
| 02 | هدف مطالعه                                                            | 05 | نتیجه گیری                     |
|    | معرفی هدف کلی و اهداف فرعی مقاله                                      |    | بررسی نمودار ها و جداول مقاله  |
| 03 | روش اجرا                                                              | 06 | پیشنهادهات من                  |
|    | بررسی بهترین روشهای نمونه گیری و انتخاب بهترین الگوریتم یادگیری ماشین |    | ارائه پیشنهاداتی برای آینده    |

# 01

## مقدمه

اهمیت موضوع و بیان مسئله ، ضرورت انجام تحقیق



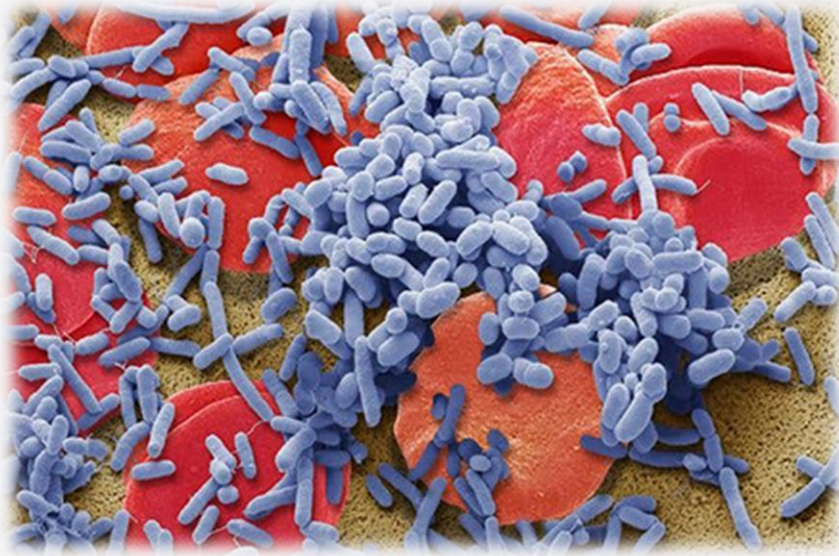
# اهمیت موضوع: پیش زمینه

بیماری Sepsis 

اهمیت بیماری 

روش های تشخیص و پیشگیری 

# Sepsis OR Septicemia



## معرفی بیماری:

- نوعی عفونت خون
- علائم شبیه به آنفلانزا و مشکلات گوارشی ساده
- واگیر نیست ولی بسیار کشنده
- پاسخ بیش از حد به سیستم ایمنی

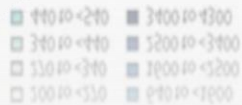
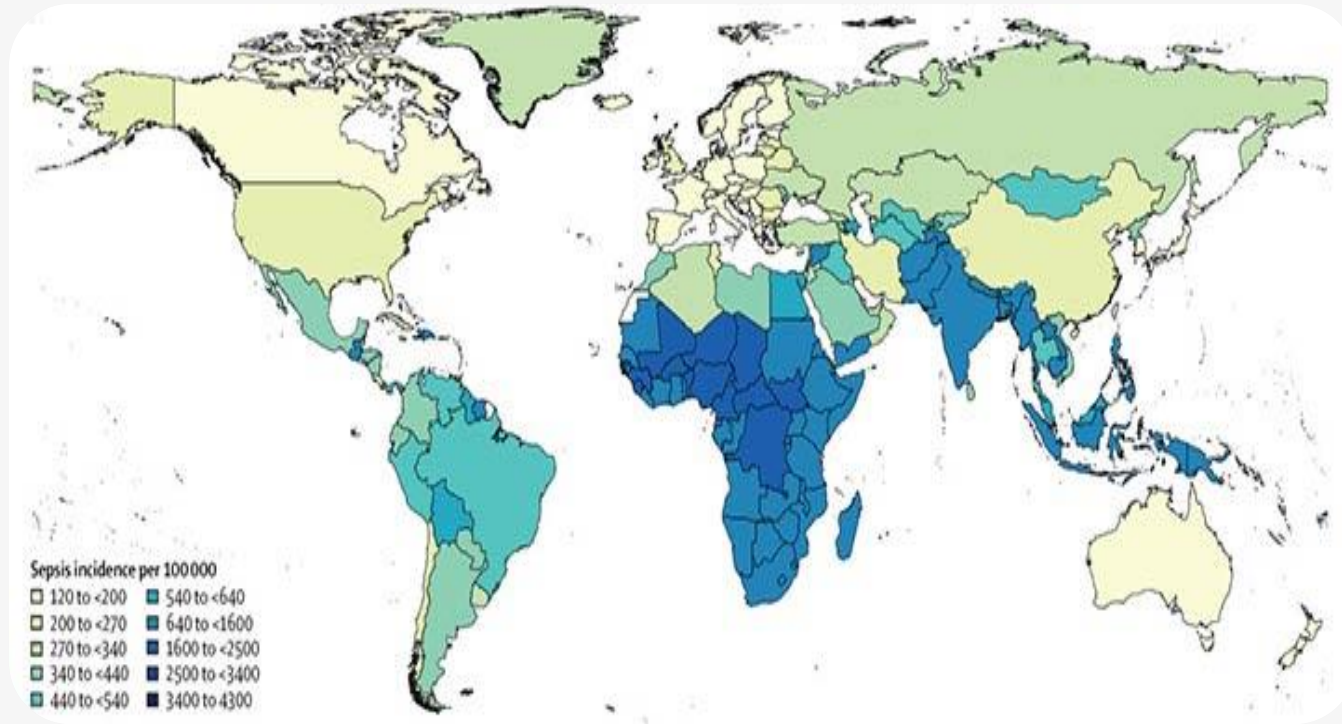
# Sepsis Importance



## اهمیت بیماری :

- WHO: یک تهدید جدی سلامت
- سرعت پیشرفت بالا، پتانسیل کشندگی، و عوارض ناتوان کننده
- در موارد حاد افت شدید فشار خون اندام ها
- آمپوتاسیون اندام از دست رفته
- هزینه های سنگین مراقبت های ویژه

# Sepsis Incidence Map



# World Sepsis Day



48.9 MILLION CASES  
11 MILLION DEATHS

1 IN EVERY 5  
DEATHS WORLDWIDE  
ARE ASSOCIATED  
WITH SEPSIS

85% OCCUR IN  
LOW- OR  
MIDDLE-INCOME  
COUNTRIES

2 OUT OF  
EVERY 5 CASES  
ARE IN CHILDREN  
UNDER 5

# Diagnosis and prevention methods



## ضرورت تشخیص زود هنگام

- تشخیص اشتباه با سایر وضعیت های پزشکی
- بسیار پیشرونده
- باعث شوک سپتیک و مرگ
- امکان سرایت به سایرین
- بروز بیشتر در گروه های سنی پر خطر و بیماری زمینه ای

# ضرورت انجام مطالعه

وجود ارتباط بین طول مدت بستری و بقا  
revealing a direct correlation between the length of hospital stay and survival,

- میانگین بستری ۱۰ روز نشان از کاهش بقاست (Suveges and other)
- سیستم امتیازدهی SOFA برای پیش بینی مرگ و میر در بیمارستان
- گرچه موثر ولی محدود به چند ویژگی خاص
- مطالعات دیگر، مانند تحلیل گذشته نگر (retrospective analysis)
- همبستگی بین سطح  $PaO_2/FiO_2$  و مرگ و میر در بیمارستان (Bi's study)
- تاثیر کمتر و محدود شدن به یک جفت پارامتر

# ضرورت انجام مطالعه

## مشکل مطالعات گذشته :

- ۱- عدم توجه به ماهیت چند عاملی سیستم
- ۲- ناتوانی در ادغام و تجزیه و تحلیل همزمان چندین ویژگی
- ۳- اتکا به داده های گذشته نگر

## برای غلبه بر محدودیت های روش های سنتی

- اتکا به رویکرد های یادگیری ماشین (ML) و یادگیری عمیق (DL)

# بررسی مطالعات اخیر با رویکرد جدید

۱- اثربخشی الگوریتم **Light GBM** در پیش‌بینی مرگ و میر بیماران سپسیس و ادغام آن را در ابزارهای بالینی (Bao's study)

۲- پتانسیل شبکه‌های عصبی مصنوعی **ANN** در تشخیص زودهنگام بیماران پرخطر در طیف وسیعی از موضوعات مرتبط با پزشکی (Shifang et al)

# مشکلات مطالعات اخیر با رویکرد جدید

"عدم دستیابی به نتایج رضایت بخش با وجود استفاده از روش های تحلیلی پیشرفته"

## عوامل ناکارآمدی مطالعات گذشته :

- فاقد استراتژی جامع برای انتخاب ویژگی
- عدم بکارگیری از تکنیک های پیشرفته پیش پردازش داده
- استفاده از ویژگی های زیاد و متعدد منجر به Overfitting
- ناکارآمدی این روش ها در محیط های بالینی real-time

٢٠

هدف مطالعه

## هدف اصلی :

توسعه مدلی با دقت بالا و تفسیرپذیر برای پیش‌بینی مرگ‌ومیر بیماران سپتیک در ICU با استفاده از تعداد کمتری ویژگی نسبت به مطالعات قبلی

- ✓ انتخاب الگوریتم جنگل تصادفی با دقت پیش‌بینی بالاتر ( $AUROC=0.94$ )
- ✓ استفاده از 38 ویژگی (نصف مطالعات قبلی) به منظور عملی‌تر و تفسیرپذیرتر شدن نتایج

## اهداف فرعی :

استفاده از تکنیک‌های پیش‌پردازش پیشرفته داده‌ها  
✓ افزایش کیفیت داده‌ها و عملکرد مدل

بررسی کامل متون دانشگاهی و توصیه‌های متخصصان بالینی  
✓ پیش‌بینی‌های دقیق‌تر با استفاده از مجموعه ویژگی کوچک‌تر و مرتبط‌تر

استفاده از تحلیل (SHAP)  
✓ افزایش تفسیر پذیری نتایج

مدل پیشنهادی جنگل تصادفی (Random Forest)  
✓ بهبود ۶,۳ درصدی در مقایسه با بهترین مطالعه موجود

تکنیک نمونه‌گیری (SMOTE)  
✓ رفع عدم تعادل داده‌های مورد استفاده imbalance data

۰۳

# روش اجرا

بررسی بهترین روشهای نمونه گیری و  
انتخاب بهترین الگوریتم یادگیری ماشین

## پایگاه داده‌ها و معیارهای ورود

### MIMIC-IV

Medical Information Mart for  
Intensive Care

### انتخاب منبع داده‌ها از پایگاه MIMIC-IV

- ✓ تجزیه و تحلیل ما مبتنی بر جدیدترین داده‌های موجود
- ✓ شامل بیش از ۴۰۰۰۰ بیمار منحصر به فرد از بخش‌های مراقبت‌های ویژه
- ✓ اطلاعات پذیرش در جداول مختلف مانند اطلاعات جمعیت‌شناختی، نتایج آزمایشگاهی و اطلاعات ICU

## برای محدود کردن بیماران هدف

۱- فقط بیماران بزرگسال (۱۸ سال و بالاتر)

۲- با حداقل مدت بستری ۲۴ ساعت در بخش ICU

۳- با نمره ارزیابی نارسایی متوالی اندام (SOFA) ۲ یا بالاتر و مشکوک به عفونت

استفاده از **BigQuery** به عنوان ابزار استخراج داده برای انتخاب بیماران هدف از مجموعه داده‌ها

# انتخاب ویژگی و پیش پردازش داده‌ها

مجموعه داده‌های نهایی شامل ۳۸ ویژگی متمایز

|                        |                          |
|------------------------|--------------------------|
| اطلاعات جمعیت‌شناختی   | متغیرهایی مانند SOFA     |
| مصرف آنتی‌بیوتیک       | میانگین برون‌ده ادرار    |
| سابقه پزشکی بیمار      | حداقل و حداکثر سطح گلوکز |
| نتایج آزمایشگاهی مختلف | سطح سدیم                 |
|                        | ضربان قلب                |
|                        | فشار خون SBP و DBP       |
|                        | تعداد تنفس               |
|                        | اشباع اکسیژن SPO2        |
|                        | سطح آل‌بومین             |

## انتخاب ویژگی و پیش پردازش داده‌ها

- برای بهبود بیشتر مجموعه داده‌ها
- تعیین آستانه برای  $\text{PaO}_2/\text{FiO}_2$  برابر با 200  
به پیشنهاد یک متخصص بالینی
- لحاظ کردن نمره کمای بالاتر از ۸ برای بیماران کمایی  
محدود شدن داده‌ها به 6401 پرونده پذیرش

## انتخاب ویژگی و پیش‌پردازش داده‌ها

### پاکسازی مجموعه داده‌ها

- ۱- پرداختن به مقادیر تهی و تکراری در داده‌های عددی
- ۲- گروه‌بندی متغیرهای طبقه‌بندی شده موجود (نژاد و آنتی‌بیوتیک‌ها) در ویژگی‌های جدید
  - چهار گروه نژادی: سیاه‌پوست یا آمریکایی آفریقایی تبار، اسپانیایی یا لاتین تبار، سفیدپوست و سایر نژادها
  - برای آنتی‌بیوتیک‌ها، ۲۵ دسته در ۷ گروه : آمینوگلیکوزیدها، کارباپنم‌ها، گلیکوپپتیدها، اکسازولیدینون‌ها، پنی‌سیلین‌ها، سولفونامیدها و تتراسایکلین‌ها.

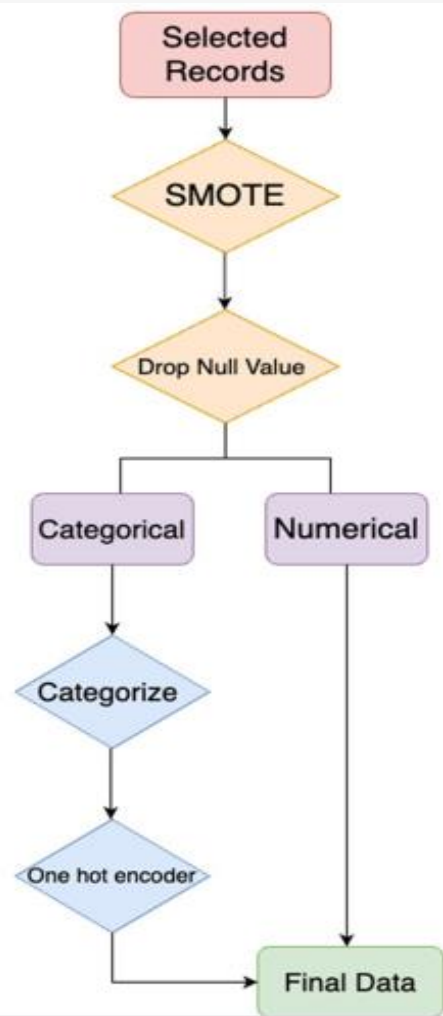
## انتخاب ویژگی و پیش پردازش داده ها

جهت رفع مشکل عدم تعادل بین داده های **train**

✓ برخلاف روش مراکز خوشه ای **cluster centroids method** استفاده از روش جدید

**SMOTE**

✓ گسترش نقاط داده از ۱۶۴۰۱ نمونه به ۴۷۳۰۴



## مدل سازی

دو روش ارزیابی کامل عملکرد مدل های مختلف طبقه بندی یادگیری ماشین

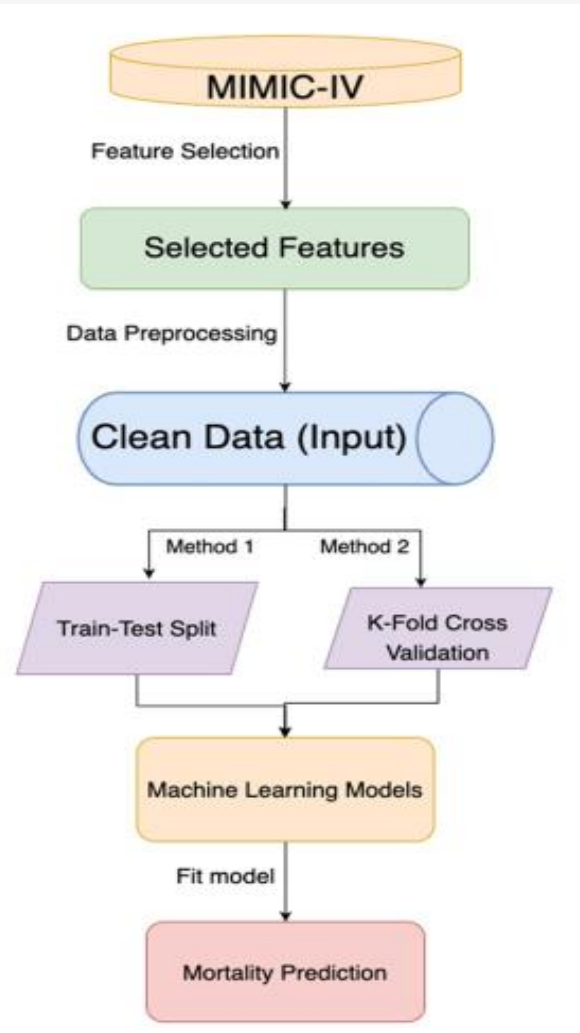
۱- Test and Train به روش ۸۰ به ۲۰ درصد

۲- 5-Fold cross validation

۳- الگوریتم نصف کردن متوالی و طبقه بندی (SHAC) به جای استفاده از روش قدیمی Grid Search به منظور کاهش overfitting و تنظیم hyperparameter

در نهایت استفاده از ۷ روش یادگیری ماشین و مقایسه عملکرد آنها

Decision Trees , Gradient Boosting , XGBoost , LightGBM , MLP, SVM ,  
Random Forest



# تحلیل آماری بین گروه های Test & Train

بررسی گروه ها با استفاده از روش های T-Test و کای دو

- استفاده از روش t-test دو طرفه برای ویژگی های عددی
- استفاده از روش Chi-square برای ویژگی های دسته بندی شده

همچنین استفاده از Python version 3.6 به عنوان ابزار تحلیل



04

یافته ها

## میانگین مدت اقامت در ICU

عدم وجود تفاوت معناداری  
در مدت اقامت در بخش

$$P-V = 0.988$$

- در گروه آموزش ۶/۹۴۷ روز
- در گروه تست ۶/۹۹۷ روز

## میانگین سنی افراد در ICU

- در گروه آموزش 65.160 سال
- در گروه تست ۶۶/۰۳۹ سال

$$P-V = 0.083$$

عدم وجود تفاوت معناداری  
در میانگین سنی افراد بخش

**Table 2** Statistical analysis results of train and test cohorts for numerical and categorical variables

| Feature                      | Train   | Test    | P-value |
|------------------------------|---------|---------|---------|
| <b>Admission Information</b> |         |         |         |
| los_icu                      | 6.974   | 6.977   | 0.989   |
| <b>Demographics</b>          |         |         |         |
| Age                          | 65.160  | 66.039  | 0.086   |
| gender_F                     | 42.188  | 41.842  | 0.848   |
| <b>Lab Results</b>           |         |         |         |
| SOFA_score                   | 4.317   | 4.393   | 0.305   |
| avg_urineoutput              | 160.971 | 153.108 | 0.007   |
| temperature_min              | 36.663  | 36.632  | 0.478   |
| temperature_max              | 37.149  | 37.145  | 0.920   |
| temperature_avg              | 36.906  | 36.888  | 0.652   |
| glucose_min                  | 110.363 | 111.451 | 0.414   |
| glucose_max                  | 211.147 | 214.524 | 0.453   |
| glucose_average              | 154.669 | 156.739 | 0.332   |
| sodium_min                   | 135.376 | 135.385 | 0.961   |
| sodium_max                   | 141.558 | 141.768 | 0.222   |
| sodium_average               | 138.499 | 138.595 | 0.521   |
| heart_rate_min               | 70.250  | 69.694  | 0.271   |
| heart_rate_max               | 114.752 | 114.909 | 0.819   |
| heart_rate_mean              | 89.384  | 89.530  | 0.763   |

|                      |         |         |       |
|----------------------|---------|---------|-------|
| heart_rate_mean      | 89.384  | 89.530  | 0.763 |
| sbp_min              | 81.941  | 81.595  | 0.489 |
| sbp_max              | 154.473 | 153.932 | 0.475 |
| sbp_mean             | 114.334 | 114.630 | 0.495 |
| dbp_min              | 41.534  | 41.300  | 0.490 |
| dbp_max              | 93.673  | 93.545  | 0.849 |
| dbp_mean             | 61.570  | 61.466  | 0.725 |
| resp_rate_min        | 11.932  | 11.804  | 0.307 |
| resp_rate_max        | 31.218  | 31.304  | 0.706 |
| resp_rate_mean       | 20.510  | 20.529  | 0.881 |
| spo2_min             | 89.180  | 89.023  | 0.586 |
| spo2_max             | 99.732  | 99.755  | 0.362 |
| spo2_mean            | 96.887  | 96.868  | 0.801 |
| hospital_expire_flag | 28.242  | 30.445  | 0.127 |

#### Comorbidities

|                      |        |        |       |
|----------------------|--------|--------|-------|
| diabetes_without_cc  | 31.641 | 33.880 | 0.133 |
| diabetes_with_cc     | 10.117 | 9.446  | 0.506 |
| severe_liver_disease | 10.449 | 9.758  | 0.499 |
| aids                 | 1.035  | 0.937  | 0.874 |
| renal_disease        | 26.738 | 27.244 | 0.741 |

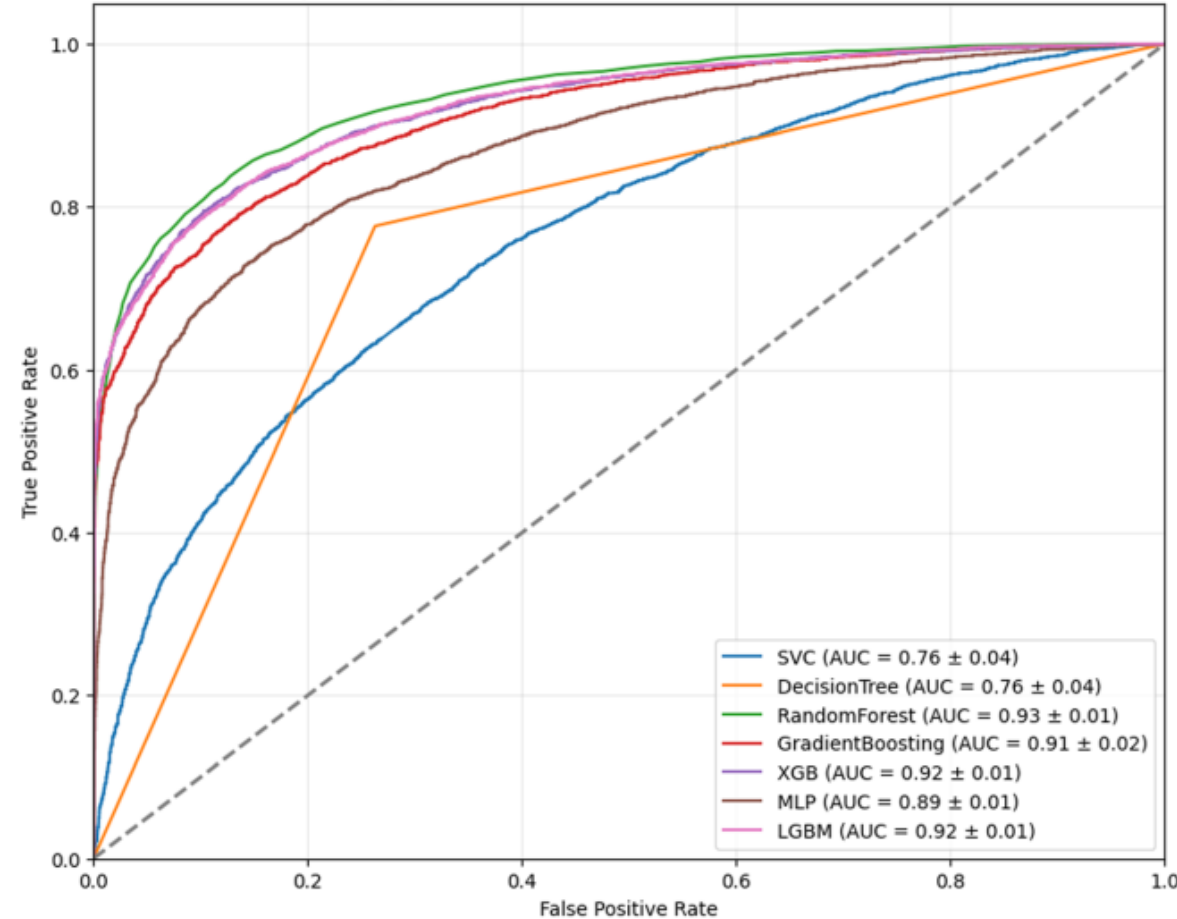
#### Medications

|                           |        |        |       |
|---------------------------|--------|--------|-------|
| antibiotic_Carbapenem     | 15.664 | 14.910 | 0.533 |
| antibiotic_Aminoglycoside | 7.168  | 7.962  | 0.360 |
| antibiotic_Glycopeptide   | 65.039 | 66.042 | 0.521 |
| antibiotic_Oxazolidinone  | 0.195  | 0.156  | 1.000 |
| antibiotic_Penicillin     | 0.000  | 0.078  | 0.453 |
| antibiotic_Sulfonamide    | 0.723  | 1.093  | 0.247 |
| antibiotic_Tetracycline   | 0.000  | 0.078  | 0.453 |

## بهترین مدل

بهترین عملکرد مدل جنگل تصادفی  
(Random Forest) با استفاده از  
هر دو روش Test and Train و  
cross-validation، از سایر مدل‌ها.

Receiver Operating Characteristic (ROC) Curves with AUC Confidence Intervals



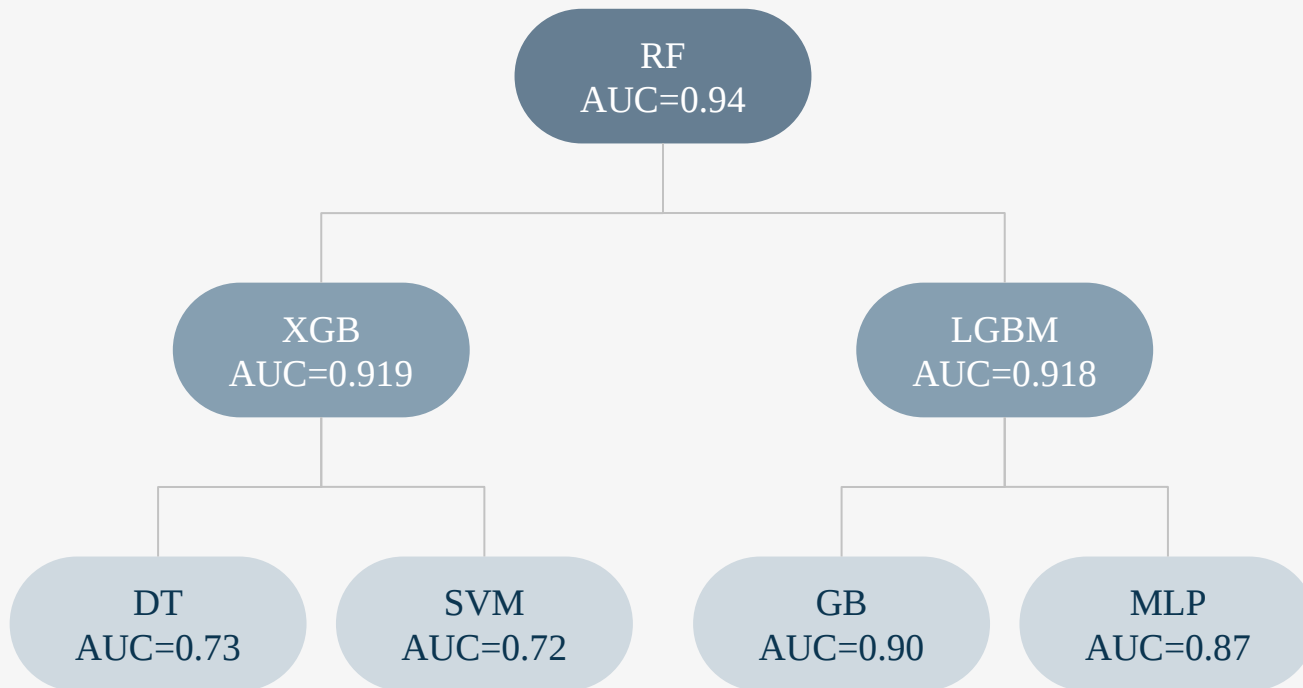
## بهترین مدل

**Table 3** Metrics for baseline models on test set and random forest on different sets

| Model                 | AUROC  | Precision | Sensitivity | Accuracy | F-score |
|-----------------------|--------|-----------|-------------|----------|---------|
| Decision Tree         | 0.7329 | 0.7219    | 0.7475      | 0.7327   | 0.7345  |
| SVM                   | 0.7277 | 0.6772    | 0.6179      | 0.6654   | 0.6462  |
| GradientBoosting      | 0.906  | 0.8469    | 0.7597      | 0.8133   | 0.8009  |
| XGB                   | 0.9192 | 0.8491    | 0.8228      | 0.8401   | 0.8358  |
| MLP                   | 0.8763 | 0.7495    | 0.835       | 0.7804   | 0.7899  |
| LGBM                  | 0.9185 | 0.8521    | 0.7973      | 0.8313   | 0.8238  |
| RF (Test Set)         | 0.9388 | 0.876     | 0.8372      | 0.8631   | 0.857   |
| RF (Trainig Set)      | 1.0000 | 1.0000    | 1.0000      | 1.0000   | 1.0000  |
| RF (Cross Validation) | 0.9293 | 0.8599    | 0.8312      | 0.8475   | 0.8453  |

برای هر دو تقسیم آزمون آموزش و اعتبارسنجی متقابل، جنگل تصادفی بهترین نتیجه را نشان داد. از این رو، جنگل تصادفی (RF) مدل پیشنهادی در این مقاله است که به امتیاز **AUC 0.94** با فاصله اطمینان ۰/۰۱ دست یافته است

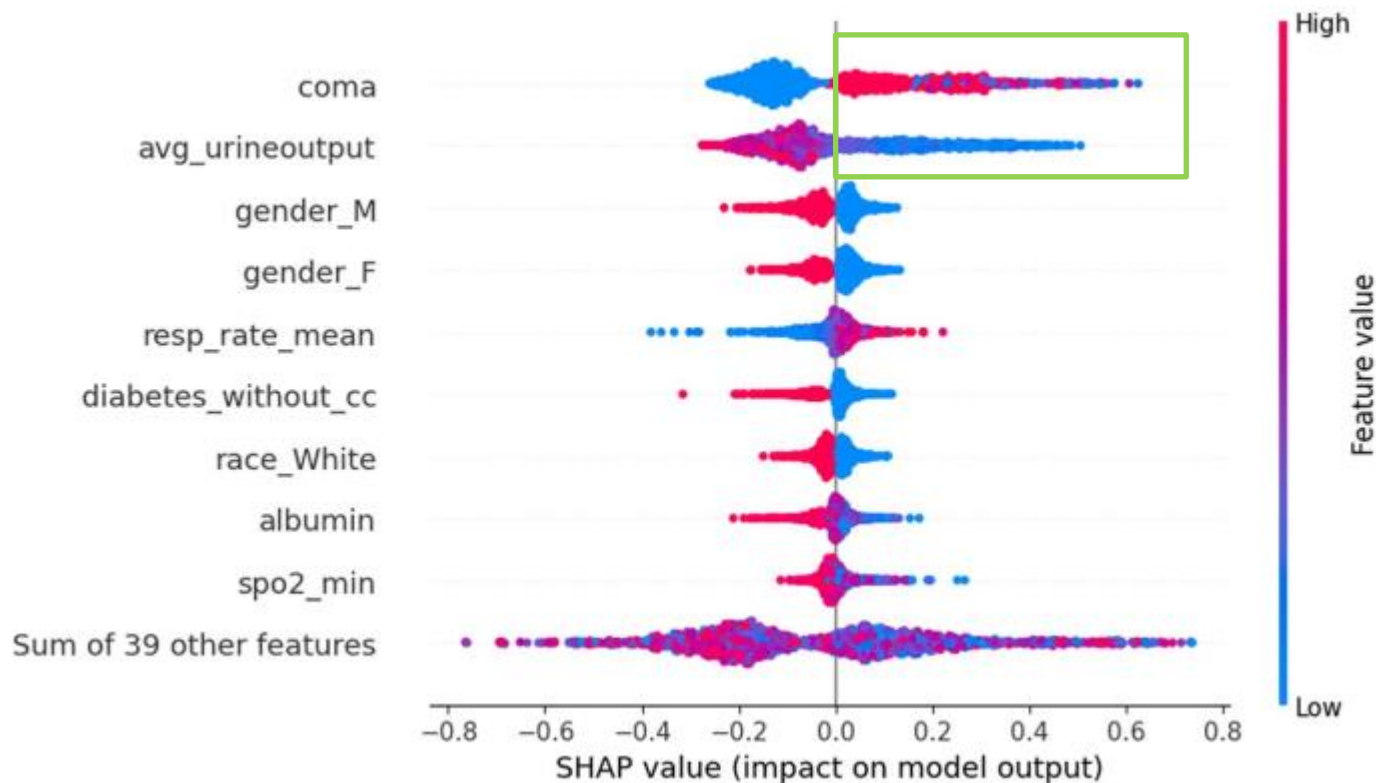
# Machine Learning Algorithms Performance League !



## تحلیل مقادیر شپلی (Shapely)

- ✓ برای ارزیابی اهمیت یک ویژگی در چارچوب جنگل تصادفی
- ✓ ۳ یافته مهم در نتایج نمودار SHAP
- ۱- امتیاز کما بیشترین تأثیر را بر پیش‌بینی مرگ و میر ناشی از sepsis (افزایش احتمال مرگ و میر با افزایش نمره کما نیز افزایش می‌یابد)
- ۲- میانگین خروجی ادرار (افزایش مرگ و میر با میانگین خروجی ادرار پایین‌تر)
- ۳- جنسیت مرد (مرد بودن با خطر مرگ و میر کمتر در مقایسه با زنان)

# Shapley Values Analysis





05

نتیجه گیری



## نتایج اصلی مطالعه

- ✓ ارائه پیشنهادهای شخصی سازی شده برای هر بیمار از طریق fitting مدل
- ✓ گروه بندی داده ها و راه حل های مؤثر برای مشکلات عدم تعادل داده ها
- ✓ تکیه بر تعداد محدودی از ویژگی ها برای تولید پیش بینی های بسیار دقیق
- ✓ ویژگی های کمتر در نتیجه تفسیرپذیرتر بودن داده ها
- ✓ کمک به پزشکان و متخصصان بالینی برای تمرکز بیشتر بر جنبه های دیگر بیماری

بیمارانی که نمره کما بالاتر، میانگین خروجی ادرار پایین تر و جنسیت زن دارند، بیشتر در معرض خطر مرگ و میر ناشی از سپسیس قرار می گیرند

## محدودیت ها

- MIMIC-IV تنها منبع داده
- پیچیدگی و تفسیر پذیری مشکل در الگوریتم های یادگیری ماشین
- استفاده از مجموعه داده های گذشته نه به هنگام

## کارهای آینده

- مجموعه داده های اضافی مانند eICU
- حل مشکل تفسیر پذیری مدل
- ایجاد یک جریان داده **real-time**
- ابزارهای ساده سازی داده ها، مانند **Google Cloud Dataflow**



06

پیشنهادات من!



# پیشنهادهای من

همکاری با موسسات تخصصی مرتبط از جمله  
NIH , BIDMC , CSR



طراحی یک پلتفرم با حفظ دقت بالا و  
کارایی بیشتر برای پزشکان



ایجاد سیستم‌های پیش‌بینی به صورت جریان  
داده‌های زنده برای بیماران بستری در ICU



# Thanks!

Do you have any questions?