Machine Learning Solutions

for early sepsis detection

1

Introduction

Sepsis is one of the leading causes of morbidity and mortality in hospitals.

The fundamental need for early detection and treatment remains unmet.

This resume is the first draft to outline how one could proceed in developing an MLA solution for questions like: early detections, survival or length of stay.

2

What is sepsis and why would a ML solution be helpful

Sepsis is a major health crisis. It is one of the leading cause of hospital admissions and mortality world wide.

→ Play a role in 50% of hospital deaths.
→ 270.00 fallacies each year US alone.
→ 6 Millions world wide.
→ 24 Billion each year in the US, 13% of hospital budgets.
→ 4.2 million newborns and children are each year affected.

- Sepsis is a life threatening condition caused by your body' s response to an infection.
- In cases of septic shock, the risk of dying increases by approximately 10% for every hour of delay in receiving antibiotics.
- Early detection of sepsis events is essential improving sepsis management and mortality rates in the ICU.
- Few electronically monitoring methods of patients provide predictive capabilities to enable early intervention.

3

Objetive

The goal of this work is the early detection of sepsis using physiological data. The early prediction of sepsis is potentially life-saving, and we aim to predict sepsis at least 6 hours before the clinical prediction of sepsis.

- Q1: Early sepsis prediction
- Q2: Predict patient admission
- Q3: Predict severeness and survival
- Q4: Predict hospital length of stay (LOS)

4

What are sepsis Indicator

- Common scores are such as the Modified Early Warning Score (MEWS), the SIRS criteria and the Sequential Organ Failure Assessment (SOFA)

- Sepsis: suspicion of infection with at least two SIRS criteria. SIRS criteria are defined as:

  →heart rate >90 beats per minute
  →body temperature >38°C or <36°C
  →respiratory rate >20 breaths/min or PaCO2 (alveolar
  →carbon dioxide tension) <32 mm Hg
  →white cell count >12 x 109 cells/L or <4 x 109 cells/L.10

- Designed to predict patient risk, rather than specifically diagnose sepsis

5

DataSet

- The date obtained is a 40.000 patient strong dataset which was made public for the 2019 Physionet Challenge

- The data has 40 features which can broadly be classified into

  → 8 Vitals Signs – Heart Rate, Temperature, MAP, …
  → 28 Laboratory Values – FiO2, Lactate, Bilirubin, …
  → 6 Demographics – Age, Gender, Hospital Unit, …

- There are two ways in which one can approach this problem:

  1. Temporal Approach: Take into the account the time component for the data. Sepsis is diagnosed for each patient at each hour using the past data.
  2. Non-temporal Approach: Ignore the time component and treat record as independently and identically distributed. This approach would help in predicting Sepsis at each hour for any patient(with or without patient past data).

6

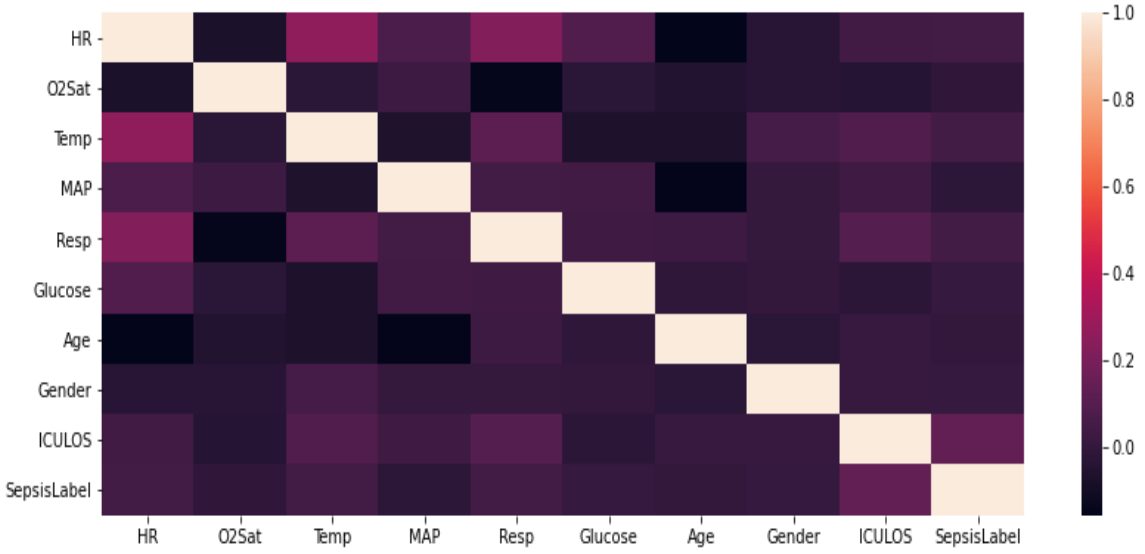
Data processing pipeline and feature selection

- We eliminate variables which have greater than 82% missing values, except Bilirubin direct, Lactate, Partial thromboplastin time (PTT), Creatinine, Leukocyte count (WBC) and Glucose which are known as significant variables for detecting sepsis.

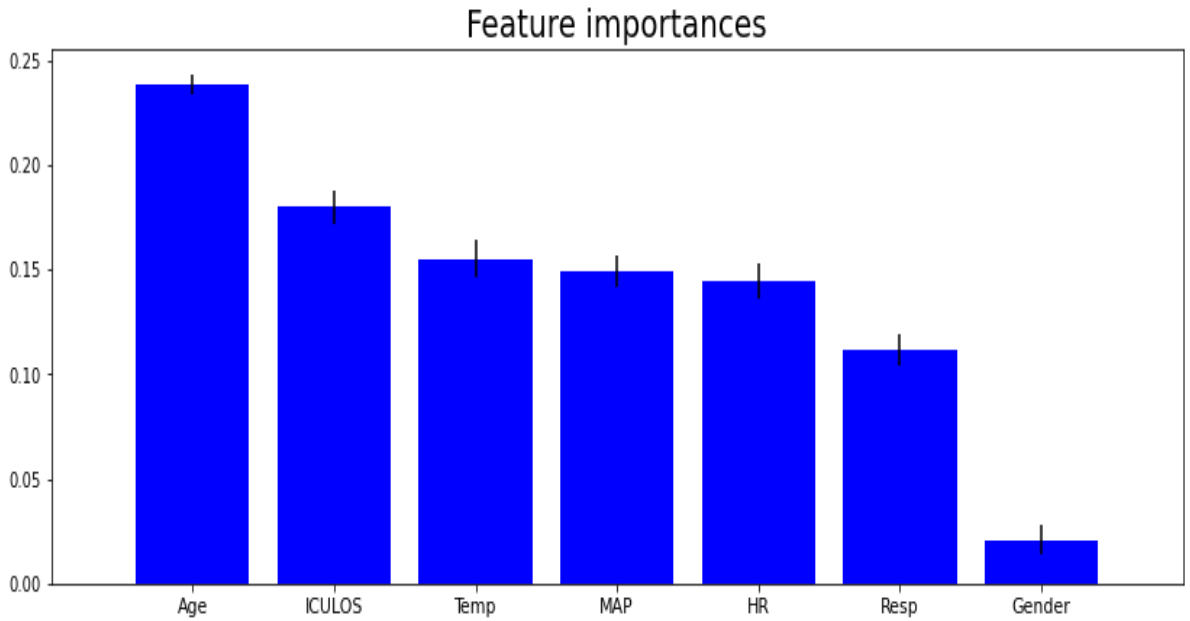
- Eliminate SBP and DBP, since MAP = (SBP + 2*DBP) / 3

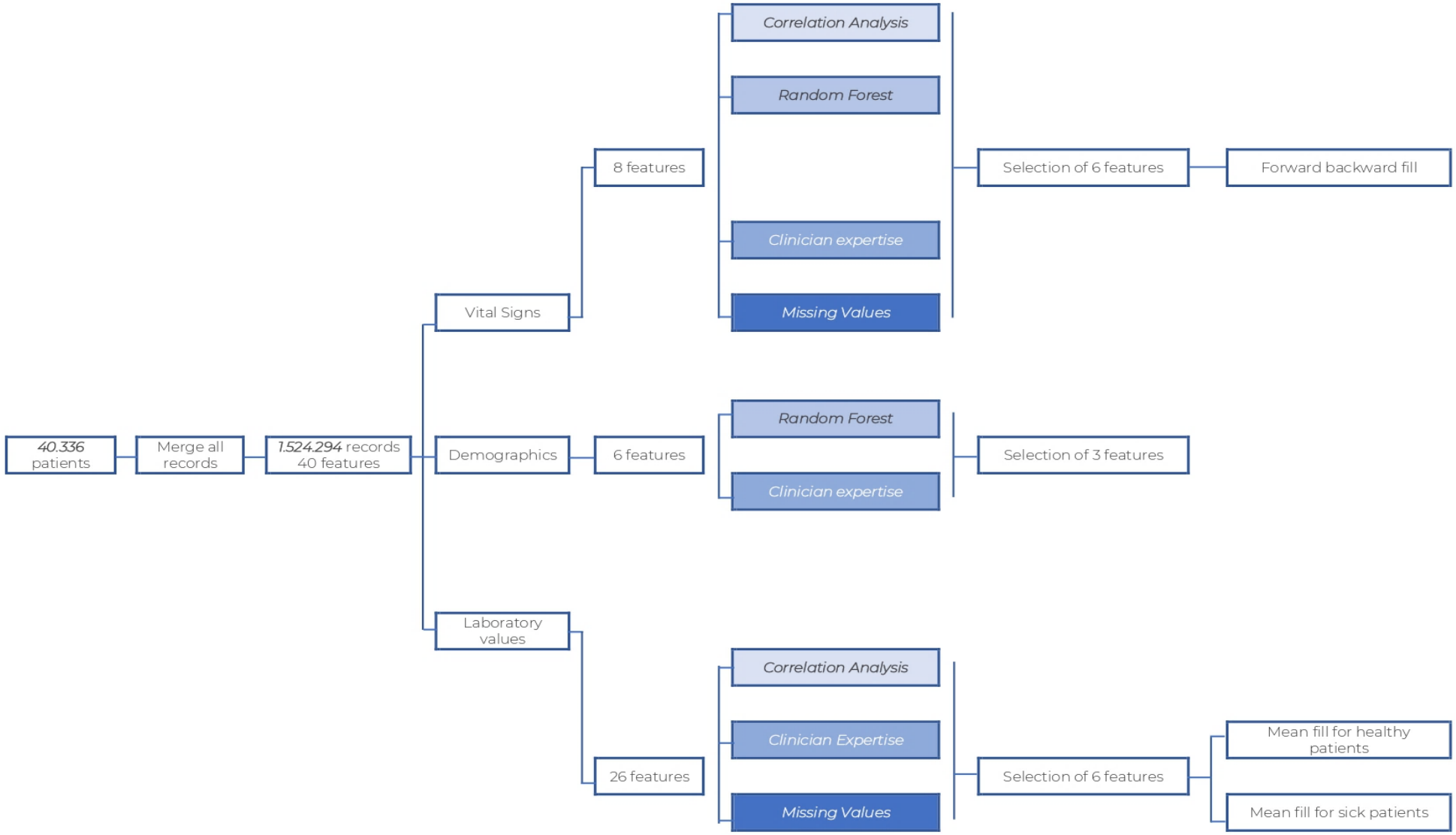
- For the vital sign’ s variables, if a patient did not have a measurement in a given hour, the missing measurement was filled in using carry-forward imputation

- Correlation heatmap of selected features:

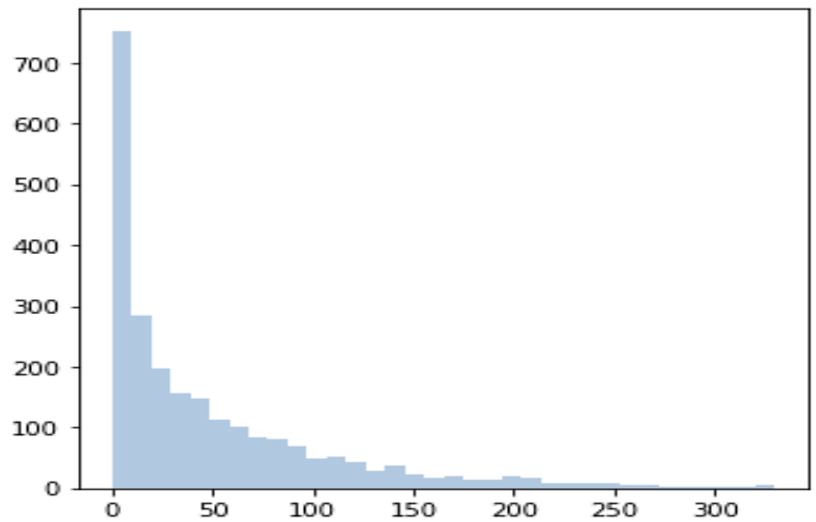


- Random Forest importance features computation:

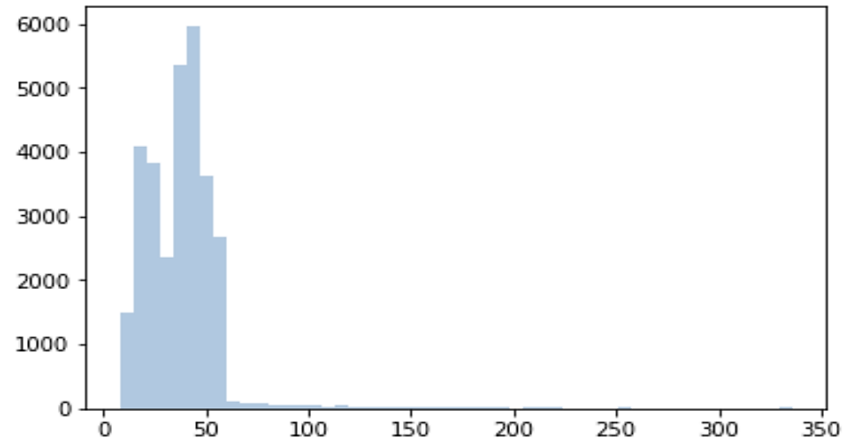


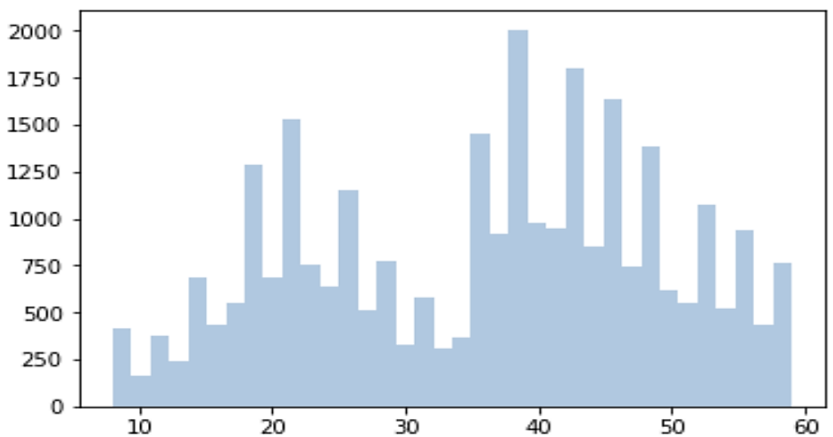

40.336 patients
Merge all records
1.524.294 records 40 features
Vital Signs
8 features
Correlation Analysis
Random Forest
Clinician expertise
Missing Values
Selection of 6 features
Forward backward fill
Demographics
6 features
Random Forest
Clinician expertise
Selection of 3 features
Laboratory values
26 features
Correlation Analysis
Clinician Expertise
Missing Values
Selection of 6 features
Mean fill for healthy patients
Mean fill for sick patients

- Time at which a patient contracts Sepsis:



- Number of hours a patient spends in Hospital





7

Models

- To get an Baseline one can quickly run a mix of standard ML algorithms to get familiarized the the dataset and its limits, such as

  → Logistic Regression
  → AdaBoost
  → Gradient Boosting
  → Random Forest

- Autoencoders / anomalies detection
- DNN or GXBOOST

SIGNATURE APROCHE

1. Hand-Crafted features

   → ShockIndex : Heart Rate / Systolic Blood Pressure

   → BUN/CR: Bilirubin / Creatinine

   → PartialSOFA Score of the SOFA components that are found in the challenge data

2. Signature Features

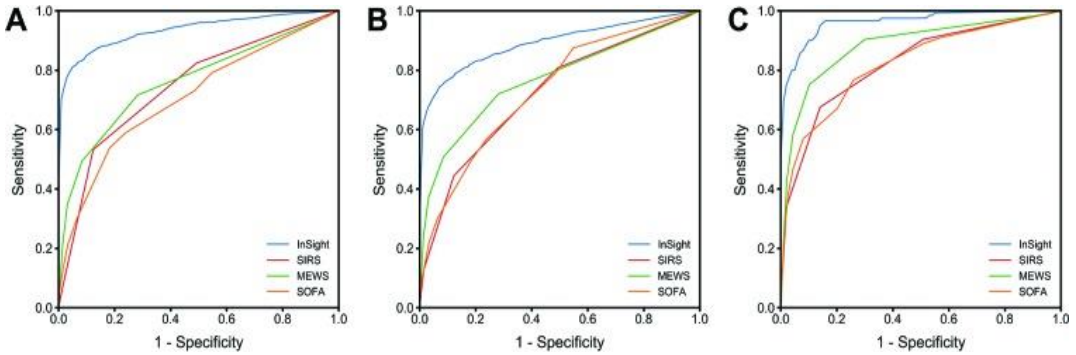
   The signatures of 'PartialSOFA, 'MAP' and 'BUN/CR' are computed with a time dimension and the lead-lag transformation and then signatures of all non-stationary columns are computed after first applying the cumulative sum followed by the lead-lag transformation.

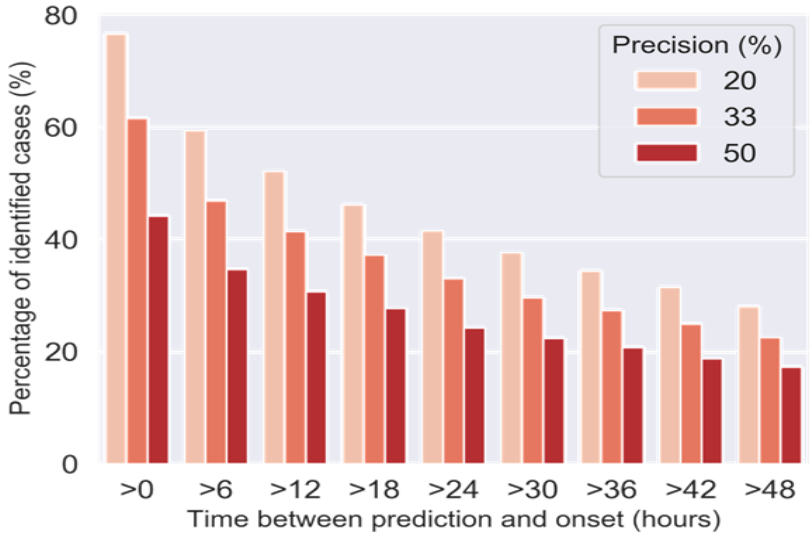
Train an LSTM and use the output as a feature in the LGBM Regressor with a stratiified 5-fold croos validation method.

8

Monitization

- An market ready MLA can be integrated with all major EHR systems.
- Gold standards test SOFA, SIRS, SEWS and InSight with AUROC curve of 0.725, 0.609, 0.803 and over 0.90, respectively.
- The Signature model often detects sepsis much further in advance, regularly over 24 hours





Jonas Grabbe

grabbe.jonas@gmail.com

+34 688 90 89 62