# Sepsis Prediction using MIMICIII database

**Viral Doshi(vdoshi30), Ross Broderson (rbroderson3), Gauthier Husson (ghusson3)**

**Abstract**

***Sepsis is one of the leading causes of mortality in hospitalized patients. Despite this fact, a reliable means of predicting sepsis onset remains elusive. Early and accurate sepsis onset predictions could allow more aggressive and targeted therapy while maintaining antimicrobial stewardship. Prior models for the early detection of sepsis have typically relied on either manual chart review or a small number of hand-selected features. The use of vast amounts of digital medical information can help in predicting the best course of action for the diagnosis and treatment of patients. Existing detection methods suffer from low performance and often require time-consuming laboratory test results. We propose utilizing the full numeric and categorical entries in the MIMIC-III database to extract relevant features for sepsis classification. The proposed technique investigates using different models and structured data to retrospectively identify sepsis cases with high performance.***

## Introduction

Sepsis is a complication of severe infection that arises when the body responds to a systemic inflammatory response caused by an injury to its own tissues and organs. It is a well-recognized worldwide healthcare issue, ultimately resulting in significant mortality, morbidity and resource utilization during and after critical illness[3]. It is important to know as soon as possible if patients are affected by sepsis, as it will be easier to treat them. Sepsis affects over a million patients annually and remains one of the largest contributors to mortality in the ICU, accounting for more than $23.6 billion (6.2%) of total US hospital costs in 2013.Early and accurate prediction of the onset of sepsis could facilitate effective and targeted treatment, which could, in turn, reduce the patient death rate and lower the risk of organ damage for each one-hour delay in the administration of antibiotic treatment in a case of sepsis, the mortality rate increases by 7%[5]. We can help physicians predict and detect sepsis at an early stage, using the data regularly collected on each patient in the ICU**.**

## Related Work

Early and accurate prediction of the onset of sepsis could facilitate effective and targeted treatment, which could, in turn, reduce the patient death rate and lower the risk of organ damage. E. Sheetrit et al[4] used temporal data mining methods to predict and detect sepsis at an early stage using the data regularly collected on each patient in the ICU.

T. Desautels et al[10] developed InSight, for the new Sepsis-3 definitions in retrospective data, make predictions using a minimal set of variables from within the electronic health record data. However, they recognized several limitations. They chose only a subset of the patients in the MIMIC-III database, the study was performed exclusively on ICU data and at a single center. They used the definition of onset from Singer et al which could be problematic. Early and accurate sepsis onset predictions could allow more aggressive and targeted therapy while maintaining antimicrobial stewardship. Existing detection methods suffer from low performance and often require time consuming laboratory test results[10]. The aim of this project is to develop a classification model for retrospectively identifying hospital admissions that result in the development of Sepsis. To achieve this, we process clinical data from the MIMIC-III database to form a set of features for each hospital admission that can be used for classification.

## Modeling Overview

In order to model for Sepsis prediction, we developed a classification model for retrospectively identifying hospital admissions that result in the development of Sepsis. To achieve this, we process clinical data from the MIMIC-III database to form a set of features for each hospital admission that can be used for classification.

## Database Description

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely available database comprising de identified health related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the

bedside ( 1 data point per hour), laboratory test results, procedures, medications, caregiver motes, imaging reports, and mortality ( both in and out of the hospital). MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors: it is freely available to researchers worldwide, it encompasses a diverse and very large population of ICU patients, it contains high temporal resolution data including lab results, electronic documentation and bedside monitor trends and waveforms.

The key tables from the MIMIC III database that have been used for this study include:

ADMISSIONS - The ADMISSIONS table gives information regarding a patient's admission to the hospital.

Number of rows: 58976

DIAGNOSES_ICD - Contains ICD diagnoses for patients, most notably ICD-9 diagnoses.

Number of rows: 651,047

PRESCRIPTIONS - Contains medication related order entries, i.e. prescriptions.

Number of rows: 4,156,450

CPTEVENTS - The CPTEVENTS table contains a list of which current procedural terminology codes were billed for which patients. This can be useful for determining if certain procedures have been performed (e.g. ventilation)

Number of rows: 573146

LABEVENTS- Contains all laboratory measurements for a given patient, including out patient data

Number of rows: 27,854,055

PROCEDURES_ICD - Contains ICD procedures for patients, most notably ICD-9 procedures.

Number of rows: 240,095

**Cohort Definition**

To perform classification, we have labelled hospital admissions with positive Sepsis cases as 1 and negative Sepsis cases as 0. This selection is based on the criteria laid out by Angus et. Al [9] as their criteria was used in similar work to identify cases for sepsis. Angus et. Al labels cases as sepsis if the ICD-9 billing codes for the case contains a bacterial or fungal infection and either an acute organ dysfunction or if the patient required mechanical ventilation. We used the following ICD-9 diagnoses codes to label the cases as positive for sepsis:

99592 – severe sepsis

78552 – septic shock

In order to label negative cases for sepsis, we used the definition of sepsis from Angus et. Al as stated above and used ICD-9 codes for neither bacterial or fungal infectious process nor diagnoses of acute organ dysfunction or mechanical ventilation.

From the MIMIC III dataset 4085 cases have been identified as positive cases for sepsis on admission and 46535 cases as negative cases for Sepsis.

**Technical Approach**

The entire MIMIC III dataset in the form csv files was loaded into AWS S3 buckets for analysis and later for modeling. I developed some python scripts to do an initial analysis of the data by loading the csv files into Pandas dataframe and running several Sql queries on it. Feature generation and modeling was done using Apache Spark which is an open source distributed computing framework. Spark was also the ideal choice since it provides a general machine learning library MLlib that is designed for simplicity, scalability and easy integration with other tools. It also provides a Python based API called PySpark. Pyspark makes it very simple to write parallelized code and provides implementation and evaluation of algorithms like Logistic Regression and Random Forest. The feature vectors are stored in SVMLite format and Spark ML's Multiclass Classification Evaluator was used to compare the performance of the machine learning algorithms on the given feature vectors. The experiments using spark were carried out on a small subset of the data on a Macbook Pro with 16GB of RAM and the complete dataset used an EMR cluster with a master node and 2 slave nodes. The nodes were AWS EC2 m4.xlarge instance with 4 cores and 16 GB of RAM.

**Feature Generation**

To construct the feature vectors, we used admissions data to extract general patient and admission information. Along with this, we listed the most common medications from the prescription table, procedures from the CPT events table and lab tests from the lab event table. These features were limited to those within 24 hours after admission. To limit the features to the 24 hours window we calculate the 24 hour admission time and added it as a feature. For admissions table, which contains demographic information of patients, we convert categorical variables such as admission type, insurance and location to counts and append these to the feature set. For medications feature, we use the prescriptions table and extracted the drug information which was prescribed to the patients within the first 24 hours of admission. We also filter to 100 most common prescriptions made. For the lab feature, we selected laboratory tests information from the lab events table for the tests which were performed in the first 24 hours of admission and selected 100 most common lab tests. For procedure features, we extracted CPT

information from the CPTEVENTS table. This was limited to the 100 most common procedures. The above set of features formed our final feature set which was then cleaned up to remove special characters and filled missing values(NA) with 0.

**Modeling:**

a) Split into training and validation sets

We randomly select 80% as the training data, and the remaining 20% as test set or validation set. It is important to set seed for the randomSplit() function in order to get same split for each run. This is the crucial step for the success of the subsequent tests later on.

b) Classification Methods

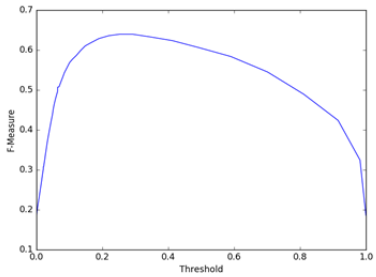
i) Logistic Regression

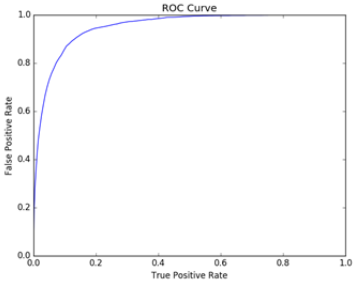
For the purpose of classification, we applied logistic regression with L1 regularization. The model is fit by minimizing the following loss function:

$$L(\theta) = \sum_i (y^{(i)} - \theta^T x^{(i)})^2 + \lambda ||\theta||_1$$

The performance curves generated from running Logistic regression are show below

AUC: 0.929317474123



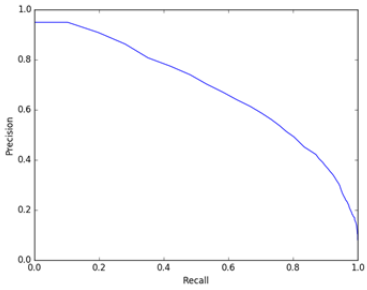




Figure 1: ROC, F-measure and Precision curves

ii) Random Forest

We also used Random Forest to perform classification on the model. Random forest is an ensemble of decision trees where each decision tree is constructed on a random bootstrap sample of the data set where at each split point in the tree a random bootstrap sample of the features are considered as candidate features and the optimal feature and decision rule are determined such that the heterogeneity of the class distribution in the parent node is optimally decreased upon assigning samples to the child nodes. Predictions are made by evaluating the sample on each of the individual trees and then using the aggregate of the predictions across all trees in the ensemble.

**Results and discussions:**

| ML Algorithm | F1 Score | Precision | Recall |
|---|---|---|---|
| Logistic Regression | 0.937319695095 | 0.936177722787 | 0.94274393264 |
| Random Forest | 0.897491762908 | 0.916133437225 | 0.925705794948 |

Table 1: Comparison of Machine learning algorithms

The performance results are summarized in table above. We can see that the logistic regression model performs better than the random forest model on the test set for each of the feature sets considered. Based on the results, this happens because f(x)=Pr[Y=1 ‖ X=x]f(x)=Pr[Y=1 ‖ X=x] is very smooth in x, and furthermore approximately obeys the logistic curve, and hence the logistic model outperforms Random Forest. Based on the performance it looks like there is a nearly linear dependence on the covariates and hence the logistic model performs better than random

forest. Random forest is not very efficient to approximate a high dimensional linear relationship with a series of step functions.

## Conclusion

We developed a pipeline that extracts informative features to characterize sepsis patients from EHR in an unbiased manner. We speculate that random forest performs better compared to logistic regression in the classification task because it can capture nonlinear relationship among features. In this project, we focused on numeric and categorical data in MIMIC-III database. In the future, we could expand our feature space so that that we handle images and clinical notes. Additionally, we could further optimize the parameters in our models, such as the number of trees in random forest model, to avoid overfitting to the training data set. While our classification pipeline is useful for early detection of Sepsis, one of the limitations of this work is that the classification models are only valid for retrospective classification. However, with this classification model in place, it is likely that similar unbiased methods of deriving features from the EHR may perform well in the real-time risk prediction setting with the help of some algorithm to determine the precise time of septic event.

## Supplemental Material

Presentation Video link:
**https://youtu.be/_XRwh_jdE9M**

## References


1. M. Rothman, M. Levy, R. Philip Dellinger, S. L. Jones, R. L. Fogerty, K. G. Voelker, Barry G., A. Marchetti, J. Beals IV, Sepsis as 2 problems: Identifying sepsis at admission and predicting onset in the hospital using an electronic medical record–based acuity score 2016
2. Y. Arabi, N. Al Shirawi, Z. Memish, S. Venkatesh, A. Al-Shimemeri (2003, June) Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study 2003
3. B. Adler Maccagnan Pinheiro Besen, T. Gomes Romano, A. Paulo Nassar, Jr., L. Utino Taniguchi, L. Cesar Pontes Azevedo, P. Vitale Mendes, F. Godinho Zampieri, Marcelo Park, Sepsis-3 definitions predict ICU mortality in a low–middle-income country, 2016
4. E. Sheetrit, Nir Nissim, D. Klimov, L. Fuchs, Y. Elovici, Y. Shahar, Temporal Pattern Discovery for Accurate Sepsis Diagnosis in ICU Patients 2017
5. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG, MIMIC-III, a freely accessible critical care database 201
6. K.E.Henry, D.N.Hager, P.J.Pronovost, S.Saria, A targeted real-time early warning score (TREWScore) for septick shock
7. F.Gwadry-Sridhar, A.Hamou, B.lewden, C.Martin, M.Bauer, Predicting Sepsis, A Comparison of Analytical Approaches
8. J.Guillen, J.Liun, M.Furr, T.Wang, S.Strong, C.C.Moore, A.Flower, L.E.Barnes, Predictive Models for Severe Sepsis in Adult ICU Patients
9. DC.Angus, WT.Linde-Zwirble, J.Lidicker, G.Clermont, J.Carcillo, MR.Pinsky, Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care
10. T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D Feldman, C. Barton, D. J Wales, and R Das, Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach