Sepsis is a life-threatening disease, which claims millions of people's life each year. Early detection and antibiotic treatment of sepsis are critical for improving survival rate of sepsis. For the missing data, if there is the previous term, we impute the missing data using the previous term. And if there is no previous term, we will impute the missing data using the post term. And if neither the pre-term nor the post-term, we will impute the missing data using the global mean. Then, we designed 280 features for each time step, where each variable provide 6 statistics features, including minimum, maximum, mean, standard deviation, skew and number of measurements until the current time step. In total, we obtain 40 + 40*6 = 280 features for each time step. Finally, we used XGBoost (eXtreme Gradient Boosting) to build a Gradient Boosting Decision Tree (GBDT) model to predict the definitions for sepsis.

In our experiment, the original data (5000) was divided into training set (3200), validation set (800), and testing set (1000). In order to better evaluate the result, the K-fold cross-validation and the grid search strategies are used to find the best hyper parameter. We calculated the top 10 importance feature from XGBoost model, including HospAdmTime, ICULOS, EtCO2, Age, Temp, HR, MAP, respiration, O2Sat, Unit1. Our results show that HospAdmTime (Hours between hospital admit and ICU admit) and ICULOS (ICU length-of-stay) are the two most important features for sepsis detection. Moreover, we also find the significance of vital signs is larger than the significance of laboratory values.

In conclusion, the AUC value of the XGBoost model can reach 0.82. And the utility value, which is the official metrics score, achieves 0.93.