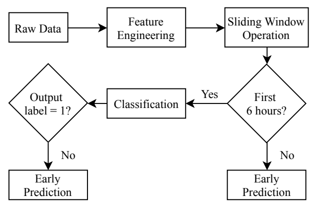
Early Prediction of Sepsis: Using State-of-the-art Machine Learning Techniques on Vital Sign Inputs

Manmay Nakhashi*, Anoop Toffy, Achuth PV, Lingaselvan Palanichamy

Tricog Health Services Private Limited, Bengaluru, Karnataka, India

We created a training dataset of 40336 from 45336 patient files provided by the Physionet challenge. Missing samples issue within the dataset was handled using column-wise forward-backward filling technique. We used synthetic minority oversampling technique to handle the class imbalance problem.



Our proposed approach for early prediction of sepsis.

We identified extreme gradient boosting (XGBoost) algorithm (objective: binary-logistic, n_estimators: 1000, learning_rate: 0.01) as the preferred option for classification and early-prediction based upon the data. A set of time series features were generated using *tsfresh* from all the vital signs. From this, 18 relevant features were selected based on their low cross-correlation values.

We trained a model each for the classifier and the early-predictor. The input data to the early-predictor was obtained using a three-hour long (two-hour overlap) sliding window every hour. Training label for early-predictor was from the label at $(n+6)^{\mathrm{th}}$ hour, where n is the current hour. We trained the classifier over each one-hour data.

Until the first six hours, data given by the sliding window will be insufficient for the early-predictor. So, both the classifier and the early-predictor were used to confirm the presence of sepsis. After six hours, only the early-predictor is used to predict sepsis six hours before its clinical recognition.

The models were evaluated using k-fold cross-validation ($k = 5$). We obtained an area under the curve (AUC) of 0.81 and a Physionet utility score of 0.12. The results indicate that it is possible to early-predict sepsis with moderate accuracy using vital sign inputs.