[BIOINFO 2023]

Development of Data/Model-based Ensemble Machine Learning Model for Prediction of Sepsis Severity Using Sepsis Genomic Synthetic Data

Sooyoung Jang[1], Chanyoung Ko[1], JongHyun Kim[1], Yu Rang Park[1,*]

[1] *Department of Biomedical systems informatics, Yonsei University College of Medicine*

**Corresponding author: yurangpark@yuhs.ac*

Background:
Sepsis denotes the body's inflammatory response to infection. Sepsis may progress to severe sepsis, resulting in multi-organ failures. Therefore, early recognition of sepsis and predicting sepsis severity are clinically valuable. Machine learning(ML) models trained on health data have shown good performance in predicting diagnostic status, treatment response, and disease severity. ML model training requires large volumes of data but privacy issues when accessing patient data proves to be a major obstacle. One method for overcoming such limitations is the use of synthetic data. Previous study suggested that early biomarkers of sepsis may be extracted from single cell RNA sequencing(scRNA-seq) data. We aim to develop ML models for predicting sepsis severity using genomic synthetic data.
Method:
The study was conducted using scRNA-seq data collected at 3 time points from the sepsis cohort (n=8) of Severance Hospital. We selected 136 genes that were highly expressed in the scRNA-seq of CD8 cells (2624 cells). Synthetic genomic data was generated using the REaLTabFormer model based on the expression values of the genes. The real data was split 8:2 into a train set and a test set. We developed the model to estimate SOFA scores over 2, at the cellular level using real and synthetic genomic data. We conducted the model ensemble, which aggregates the results of both XGB(xgboost) and LGBM(LightGBM) with soft-voting, and the data ensemble, which uses both real and synthetic data for training. The data/model-based ensemble was conducted, which performs the model ensemble and data ensemble simultaneously.
Result:
When training with XGB and LGBM based on real data, the AUROC(Area Under the Receiver Operating Characteristics) was found to be 0.815 and 0.837, respectively. Using synthetic data, the AUROCs were 0.742 and 0.784. The AUROC was 0.837 for the model ensemble using real data and 0.786 for the model ensemble using synthetic data. The model using the data/model-based ensemble had the highest AUROC of 0.839.
Discussion:
The study's focus was to predict sepsis severity via developing ML models trained on synthetic data. Current study's results support the feasibility of using synthetic genomic data for training ML models for prediction of sepsis severity. Synthetic data-trained ML model using the data/model-based ensemble method yielded the best performance, comparable to that of ML model trained on real world data.