GSOC final report

Title: **Integrating genomics and high-frequency physiologic data for sepsis detection**

Github link: [Link to codes]

Motivation: Sepsis is a life-threatening condition that occurs as a result of the body's response to infection causing tissue damage and even death. Internationally, an estimated 30 million people develop sepsis and 6 million die from sepsis each year; an estimated 4.2 million newborns and children are affected. Hospital-acquired sepsis infection is even worse as patients can contract the disease even after the primary ailment is cured. In India, the mortality rate of ICU acquired sepsis is around 50% [1].

Objective: Our objective was three folds:

1. Analyze gene expression data and identify differentially expressed genes to differentiate between the complicated and uncomplicated course patients.
2. Perform feature selection from clinical and genomic features.
3. Develop machine learning models using combinations of genomic and clinical features to predict the complicated course in sepsis

Method: We used microarray gene expression data (GSE66099) from 229 pediatric intensive care unit (PICU) patients . We preprocessed and normalized the data using Bioconductor packages. Differential gene expression analysis was performed using the limma package in R. We used only the PRISM (Pediatric Risk of Mortality Score) as mentioned in the Pollack MM et al paper [2] for building the clinical model. We also explored different feature selection methods to extract relevant features from the initial list of 21,724 genes. LASSO-based feature selection, minimum redundancy feature selection and the differentially expressed genes were pooled into the final feature list. 50 times repeated stratified cross validation with feature selection was performed to get a list of consistently chosen features (stability analysis). These were then used to build the final model

Results: **Different machine learning models were constructed to predict the complicated course outcomes and the results are summarized below using Area Under the ROC curve (AUROC)**

Features	AUROC
Severity alone (PRISM)	0.7697
Gene features alone	**0.863**
Severity + Gene features	0.845
Using severity and other clinical features	0.651
Using severity and other clinical features + gene features	0.855

Conclusion: The machine learning models predict the complicated course in sepsis, septic shock pediatric patients using a combination of clinical and genomics features and shown to perform better than the individual models (clinical or genomic). A detailed presentation discussing the methodology and results is available here.