Initiation to research: Practical course

Farida Zehraoui

November 2025

Abstract

This project investigates the prediction of sepsis outcomes using advanced classification pipelines applied to omics data.

1 Introduction

1.1 Definition of Sepsis

Sepsis is defined as a life-threatening organ dysfunction resulting from a dysregulated host response to infection (Singer et al., 2016). Despite advances in medicine, Sepsis remains a leading cause of mortality worldwide, largely due to its complex and heterogeneous biological mechanisms. Accurate early prediction of Sepsis outcomes can guide clinical decisions and improve survival rates.

1.2 Motivation

Omics technologies such as genomics, transcriptomics, and proteomics enable a comprehensive exploration of the molecular processes underlying Sepsis.

1.3 Objectives

The objective of this project is to compare a new method from the literature to traditional ML models such as XGBoost, Support Vector Machines (SVM), or Neural Networks.

2 Datasets

The primary dataset used in this study is GSE54514 (Parnell et al., 2013), titled Whole blood transcriptome of survivors and non-survivors of Sepsis. The dataset consists of expression profiling by array using Illumina HT-12 gene expression microarrays with 48,804 probes. Whole blood samples were collected daily for up to 5 days from patients admitted to the intensive care unit with sepsis. The cohort includes 26 Sepsis survivors, 9 Sepsis non-survivors, and 18 healthy controls. The different datasets are described below:

2.1 GSE65682

The **GSE65682** dataset, generated by the MARS (Molecular Diagnosis and Risk Stratification of Sepsis) consortium, includes genome-wide blood transcriptional profiles from critically ill patients. The study primarily focuses on differentiating Sepsis endotypes, analyzing gene expression patterns associated with various infection sources such as abdominal infections and pneumonia (Scicluna et al., 2015).

2.2 GSE54514

The study aims to identify differential gene expression patterns between sepsis survivors and non-survivors, providing insights into molecular mechanisms that influence sepsis outcomes (Parnell et al., 2013).

2.3 GSE76293

This dataset examines gene expression profiles in neutrophils obtained from ventilated ARDS patients and healthy volunteers, including additional samples from healthy individuals treated with phosphoinositide 3-kinase (PI3K) inhibitors (Juss et al., 2016).

2.4 GSE40012

This dataset explores gene expression profiles in whole blood samples from critically ill patients with influenza A pneumonia, bacterial pneumonia, mixed infections, and systemic inflammatory response syndrome (SIRS), as well as healthy individuals. It identifies specific gene expression patterns that help distinguish between different causes of severe respiratory illness, including distinct influenza A signatures related to cell cycle regulation, apoptosis, and DNA damage response (Parnell et al., 2012).

3 Methodology

- Select and analyze a paper from the literature related to the research problem;
- Choose one dataset;
- Define an experimental based the following stages:
    1. Data preprocessing: Normalize and filter the transcriptomic data.
    2. Model evaluation: Apply the target model (identified from the literature) to the data and compare its results with those of traditional machine learning to perform the classification task.

4 Expected Outcomes

At the end of the project, you must:

- Critically analyze a selected research paper.
- Analyze and interpret the performance across all models;
- Provide a final conclusion based on the experimental evidence, identifying which model performs best and explaining why.

5 Conclusion

This project aims to evaluate your ability to effectively conduct a research project from inception to conclusion.

References

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Juss, J. K., House, D., Amour, A., Begg, M., et al. (2016). Acute respiratory distress syndrome neutrophils have a distinct phenotype and are resistant to phosphoinositide 3-kinase inhibition. *American Journal of Respiratory and Critical Care Medicine*, 194(8):961–973.

Parnell, G. P., McLean, A. S., Booth, D. R., Armstrong, N. J., et al. (2012). A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. *Critical Care*, 16(4):R157.

Parnell, G. P., Tang, B. M., Nalos, M., Armstrong, N. J., et al. (2013). Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. *Shock*, 40(3):166–174.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *The Semantic Web – ESWC 2018*, pages 593–607.

Scicluna, B. P., Klein Klouwenberg, P. M., van Vught, L. A., Wiewel, M. A., et al. (2015). A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission. *American Journal of Respiratory and Critical Care Medicine*, 192(7):826–835.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2071–2080.