Comparing Traditional and Graph-Enhanced Pipelines for Sepsis Prediction Using Omics Data

Laura Forero

October 15, 2025

Abstract

This project investigates the prediction of sepsis outcomes using advanced classification pipelines applied to omics data. Two distinct approaches are compared: (1) a traditional machine learning pipeline trained solely on transcriptomic data, and (2) a graph-augmented pipeline that incorporates pretrained graph embeddings representing biological relationships among genes, proteins, and pathways. These embeddings, generated using **Complex (Transactional)** and **Relational Graph Convolutional Network (RGCN)** models, encode rich relational information from curated biological knowledge graphs. The goal is to evaluate whether integrating graph-based biological context improves the predictive performance of sepsis outcome models.

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. However, traditional analytical methods often treat genes and proteins as independent features, ignoring the network of biological relationships that drive disease behavior. Graph-based methods overcome this limitation by representing biological systems as networks of interacting entities, such as genes, proteins, and pathways.

Graph embeddings transform these networks into continuous vector representations that preserve biological structure and relationships. This enables the use of advanced classification models that leverage both molecular measurements and biological knowledge.

1.3 Objective

The objective of this project is to **compare two advanced classification pipelines for sepsis prediction**:

1. A **traditional pipeline**, which uses normalized transcriptomic data from the GSE54514 dataset as input to advanced models such as XGBoost, Support Vector Machines (SVM), or Neural Networks.
2. A **graph-augmented pipeline**, which uses the same classification models but enriches the input space with pretrained graph embeddings that encode biological relationships among genes and proteins.

Both pipelines will be evaluated under the same conditions to assess whether the inclusion of pretrained graph embeddings improves predictive performance.

2 Dataset

The primary dataset used in this study is **GSE54514** (Parnell et al., 2013), titled *Whole blood transcriptome of survivors and nonsurvivors 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.

3 Graph Embeddings

3.1 Definition

Graph embeddings map nodes, edges, or entire graphs into a continuous vector space while preserving structural and semantic relationships (Cai et al., 2018). In this project, pretrained graph embeddings derived from biological knowledge graphs are used to capture interactions that cannot be inferred from omics data alone.

3.2 Embedding Models

Two pretrained embedding models are included:

- **Complex (Transactional) embeddings** (Trouillon et al., 2016): Represent entities and relations in a complex-valued vector space, capturing asymmetric and multi-relational biological connections.
- **Relational Graph Convolutional Networks (RGCN)** (Schlichtkrull et al., 2018): A neural network architecture that learns node representations by aggregating relational information from neighboring nodes across multiple edge types.

3.3 Knowledge Graph Construction

We constructed a knowledge graph to integrate omics data with curated biological knowledge. Key concepts in the graph include:

- **Patient Sample**: Biological specimen from which molecular data is obtained.
- **Protein**: Molecules responsible for cellular functions, derived from omics measurements.
- **Reaction**: Biochemical processes converting molecular entities.
- **Pathway**: Sequences of molecular interactions supporting biological functions.
- **GoTerm**: Categories from the Gene Ontology describing gene product functions, processes, and cellular components.

Relationships among these entities, such as *participates_in* or *interacts_with*, were instantiated to capture biological mechanisms, as illustrated in Figure 1.

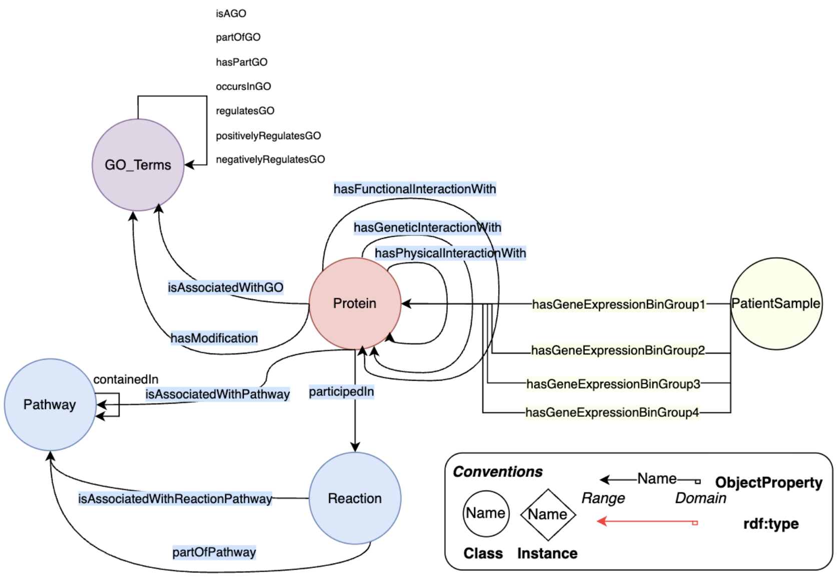

Figure 1: Schematic representation of the constructed knowledge graph integrating omics data and ontologies.

The pretrained embeddings were generated from large-scale biological knowledge graphs integrating curated data from multiple ontologies and databases (Table 1).

4 Methodology

The experimental workflow consists of the following stages:

1. **Data preprocessing:** Normalize and filter the transcriptomic data from GSE54514. Align each patient's gene expression profile with corresponding pretrained graph embeddings.

Resource	**Type**	**Information Provided**
Gene Ontology (GO)	Ontology	Gene functions, processes, and cellular components
Pathway Ontology	Ontology	Biological pathways
PPI Ontology	Ontology	Protein–protein interactions
Entrez	Database	Gene annotations
BioGRID	Database	Protein interactions
UniProt	Database	Protein functions and sequences
STRING	Database	Protein associations
Reactome	Database	Pathways and biochemical reactions

Table 1: Biological databases and ontologies integrated into the pretrained knowledge graph embeddings.

2. **Pipeline 1 – Traditional Classification:** Train advanced models such as XGBoost, Deep Neural Networks (DNNs), or SVMs using the transcriptomic data to predict sepsis outcomes.

3. **Pipeline 2 – Graph-Augmented Classification:** Use the same models as in Pipeline 1, but enhance the feature space by concatenating the pretrained graph embeddings (Complex and RGCN) with the omics features.

4. **Evaluation:** Compare the two pipelines using standard metrics (accuracy, AUROC, F1-score, precision, recall) to evaluate the impact of graph embeddings on predictive performance.

5 Expected Outcomes

At the end of the project, students must:

- Analyze and interpret the performance of both pipelines across all models and metrics.

- Discuss whether incorporating pretrained graph embeddings leads to measurable improvements in sepsis prediction.

- Provide a final conclusion based on the experimental evidence, identifying which pipeline performs best and why.

6 Conclusion

This project evaluates whether adding biological network information improves predictive models for sepsis. The main goal is to compare two classification pipelines: one using only omics data, and another enhanced with pretrained graph embeddings. This comparison reveals how biological context affects model performance. The analysis will determine whether graph-based representations deliver meaningful improvements for clinical prediction tasks.

References

Parnell, G. P., Tang, B. M., Nalos, M., Armstrong, N. J., et al. Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions. *Shock*, 2013; 40(3):166–174. doi:10.1097/SHK.0b013e31829965c2.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 2016; 315(8):801–810. doi:10.1001/jama.2016.0287.

Cai, H., Zheng, V. W., Chang, K. C.-C. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 2018; 30(9):1616–1637.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G. Complex Embeddings for Simple Link Prediction. *Proceedings of ICML*, 2016.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., Welling, M. Modeling Relational Data with Graph Convolutional Networks. *The Semantic Web – ESWC*, 2018.