Sepsis Prediction using Structure Data

CS371-Final Project Report Document

Session: 2022-2026

Project Supervisor

Dr. Samyan Qayyum Wahla
Mr. Nazeef ul Haq

Project Members (G4)

Muhammad Wali Ahmad	2022-CS-65
Shammas Ahmad	2022-CS-83

Department of Computer Science

University of Engineering and Technology, Lahore Pakistan

Contents

1 Overview 3

2 Objectives 3

3 Features 3

4 Architecture 3
4.1 Data Preprocessing . . . 3
4.2 Model Architecture . . . 3
4.3 Backend . . . 4
4.4 Frontend . . . 4

5 Class Imbalance 4
5.1 Source of Class Imbalance . . . 4
5.2 Resampling Technique: SMOTE . . . 4

6 Metrics and Model Performance 4
6.1 How Our Model Fits . . . 5

7 Sustainable Development Goal (SDG) Alignment 5

8 Results and Evaluation 5

9 Challenges and Limitations 5

10 Conclusion 6

List of Figures

1 Overview

Sepsis is known as a life-threatening organ dysfunction caused by a dysregulated host response to infection. Early detection and treatment reduce mortality and improve outcomes among patients. The project based on the proposed utilizes artificial intelligence to predict sepsis in ICU patients by using structured clinical data inside the MIMIC-III database. It uses deep learning technique, different modeling, to capture temporal relationships in patient data in making robust prediction. Hence, the system thus aims at allowing healthcare specialists to make timely decisions that are likely to avert avoidable sepsis-related deaths.

2 Objectives

Following are the objectives of our project:

- Use the MIMIC-III dataset to extract features and then train models.
- Design an artificial intelligence models which predicts the appearance of sepsis.
- To provide useful information to healthcare providers for earlier detection of sepsis.
- The model is tested through metrics like accuracy, precision, recall, and F1 score.

3 Features

- **Data Utilization:** Structured clinical data, such as vital signs, lab results, and demographic details, are processed to extract meaningful features.
- **Sepsis Prediction:** A different models predict the risk of sepsis based on temporal patient data.
- **Visualization:** Intuitive visualizations of sepsis probabilities and critical factors influencing predictions.

4 Architecture

4.1 Data Preprocessing

Extract the Structured Data of a Nature Such as Heart Rate, Blood Pressure, and Laboratory Test Results of MIMIC-III Database. Missing values are done with imputation technique, and outliers are removed for the quality of data. The features derived include components of SOFA score, temporal trends and clinical thresholds for sepsis indicators.

4.2 Model Architecture

The following machine learning models were employed to predict the probability of sepsis using structured data from the MIMIC-III dataset. Each model has been carefully chosen for its ability to handle structured tabular data, with varied approaches to learning and decision-making:

- **Support Vector Classifier (SVC):** Constructs hyperplanes in a high-dimensional space to classify sepsis and non-sepsis cases, leveraging kernel functions like linear or RBF for non-linear decision boundaries.
- **Decision Tree:** Creates a tree structure where data is split at each node using metrics like Gini impurity or entropy, leading to a predicted class label at the leaf nodes.
- **Random Forest:** Combines predictions from multiple decision trees trained on different data subsets to improve classification accuracy and reduce overfitting.

- **Gradient Boosting:** Sequentially builds weak learners (decision trees), with each tree correcting the errors of the previous one, optimizing for a loss function like log-loss.
- **XGBoost:** An optimized gradient boosting model that uses regularization techniques and parallel processing for faster training and improved generalization.
- **MLP (Multi-Layer Perceptron):** A neural network architecture with fully connected layers that learns non-linear relationships in the data through backpropagation.
- **LGBM (LightGBM):** A gradient boosting framework that uses histogram-based learning and leaf-wise tree growth to handle large datasets efficiently while maintaining high prediction accuracy.

4.3 Backend

The backend is developed in Django, managing data processing, model integration, and API endpoints to make the predictions.

4.4 Frontend

A simple web-based interface in NextJS is designed for healthcare professionals to input patient data and visualize predictions.

5 Class Imbalance

5.1 Source of Class Imbalance

There is class imbalance in this MIMIC-III dataset where the far majority is that of non-sepsis while relatively fewer cases occur in sepsis. In general, sepsis occurs not as frequently as it might regarding the cases seen in patients in an ICU and, as such, training an AI algorithm on this imbalanced class set will make the model to be biased towards the majority class and will eventually misclassify this class that's critical to this project.

5.2 Resampling Technique: SMOTE

To handle the class imbalance problem, we used the Synthetic Minority Over-sampling Technique, SMOTE. SMOTE is the most widely applied resampling method of all others; it creates synthetic examples of the minority class (sepsis cases) by interpolating between existing instances. Key steps of SMOTE:

1. In the feature space, SMOTE finds the nearest neighbors for every instance in a minority class.
2. The new synthetic samples are generated at intervals along the line segments that connect the chosen instances with their neighbors.
3. The dataset is now balanced and more evenly distributed because it generates synthetic samples of the minority class, thus enabling the model to learn sepsis-specific patterns better.

Such approach tends not to overfit the model given on available instances of the minority classes in that it maintains a lot of capacity to generalize properly based on unseen data. As well, SMOTE negates possible overfitting from trivial class duplication.

6 Metrics and Model Performance

To evaluate our model's performance, we used key metrics to measure the predictive quality:

1. **Accuracy:** Measures the proportion of correctly predicted instances (sepsis and non-sepsis) to the total instances. Achieved 84.27%, indicating robust generalization.

2. **Precision:** Focuses on the proportion of true sepsis cases among all cases predicted as sepsis. Achieved 87.11%, reflecting fewer false positives.
3. **Recall:** Indicates the ability to identify true sepsis cases from all actual sepsis cases. Achieved 79.54%, highlighting the model's reliability in capturing sepsis instances.
4. **F1 Score:** The harmonic mean of precision and recall, balancing false positives and false negatives. Achieved 83.16%, showcasing overall effectiveness.

6.1 How Our Model Fits

The XGBoost (Extreme Gradient Boosting) model demonstrated superior performance in predicting sepsis cases. Its ability to effectively handle structured data and focus on important features made it the optimal choice. The model leverages gradient boosting techniques to iteratively correct errors of previous trees, capturing complex patterns indicative of sepsis. Additionally, its built-in regularization mechanisms helped prevent overfitting, ensuring robust generalization.

The use of the SMOTE resampling technique addressed the class imbalance by generating synthetic sepsis cases, providing the model with a balanced dataset for training. XGBoost's high recall reflects its capability to prioritize identifying sepsis cases, which is critical for early detection in clinical settings. This aligns with healthcare priorities, where timely and accurate identification of sepsis can save lives and improve patient outcomes.

7 Sustainable Development Goal (SDG) Alignment

This project aligns with SDG 3: Good Health and Well-Being. By facilitating early detection of sepsis, the system enhances patient care, reduces the burden on ICU resources, and contributes to improved health outcomes in underserved regions.

8 Results and Evaluation

The model's performance was evaluated on a subset of the MIMIC-III dataset, achieving the following metrics:

- AUC: 91.78%
- Accuracy: 84.27%
- Precision: 87.11%
- Recall: 79.54%
- F1 Score: 83.16%

Visualization tools validated the model's predictions, ensuring transparency and clinical relevance.

9 Challenges and Limitations

1. **Data Imbalance:** Addressed using resampling and class weighting.
2. **Interpretability:** Enhancements like SHAP (SHapley Additive exPlanations) provided insights into predictions.
3. **Resource Constraints:** Training the XGBoost model was computationally efficient; however, tuning its hyperparameters, especially with a large dataset, required substantial time and computational resources to achieve optimal performance.

10 Conclusion

The "Predicting Sepsis using Structured Data" project illustrates the potential of applying AI in critical healthcare tasks. Using the XGB model and the MIMIC-III dataset, the system provides healthcare experts with a reliable tool for detecting early sepsis onset so that life-saving interventions may be made. Future versions can include the integration of unstructured data, as is the case with clinical notes, to further narrow down predictions.