Deep Learning Mini-Project 2

Early Sepsis Prediction using MIMIC-III Dataset

Lise ABI RAFEH
Walid SATI

Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection. It occurs when the body's immune response to an infection becomes uncontrolled, leading to widespread inflammation, tissue damage, and potential organ failure.

The Sepsis-3 definition (2016) by the Third International Consensus Definitions for Sepsis and Septic Shock states:

- **Sepsis**: A life-threatening organ dysfunction caused by a dysregulated host response to infection. Organ dysfunction is identified by an increase of 2 or more points in the Sequential Organ Failure Assessment (SOFA) score.
- **Septic Shock**: A subset of sepsis characterized by circulatory and cellular/metabolic dysfunction associated with a higher risk of mortality. It is clinically identified by:
    - Persistent hypotension requiring vasopressors to maintain a mean arterial pressure (MAP) ≥ 65 mmHg.
    - Serum lactate > 2 mmol/L despite adequate fluid resuscitation.

Early recognition and prompt treatment with antibiotics, fluid resuscitation, and organ support are crucial to improving outcomes. Early predictors of sepsis involve a combination of clinical, laboratory, and physiological markers that indicate an escalating inflammatory response and organ dysfunction. Key early indicators include:

Clinical Signs and Symptoms

- Fever or Hypothermia (Temperature >38.3°C or <36°C)
- Tachycardia (HR >90 bpm in adults)
- Tachypnea or Respiratory Distress (RR >22/min)
- Altered Mental Status (Confusion, disorientation, or lethargy)
- Hypotension (Systolic BP <100 mmHg)

- Decreased Urine Output (Oliguria <0.5 mL/kg/h)

Laboratory Biomarkers

- Elevated White Blood Cell Count (WBC) (>12,000/mm³ or <4,000/mm³)
- Elevated Procalcitonin (PCT) (>0.5 ng/mL; >2 ng/mL is highly suggestive of sepsis)
- Increased C-Reactive Protein (CRP) (>100 mg/L)
- Elevated Lactate (>2 mmol/L suggests tissue hypoxia; >4 mmol/L is severe)
- Coagulation Abnormalities (INR >1.5, aPTT >60s, or thrombocytopenia <100,000/mm³)

Scoring Systems for Early Detection

- **qSOFA (Quick SOFA) Score** (≥2 suggests a higher risk of sepsis)
  - RR ≥22/min
  - Altered mental status (GCS <15)
  - Systolic BP ≤100 mmHg
- **SOFA Score** (Sequential Organ Failure Assessment; increase by ≥2 points indicates sepsis)
- **NEWS (National Early Warning Score)** (combines vital signs to detect deterioration)

Material & Method

This project aimed to build a deep learning model for **early sepsis prediction** using the **MIMIC-III** dataset. The study focused on two approaches:

1. **LSTM-GRU Model**: Predicting sepsis using structured patient data, including vital signs, lab biomarkers, and clinical indicators.
2. **BioClinicalBERT Model**: Extracting insights from **clinical notes** (written by nurses, residents, and physicians) for sepsis prediction. Sepsis is a life-threatening condition that requires early detection and intervention. This project aimed to build a deep learning model for **early sepsis prediction** using the **MIMIC-III** dataset. The study focused on two approaches:
3. **LSTM-GRU Model**: Predicting sepsis using structured patient data, including vital signs, lab biomarkers, and clinical indicators.
4. **BioClinicalBERT Model**: Extracting insights from **clinical notes** (written by nurses, residents, and physicians) for sepsis prediction.

Sepsis Criteria Used

Data was retrieved from the dataset using the following criteria for sepsis identification:

- **Vital Signs**: Heart rate, blood pressure, respiratory rate, temperature, SpO2, etc.
- **Biological Markers**: Lactate, White Blood Cell (WBC) count, C-Reactive Protein (CRP), etc.
- **Microbiological Findings**: Bacterial cultures, infections.

- **Glasgow Coma Scale (GCS)**: Used to detect alteration of consciousness.
- **Oliguria Criteria**: Urine output to weight ratio (< 0.5 mL/h).
- **Scoring Systems**: qSOFA, SOFA, NEWS for early risk assessment.

Challenges Faced

During the project, several difficulties were encountered:

- **Restricted access to MIMIC-III**: Required a course enrollment, certificate of completion, and an access request.
- **Dataset size**: 41.3 GiB, making it difficult to process with limited computing resources.
- **Memory & Computational Constraints**: Handling high-dimensional time-series data (LSTM-GRU) and text-based clinical notes (BioClinicalBERT) in an environment with limited RAM, CPUs, and GPUs.
- **Workarounds Implemented**: (Details will be provided in the next sections)

Dataset & Feature Selection

Data was extracted from MIMIC-III, focusing on the following key features:

- **Vital Signs**: Heart rate, blood pressure, respiratory rate, temperature, SpO2, etc.
- **Biological Markers**: Lactate, White Blood Cell (WBC) count, C-Reactive Protein (CRP), etc.
- **Microbiological Findings**: Bacterial cultures, infections.
- **Glasgow Coma Scale (GCS)**: Used to detect alteration of consciousness.
- **Oliguria Criteria**: Urine output to weight ratio (< 0.5 mL/h).

2 different models were used for prediction: LSTM-GRU and BioClinicalBert

Material & Method

Model 1: LSTM-GRU for Sepsis Prediction

The LSTM-GRU model was implemented with the following steps:

4.1 Data Preprocessing

- **Handling Missing Data**: Imputed missing values using median or interpolation.
- **Feature Scaling**: Normalized continuous variables.
- **Time-Series Processing**: Reshaped the dataset for sequential learning.

4.2 Model Architecture

The LSTM-GRU model was developed to process time-series patient data for sepsis prediction. The key components included:

LSTM Model:

- **Input Layer**: Accepts sequential patient data of vital signs and laboratory markers.
- **LSTM Layers**:
  - First LSTM layer with 64 units, ReLU activation, and return sequences enabled.
  - Dropout layer (0.2) to reduce overfitting.
  - Second LSTM layer with 32 units and ReLU activation.
  - Dropout layer (0.2) to enhance generalization.
- **Dense Layers:**
  - Fully connected layer with 16 neurons and ReLU activation.
  - Output layer with a single neuron and sigmoid activation for binary classification (sepsis vs. non-sepsis).

GRU Model:

- **Input Layer**: Same as LSTM model.
- **GRU Layers:**
  - First GRU layer with 64 units, ReLU activation, and return sequences enabled.
  - Dropout layer (0.2) to mitigate overfitting.
  - Second GRU layer with 32 units and ReLU activation.
  - Dropout layer (0.2) to improve robustness.
- **Dense Layers:**
  - Fully connected layer with 16 neurons and ReLU activation.
  - Output layer with sigmoid activation to classify sepsis cases.

Both models were trained separately and their performance was compared to determine the optimal architecture for sepsis prediction.

- **Input Layer**: Sequential input of patient data.
- **LSTM & GRU Layers**: Captured temporal dependencies.
- **Dense Layers**: Used to output probability scores for sepsis risk.

4.3 Training Process

LSTM Model Training:

- **Optimizer**: Adam with a learning rate of 0.001.
- **Loss Function**: Binary cross-entropy.
- **Batch Size**: 32.
- **Epochs**: 20.
- **Validation Split**: 10% of training data used for validation.
- **Performance Metrics**: AUC-ROC, Precision, Recall, and Accuracy.

GRU Model Training:

- **Optimizer**: Adam with a learning rate of 0.001.
- **Loss Function**: Binary cross-entropy.
- **Batch Size**: 32.
- **Epochs**: 20.
- **Validation Split**: 10% of training data.
- **Performance Metrics**: AUC-ROC, Precision, Recall, and Accuracy.

Additionally, hyperparameter tuning was performed using grid search on:

- **LSTM/GRU units**: 32, 64.
- **Dropout rates**: 0.2, 0.3.
- **Learning rates**: 0.001, 0.0001.

The best hyperparameters were selected based on the highest AUC-ROC score.

- **Loss Function**: Binary Cross-Entropy.
- **Optimizer**: Adam.
- **Batch Size**: Tuned to balance speed vs. memory.
- **Performance Metrics**: AUC-ROC, Precision, Recall.

4.4 Results & Performance

LSTM Model Results:

- **Test Accuracy**: 94.03%
- **AUC-ROC Score**: (49.95%

- **Precision** = (True Positives) / (True Positives + False Positives)
    - Measures how many of the predicted positive cases are actually correct.
    - For **class 1 (sepsis), precision = 0.00**, meaning the model did not predict **any** sepsis cases correctly.
- **Recall (Sensitivity)** = (True Positives) / (True Positives + False Negatives)
    - Measures how many actual sepsis cases were correctly identified.
    - For **class 1 (sepsis), recall = 0.00**, meaning the model failed to detect **any** actual sepsis cases.
- **F1-Score** = 2 * (Precision * Recall) / (Precision + Recall)
    - A harmonic mean of precision and recall.
    - Since **both precision and recall for class 1 are 0**, **F1-score is also 0.00**.

Interpretation:

- **Class 0 (Non-Sepsis)**
    - Precision: **0.94** → 94% of predicted non-sepsis cases were correct.
    - Recall: **1.00** → The model correctly identified all non-sepsis cases.
    - F1-Score: **0.97** → A high balance between precision and recall.
- **Class 1 (Sepsis)**

- Precision: **0.00** → No predicted sepsis cases were correct.
- Recall: **0.00** → The model **never** detected actual sepsis cases.
- F1-Score: **0.00** → Complete failure in predicting sepsis.

The Problem: Model is Ignoring Sepsis Cases

- **Our model is highly imbalanced**: It is predicting **only class 0 (non-sepsis)** and completely ignoring **class 1 (sepsis)**.
- This is confirmed by the confusion matrix:

```
[[15677     0]
 [  995     0]]
```

- **15677** non-sepsis cases correctly classified.
- **995** sepsis cases all misclassified as non-sepsis.

GRU Model Results:

- **Test Accuracy**: 94.03%
- **AUC-ROC Score**: 66.27%
- **Precision, Recall, F1-score**:
  - **Classification Report:**

```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     15677
           1       0.00      0.00      0.00       995

    accuracy                           0.94     16672
   macro avg       0.47      0.50      0.48     16672
weighted avg       0.88      0.94      0.91     16672
```

- **Confusion Matrix**: (`array([[15677,     0],[  995,     0]])`

Addressing Class Imbalance and Model Improvement

To improve the performance of the LSTM-GRU model, several strategies can be applied:

➔ Class Imbalance Handling:

- Use **oversampling (SMOTE)** to increase the number of sepsis cases in training.
- Use **undersampling** to balance the dataset.
- Try **class weighting** in model training: `class_weight={0:1, 1:10}`.

➔ Better Evaluation:

- Instead of **accuracy**, focus on **ROC-AUC score** and **F1-score** to better assess model performance.
- Use the **Precision-Recall Curve** instead of relying only on the confusion matrix.

➔ Hyperparameter Tuning:

- Adjust the **threshold** of model predictions (`y_pred_proba > 0.5` might be too high, leading to missed sepsis cases).
- Try different architectures such as **Bidirectional LSTMs** or **attention mechanisms** to enhance model learning and sensitivity to sequential data.

Not tried for the sake of time and lack of resources…

Model 2: BioClinicalBERT for Clinical Notes Analysis

5.1 Preprocessing

- Extracted sepsis-related clinical notes.
- Tokenized text using **BioClinicalBERT tokenizer**.
- Converted text into **word embeddings**.

5.2 Model Implementation

- Used a **pretrained BioClinicalBERT model**.
- Fine-tuned on **annotated clinical notes**.
- Outputted probability of **sepsis presence**.

5.3 Training Progress

the BioClinicalBERT model training is done, but performance results indicate significant class imbalance issues, leading to poor sepsis detection.

5.4 Model Performance

- **Accuracy**: 77.57%
- **Precision**: 0.0000
- **Recall**: 0.0000
- **F1-score**: 0.0000

5.5 Addressing Model Issues

To improve BioClinicalBERT's ability to detect sepsis cases, the following adjustments could be implemented:

Threshold Adjustment:

- The classification threshold (`y_pred_proba > 0.5`) might be too high and will be lowered to **0.3 or optimized dynamically**.

Class Imbalance Handling:

- **Oversampling sepsis cases** to improve model training.
- **Applying class weighting** to ensure sepsis cases have a greater impact.

Feature Engineering & Preprocessing:

- **Fine-tuning on a larger clinical note subset**.
- **Exploring contextual embeddings** to improve sepsis-related text recognition.

Further evaluations will determine the best adjustments to enhance performance. Currently ongoing (results to be added later).

6. Conclusion

Despite resource limitations, multiple optimizations allowed successful training. Next steps could include:

- Fine-tuning BioClinicalBERT.
- Optimizing LSTM-GRU hyperparameters.
- Fine-tuning BioClinicalBERT.
- Optimizing LSTM-GRU hyperparameters.
- Deploying models for real-time clinical use. This project demonstrated the feasibility of deep learning for early sepsis prediction using structured time-series data (LSTM-GRU) and unstructured clinical notes(BioClinicalBERT).

Our implemented models showed limited performance due to several factors, including computational constraints, the large number of datasets, and the inherent limitations of the MIMIC-III dataset. Despite containing data from over 40,000 ICU patients, MIMIC-III is not a recent dataset (2001-2012) and lacks modern advancements in sepsis diagnosis and treatment. Additionally, data gaps and inconsistencies—due to different hospital systems (CareVue vs. MetaVision)—resulted in missing or inconsistent data, affecting model predictions.

Another challenge was the limited continuous monitoring available in the dataset. Unlike modern ICUs, not all patients had high-frequency vitals recorded, reducing the effectiveness of time-series modeling. Furthermore, to comply with HIPAA regulations, MIMIC-III underwent deidentification and timestamp shifting, which may have disrupted temporal relationships, making time-dependent models like LSTM and GRU less effective.

An optimal approach would involve merging multiple available datasets, such as MIMIC-IV or eICU, integrating structured clinical data with textual notes, and processing them into a unified, high-quality dataset. Training a hybrid model combining time-series patient data with contextual embeddings from clinical notes could significantly improve sepsis detection.

With better data integration, enhanced preprocessing, and a more advanced deep learning framework, the results could be vastly different. Perhaps this idea could serve as the foundation for a future large-scale project in sepsis prediction.