Predicting Sepsis Onset Using Recurrent Neural Networks and the MIMIC-III Database

Group AJA:
Jessica Becerra, Ashley Cortez, Angelica Verduzco
BANA 650
December 4, 2024

Table of Contents

Abstract

Objective: Early prediction of sepsis is critical for improving patient outcomes in intensive care units (ICUs). This study replicates and extends prior research by evaluating the predictive capabilities of a recurrent neural network (RNN) for detecting sepsis onset and comparing its performance to the InSight algorithm. The study investigates the impact of varying look-back and prediction windows on model performance, leveraging a robust data preprocessing pipeline to enhance accuracy.

Methods: Data were sourced from the MIMIC-III database, including over 58,000 ICU admissions between 2001 and 2012. The study focused on adult patients (≥18 years) without sepsis upon admission, identifying sepsis onset using SIRS criteria and ICD-9 codes. Vital signs and laboratory measurements, including systolic/diastolic blood pressure, pH, and oxygen saturation, were used as predictive features. Data preprocessing involved linear interpolation, carry-forward/backward imputation, hourly aggregation, and normalization. Time-series datasets were generated for prediction windows of 3, 6, and 12 hours prior to sepsis onset, with look-back windows of 5, 10, 15, and 20 hours to evaluate their impact on model performance. A 4-fold cross-validation strategy ensured robust evaluation, with performance measured using AUROC, sensitivity, and specificity.

Results: The RNN demonstrated strong predictive performance across all prediction windows, with AUROC scores of 0.982, 0.973, and 0.997 for 3, 6, and 12-hour windows, respectively. Longer look-back windows significantly improved accuracy, with the 20-hour window outperforming shorter intervals. The RNN also consistently outperformed the InSight algorithm, achieving higher sensitivity and specificity.

Conclusion: This study confirms the potential of RNNs for early sepsis detection by effectively analyzing time-series data. Findings underscore the importance of look-back duration, preprocessing techniques, and data quality in predictive modeling. Further research is recommended to validate these findings across diverse populations and clinical settings, as well as to explore their integration into real-time clinical decision-making systems.

Introduction

Sepsis is a life-threatening condition characterized by a dysregulated immune response to infection, leading to organ dysfunction and high mortality rates. Early detection and intervention are crucial for improving patient outcomes, as timely administration of treatment significantly reduces sepsis-related complications. However, identifying sepsis onset remains a challenging task due to its complex and variable clinical presentation. In recent years, advancements in machine learning have shown promise in addressing these challenges by leveraging vast amounts of clinical data to predict sepsis onset with greater precision.

Traditional predictive algorithms, such as the InSight algorithm, have been widely studied for sepsis detection and prediction. These methods typically rely on fixed thresholds for physiological parameters, limiting their ability to capture temporal trends in patient data. In contrast, deep learning approaches, particularly recurrent neural networks (RNNs), offer a robust framework for analyzing time-series data. By incorporating temporal dependencies, RNNs can model the dynamic changes in a patient's clinical state, providing an opportunity to improve early sepsis detection.

This paper builds upon prior research by replicating and extending a comparative analysis of RNNs and the InSight algorithm for predicting sepsis onset. Using data from the MIMIC-III database, we aim to explore the potential of RNNs in capturing time-series patterns and enhancing predictive performance. Specifically, the study evaluates the prediction accuracy at different time intervals prior to sepsis onset (3, 6, and 12 hours) and examines the effect of varying look-back windows on model performance. Furthermore, we compare the RNN's performance to that of the InSight algorithm, focusing on key evaluation metrics, including the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity.

Objectives of this paper are to replicate the following steps from the article:

1. Assess the performance of the RNN in predicting sepsis onset at 3, 6, and 12 hours before diagnosis.
2. Analyze the impact of varying look-back windows (5, 10, 15, and 20 hours) on model accuracy.
3. Compare the RNN's performance to the InSight algorithm in terms of AUROC, sensitivity, and specificity.

Through this, we aim to provide insights into the advantages and limitations of RNNs for sepsis prediction and identify key factors influencing model performance.

Methodology

Data Collection and Extraction

For this project, data was sourced from the MIMIC-III database which contains over 58,000 ICU admissions recorded between the years 2001 and 2012. The dataset provides comprehensive patient records, including the demographics, vital signs, and laboratory measurements.

To emulate our selected research paper, the data extraction process focused on patients aged 18 years or older and excluded those diagnosed with sepsis upon admission. SIRS (Systemic Inflammatory Response Syndrome) parameters, including vital signs and laboratory results were extracted to identify patients meeting the criteria of sepsis onset. Additionally, the data extractions used for this project filtered for confirmed sepsis cases, which were identified using ICD-9 codes (99591, 99592, 78552).

As outlined in the original paper, sepsis onset is identified when at least two of the four following criteria are met for SIRS parameters:

- Temperature <36℃ or >38℃
- Heart Rate > 90 BPM
- Respiratory rate > 20/min, or PaCO2 <32mmHg
- White Blood Cell Count <4k/μL or >12k/μL

In addition to SIRS parameters, the following vital signs were also used for prediction in our replication project:

- Systolic blood pressure
- Diastolic blood pressure
- pH value
- Blood Oxygen Saturation

Data extractions from the MIMIC-III database were created via Google BigQuery. In order to optimize processing efficiency, pre-filtered tables were created, ultimately reducing processing times significantly during table joins. The steps taken throughout the extraction process are outlined below *(exact queries and previews of their results can be found in the appendix)*:

1. Extract ICU Admissions with Patients 18+ (hadm_id, subject_id, admittime, dischtime)
2. Extract Vital Signs/Lab Results for SIRS Parameters (hadm_id, itemid, charttime, value)
3. Extract Confirmed Sepsis Diagnoses using ICD-9 Codes (hadm_id, icd9_code, short_title)
4. A unique table was created for each of the first three extractions. This minimized overall processing time for joins on the larger MIMIC-III clinical tables
    a. Join Vital Signs/Lab Results and Confirmed Sepsis Diagnoses Only
    b. Use the 18+ table to conduct a final filter of all data

Once the initial data extraction was complete, we then moved on to data preprocessing to clean and prepare the data for the RNN model.

Data Preprocessing and Model Preparation

To ensure data quality and consistency for our replication we completed our preprocessing steps by handling missing data using a combination of linear interpolation and "carry forward/backward" fill methods to ensure the completeness and consistency in the dataset. Additionally, a binary feature to indicate sepsis onset was added to the data. The data was grouped by hadm_id and sorted by chart time - once a patient met the minimum requirements for SIRS, they were labeled with a 1, otherwise each row for a chart event carried a zero until SIRS parameter requirements were met.

Interpolations and Prediction Windows: As outlined by our selected research paper, we conducted 6 different interpolations of the data (0/1/2/3/4/5). In the context of this paper, interpolation is defined as the process of estimating missing values, and the levels being the amount of times interpolation was applied. Following interpolation, our team then generated the time-series data sequences for 3

different prediction windows (3h/6h/12h) resulting in 18 different versions of the initial extraction. File names were named systematically (e.g., sepsis_interp_1_window_12.csv) to aid with organization prior to model implementation.

Normalization: Each feature was aggregated into hourly intervals, and mean values were calculated to standardize the inputs. Lastly, all parameters were normalized by scaling them into a range between 0 and 1 to ensure comparable scales for RNN model implementation. The normalization was performed per parameter using a Min-Max scaling technique. Finally, Look-back windows were evaluated at 5, 10, 15 and 20 hours to assess the impact of the prediction accuracy.

K-Fold Cross-Validation: After cleaning and normalization, we replicated the 4-fold cross validation approach used in the article to ensure a robust evaluation of the model. This method also verified the balance of sepsis onset and non-sepsis onset cases was maintained across the training and test sets in each fold, by reducing the risk of bias and overfitting. This approach also helped by providing a reliable estimate of the model's generalizability to any unseen data.

Synthetic Minority Oversampling Technique (SMOTE)*: This augmentation technique was applied to address imbalances between training and testing data. After splitting the data into the k-fold cross validation, SMOTE is only applied to the training data of each fold. This was ultimately implemented to optimize the RNN's ability to predict sepsis onset. In simpler terms, SMOTE selects a real example from the minority class (sepsis onset), finds similar example neighbors, and then creates synthetic data points, creating a new realistic data point instead of just duplicating existing data so that the model learns better patterns for sepsis onset.
**This method was not explicitly outlined in the original research paper, however we found that its addition provided optimized performance of the final model.*

RNN Model Implementation and Evaluation

To replicate and evaluate the predictive models, various techniques were implemented. These include the application of traditional and deep learning algorithms to predict sepsis onset within specified prediction windows. A key approach includes the development of the InSight algorithm as a baseline and the use of recurrent neural networks (RNNs) for identifying patterns in patient data. The model we replicated was evaluated based on AUROC score.

Insight Algorithm and Recurrent Neural Networks (RNN)

The article compared the performance of the RNN model with InSight, a previously established predictive model. Similar to the article we created a Recurrent Neural Network and evaluated the performance against the InSight model already established. The evaluation focused on the AUROC, sensitivity and specificity across prediction windows of (3/6/12) hours prior to sepsis onset. Like the article, our models comparison demonstrated the advantages of the RNN by having higher predictive accuracy and robustness.

Evaluation Metrics: Similar to the article we evaluated the model performance by analyzing the AUROC (Area Under the Receiver Operating Characteristic Curve) to distinguish between patients who would develop sepsis and those who would not. We also evaluated the models performance by the sensitivity and specificity. The sensitivity measured the proportion of correctly identified sepsis cases and specificity identifying the proportion of correctly identified non-sepsis cases. Look-Back Windows was used to evaluate the predictive accuracy by highlighting the importance in the time-series data by window lengths of (5/10/15/20) hours.

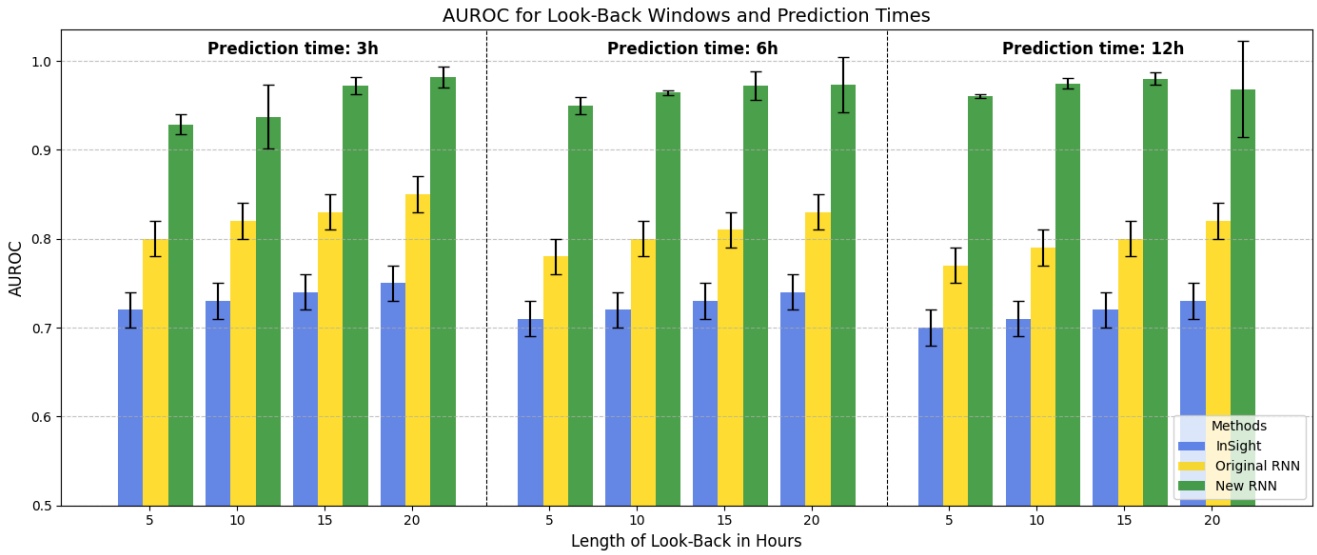
Analysis and Findings

Compare the performance of the RNNs in predicting sepsis onset at 3, 6, and 12 hours before diagnosis

The Recurrent Neural Network (RNN) in the original study demonstrated robust performance in predicting sepsis onset across varying prediction windows (3h/6h/12h) hours before diagnosis. Our replicated model showed similarly strong results, achieving consistently high AUROC scores across all levels of interpolation. At the 3-hour prediction window within a 20-hour look-back period, the model achieved a mean AUROC of 0.982, highlighting its precision in early detection. For the 6-hour and 12-hour windows, the mean AUROCs were 0.973 and 0.997, respectively. These results highlight the RNN's ability to effectively utilize historical data to predict sepsis onset with high accuracy.

Compare the impact of varying look-back windows (5, 10, 15, and 20 hours) on model accuracy

The length of the look-back window had a significant influence on the RNN's predictive accuracy, seen in both the original article and our replication. Longer look-back windows consistently improved the AUROC, likely due to additional historical data aiding the model in identifying sepsis onset trends. In our replication, the AUROC for the 3-hour prediction window increased from 0.905 with a 5-hour look-back to 0.982 with a 20-hour look-back. This trend was even more pronounced compared to the original study, where AUROC for the same look-back lengths improved from 0.72 to 0.81. Similarly, for the 6-hour prediction window, the model's AUROC rose from 0.947 (5-hour look back) to 0.973 (20-hour look-back). These results highlight the effectiveness of our model.



Compare our RNN's performance to both the InSight algorithm and RNN from the original study, focusing on AUROC, sensitivity, and specificity.
Our RNN model outperformed both the InSight algorithm and the original study's RNN across all evaluation metrics. While the original RNN's AUROC increased from 0.72 (5-hour look-back) to 0.81 (20-hour look-back), our model achieved higher scores ranging from 0.905 to 0.982 for the same look-back periods at the 3-hour prediction window. At a fixed sensitivity level of 90%, the original RNN's specificity was at 31.4% at 3 hours and 23.9% at 12 hours. In contrast, our replication showed improvements by achieving 47% at 3 hours and 46% at 12 hours, indicating fewer false positives while maintaining high sensitivity.

Conclusion

As we replicated the original study, we reconfirmed their findings that a Recurrent Neural Network can effectively and efficiently predict sepsis onset based on vital signs and SIRS parameters recorded throughout a patient's stay in the ICU. This approach provides insight into how sepsis can develop overtime by pinpointing the estimated start of sepsis onset.

The accuracy of detection is affected by how much missing data is filled in via interpolations and is also highly reliant on vital sign and lab test results recorded

throughout a patient's stay. Additionally, we found that there are additional preprocessing and model preparation steps (e.g., k-fold Cross-Validation and SMOTE) that can be utilized in order to improve the AUROC performance of an RNN.

The original paper outlines that the filtering done in the beginning of the data extraction may affect overall performance of the RNN. We agree that there could be certain special cases where sepsis onset doesn't meet all of the requirements (i.e., manifesting certain vital signs for 5+ hours). In these cases the RNN would not be able to predict sepsis onset due to its structure. Additionally since RNNs are "black boxes" there is limited interpretability on how the model formulates its predictions/conclusions. It is ultimately recommended to further investigate the functionality of the RNN against non-ICU data given that it has less time series data to predict sepsis onset, which may result in an increased amount of false alarms. At this time, it is recommended to utilize RNN models as a starting point for predicting sepsis onset and fine tuning them to be utilized in various environments outside of what was explored in this project.

Lessons Learned

This project allowed us to apply our analytical skills in a field that we are unfamiliar with. With this we were able to utilize real-world data and experienced the in-depth process of data extraction, preprocessing, cleaning and modeling in analytics.

Healthcare data is especially complex, particularly when the data source is a large database that has records from many different hospitals and institutions. We learned that you cannot assume that all hospitals record vital signs and other lab events the same way. For example, some sources may record temperature in Fahrenheit while others use Celsius. These kinds of nuances require meticulous attention to detail, otherwise predictive models and other analysis developed may be skewed if data is not cleaned and standardized correctly.

As the data and the processes are complex, we also learned that it is important to create a timeline to outline project milestones. Creating a timeline was very useful as we were able to keep track of remaining tasks as well as workload amongst team members.

Overall we have many takeaways following the completion of this project that have allowed us to refine our analytical skills to be able to apply them in various fields and situations.

References

Matthieu Scherpf, Felix Gräßer, Hagen Malberg, Sebastian Zaunseder,
Predicting sepsis with a recurrent neural network using the MIMIC III database,
Computers in Biology and Medicine, Volume 113, 2019, 103395,ISSN 0010-4825,
https://doi.org/10.1016/j.compbiomed.2019.103395.
(https://www.sciencedirect.com/science/article/pii/S0010482519302720)

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Appendix

SQL Queries

1. **ICU Admissions**

```
SELECT
   a.hadm_id,
   a.subject_id,
   a.admittime,
   a.dischtime,
   CAST(TIMESTAMP_DIFF(a.admittime, p.dob, DAY) / 365.242 AS INT) AS age
FROM
   `physionet-data.mimiciii_clinical.admissions` a
JOIN
   `physionet-data.mimiciii_clinical.icustays` i
   ON a.hadm_id = i.hadm_id
```

```
JOIN
   `physionet-data.mimiciii_clinical.patients` p
   ON a.subject_id = p.subject_id
WHERE
   TIMESTAMP_DIFF(a.admittime, p.dob, DAY) / 365.242 >= 18;
```

2. Prediction and SIRS Parameters

```
SELECT
   ce.hadm_id AS hadm_id,
   ce.itemid AS itemid,
   ce.charttime AS charttime,
   -- Convert temperature values from Celsius to Fahrenheit for itemid 676 and
223762
   CASE
       WHEN ce.itemid IN (676, 223762) THEN SAFE_CAST(ce.value AS FLOAT64) *
9.0 / 5.0 + 32.0
       ELSE SAFE_CAST(ce.value AS FLOAT64)
   END AS value,
   -- Update label, category, and unit for Fahrenheit-converted itemids
   CASE
       WHEN ce.itemid IN (676, 223762) THEN 'Temperature Fahrenheit'
       ELSE di.label
   END AS label,
   CASE
       WHEN ce.itemid IN (676, 223762) THEN 'Routine Vital Signs'
       ELSE di.category
   END AS category,
   CASE
       WHEN ce.itemid IN (676, 223762) THEN '?F'
       ELSE di.unitname
   END AS unit,
   'chartevent' AS source
FROM
   `physionet-data.mimiciii_clinical.chartevents` ce
LEFT JOIN
```

```
    `physionet-data.mimiciii_clinical.d_items` di
    ON ce.itemid = di.itemid
WHERE
    ce.itemid IN (
        220050, 225309, 220179, 51, 455, 6701,       -- Systolic Blood Pressure
        220051, 220180, 225310, 8368, 8441, 8555,    -- Diastolic Blood Pressure
        223762, 223761, 676, 678,                    -- Temperature
        220045, 211,                                 -- Heart Rate
        220210, 224422, 224689, 224690, 618, 651,    -- Respiratory Rate
        615, 614,                                    -- Additional Respiration
Rate
        220227, 220277, 834, 646,                    -- Blood Oxygen Saturation
(SO2)
        220235, 778                                  -- CO2 Partial Pressure
    )

UNION ALL
SELECT
    le.hadm_id AS hadm_id,
    le.itemid AS itemid,
    le.charttime AS charttime,
    SAFE_CAST(le.valuenum AS FLOAT64) AS value,
    dli.label AS label,
    dli.category AS category,
    le.valueuom AS unit,
    'labevent' AS source
FROM
    `physionet-data.mimiciii_clinical.labevents` le
LEFT JOIN
    `physionet-data.mimiciii_clinical.d_labitems` dli
    ON le.itemid = dli.itemid
WHERE
    le.itemid IN (
        50820,                                       -- pH Value
        51301                                        -- WBC Count
    );
```

3. **Confirmed Sepsis ICD-9 Codes**

```
SELECT DISTINCT
   a.hadm_id,
   icd.ICD9_CODE,
   icd.short_title
FROM
   `physionet-data.mimiciii_clinical.admissions` a
JOIN
   `physionet-data.mimiciii_clinical.diagnoses_icd` d
   ON a.hadm_id = d.hadm_id
JOIN
   `physionet-data.mimiciii_clinical.d_icd_diagnoses` icd
   ON d.icd9_code = icd.icd9_code
WHERE
   d.icd9_code IN ('99591', '99592', '78552');
```

4. **Merge Sepsis_ICD-9 and Parameters**

```
SELECT s.*,d.ICD9_CODE,d.short_title
FROM `bana650-research-project.BANA650_GroupAJA.Prediction and SIRS
Parameters2` s
JOIN `bana650-research-project.BANA650_GroupAJA.Confirmed Sepsis_ICD9` d
ON s.hadm_id = d.hadm_id;
```

5. **Merge Confirmed Sepsis, Parameters and 18+**

```
SELECT DISTINCT f.*, i.subject_id, i.admittime, i.dischtime, i.age
FROM `bana650-research-project.BANA650_GroupAJA.Confirmed Sepsis and
Parameters2` f
JOIN `bana650-research-project.BANA650_GroupAJA.ICUAdmissions_18plus` i
ON f.hadm_id = i.hadm_id;
```

Raw Data Python Code

Submitted separately as .csv and .html files