**Early Prediction of Sepsis**

**using the Clinical data**

Sree Sarada Alapaty

HAP 780, G01440615

Dr. Sanja Avramovic

**Introduction:**

Sepsis is an emergency condition marked by the body's excessive and sometimes lethal response to infection, which can cause tissue damage, organ failure, or death. This report explores the creation of a predictive model that is designed to detect sepsis at an early stage, a critical need that is underscored by the alarming global and national statistics of the condition.  
  
In the United States alone, approximately 270,000 fatalities occur annually, and nearly 1.7 million individuals are diagnosed with sepsis. The critical impact of sepsis on healthcare outcomes is underscored by the fact that it is documented in more than one-third of all hospital fatalities in the United States (CDC). Internationally, the situation is significantly worse, with an estimated 30 million cases and 6 million deaths per year, including a high incidence among infants and children (WHO). From a financial perspective, sepsis is the most expensive condition for U.S. hospitals, costing $24 billion annually or 13% of total U.S. healthcare expenditures. Many of these costs are attributed to cases that are not diagnosed at admission (Paoli et al., 2018).

A white and purple text on a white background

Description automatically generated A screen shot of a white background

Description automatically generated

Figure 1. Facts about Sepsis

The necessity of early detection is further underscored by the fact that mortality rates increase by approximately 4-8% for each hour of delay in sepsis treatment (Kumar et al., 2006; Seymour et al., 2017). Despite the development of new definitions and criteria to improve diagnosis and management (Singer et al., 2016), the timely detection of sepsis continues to pose significant challenges, a fundamental void that this project aims to overcome.  
  
This issue will be addressed by examining the technique, implementation, and potential effects of a machine learning-based sepsis detection algorithm. This method reduces treatment lag and improves diagnostic precision, enabling faster clinical intervention and much higher patient survival rates. This research uses data integration, algorithmic innovation, and clinical insights to develop new sepsis management standards and reduce morbidity and death.

**Data Source and Data Collection:**

The dataset utilized in this study was obtained from the most recent edition of the MIMIC data, MIMIC-III, which includes more than 15,52,210 patients. The data was loaded into a Microsoft SQL server database to facilitate faster processing and enhanced visualization.

MIMIC-III is a vast, publicly accessible database that contains deidentified health-related information about more than forty thousand patients who were hospitalized in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Researchers are required to have finished the CITI course and obtain approval before they can receive access to the MIMIC-III dataset. This is because the researchers' study will involve human subjects.

A screenshot of a computer

Description automatically generated

Figure 2. Data loaded in Microsoft SQL Server

The dataset is a clinical time-series dataset specifically designed for the purpose of detecting sepsis in ICU patients. It includes vital signs, laboratory values, demographic, and outcome data, with repeated measurements over time for each patient. Imputation is done per-patient to retain temporal and patient-specific data integrity, which is essential for reliable sepsis prediction due to the high prevalence of missing values. The dataset contains a total of 40336 distinct patients.

A screenshot of a computer

Description automatically generated

Figure 3. Unique patients in the dataset

Early sepsis detection and prediction in intensive care unit (ICU) patients is the main result I am aiming for from this dataset. I aim to detect sepsis-onset patterns by examining vital signs, laboratory, and patient demographic data over time. Sepsis mortality can be reduced by early detection and timely treatment. The selection of predictor variables in this project was based on careful consideration of the factors most relevant to sepsis detection in ICU patients. While the dataset contains a total of 44 variables, I focused on 24 key variables that are most indicative of sepsis onset and progression. These variables include vital signs, laboratory results, and demographic data, which were chosen based on their relevance in clinical research and their impact on patient outcomes. The selected variables are outlined in the table below:

Table 1. Columns in dataset

|  |  |  |
| --- | --- | --- |
| **Category** | **Variable** | **Description** |
| Patient Info | Patient\_ID | Unique identifier for each patient |
|  | Hour | Time in hours since ICU admission |
| Vital Signs | HR | Heart rate (beats per minute) |
|  | O2Sat | Pulse oximetry (%) |
|  | Temp | Temperature (°C) |
|  | MAP | Mean arterial pressure (mm Hg) |
|  | Resp | Respiration rate (breaths per minute) |
|  | FiO2 | Fraction of inspired oxygen (%) |
| Lab Values | PaCO2 | Partial pressure of carbon dioxide (mm Hg) |
|  | BUN | Blood urea nitrogen (mg/dL) |
|  | Chloride | Chloride (mmol/L) |
|  | Creatinine | Creatinine (mg/dL) |
|  | Glucose | Serum glucose (mg/dL) |
|  | Hct | Hematocrit (%) |
|  | Hgb | Hemoglobin (g/dL) |
|  | WBC | White blood cell count (count \* 10^3/µL) |
|  | Platelets | Platelet count (count \* 10^3/µL) |
| Demographics | Age | Age in years (100 for patients aged 90 or above) |
|  | Gender | Female (0) or Male (1) |
|  | Unit1 | ICU unit identifier (MICU) |
|  | Unit2 | ICU unit identifier (SICU) |
| Admin Info | HospAdmTime | Hours between hospital and ICU admission |
|  | ICULOS | ICU length-of-stay (hours since ICU admission) |
| Outcome | SepsisLabel | Indicates sepsis presence (1) or absence (0) |

**Data Preprocessing:**

The sepsis dataset used in this study is a part of a comprehensive healthcare database that specifically focuses on patient outcomes, with a particular emphasis on the diagnosis of sepsis. The dataset was provided in formats that can be directly imported into a Relational Database Management System (RDBMS).  
  
Microsoft SQL Server 2020 was chosen for this project because of its strong performance in processing huge datasets and running complicated queries, both of which are critical for medical data analysis. The decision to choose SQL Server for this task was based on its compatibility with the dataset's structure and its capacity to handle advanced analytical capabilities that are crucial for processing and evaluating healthcare data.  
  
Microsoft SQL Server Management Studio (SSMS), version 20, was utilized to oversee the database and streamline the presentation of the data structure. SSMS provides a full interface for conducting SQL queries, scripting, and overall database management. One of the first jobs was to generate an Entity-Relationship Diagram (ER Diagram) for the sepsis\_data table in SSMS.

The ER Diagram played a crucial role in elucidating the organization and interconnections of patient data, facilitating the implementation of more efficient query and data extraction techniques that are specifically geared to assess sepsis indicators and outcomes. This representation was crucial in simplifying the intricacy of the data, guaranteeing that all project participants could comprehend the data structure without having to manually examine the database design.

Figure 4. ER diagram of dbo.Sepsis\_data

A screenshot of a computer

Description automatically generated

To determine the extent of the dataset and the quality of the preliminary data, the project's data exploration phase started with obtaining every record from the sepsis\_data database. This stage was crucial in understanding the sheer size of the dataset and strategizing further analyses. In addition, all the column names in the table were methodically listed to identify the variables that are available, which serves as a basis for selecting the relevant data for further analysis. The dataset's granularity was assessed by counting unique Patient\_ID values, which was necessary for precise patient-level analyses.

Next, I performed an in-depth analysis of missing values, calculating the proportion of missing data in each column and ranking them from highest to lowest. This step was essential as it enabled me to promptly ascertain which columns were most impacted by the absence of data. To maintain the accuracy and effectiveness of my study, I made the decision to eliminate any columns that included above 95% missing data. This method facilitated the optimization of the dataset, enabling me to concentrate on the most dependable and important variables for subsequent analysis.

A screenshot of a computer

Description automatically generated

Figure 5. Missing percentages of the dataset

To assure the dataset's cleanliness and usefulness, I eliminated these columns and created a new table called dbo.sepsis\_final\_cleaned. This table eliminates the very incomplete columns, allowing me to concentrate on more accurate and valuable data.

Figure 6. dbo.sepsis\_final\_cleaned table

A screenshot of a computer

Description automatically generated

In this processed dataset, I resolved the remaining instances of missing values. I used median imputation to fill missing values in continuous variables like heart rate (HR), temperature (Temp), and mean arterial pressure (MAP). This method was chosen because the median is less influenced by outliers, which is especially useful in medical data where extreme values can have significant effects on the study.  
  
I used mode imputation for the categorical variables Unit1 and Unit2, which correspond to different types of hospital units, especially MICU and SICU. This guaranteed that the categories with the highest frequency, specifically MICU or SICU, were selected to replace any missing values. This approach maintained the original distribution of the data and ensured consistency across the whole dataset.A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Figure 7. Median Imputation of HR

A screenshot of a computer

Description automatically generated

dbo.sepsis\_final\_cleaned

Creating Analytical Features for Sepsis Detection:

During the data enrichment step, I added significant columns to the dbo.sepsis\_final\_cleaned dataset to capture critical signs using the Systemic Inflammatory Response Syndrome (SIRS) criteria. For instance, the condition\_temp variable was marked as '1' when temperatures fell outside the normal range (below 36°C or above 38°C), indicating the possibility of fever or hypothermia, which are both important for diagnosing sepsis. Condition\_HR was identified for heart rates above 90 beats per minute, condition\_Resp for respiratory rates above 20 breaths per minute or PaCO2 levels below 32 mmHg, and condition\_wbc for aberrant white blood cell counts above 12,000 or below 4,000 cells per microliter. The flags were aggregated to compute a comprehensive SIRS score, which served as a measurable indicator of systemic inflammation among the patients.  
  
After analyzing the distribution of SIRS scores, it was found that most patients had lower scores and that a significant portion of them had more severe inflammatory reactions. The **SIRS scores** were distributed as follows: **1** (820,654 patients), **2** (517,795 patients), **3** (199,965 patients), and **4** (13,896 patients). These steps not only improved the dataset but also established a solid basis for subsequent statistical analysis and predictive modeling, with the goal of enhancing sepsis detection and improving patient outcomes.

A screenshot of a computer

Description automatically generated

Figure 8. SIRS Score

After enriching the dataset with SIRS criteria and calculating the SIRS scores, I further classified the patients based on their sepsis status at the time of ICU admission into a new column called **SepsisType**. Patients were categorized into three groups: **NonSepsis** for those who did not develop sepsis, **SepsisBeforeAdm** for those admitted with sepsis, and **SepsisAfterAdm** for those who developed sepsis after admission. The distribution of these categories was as follows: 1,379,800 patients were classified as NonSepsis, 3,646 as SepsisBeforeAdm, and 168,764 as SepsisAfterAdm. This classification allowed for a more targeted analysis of sepsis progression and outcomes, laying the groundwork for further in-depth study and predictive modeling.

A screenshot of a computer

Description automatically generated

Figure 9. SepsisType

Exploratory Data Analysis**:**

In this analysis, I calculated descriptive statistics for key physiological and demographic variables, including Age, Heart Rate (HR), ICU Length of Stay (ICULOS), Respiration Rate (Resp), Temperature (Temp), and White Blood Cell Count (WBC) across three sepsis classifications: NonSepsis, SepsisBeforeAdm, and SepsisAfterAdm. The results showed that patients in the sepsis groups, particularly SepsisAfterAdm, exhibited higher heart rates, longer ICU stays, and increased respiration rates compared to NonSepsis patients, reflecting greater physiological stress and severity of their condition. The temperature remained relatively stable across all groups, with slight elevations in the sepsis categories. These findings provide essential insights into how vital signs and patient characteristics vary with different sepsis conditions, informing further analysis and predictive modeling.

A screenshot of a computer

Description automatically generated

Figure 10. Descriptive Statistics for Sepsis Classifications (NonSepsis, SepsisBeforeAdm, SepsisAfterAdm)

To prepare the data for export, I first converted the important variables to nvarchar(max) format and made the sepsis\_final\_cleaned\_export table, which had important fields as sepsisType and SIRS Score. A subset of this data, sepsis\_sub, was then created and exported to Weka for additional analysis, allowing me to use machine learning algorithms to discover trends and develop predictive models for sepsis outcomes.

A screenshot of a computer

Description automatically generated

Figure 11. sepsis\_sub dataset

The last step of the data preprocessing was done in Weka. After loading the csv file in Weka I removed the attributes that were not relevant for the machine learning process: Patient\_ID, Unit1, Unit2. The first step was to convert certain features from numeric to nominal after exporting the sepsis\_sub dataset to Weka. Columns like SepsisLabel, Condition\_temp, condition\_HR, condition\_Resp, condition\_wbc, SIRS\_Score, and SepsisType were among those for which this was done. This conversion guaranteed that the binary and category variables were handled correctly in the analysis. Subsequently, I implemented a normalization filter on the dataset, which standardized the numerical attributes while keeping the nominal (categorical) columns unchanged. This method guaranteed that the binary categorical columns were kept unchanged, maintaining their categorical nature, while the numerical features were normalized to enhance the effectiveness of modeling.

A screenshot of a computer

Description automatically generated

Figure 12. Weka Preprocessing

**Feature Selection:**

Attribute visualization was useful in determining the effect of various attributes on sepsis classification. The scatter plot displays the correlation between Mean Arterial Pressure (MAP) and ICU Length of Stay (ICULOS), with data points categorized by sepsisType. This visualization shows that patients who developed sepsis after ICU admission (SepsisAfterAdm) have more variable MAP values and longer ICU stays than non-sepsis patients. The data indicates that patients with lower mean arterial pressure (MAP) values and longer stays in the intensive care unit (ICU) are more likely to acquire sepsis after being admitted. This information is crucial for understanding patient outcomes and emphasizes the significance of these factors in predicting sepsis.

A screenshot of a computer

Description automatically generated

Figure 13. Attribute Visualization - MAP & ICULOS

Following the visualization of the data, I experimented with a few different attribute selection methods to find out the ranking of the attributes. I choose to retain all the initially chosen attributes for the classification phase.

A screenshot of a computer program

Description automatically generated

Figure 14. Attribute selection - GainRatioAttributeEval

In the classification part of the project, I used several well-known machine learning algorithms to find the best model for predicting early sepsis. The models I concentrated on were Logistic Regression, Random Forest, and Naïve Bayes. The selection of each of these models was based on its distinctive strengths and its ability to deal with various aspects of the dataset. The RandomForest model, one of the models that demonstrated excellent predictive potential in this study, is shown with its performance metrics below. I have split the data into 60% training and 40% test.

A screenshot of a computer

Description automatically generated

Figure 15. RandomForest Model

In the Weka Experiment Environment, I compared three models: Logistic Regression, Random Forest, and Naive Bayes, utilizing the Area\_under\_ROC as a comparison metric. The findings demonstrate that Random Forest had superior performance compared to the other models, with an Area\_under\_ROC of 0.99. In contrast, Logistic Regression and Naive Bayes Area\_under\_ROC achieved 0.80 and 0.79, respectively. Given the results of this study, the Random Forest model is the best suited option among the models that were evaluated for this project because it can accurately forecast the onset of early sepsis.

A screenshot of a computer

Description automatically generated

Figure 16. Weka Experimenter – model comparison on Naïve Bayes, RandomForest, Logistic Regression (base)

I have also included tables with the results of the top two performing classification models, Random Forest and Logistic Regression.

Table 2. Logistic Regression

**Logistic Regression:**

=== Summary ===

Correctly Classified Instances 39489 98.7225 %

Incorrectly Classified Instances 511 1.2775 %

Kappa statistic 0.2289

Mean absolute error 0.022

Root mean squared error 0.1045

Relative absolute error 75.2725 %

Root relative squared error 86.9682 %

Total Number of Instances 40000

=== Detailed Accuracy by Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.869 0.987 1.000 0.994 0.355 0.970 1.000 0

0.131 0.000 0.975 0.131 0.232 0.355 0.970 0.374 1

Weighted Avg. 0.987 0.856 0.987 0.987 0.982 0.355 0.970 0.990

=== Confusion Matrix ===

a b <-- classified as

39412 2 | a = 0

509 77 | b = 1

**Random Forest:**

=== Summary ===

Correctly Classified Instances 39644 99.11 %

Incorrectly Classified Instances 356 0.89 %

Kappa statistic 0.5704

Mean absolute error 0.0152

Root mean squared error 0.077

Relative absolute error 52.0123 %

Root relative squared error 64.0795 %

Total Number of Instances 40000

=== Detailed Accuracy by Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.590 0.991 1.000 0.996 0.624 0.997 1.000 0

0.410 0.000 0.960 0.410 0.574 0.624 0.997 0.867 1

Weighted Avg. 0.991 0.582 0.991 0.991 0.989 0.624 0.997 0.998

=== Confusion Matrix ===

a b <-- classified as

39404 10 | a = 0

346 240 | b = 1

Table 3. RandomForest

In addition to the primary analysis, I conducted classification using the SepsisType variable. This allowed me to investigate how well different models could distinguish between various sepsis types, including SepsisBeforeAdm, SepsisAfterAdm, and NonSepsis. By including SepsisType in the classification process, I gained deeper insights into the model's performance across different sepsis categories, thereby enhancing the overall robustness and applicability of the predictive analysis.

During the classification analysis of sepsis prediction utilizing the SepsisType variable, three models—Logistic Regression, Random Forest, and Naive Bayes—were examined for their abilities to accurately predict sepsis. Logistic Regression had significant precision, especially in detecting NonSepsis instances, although it exhibited limits in predicting the SepsisAfterAdm and SepsisBeforeAdm categories. The Random Forest model demonstrated superior performance in addressing class imbalance, particularly in regard to recall and F-measure for the SepsisBeforeAdm category, highlighting its appropriateness for early sepsis identification. In contrast, the Naive Bayes model, while efficient, had lower overall accuracy and faced greater difficulties with the minority classes, rendering it less suitable for this dataset. The models were evaluated by comparing important metrics such as precision, recall, and F-measure to identify the most efficient technique for early sepsis detection.

A screenshot of a computer

Description automatically generated

Figure 18. RandomForest -SepsisType

**Results:**

The table below presents the weighted average results for ROC area, precision, and recall across the different models for the class SepsisType. Among these, the Random Forest model demonstrated the highest performance with a ROC area of 0.989, indicating its exceptional capability to distinguish between different sepsis types. Logistic Regression also performed reliably, with a ROC area of 0.790, showing solid performance in this classification task. Naïve Bayes, while slightly lower, still showed a reasonable performance with a ROC area of 0.788. These metrics provide valuable insights into each model's predictive power and generalization ability when classifying sepsis types, with Random Forest clearly emerging as the best-performing model in this comparison.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **ROC Area** | **Precision** | **Recall** |
| Logistic Regression | 0.790 | 0.905 | 0.917 |
| RandomForest | **0.989** | 0.967 | 0.966 |
| Naïve Bayes | 0.788 | 0.889 | 0.878 |

Table 4. Classification Model Results – Class: SepsisType

The models showed different amounts of performance for the SepsisLabel class 1. The Logistic Regression model demonstrated a precision of 0.975, but its recall was only 0.131. Additionally, it had a ROC area of 0.970, suggesting that although it accurately classified a large number of non-sepsis cases, it faced difficulties in correctly identifying all positive sepsis instances. However, RandomForest demonstrated superior performance in terms of precision (0.960), recall (0.410), and ROC area (0.997). This indicates that it was more successful in accurately identifying sepsis patients while minimizing the number of false negatives. Naïve Bayes, on the other hand, exhibited a significantly lower accuracy of 0.088 and a recall of 0.374, along with an ROC area of 0.924. This suggests that although it detected a greater number of sepsis cases, it also generated a higher proportion of false positives. In general, RandomForest proved to be the most efficient model for identifying sepsis cases (SepsisLabel class 1), whereas Logistic Regression and Naïve Bayes demonstrated their own strengths in precision and recall.

Table 5. Classification Model Results - Class: SepsisLabel - 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **ROC Area** | **Precision** | **Recall** |
| Logistic Regression | 0.970 | 0.975 | 0.131 |
| Random Forest | **0.997** | 0.960 | 0.410 |
| Naive Bayes | 0.924 | 0.088 | 0.374 |

**Future Work:**

To make sepsis prediction even more accurate, more work needs to be done to fix data imbalances and investigate advanced imputation methods. Improving the generalizability of models by testing them on a wider range of patients will help make sure they work for everyone. Furthermore, for practical deployment in healthcare settings, improving the interpretability of the models and incorporating real-time prediction capabilities into clinical procedures would be essential.

**Conclusion:**

This project focused on developing and evaluating machine learning models for the early detection of sepsis, a critical condition with significant implications for patient outcomes. The models tested included Logistic Regression, Random Forest, and Naive Bayes, each bringing distinct strengths and challenges to the task.

Among the models, the Random Forest algorithm had the best overall performance, with a ROC area of 0.997 and high scores in both accuracy and recall. This made it the best model for telling the difference between sepsis and non-sepsis cases. Its ability to process complicated, high-dimensional data enabled it to more accurately identify patients at risk of developing sepsis, particularly in cases when sepsis occurred after admission. Logistic Regression had high performance, with an ROC area of 0.970. It excelled in precision with a score of 0.975, but encountered difficulties in recall, specifically when distinguishing between SepsisAfterAdm and SepsisBeforeAdm cases. Naive Bayes, despite its computational efficiency, exhibited limits in managing the intricacy of sepsis prediction, as evidenced by its trouble with recall. Its ROC area was measured at 0.924.  
  
Although Random Forest was found to be the most successful model overall, there is still potential for improvement, especially in terms of improving the models' capacity to accurately identify all instances of sepsis, even those that are more complex or uncommon. Based on what this study showed, it seems like the performance and dependability could be even better in real-world settings if it were further improved. This could be done by adding more clinical features or using more advanced methods.  
  
To summarize, this effort showcases the capabilities of machine learning models, specifically Random Forest, in detecting sepsis at an early stage. These models give healthcare clinicians with significant tools that allow for earlier intervention and have the potential to enhance patient outcomes. Nevertheless, continuous progress and verification in real-world medical environments are crucial to guarantee that these models can be efficiently implemented to assist in crucial decision-making regarding the treatment of sepsis.

**Limitations:**

The project had several challenges that could affect how well and how widely the models could be used.

* One major problem with the dataset was that some numbers were missing. Imputation methods were used to fix this problem. Imputation maintains dataset completeness, but if missing data is not random, it might add bias and impact model accuracy.
* Data imbalance, especially between NonSepsis and Sepsis classes, may have biased predictions, as Logistic Regression and Naive Bayes struggle with sepsis recall.
* While effective, the Random Forest model's complexity creates interpretability difficulties, which are critical in clinical contexts.
* Based on available data, feature selection may have missed clinical variables that could improve model accuracy.
* The dataset used to train the algorithms may limit their application to other patient populations or clinical settings.
* The computational intensity of models like Random Forest may also hinder real-time implementation. Finally, overfitting and parameter tuning may impact model performance on unseen data.

Future study must address these constraints to improve sepsis prediction models' reliability and effectiveness.

**References:**

* Reyna, Matthew, et al. "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019" (version 1.0.0). *PhysioNet* (2019), <https://doi.org/10.13026/v64v-d857>.
* 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.
* <https://journals.lww.com/ccmjournal/fulltext/2020/02000/early_prediction_of_sepsis_from_clinical_data__the.10.aspx>
* <https://physionet.org/content/challenge-2019/1.0.0/#description> --Dataset
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6964870/>
* <https://www.mdpi.com/2077-0383/12/17/5658>
* <https://www.beckmancoulter.com/products/hematology/early-sepsis-detection>
* <https://mimic.mit.edu/docs/gettingstarted/>
* Weka - <https://www.geeksforgeeks.org/how-to-use-classification-machine-learning-algorithms-in-weka/>
* <https://physionet.org/content/challenge-2019/1.0.0/physionet_challenge_2019_ccm_manuscript.pdf>