Fuzzy Inference System for Early Sepsis Detection

Introduction

a. Definition of the Problem

Sepsis is a critical medical condition that arises when the body's response to an infection triggers widespread inflammation. If not treated promptly, it can lead to organ failure, septic shock, and even death. Diagnosing sepsis is challenging due to its nonspecific symptoms, which can overlap with other medical conditions. This project addresses the problem of early and accurate sepsis detection by analyzing patient data from the first 12 hours of monitoring to predict the risk of sepsis.

b. Importance of the Solution

Early detection of sepsis is essential to improving patient outcomes and reducing mortality rates. Delayed or inaccurate diagnoses can lead to catastrophic consequences, including increased healthcare costs and prolonged hospital stays. Current diagnostic methods often rely on manual assessments, which are prone to subjectivity and delays. Automating the risk assessment of sepsis using a data-driven approach enhances consistency, speed, and accuracy, making it a crucial step in modern healthcare. By leveraging advanced techniques like Fuzzy Inference Systems (FIS), we can create a system that supports clinicians in making timely and informed decisions.

c. Suitability Assessment of FIS Application on the Problem

Fuzzy Inference Systems are particularly suitable for solving the difficulty of sepsis detection, as they are designed to address uncertainty and imprecise data. Clinical parameters such as heart rate, blood pressure and hemoglobin levels usually do not follow strict thresholds and can vary significantly between patients. The FIS uses logical rules based on expert knowledge to interpret these variables in a way that mimics human reasoning.

Additionally, FIS can integrate medical expertise directly into the system, ensuring that its decisions align with established clinical practices. Its flexibility allows it to adapt to different datasets and clinical environments, making it a versatile tool for healthcare applications. This project demonstrates the potential of FIS to provide interpretable and effective solutions for one of the most urgent problems in modern medicine.

2. Data

a. Data Selection Approach

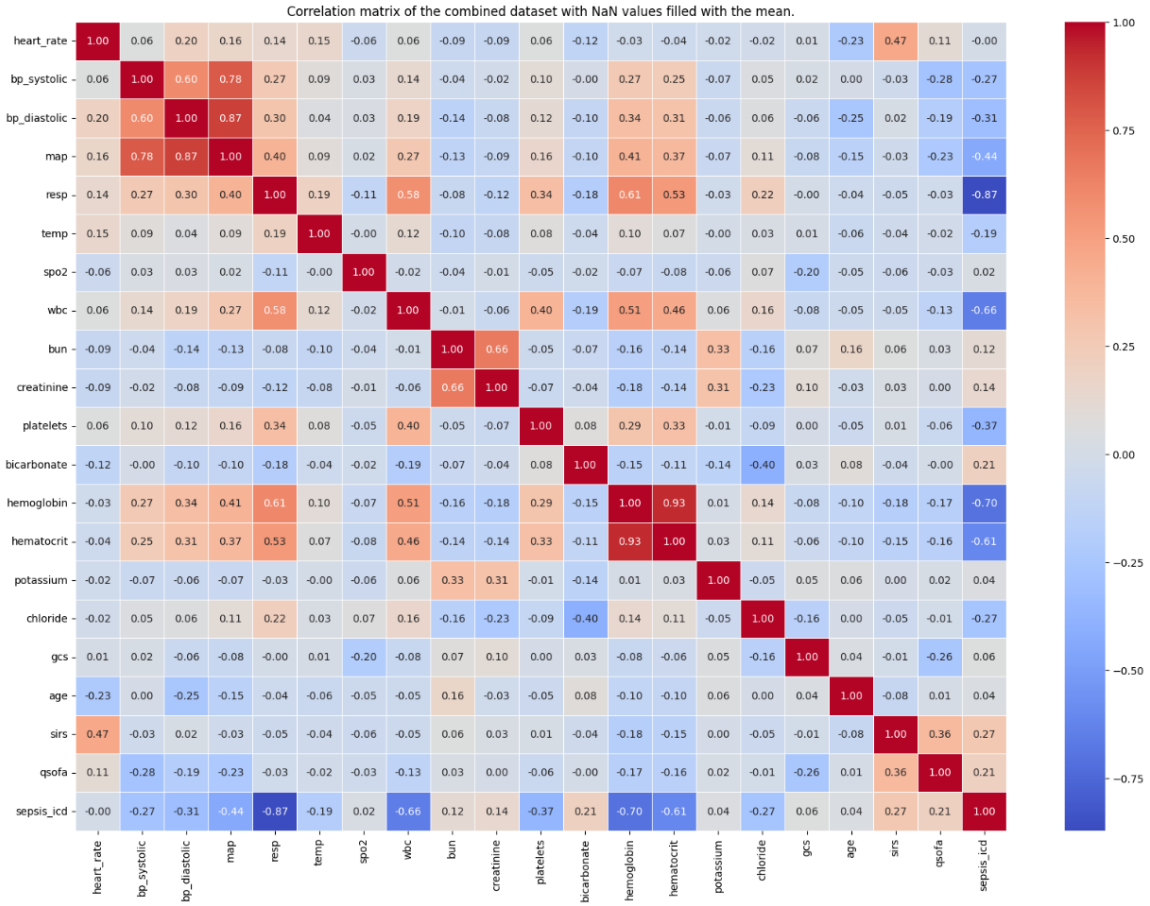
The data employed in this project was sourced from clinical records containing vital signs and laboratory results of patients with and without sepsis. To ensure balanced and representative samples, 200 records each were randomly selected from both the sepsis and non-sepsis datasets. The selection process focused on capturing a diverse range of patient profiles while maintaining equal representation of positive and negative instances. Only data from the first 12 hours of monitoring was considered, as early intervention is crucial in sepsis management.

b. Data Distribution of the Employed Set

i. Correlation Coefficients of the Inputs and the Output

The correlation matrix for the combined dataset reveals the relationships between input variables and the output variable (sepsis_icd). Key insights include:

- There is no significant positive correlations. The highest correlation observed is with SIRS at 0.27, which is considered relatively low.
- While the correlation matrix for the combined dataset does not exhibit significant positive correlations, several high negative correlations are present. Notably, respiratory rate (RESP) shows a correlation of -0.87, hemoglobin (HGB) -0.70, and white blood cell count (WBC) -0.66. Despite being negative, these strong correlations may provide valuable insights for further analysis



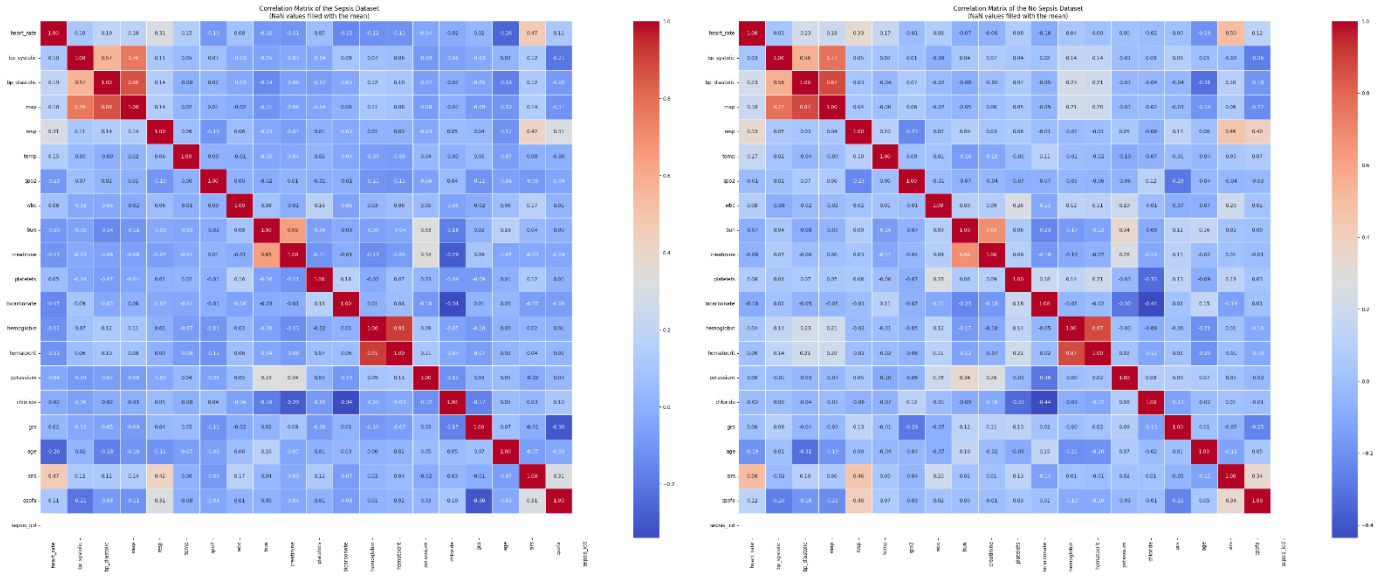
ii. Correlation Coefficients of Each Input and the Resting Inputs

Separate correlation matrices for the sepsis and no-sepsis datasets were analyzed:

- In the sepsis dataset, features such as MAP and BP_Diastolic exhibit a strong correlation (0.86), and Hemoglobin and Hematocrit also have a very high correlation (0.91). These high correlations suggest that these features may provide redundant

information. In feature selection for rule determination, highly correlated features can be reduced to avoid multicollinearity, which can lead to overfitting and unnecessary complexity. For instance, instead of including both Hemoglobin and Hematocrit, it may be more efficient to include only one of them to maintain model simplicity while retaining most of the useful information. This approach helps in improving model performance and interpretability

- In the no-sepsis dataset, similar interdependencies were observed, albeit with slightly different correlation strengths. For instance, MAP and BP_Diastolic exhibit a correlation of 0.77, which, while lower than in the sepsis dataset, still indicates a strong relationship. Additionally, a notable correlation of 0.68 was found between Bun and Creatinine. These correlations should be considered in feature selection for rule determination, as features with high correlations may contribute redundant information. As in the sepsis dataset, reducing such redundancies can help simplify the model and improve its efficiency



iii. visualization of the data distributions

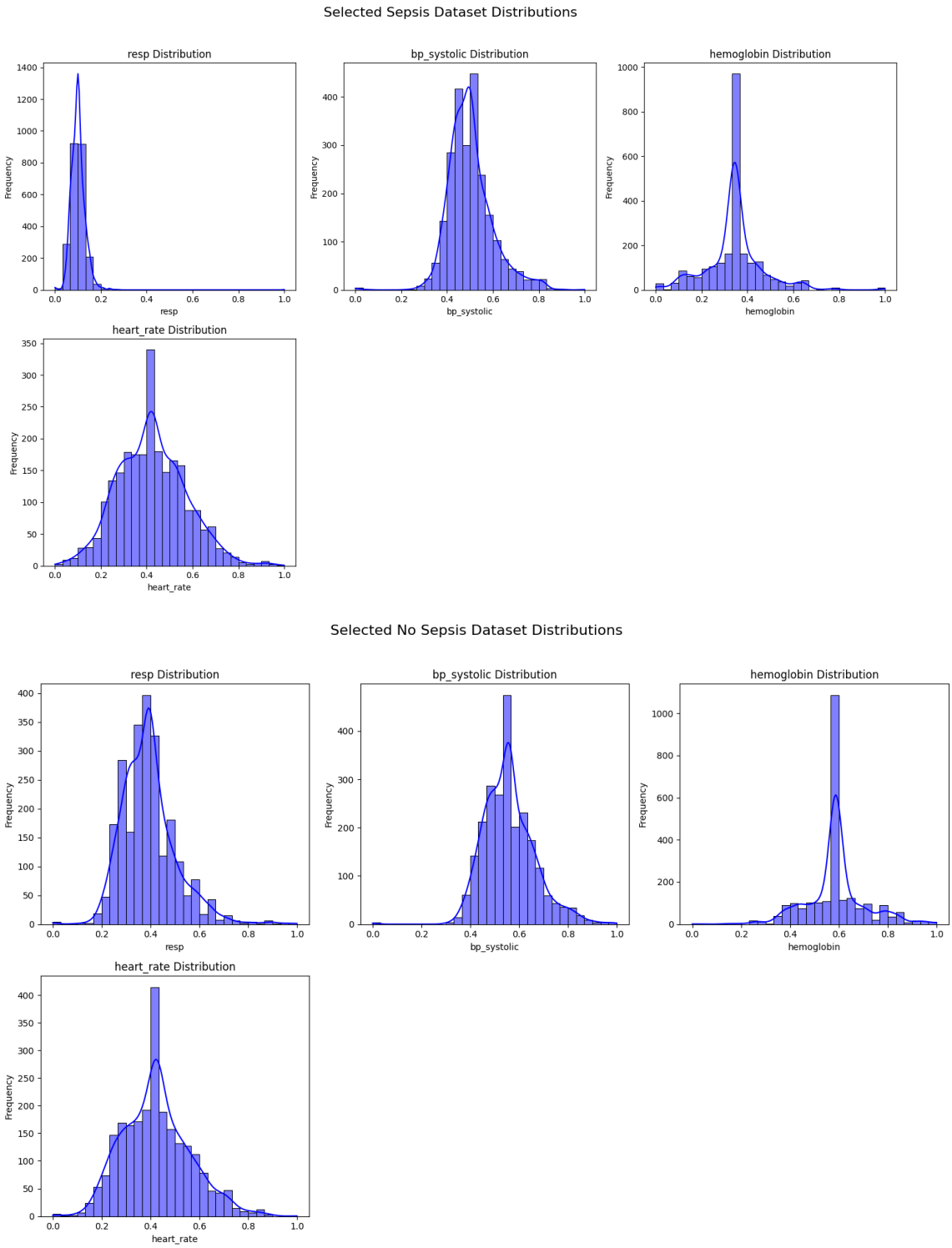
The selected features—respiratory rate (resp), systolic blood pressure (bp_systolic), hemoglobin, and heart rate—are visualized here to better understand their distributions. These visualizations played a crucial role in the development of fuzzy rules, as they helped reveal patterns and value ranges that informed the creation of membership functions.

For example, respiratory rate (resp) shows a tendency to have smaller values in the sepsis dataset, whereas the no-sepsis dataset exhibits more average or higher values. A similar trend is observed for hemoglobin levels, where the sepsis dataset tends to have lower values compared to the no-sepsis dataset. These insights helped identify the ranges where membership functions would be most effective.

While features like respiratory rate, systolic blood pressure, and hemoglobin were chosen for their strong correlation with the target variable, heart rate was also included due to its clinical significance, despite a relatively lower correlation. Heart rate is an important clinical

indicator, and its inclusion ensures that the fuzzy rules account for its relevance in sepsis prediction.

By incorporating both statistically significant features and clinically important variables, the fuzzy rules were designed to effectively capture the nuances of sepsis prediction. This balanced approach highlights the importance of using both data-driven insights and domain knowledge in feature selection.



c. Statistical Summarization of Data

Sepsis Dataset:

- All features except temp have only positive instances
- Temp has 4 negative instances, meaning 4 rows have negative values for temperature, likely due to measurement errors or unconventional units.

No-Sepsis Dataset:

- Similar to the sepsis dataset, most features have exclusively positive instances.
- Temp again has 1 negative instance, which may point to the same issue.

General information:

Sepsis Dataset:

1. **Heart Rate**: Values range from 40 to 169 bpm, with a mean of 94.55 and a standard deviation of 19.11. All values are positive, indicating a consistent distribution.
2. **Systolic Blood Pressure**: The range spans from 0 to 221 mmHg, with a mean of 110.83 and a standard deviation of 21.45. There are no negative values, but the wide range suggests significant variability among patients.
3. **Diastolic Blood Pressure**: Ranges from 0 to 160 mmHg, with a mean of 58.17 and a standard deviation of 13.91. All values are positive.
4. **Mean Arterial Pressure (MAP)**: Values range from 21 to 151 mmHg, with a mean of 73.34 and a standard deviation of 14.28. All values are positive.
5. **Respiratory Rate**: Ranges from 0 to 211 breaths per minute, with a mean of 21.36 and a standard deviation of 7.28. All instances are positive.
6. **Temperature**: The minimum is -17.78°C (likely an error), and the maximum is 40.22°C. The mean is 36.74°C, with a standard deviation of 2.71. Four negative values are observed, indicating potential anomalies.

No-Sepsis Dataset:

1. **Heart Rate**: Values range from 30 to 157 bpm, with a mean of 83.85 and a standard deviation of 17.60. All instances are positive.
2. **Systolic Blood Pressure**: Ranges from 0 to 215 mmHg, with a mean of 120.22 and a standard deviation of 22.62. No negative values are present.
3. **Diastolic Blood Pressure**: Values range from 0 to 141 mmHg, with a mean of 60.07 and a standard deviation of 14.48. All values are positive.
4. **Mean Arterial Pressure (MAP)**: The range is 1 to 154 mmHg, with a mean of 78.26 and a standard deviation of 15.42. All values are positive.
5. **Respiratory Rate**: Ranges from 0 to 47 breaths per minute, with a mean of 18.48 and a standard deviation of 5.25. No negative values are present.

6. **Temperature**: The minimum is -17.78°C, and the maximum is 39.33°C. The mean is 36.71°C, with a standard deviation of 1.34. One negative value is observed, which could be a data entry error.

Insights:

- Negative values, particularly in temperature, highlight the need for data cleaning to ensure reliability.
- All other parameters, such as respiratory rate and heart_rate, show consistent distributions with no negative instances, making them reliable for use in the fuzzy inference system.

3. Method

a. Definition of FIS

A Fuzzy Inference System (FIS) is a rule-based system that processes input variables using fuzzy logic to produce an output. It incorporates fuzzy sets and logical rules to model complex systems where precise mathematical models are difficult to define. In this study, the FIS evaluates patient data to predict the risk of sepsis, leveraging fuzzy sets to capture the inherent uncertainty in medical data.

b. Implementation Details

The implementation is built using the **scikit-fuzzy** library in Python, which provides tools to define fuzzy variables, membership functions, and rules.

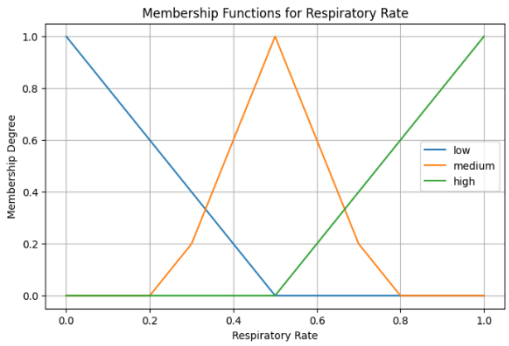
- **Libraries Used**:
    - scikit-fuzzy: To implement fuzzy logic, define rules, and perform defuzzification.
    - pandas and numpy: For data processing and handling large datasets.
    - matplotlib and seaborn: For data visualization and distribution analysis.
    - sklearn: To calculate performance metrics like AUC, ROC curves, and confusion matrices.
- **Ready-to-Use Methods**:
    - Membership function generation (trimf for triangular functions).
    - Fuzzy rule definition (ctrl.Rule).
    - Inference computation (ctrl.ControlSystemSimulation).
- **Self-Coded Components**:
    - Data preprocessing to handle missing values.
    - Selection of important features based on correlation.

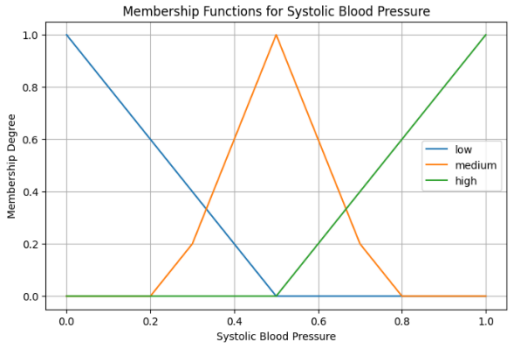
- Evaluation pipeline for the FIS using custom thresholds and comparison with ground truth labels.

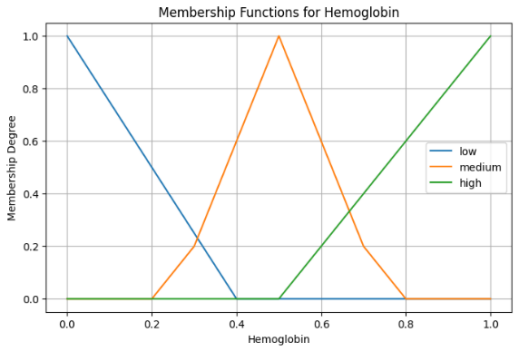
c. Rule Number and Fuzzy Set Specifications

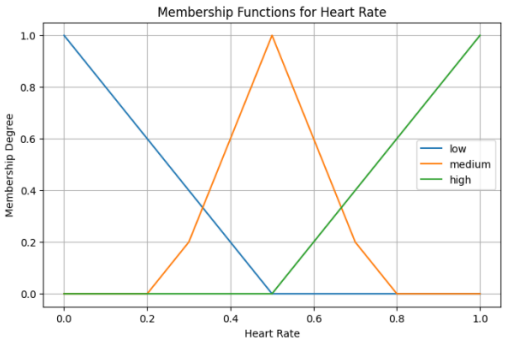
- **Fuzzy Sets**:
  - Input Variables:
    - Heart Rate, Respiratory Rate, Systolic BP, and Hemoglobin.
  - Membership Functions: Low, Medium, and High for each variable, defined using triangular functions.
    - Membership Function Visualization

For normalized data:

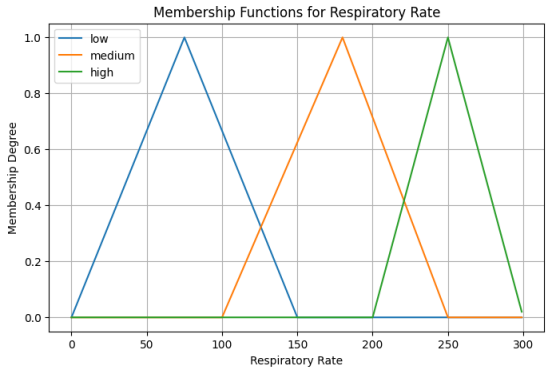


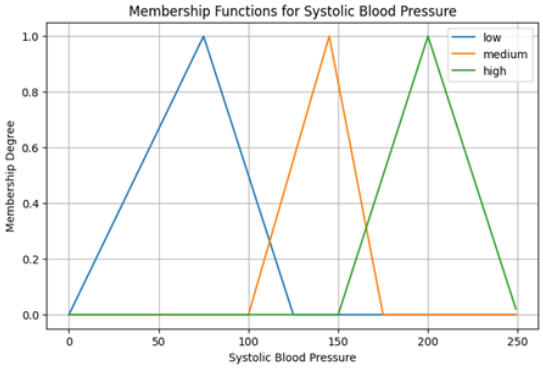


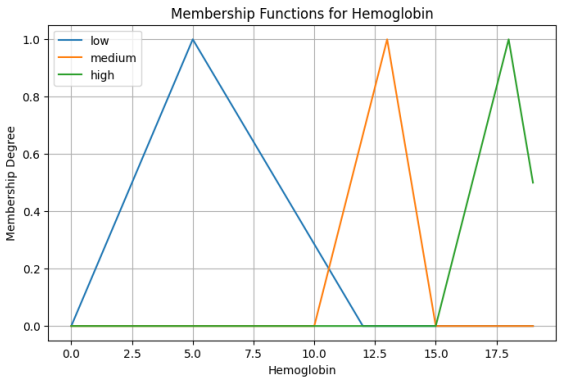


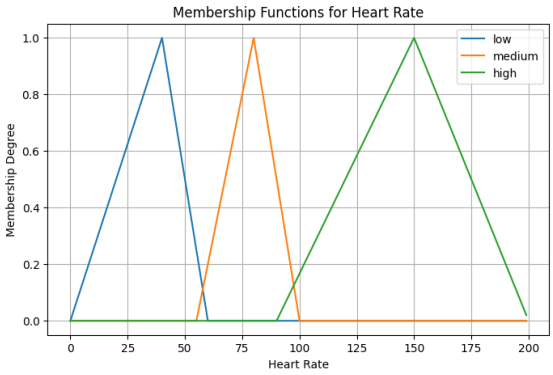


For no normalized data



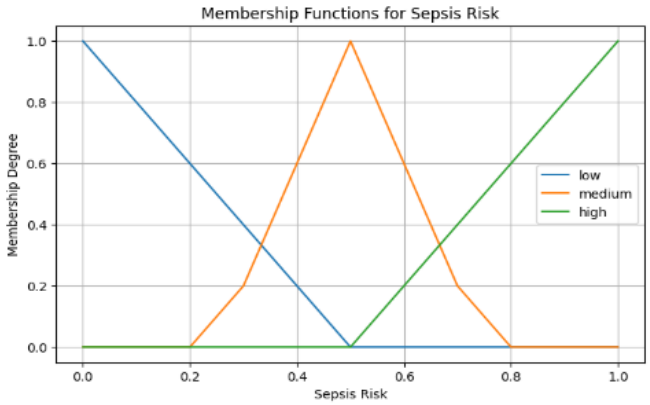




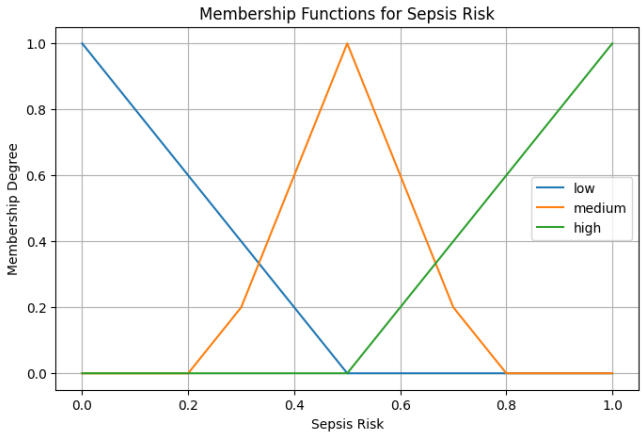


- Output Variable: Sepsis Risk, with Low, Medium, and High fuzzy sets.
  - Membership Function Visualization

For normalized data:



For no normalized data:



In normalized data for Hemoglobin, the membership function ranges (low, medium, high) were adjusted due to the skewed data distribution observed during analysis. This ensures better representation and separation of the fuzzy sets. The rules were tested with both normalized and non-normalized data. For normalized data, boundaries were established based on normalized values to maintain consistency and ease of computation. For non-normalized data, the membership function ranges were manually defined using the feature's maximum

and minimum values to better capture the distribution of the raw data. This dual approach helped evaluate the impact of normalization on rule definition and model performance.

- **Rules**: Six rules were defined to capture the relationships between input variables and sepsis risk:

1. If Heart Rate is low and BP Systolic is low or Respiratory Rate is low, then Sepsis Risk is high.
2. If Respiratory Rate is medium and Hemoglobin is low, then Sepsis Risk is medium.
3. If Hemoglobin is high, BP Systolic is high, and Heart Rate is medium, then Sepsis Risk is low.
4. If Heart Rate is medium and Hemoglobin is low, then Sepsis Risk is high.
5. If Heart Rate is medium and BP Systolic is medium, then Sepsis Risk is medium.
6. If Respiratory Rate is medium or Hemoglobin is medium, then Sepsis Risk is low.

- **Defuzzification Method**:
    - The Center of Area (COA) method is used to calculate the crisp output for sepsis risk.

d. Definition of Performance Evaluators

The performance of the FIS was evaluated using the following metrics:

- **Sensitivity (TPR)**: Measures the ability to correctly identify sepsis cases.
- **Specificity (TNR)**: Measures the ability to correctly identify non-sepsis cases.
- **Precision**: Evaluates the accuracy of sepsis predictions.
- **F1 Score**: A balance of precision and recall.
- **Accuracy**: The overall correctness of the predictions.
- **ROC Curve and AUC**: Visualize the trade-off between sensitivity and specificity and measure the classifier's ability to distinguish between classes.

These metrics ensure a robust evaluation of the FIS and its suitability for sepsis risk prediction. Adjustments made to the membership functions, particularly for Hemoglobin, were validated against these performance metrics.

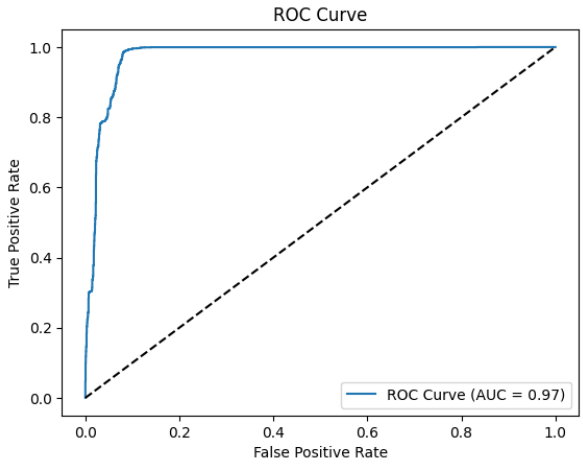
4. Results and Discussion

a. Obtained Result Tables

For the normalized dataset, the model achieved impressive results, as reflected in the metrics:

- **True Positive Rate (Sensitivity):** 0.92
- **False Positive Rate (FPR):** 0.07

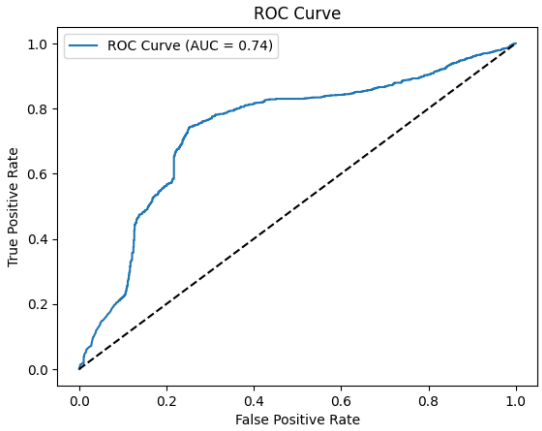
- **True Negative Rate (Specificity):** 0.93
- **False Negative Rate (FNR):** 0.08
- **Precision:** 0.93
- **F1 Score:** 0.93
- **Accuracy:** 0.93
- **AUC:** 0.97



The ROC curve demonstrates excellent performance, with a steep rise toward the top-left corner and an AUC of 0.97, indicating outstanding discriminative ability between sepsis and non-sepsis cases. The curve's shape suggests high sensitivity with minimal false positives at lower thresholds, which is crucial in medical scenarios to avoid missed diagnoses. The significant distance from the diagonal baseline highlights the model's reliability compared to random guessing. Overall, the Fuzzy Inference System (FIS) is well-calibrated, effectively balancing sensitivity and specificity, showcasing its potential for accurate and early sepsis detection.

On the other hand, the non-normalized dataset produced less robust results:

- **True Positive Rate (Sensitivity):** 0.79
- **False Positive Rate (FPR):** 0.34
- **True Negative Rate (Specificity):** 0.66
- **False Negative Rate (FNR):** 0.21
- **Precision:** 0.70
- **F1 Score:** 0.74
- **Accuracy:** 0.73
- **AUC:** 0.74



The ROC curve shows moderate performance, with an AUC of 0.74, indicating the model's ability to differentiate between sepsis and non-sepsis cases is acceptable but not optimal. The curve is less steep and closer to the diagonal baseline compared to a high-performing model, suggesting weaker discriminative power. While the model achieves a balance between sensitivity and specificity, the relatively flat sections indicate limitations in handling false positives and false negatives. This performance highlights the need for further improvement, such as refining membership functions, adjusting fuzzy rules, or normalizing the data, to enhance the model's predictive accuracy.

Additionally, certain rules could not generate output for specific rows (e.g., row 2525 and row 4301), leading to default values being used in non-normalized cases. This highlights challenges in defining universal membership ranges based on raw data distributions.

b. Subjective Interpretation on FP and FN Values and Cases

The performance discrepancy between normalized and non-normalized datasets emphasizes the importance of consistent data preprocessing. The significantly higher False Positive Rate (FPR) and False Negative Rate (FNR) in the non-normalized dataset indicate reduced reliability in distinguishing between sepsis and non-sepsis cases.

In normalized data:

- Low FNR suggests the model accurately identifies patients with sepsis.
- Minimal FPR shows it rarely misclassifies non-sepsis cases.

In non-normalized data:

- Higher FPR might lead to unnecessary medical interventions, while the elevated FNR could result in missed sepsis diagnoses, which can be life-threatening.

c. Suggestions for Further Studies

1. **Normalization Importance:** The stark difference in performance demonstrates the necessity of normalizing data before applying fuzzy inference systems, especially for datasets with varied scales.

2. **Rule Refinement for Non-Normalized Data:** Future work should explore adaptive methods for defining membership functions dynamically based on raw data characteristics. For instance, automatic adjustment of ranges instead of manual assignments can enhance rule coverage.
3. **Hybrid Systems:** Combining fuzzy logic with machine learning could improve decision boundaries and handle complex cases with unclear membership overlaps.
4. **Data Augmentation:** Expanding the dataset with diverse cases may aid in refining membership functions and fuzzy rules, ensuring broader applicability.
5. **Explainability Analysis:** Further investigation into False Positive and False Negative cases can provide deeper insights into model limitations, aiding rule refinement.

By addressing these areas, future studies can achieve more reliable results, even with non-normalized datasets or varying data distributions.

Conclusion

In conclusion, this study used a Fuzzy Inference System (FIS) to detect sepsis early by analyzing patient data from the first 12 hours of monitoring. Normalizing the data greatly improved performance, with the model achieving 93% accuracy, an F1 Score of 0.93, and an AUC of 0.97. These results show that FIS can effectively identify sepsis cases. Key findings include the importance of normalization in enhancing accuracy and reducing errors, the need for flexible membership functions to handle raw data challenges, and the value of using clinically significant features. The results were obtained using a random selection of data, controlled by setting the random state to 42. It's important to note that changing this random state could lead to variations in the achieved performance, highlighting the need for careful consideration of reproducibility.

References:

https://oleg-dubetcky.medium.com/mastering-fuzzy-logic-in-python-c90463bf1135

https://towardsdatascience.com/fuzzy-inference-system-implementation-in-python-8af88d1f0a6e