

## **SPECIFIC AIMS**

The project aims to tackle the critical healthcare challenge posed by sepsis, a condition with high mortality and morbidity rates globally. Current predictive models for key ICU outcomes such as mortality, length of stay, and readmission rates fall short, highlighting the need for improved methodologies. Utilizing the MIMIC-III database, this research intends to enhance sepsis outcome prediction by integrating both structured (e.g., vital signs, lab results) and unstructured data (e.g., clinical notes). This comprehensive approach is anticipated to reveal new insights into sepsis progression and outcomes, filling a significant gap in existing research and potentially influencing clinical practices and patient management strategies effectively.

Innovation within this project stems from addressing the notable research gap in ICU outcome predictions for sepsis patients. By moving beyond traditional NLP methods and adopting a Fusion-based LSTM model, this research is set to provide a more detailed analysis that captures the complexity of sepsis patient data. The methodology not only aims to predict the severity and risks associated with sepsis more accurately but also seeks to revolutionize the way ICU patient care is approached, particularly in understanding and anticipating patient trajectories.

Our project aims on establishing a comprehensive sepsis patient cohort using the MIMIC-III database as a pivotal first step. Various sepsis identification criteria, including the Angus criteria, specific ICD-9 codes for severe sepsis and septic shock, and broader sepsis-related ICD-9 codes validated by Martin, will be evaluated and implemented to ensure accurate patient classification.

### **Specific Aim 1: To develop a comprehensive predictive model for ICU mortality integrating structured clinical data and unstructured clinical notes**

Utilize advanced deep learning models, integrating structured clinical data with unstructured notes to predict ICU mortality at different time intervals. By preprocessing data for normalization and addressing missing values, then applying Fusion-CNN and Fusion-LSTM architectures, the project aims to provide a detailed representation of patients, combining the temporal dynamics of clinical metrics with the depth of narrative information.

### **Specific Aim 2: Leveraging Multi-Modal Data Fusion for Predicting Hospital Readmissions Among Sepsis Survivors**

Extends the analytical methods from Specific Aim 1 to predict hospital readmission rates for sepsis survivors, focusing on identifying key post-discharge risk factors. Through refined feature engineering and modifications to Fusion-CNN and Fusion-LSTM models, the project seeks to identify long-term health indicators and potential readmission signals from the integrated clinical data.

## **RESEARCH STRATEGY**

### **1. Significance**

Sepsis is a critical condition that poses a significant health challenge worldwide, leading to high rates of mortality and morbidity. Despite its severity, the existing methods for predicting patient outcomes like mortality, the length of ICU stay, and rates of hospital readmission are insufficient. This project aims to leverage the MIMIC-III database to enhance the predictive models for sepsis by using both structured and unstructured data. Focusing on the cohort of patients diagnosed with sepsis fills a crucial research gap, especially since current studies do not fully address these key ICU outcomes for our defined cohort.

The MIMIC-III (Medical Information Mart for Intensive Care III) database is an invaluable asset in medical research, containing anonymized data related to over forty thousand patients who have been admitted to critical care units. It encompasses a broad spectrum of information, such as vital signs, medications, lab results, and clinical notes. In this project, we utilize the MIMIC-III database to conduct an in-depth analysis of sepsis, aiming to uncover the intricate factors that contribute to its progression and outcomes. By combining structured data like laboratory test results and medication records with unstructured data such as free-text nursing notes and physician comments, we anticipate uncovering new insights into how sepsis outcomes are influenced. This comprehensive approach is expected to identify crucial predictors for sepsis-related mortality, ICU stay length, and rehospitalization rates. The findings from this study could significantly impact clinical practices and policies by providing a more detailed understanding of sepsis and informing better patient management strategies.

This approach is particularly vital given the heterogeneous nature of sepsis, which necessitates sophisticated, data-driven models to accurately predict patient trajectories and inform treatment decisions. Furthermore, the integration of structured and unstructured data from the MIMIC-III dataset will allow for a more comprehensive analysis than previously possible, addressing a notable gap in current research methodologies as highlighted by recent systematic reviews in 2022 (Yan et al., 2022).

### **2. Innovation**

Recent advancements in sepsis research have primarily focused on utilizing structured and unstructured data to either predict the severity and risk of sepsis based on laboratory outcomes and clinical notes or to identify patients with suspected infections. However, our project breaks new ground by concentrating on predicting ICU-related outcomes, specifically mortality related to ICU length of stay and hospital readmission rates. This focus is particularly crucial as it addresses a significant gap in current research: the dynamic and multifaceted nature of ICU

patient care and the critical need for early and accurate predictions of patient trajectories within the ICU context.

In the realm of sepsis studies, traditional NLP approaches have included methods like Bag of Words (BoW), GloVe, Latent Dirichlet Allocation (LDA), or ClinicalBERT. While these methods have provided valuable insights, they often fall short in capturing the complex interplay between structured clinical data and the rich, nuanced information contained in unstructured textual data. Our project innovates beyond these conventional techniques by implementing a Fusion-based Long Short-Term Memory (LSTM) model, which integrates both structured and unstructured data sources. This approach allows for a more comprehensive and nuanced analysis, capturing the subtleties of patient data and clinical narratives that are often missed by traditional models.

### **3. Research Plan**

Creating a comprehensive patient cohort is a critical initial step in our research into sepsis outcomes using the MIMIC-III database. Our approach to patient cohort creation will involve evaluating and implementing various criteria to identify patients with sepsis accurately. We plan to explore different methods to define our sepsis cohort, including:

1. Utilizing ICD-9 codes based on the Angus criteria (Angus, 2001), which has been widely adopted for sepsis identification in administrative data.
2. Extracting explicit sepsis cases using ICD-9 diagnosis codes specifically for severe sepsis (995.92) and septic shock (785.52), providing a more direct method of identifying patients with these conditions.
3. Applying the sepsis identification criteria validated by Martin (Martin et al., 2003), which include a broader set of ICD-9 codes associated with sepsis, potentially capturing a wider range of sepsis patients within the dataset.

By implementing these different criteria, comparing their effectiveness in accurately identifying sepsis patients, we can determine the most reasonable approach to proceed with our model development stage.

For Specific Aim 1, we will integrate structured clinical data and unstructured clinical notes using advanced deep learning models to predict ICU mortality across various intervals. The approach involves preprocessing data to address missing values and normalization, followed by employing Fusion-CNN and Fusion-LSTM architectures for comprehensive patient representation. This dual analysis aims to capture both the temporal dynamics of clinical measurements and the rich, contextual information within clinical narratives. The mathematical foundation of our predictive model hinges on optimizing a binary cross-entropy loss function,

guiding the learning process towards minimizing prediction errors. This methodological framework ensures a robust, data-driven strategy for enhancing ICU mortality prediction.

For Specific Aim 2, we will leverage the integrated analysis of structured clinical data and unstructured clinical notes to specifically target the prediction of hospital readmission rates for sepsis survivors. Our approach will adapt the methodologies established in Specific Aim 1, focusing on the distinct aspects of readmission risk factors. We will refine our feature engineering process to highlight variables critical to post-discharge outcomes, such as discharge summaries and follow-up care instructions. Adjustments to the Fusion-CNN and Fusion-LSTM architectures will aim to capture long-term health trajectories and indicators of potential readmission. This tailored analytical framework is designed to unearth nuanced predictors of readmission, employing a binary classification approach to model the likelihood of return to hospital within critical post-discharge windows. By concentrating on these differential components, our aim is to develop a predictive tool that offers actionable insights to mitigate readmission rates, thereby enhancing patient care continuity and reducing healthcare system burdens.

### **Specific Aim 1: To develop a comprehensive predictive model for ICU mortality integrating structured clinical data and unstructured clinical notes**

**Hypothesis:** A model that integrates structured clinical data with unstructured clinical notes using deep learning techniques will provide superior accuracy in predicting ICU mortality across short-term, mid-term, and long-term intervals compared to models using only one data type.

**Rationale:** The combination of structured data (such as demographics, vital signs, and lab tests) with unstructured data (clinical notes) provides a more holistic view of patient health status, potentially uncovering subtle patterns not visible when analyzing these data types in isolation.

### **Experimental Approach:**

1. Data Preparation: Utilize ICU data, segmenting into structured (demographics, vital signs, lab results) and unstructured (clinical notes) components. Preprocess data for normalization and handle missing values with imputation strategies.
2. Model Construction: Implement deep learning architectures, specifically Fusion-CNN and Fusion-LSTM models, to learn patient representations from combined data sources. These architectures will include:
  3. Static Information Encoder: One-hot encoding for categorical features.
  4. Temporal Signals Representation: CNNs for temporal data in Fusion-CNN; LSTMs for Fusion-LSTM.
  5. Sequential Notes Representation: Use document embeddings (Doc2Vec) for clinical notes.

6. Patient Representation: Concatenate the outputs of the above components to form a comprehensive patient vector.
7. Output Layer: Apply a sigmoid activation function for binary mortality prediction, optimizing the model using binary cross-entropy loss.

**Mathematical Formulation:** The final prediction output  $\hat{o}$  is given by  $\hat{o} = \sigma(Wz_p + b)$ , where  $W$  and  $b$  are trainable parameters,  $z_p$  is the patient representation, and  $\sigma$  denotes the sigmoid function. The loss function  $L$  for binary classification is defined as  $L = -(y \cdot \log(\hat{o}) + (1 - y) \cdot \log(1 - \hat{o}))$ .

**Interpretation of Results:** Performance will be evaluated using metrics such as AUROC, F1 score, and AUPRC, comparing the integrated model against baselines that use either structured or unstructured data alone.

**Potential Problems and Alternative Approaches:** Challenges may include overfitting due to the complexity of the model and the sparsity of unstructured data. Alternatives include experimenting with different architectures (e.g., more complex RNN structures) or incorporating additional regularization techniques.

## **Specific Aim 2: Leveraging Multi-Modal Data Fusion for Predicting Hospital Readmissions Among Sepsis Survivors**

**Hypothesis:** We hypothesize that a predictive model utilizing a combination of structured clinical data and unstructured clinical notes through advanced machine learning techniques, specifically Fusion-CNN and Fusion-LSTM models, will significantly outperform traditional models in predicting hospital readmission rates for sepsis survivors.

**Rationale:** Hospital readmissions post-sepsis represent a significant burden on patients and healthcare systems, underscoring the necessity for accurate prediction models. The intricate nature of sepsis and its treatment outcomes necessitates a nuanced analysis that can only be achieved by leveraging both the quantitative insights from structured data and the qualitative depth of unstructured clinical notes. This dual approach allows for a more holistic understanding of patient health, potentially unveiling critical predictors for readmissions not observable in structured data alone.

### **Experimental Approach:**

1. Data Segmentation: Similar to Specific Aim 1, data will be segmented into structured components (e.g., demographics, vital signs, lab results) and unstructured components (e.g., clinical notes). However, the selection criteria for features may differ to better capture factors predictive of readmission.

2. Feature Engineering: Special attention will be given to variables associated with post-discharge outcomes, such as discharge summaries, follow-up instructions, and any documented post-discharge care plans.
3. Fusion-CNN and Fusion-LSTM Architectures: While maintaining the same foundational architectures as in Specific Aim 1, modifications will be made to tailor the models to readmission prediction. This could involve adjusting the neural network layers to focus on capturing long-term health indicators and predictors of readmission.
4. Incorporating Post-Discharge Predictors: Additional model components may be introduced to specifically handle post-discharge information, such as follow-up care plans or outpatient medication adjustments, which are critical for understanding readmission risks.
5. Readmission Prediction: The outcome layer will be tailored to predict the likelihood of hospital readmission within a specified timeframe post-discharge, such as 30, 60, or 90 days, using a binary classification approach.

### **Mathematical Formulation:**

The prediction output and loss function will remain consistent with Specific Aim 1, focusing on binary outcomes (readmission or no readmission). However, the feature set and patient representation may be adjusted to better reflect predictors of readmission.

### **Interpretation of Results:**

Comparative Analysis: The performance of the integrated model will be compared against baseline models that solely utilize structured or unstructured data, with a specific focus on metrics like precision, recall, and the area under the ROC curve (AUROC) relevant to readmission prediction.

### **Potential Problems and Alternative Approaches:**

Data Sparsity and Bias: Challenges may arise from the uneven distribution of readmission cases and potential biases in clinical notes. Techniques such as oversampling minority classes or employing advanced natural language processing methods to mitigate bias could be explored.

Model Complexity: The complexity of integrating post-discharge predictors may lead to overfitting. Alternative approaches could include simplifying the model architecture or employing more sophisticated regularization techniques to enhance model generalizability.

## Reference:

- Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., & Pinsky, M. R. (2001). Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7), 1303–1310. <https://doi.org/10.1097/00003246-200107000-00002>
- Hou, N., Li, M., He, L. *et al.* Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 18, 462 (2020). <https://doi.org/10.1186/s12967-020-02620-5>
- Martin, G. S., Mannino, D. M., Eaton, S., & Moss, M. (2003). The epidemiology of sepsis in the United States from 1979 through 2000. *The New England journal of medicine*, 348(16), 1546–1554. <https://doi.org/10.1056/NEJMoa022139>
- Mahbub, M., Srinivasan, S., Danciu, I., Peluso, A., Begoli, E., Tamang, S., & Peterson, G. D. (2022). Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PloS one*, 17(1), e0262182. <https://doi.org/10.1371/journal.pone.0262182>
- Yan, M. Y., Gustad, L. T., & Nytrø, Ø. (2022). Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *Journal of the American Medical Informatics Association : JAMIA*, 29(3), 559–575. <https://doi.org/10.1093/jamia/ocab236>