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. Patients with sepsis often require extensive resources, including prolonged ICU stays and advanced monitoring. This high level of resource utilization highlights the importance of sepsis management not only for patient outcomes but also from a healthcare system perspective. However, the current predictive method in this perspective is insufficient.

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

After doing some systematic literature review, we found that before 2017, on the sepsis cohort, research papers are mainly leveraging only the structured data to build the prediction model. which limit the scope and depth of the predictive insights. And after 2017, research papers gradually tended to integrate structured clinical measurements with unstructured clinical notes. However, the primary focus remained on predicting sepsis severity rather than broader patient and ICU outcomes.

So in our project, we are trying to bridge the research gap by integrating both structured data such as lab measurements with unstructured clinical notes to predict comprehensive ICU

outcomes. 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 finally chose criteria 2 to build our cohort and proceed with further model building. Check out the *query.sql* file in our github repo to see our cohort building code. As for part of our data preprocessing, we either delete the subjects with missing values or implement mean substitution.

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.

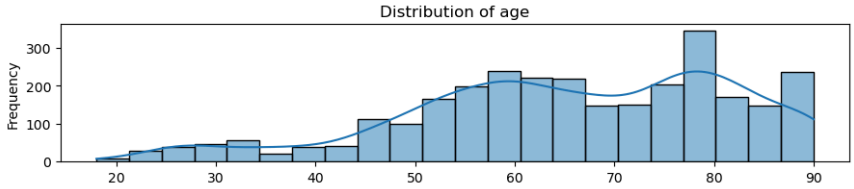
Specific Aim 1: To develop a comprehensive predictive model for ICU mortality integrating structured clinical data and unstructured clinical notes

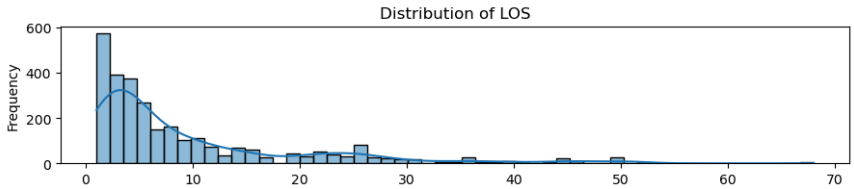
Hypothesis: A model that integrates structured clinical data with unstructured clinical notes using deep learning techniques will provide superior accuracy in predicting ICU mortality across short-term, mid-term, and long-term intervals compared to models using only one data type.

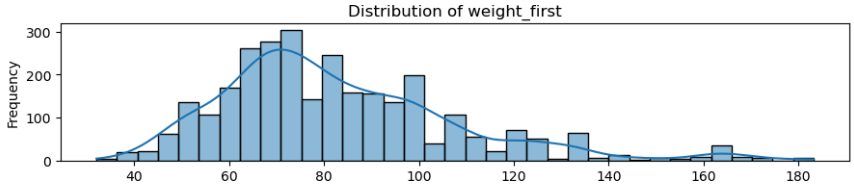
Rationale: The combination of structured data (such as demographics, vital signs, and lab tests) with unstructured data (clinical notes) provides a more holistic view of patient health status, potentially uncovering subtle patterns not visible when analyzing these data types in isolation.

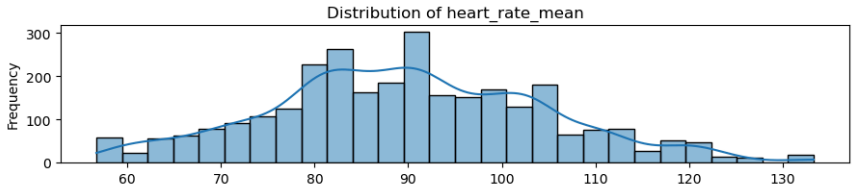
Experimental Approach:

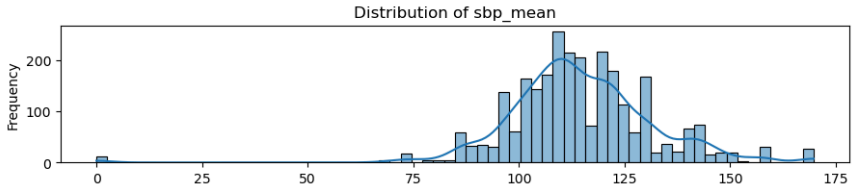
1. Data Preparation: Utilize ICU data, segmenting into structured (demographics, vital signs, lab results) and unstructured (clinical notes) components. Preprocess data for normalization and handle missing values with imputation strategies.

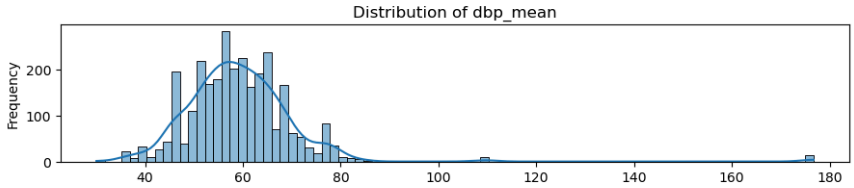

Distribution of age
Frequency
0
100
200
300
20
30
40
50
60
70
80
90

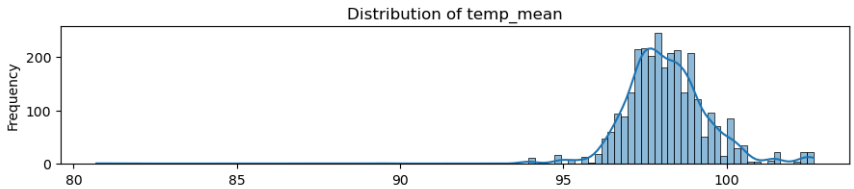

Distribution of LOS
Frequency
0
200
400
600
0
10
20
30
40
50
60
70

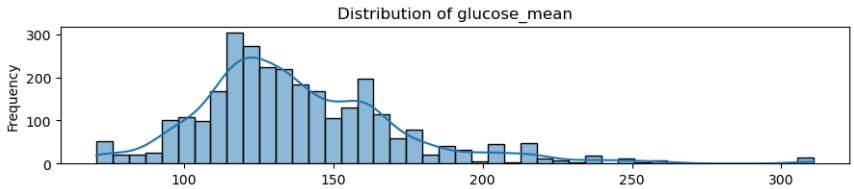

Distribution of weight_first
Frequency
0
100
200
300
40
60
80
100
120
140
160
180

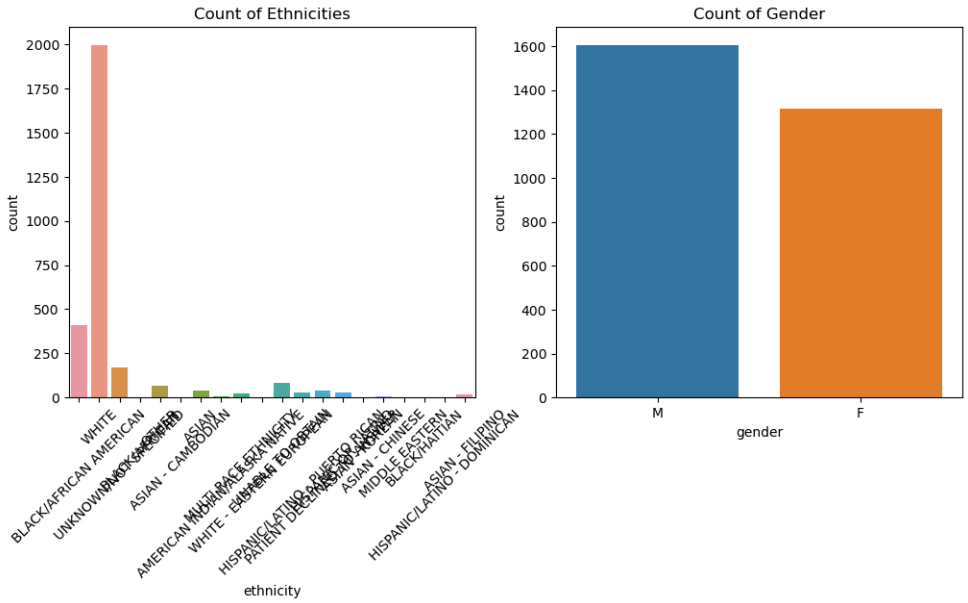

Distribution of heart_rate_mean
Frequency
0
100
200
300
60
70
80
90
100
110
120
130


Distribution of sbp_mean
Frequency
0
100
200
0
25
50
75
100
125
150
175


Distribution of dbp_mean
Frequency
0
100
200
40
60
80
100
120
140
160
180


Distribution of temp_mean
Frequency
0
100
200
80
85
90
95
100


Distribution of glucose_mean
Frequency
0
100
200
300
100
150
200
250
300



Exploratory data analysis (EDA)

Demographic and Clinical Variables: The dataset includes a diverse range of patients, as seen in the distributions of age, gender, and ethnicity. Most numerical clinical variables such as age, length of stay (LOS), and various vital signs and lab results exhibit expected distributions, with some skewness in certain variables like LOS and glucose levels, indicating a long tail of higher values.

Clinical Notes: The unstructured clinical notes are detailed and cover various aspects of patient care, including neurologic status, cardiovascular and respiratory assessments, lab results, and social information. These notes are rich in information that could complement structured data, supporting your hypothesis that integrating these data types could enhance prediction accuracy.

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.

```
array([-0.39245844, -0.77978617,  2.7352989 ,  0.17100231, -0.99174386,
        0.50364697, -2.2402072 ,  0.6450732 , -1.5534003 , -0.8011858 ,
       -0.08514636, -1.2326745 , -0.81656075, -0.54226094,  1.1032084 ,
       -1.2567371 ,  0.81210536, -0.84247357,  0.15486184,  0.05788911,
        0.43326795,  0.11756613,  1.8627218 , -0.6922509 , -0.54519564,
        0.00745075,  0.7868988 , -0.31881166, -0.2093542 , -1.031997  ,
        1.3065289 , -0.21304509,  1.1437916 ,  0.8678715 , -0.04416841,
        1.4950672 ,  0.52877164, -0.36580005,  0.40151954,  0.35578436,
        0.55291814, -0.42636123, -0.1179643 , -0.9797388 , -0.498681  ,
       -0.03713464, -0.76275593, -1.5455822 , -0.6875019 , -0.0467109 ,
        0.6224032 , -0.21929087,  0.8271678 , -1.4941013 , -0.62304646,
        1.1222001 , -0.7268441 ,  0.6109661 ,  0.51977676,  0.39917055,
        0.03379223,  0.61919695,  0.13497128,  0.7260348 , -0.87343305,
        0.7566799 , -0.13433261, -0.28248107, -1.763598  ,  0.95637894,
       -2.1370597 ,  0.1187727 ,  0.59078956,  2.0812895 ,  2.3932736 ,
        0.7271484 , -0.35342652, -0.9680992 ,  0.07990234,  2.0457795 ,
        0.63475585,  1.7488146 ,  1.2750425 ,  0.8798189 , -1.1219908 ,
        0.00715983, -0.65747666, -0.23620532,  0.4361942 , -0.00554504,
        0.6921836 , -0.01690423,  0.26235944, -0.12102018,  0.77868646,
        0.78936785,  0.99464935, -0.48817608,  0.63521487, -0.5988928 ],
      dtype=float32))
```

Sample embeddings for the first document

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 The final prediction output o is given by $o=\sigma(Wzp+b)$, where W and b are trainable parameters, zp is the patient representation, and σ denotes the sigmoid function. The loss function L for binary classification is defined as $L=-(y \cdot \log(o)+(1-y) \cdot \log(1-o))$.

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.

Translation to Scientific Knowledge and Patient Care Improvement:

This research could significantly advance our understanding of the multifaceted nature of ICU mortality risk factors. By integrating diverse data types and applying advanced machine learning techniques, the study may reveal new insights into the complex interactions between patient characteristics, treatment responses, and outcomes. Such knowledge could guide future research directions and inform the development of more effective interventions. If successful, the model could become a vital tool in the ICU, assisting clinicians in identifying patients at high risk of mortality more accurately and promptly. This early identification could enable targeted interventions, more personalized care plans, and optimized resource allocation, potentially saving lives.

Benefit to Patients and Implementation in Practice:

Patients could benefit from more personalized and timely interventions, reduced risk of adverse outcomes, and improved communication with healthcare providers about their prognosis and treatment options. Enhanced predictive capabilities might also lead to shorter ICU stays and lower healthcare costs. Implementing the model in clinical settings would involve integrating it with existing hospital information systems to ensure seamless access to real-time patient data. Clinicians would receive predictions and explanations through a user-friendly interface, aiding decision-making without adding to their workload. Ongoing training and support, along with continuous monitoring of the model's performance and impact on patient outcomes, would be crucial for successful adoption and sustained use.

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.

   Preliminary Work Under Specific Aim 1 as a Foundation for Specific Aim 2:

   In the course of developing a comprehensive predictive model for ICU mortality by integrating structured clinical data with unstructured clinical notes (Specific Aim 1), we have laid the groundwork for advancements targeted in Specific Aim 2. Our initial models included:

   Baseline Model: A preliminary approach using Doc2Vec to process unstructured clinical notes, combined with structured clinical data, establishing an initial performance benchmark.

   Enhanced LSTM Model: An advancement over the baseline, replacing Doc2Vec embeddings with LSTM layers for a dynamic analysis of clinical narratives, significantly improving our ability to capture the temporal and contextual richness of the data.

   Performance Insights:

   The baseline model demonstrated promising results, achieving a validation accuracy of up to 91.40%. This served as a proof of concept for our integrated data approach.

Transitioning to the LSTM model yielded a remarkable increase in performance, with validation accuracy reaching up to 97.83%. This underscores the effectiveness of LSTM in dealing with complex clinical narratives, setting a new standard for our predictive modeling efforts.

```
Epoch 1/10
56/56 [==============================] - 2s 11ms/step - loss: 0.5692 - accuracy: 0.7224 - val_loss: 0.5225 - val_accuracy: 0.7376
Epoch 2/10
56/56 [==============================] - 0s 4ms/step - loss: 0.4590 - accuracy: 0.7955 - val_loss: 0.4522 - val_accuracy: 0.7783
Epoch 3/10
56/56 [==============================] - 0s 4ms/step - loss: 0.3831 - accuracy: 0.8346 - val_loss: 0.3920 - val_accuracy: 0.8462
Epoch 4/10
56/56 [==============================] - 0s 4ms/step - loss: 0.3189 - accuracy: 0.8754 - val_loss: 0.4274 - val_accuracy: 0.8077
Epoch 5/10
56/56 [==============================] - 0s 4ms/step - loss: 0.2811 - accuracy: 0.8867 - val_loss: 0.2967 - val_accuracy: 0.8756
Epoch 6/10
56/56 [==============================] - 0s 3ms/step - loss: 0.2074 - accuracy: 0.9263 - val_loss: 0.2644 - val_accuracy: 0.9050
Epoch 7/10
56/56 [==============================] - 0s 3ms/step - loss: 0.1702 - accuracy: 0.9450 - val_loss: 0.2589 - val_accuracy: 0.8869
Epoch 8/10
56/56 [==============================] - 0s 3ms/step - loss: 0.1352 - accuracy: 0.9609 - val_loss: 0.2129 - val_accuracy: 0.9050
Epoch 9/10
56/56 [==============================] - 0s 3ms/step - loss: 0.1092 - accuracy: 0.9632 - val_loss: 0.2136 - val_accuracy: 0.9208
Epoch 10/10
56/56 [==============================] - 0s 3ms/step - loss: 0.0800 - accuracy: 0.9824 - val_loss: 0.1980 - val_accuracy: 0.9140
```

```
Epoch 1/10
69/69 [==============================] - 34s 448ms/step - loss: 0.5942 - accuracy: 0.6910 - val_loss: 0.4431 - val_accuracy: 0.7880
Epoch 2/10
69/69 [==============================] - 32s 459ms/step - loss: 0.3966 - accuracy: 0.8251 - val_loss: 0.2969 - val_accuracy: 0.8841
Epoch 3/10
69/69 [==============================] - 31s 449ms/step - loss: 0.2544 - accuracy: 0.8994 - val_loss: 0.2181 - val_accuracy: 0.8949
Epoch 4/10
69/69 [==============================] - 29s 424ms/step - loss: 0.1788 - accuracy: 0.9275 - val_loss: 0.1615 - val_accuracy: 0.9312
Epoch 5/10
69/69 [==============================] - 29s 415ms/step - loss: 0.1441 - accuracy: 0.9443 - val_loss: 0.1360 - val_accuracy: 0.9475
Epoch 6/10
69/69 [==============================] - 29s 416ms/step - loss: 0.1156 - accuracy: 0.9570 - val_loss: 0.1135 - val_accuracy: 0.9511
Epoch 7/10
69/69 [==============================] - 38s 548ms/step - loss: 0.0832 - accuracy: 0.9696 - val_loss: 0.1089 - val_accuracy: 0.9583
Epoch 8/10
69/69 [==============================] - 28s 405ms/step - loss: 0.0629 - accuracy: 0.9751 - val_loss: 0.0800 - val_accuracy: 0.9656
Epoch 9/10
69/69 [==============================] - 31s 445ms/step - loss: 0.0535 - accuracy: 0.9805 - val_loss: 0.0535 - val_accuracy: 0.9837
Epoch 10/10
69/69 [==============================] - 28s 408ms/step - loss: 0.0457 - accuracy: 0.9841 - val_loss: 0.0608 - val_accuracy: 0.9783
```

Adapting Our Approach for Specific Aim 2:

Building on these insights, we plan to adapt and refine our model architectures to tackle the challenges posed by Specific Aim 2—predicting hospital readmission rates among sepsis survivors. Our proposed modifications will include:

Incorporating CNN Layers for Structured Data: To enhance the processing of structured clinical data, aligning with our aim to utilize Fusion-CNN and Fusion-LSTM models for a more nuanced analysis.

Focusing on Long-term Health Indicators and Readmission Predictors: By refining our models to identify key post-discharge risk factors, we aim to unearth predictors that are

critical for readmission, leveraging the comprehensive patient data representation proven effective under Specific Aim 1.

The methodologies validated and the insights gained from Specific Aim 1 provide a robust foundation for our continued efforts under Specific Aim 2. Our aim is to extend the analytical capabilities developed thus far to more accurately predict and thereby mitigate hospital readmissions among sepsis survivors, enhancing patient care outcomes and optimizing healthcare resource allocation.

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