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 files under the query folder 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

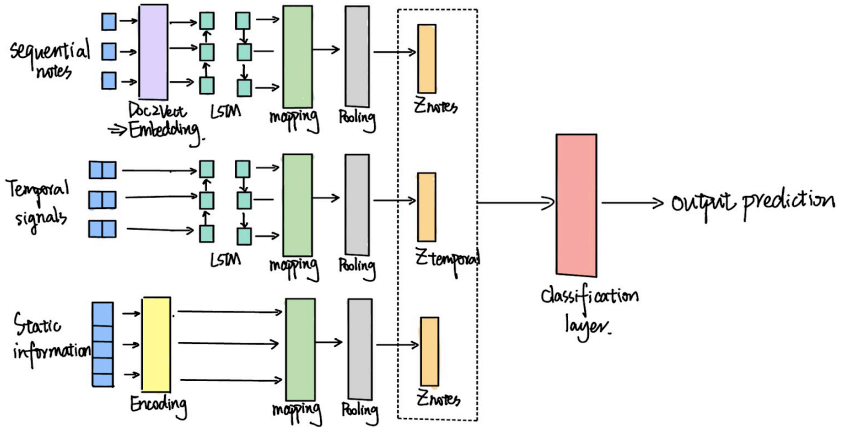
Hypothesis: A model that integrates structured clinical data with unstructured clinical notes using deep learning techniques will provide superior accuracy in predicting ICU mortality across short-term, mid-term, and long-term intervals compared to models using only one data type.

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.

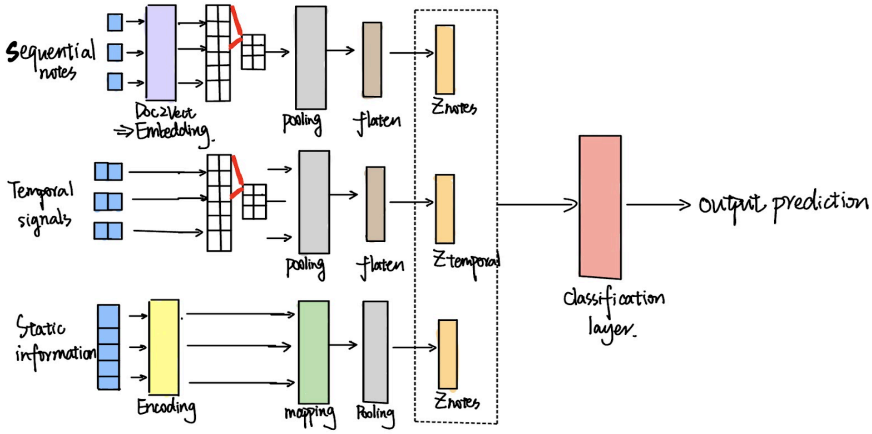
Experimental Approach:

1. Baseline Model: We tested baseline models that rely solely on unstructured clinical notes. Two classic machine learning approaches were utilized: Logistic Regression (LR) and Random Forest Classifier (RF)
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:

a. Sequential Notes Representation: By applying Doc2Vec embedding to process the text, we capture the information of the clinical narratives, which is then analyzed by LSTM networks to understand the progression over time. A pooling layer then distills this information into a fixed-size vector Z notes.
b. Temporal Signals Representation: Another LSTM network processes structured time-series data, allowing the model to learn from the temporal dynamics.
c. Static Information Encoder: Demographic data and other non-time-variant factors are encoded(one-hot) into a vector, Zstatic , preserving the crucial baseline patient information. LSTM layers are not used because these data do not exhibit temporal patterns
d. Patient Representation: Concatenate the outputs of the above components to form a comprehensive patient vector (Zhang et al., 2020).
e. Output Layer: Apply a sigmoid activation function for binary mortality prediction, optimizing the model using binary cross-entropy loss.



Fusion lstm



Fusion cnn

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 combining structured clinical data and unstructured notes, using advanced Fusion-CNN and Fusion-LSTM techniques, will substantially surpass traditional models in predicting hospital readmissions for sepsis survivors.

Rationale: Hospital readmissions after sepsis significantly strain healthcare systems, highlighting the need for precise prediction models. The complex nature of sepsis requires a detailed analysis that combines quantitative structured data with qualitative clinical notes. This approach enhances the understanding of patient health and uncovers vital readmission predictors not visible in structured data alone.

Experimental Approach:

1. Baseline Model: Similar to our mortality prediction, we started with Logistic Regression and Random Forest to establish baseline performance using static clinical data.
2. Fusion Architectures: Employed Fusion-CNN and Fusion-LSTM, integrating structured and unstructured data to better capture readmission indicators, similar to our approach in mortality prediction.
3. Data Integration: Leveraged the same dynamic analysis with LSTM networks for temporal data and static information encoding for patient demographics, tailored to readmission characteristics.
4. Patient Representation: Combined dynamic and static data outputs, enhancing model's capacity to assess readmission risks.
5. Output Layer: Implemented a sigmoid activation, optimizing with binary cross-entropy loss specifically for predicting readmission probabilities.

Potential Problems and Alternative Approaches: The uneven distribution of readmission cases and potential biases in clinical notes may affect model accuracy. Techniques like oversampling and advanced natural language processing can help mitigate these issues. Model Complexity: The integration of complex post-discharge predictors could risk overfitting; simplifying the model or using sophisticated regularization techniques may improve generalizability.

RESULT

Cohort Selection

We applied three sepsis inclusion-exclusion criteria to the MIMIC III database, followed by a comprehensive descriptive statistical analysis. The results are summarized in Table 1. This analysis included counting patient records by hospital admission ID (HADM_ID), quantifying the number of clinical notes authored by physicians, nurses, and radiologists, and evaluating the missing rate for seven commonly assessed vital signs.

Based on these analyses, we selected the explicit method for feature extraction for three primary reasons:

1. *Computational Efficiency*: Given the extensive size of the MIMIC III database and the complexity involved in joining multiple tables for feature extraction, computational efficiency was paramount. The explicit method provided a practical balance between model performance and computational resource utilization.
2. *Data Completeness*: This method exhibited a significantly lower missing rate for vital signs compared to other evaluated methods, enhancing the reliability of the subsequent analyses.
3. *Specificity and Positive Predictive Value (PPV)*: Most critically, the explicit method achieved a specificity of 100% and a PPV of 100%, with a sensitivity of 9.3%. In alignment with our research objectives, which prioritize the accurate identification of true sepsis cases (i.e., minimizing false positives), this method was deemed most appropriate. The choice to prioritize specificity and PPV justifies the acceptance of lower sensitivity in our context.

	Number of Records	Number of First Day Notes	Missingness(vital signs)
Angus	33583	137520	1.82%
Martin	30477	125778	1.23%
Explicit	3466	18552	0.89%

Table 1. Descriptive statistics on 3 sepsis cohorts

Feature Selection & Dataset Preprocessing

To construct the dataset utilized for this study, we integrated three distinct types of data: static information, temporal information, and clinical notes. The static information consisted of demographic and clinical characteristics, including age, length of stay (LOS), gender, marital status, ethnicity, and admission type.

Temporal data comprised observations of 7 common vital signs and 19 laboratory tests recorded hourly for the first 24 hours of each patient's stay. In instances of missing data within this timeframe, we employed zero imputation to maintain data integrity. Additionally, where multiple test records existed within a single hour, we calculated their average to ensure uniformity of the dataset. Each column was then normalized to facilitate comparative analysis and integration into predictive models.

Clinical notes were also collected on an hourly basis to ensure consistency with the structured temporal data. The descriptive statistics tables of each temporal variable were shown below based on mortality status (0:death, 1:alive), and readmission status (0: not readmit, 1: readmitted). This methodological approach allowed for the creation of a comprehensive dataset with sequential records, which is crucial for the development and validation of our predictive models.

	mean	std	min	25%	50%	75%	max
HeartRate	92.1	18.6	55.0	77.0	91.0	106.0	161.0
SysBP	110.6	20.8	44.0	99.0	108.0	118.0	222.0
DiasBP	60.6	13.0	26.0	53.0	59.0	66.5	122.0
MeanBP	74.1	14.5	33.0	65.0	72.0	80.5	162.0
RespRate	21.6	5.9	7.0	17.5	22.0	25.0	50.0
TempC	36.8	1.2	32.6	36.1	36.9	37.7	40.4
SpO2	97.1	3.5	50.0	95.6	98.0	100.0	100.0
ANIONGAP	15.4	3.9	5.0	13.0	15.0	17.0	28.0
ALBUMIN	2.7	0.5	1.6	2.3	2.6	3.0	3.8
BANDS	13.7	10.8	1.0	6.0	9.0	19.8	40.0
BICARBONATE	20.4	5.5	7.0	17.0	19.5	24.0	40.0
BILIRUBIN	2.2	3.0	0.1	0.4	0.8	2.7	17.8
CREATININE	2.0	1.6	0.2	1.0	1.6	2.6	11.4
CHLORIDE	107.8	7.4	90.0	103.0	108.0	112.0	140.0
GLUCOSE	144.8	66.1	47.0	103.8	123.0	165.0	511.0
HEMATOCRIT	30.7	5.0	15.2	27.2	30.2	34.0	47.3
HEMOGLOBIN	10.4	1.7	5.9	9.2	10.3	11.3	15.7
LACTATE	2.7	1.7	0.6	1.5	2.1	3.5	10.9
PLATELET	198.3	137.7	8.0	99.0	182.0	262.0	927.0
POTASSIUM	4.2	0.8	2.1	3.7	4.2	4.7	7.4
PTT	45.2	26.2	20.1	31.2	37.4	47.2	150.0
INR	2.0	1.5	0.7	1.3	1.6	2.0	10.4
PT	19.6	11.1	8.8	14.4	16.6	20.0	79.3
SODIUM	139.2	5.3	122.0	136.0	139.0	142.0	165.0
BUN	39.1	24.2	3.0	22.0	35.0	54.0	170.0
WBC	16.9	11.4	0.1	9.0	14.5	22.5	69.5

mortality 0

	mean	std	min	25%	50%	75%	max
HeartRate	93.1	18.6	56.0	78.0	91.0	107.0	150.0
SysBP	109.8	16.5	64.0	99.0	108.0	118.0	222.0
DiasBP	59.9	10.9	32.0	53.0	58.5	66.0	122.0
MeanBP	73.4	11.9	34.0	65.0	72.0	80.0	162.0
RespRate	21.3	5.7	7.0	17.0	22.0	24.6	50.0
TempC	36.8	1.3	32.6	36.1	36.9	37.7	40.4
SpO2	97.2	3.6	50.0	96.0	98.0	100.0	100.0
ANIONGAP	15.5	4.0	5.0	13.0	15.0	17.0	28.0
ALBUMIN	2.7	0.5	1.6	2.3	2.7	3.1	3.7
BANDS	14.3	11.3	1.0	6.5	10.0	22.2	40.0
BICARBONATE	19.6	4.9	7.0	16.0	19.0	23.0	35.0
BILIRUBIN	1.6	1.3	0.1	0.4	1.3	2.7	4.4
CREATININE	2.1	1.7	0.2	1.0	1.6	2.6	11.4
CHLORIDE	108.3	7.7	92.0	103.0	108.0	113.0	140.0
GLUCOSE	147.4	72.7	47.0	101.0	122.0	169.0	511.0
HEMATOCRIT	30.7	4.4	18.1	27.8	30.1	33.4	47.3
HEMOGLOBIN	10.3	1.6	5.9	9.2	10.2	11.2	15.3
LACTATE	2.8	1.8	0.6	1.6	2.2	3.4	10.9
PLATELET	193.0	126.1	17.0	96.8	183.0	250.8	806.0
POTASSIUM	4.2	0.8	2.6	3.7	4.2	4.7	7.4
PTT	43.6	24.6	20.1	30.2	37.0	44.8	150.0
INR	2.1	1.7	1.0	1.3	1.5	2.1	10.4
PT	20.7	13.0	12.1	14.4	15.9	20.3	79.3
SODIUM	139.0	5.3	128.0	136.0	138.0	141.0	165.0
BUN	35.9	19.7	3.0	20.8	34.0	47.2	98.0
WBC	16.7	10.8	0.1	9.1	15.4	21.8	69.5

mortality 1

Table 2A: Descriptive statistics of temporal variables by mortality status

readmit 0	mean	std	min	25%	50%	75%	max
HeartRate	88.2	18.1	55.0	76.0	89.0	101.0	161.0
SysBP	113.5	32.0	44.0	96.0	108.0	121.0	206.0
DiasBP	63.4	18.3	26.0	53.0	60.0	69.5	114.0
MeanBP	77.0	21.1	33.0	63.0	72.0	84.5	148.0
RespRate	22.6	6.4	10.0	18.0	21.8	27.0	39.0
TempC	36.8	1.0	34.2	36.1	36.8	37.4	39.6
SpO2	96.7	2.8	89.0	95.0	97.0	99.0	100.0
ANIONGAP	15.1	3.7	6.0	13.0	15.0	18.0	25.0
ALBUMIN	2.6	0.5	1.9	2.3	2.5	2.9	3.8
BANDS	12.9	10.4	1.0	5.8	9.0	19.0	39.0
BICARBONATE	22.1	6.3	10.0	18.0	21.0	25.0	40.0
BILIRUBIN	3.3	4.7	0.2	0.4	0.7	7.1	17.8
CREATININE	1.9	1.5	0.2	0.8	1.5	2.5	8.4
CHLORIDE	106.7	6.7	90.0	102.5	108.0	111.0	125.0
GLUCOSE	139.4	49.8	50.0	108.0	125.0	155.0	303.0
HEMATOCRIT	30.7	6.1	15.2	25.7	30.5	34.8	47.0
HEMOGLOBIN	10.6	1.9	7.2	9.2	10.6	11.8	15.7
LACTATE	2.4	1.4	0.7	1.2	1.8	3.6	5.5
PLATELET	209.7	159.7	8.0	108.0	174.0	277.0	927.0
POTASSIUM	4.2	0.9	2.1	3.7	4.2	4.8	6.5
PTT	48.5	29.1	23.9	33.6	37.5	49.2	150.0
INR	1.8	0.6	0.7	1.3	1.8	2.0	4.4
PT	17.3	3.8	8.8	14.2	17.5	19.4	36.9
SODIUM	139.6	5.2	122.0	136.0	140.0	143.0	155.0
BUN	45.4	30.5	3.0	23.5	40.0	65.0	170.0
WBC	17.2	12.7	0.1	6.5	13.2	27.0	46.1

readmit 1	mean	std	min	25%	50%	75%	max
HeartRate	92.6	18.8	55.0	78.0	91.0	107.0	161.0
SysBP	110.5	21.1	44.0	98.0	108.0	118.0	222.0
DiasBP	60.4	13.1	26.0	52.5	59.0	66.0	122.0
MeanBP	74.0	14.6	33.0	65.0	72.0	80.0	162.0
RespRate	21.6	6.0	7.0	17.0	21.0	25.0	50.0
TempC	36.8	1.2	32.6	36.1	36.9	37.6	40.4
SpO2	97.0	3.5	50.0	95.0	98.0	100.0	100.0
ANIONGAP	15.4	3.9	5.0	13.0	15.0	17.8	28.0
ALBUMIN	2.6	0.5	1.6	2.3	2.5	3.0	3.8
BANDS	13.7	10.8	1.0	6.0	9.0	19.8	40.0
BICARBONATE	20.4	5.6	7.0	17.0	19.0	24.0	40.0
BILIRUBIN	2.1	3.1	0.1	0.4	0.7	2.7	17.8
CREATININE	2.0	1.7	0.2	1.0	1.6	2.6	11.4
CHLORIDE	107.8	7.5	90.0	102.8	108.0	112.0	140.0
GLUCOSE	145.6	66.9	47.0	102.0	124.0	167.5	511.0
HEMATOCRIT	30.9	5.0	15.2	27.3	30.2	34.3	47.3
HEMOGLOBIN	10.4	1.7	6.8	9.2	10.3	11.4	15.7
LACTATE	2.7	1.7	0.6	1.5	2.1	3.5	10.9
PLATELET	199.7	140.0	8.0	98.5	181.0	263.0	927.0
POTASSIUM	4.2	0.8	2.1	3.7	4.2	4.7	7.4
PTT	45.4	26.6	20.1	31.2	37.4	46.8	150.0
INR	2.0	1.5	0.7	1.3	1.6	2.1	10.4
PT	19.7	11.3	8.8	14.4	16.6	20.0	79.3
SODIUM	139.3	5.4	122.0	136.0	139.0	142.0	165.0
BUN	39.1	24.4	3.0	22.0	35.0	52.0	170.0
WBC	17.1	11.5	0.1	9.0	14.5	23.0	69.5

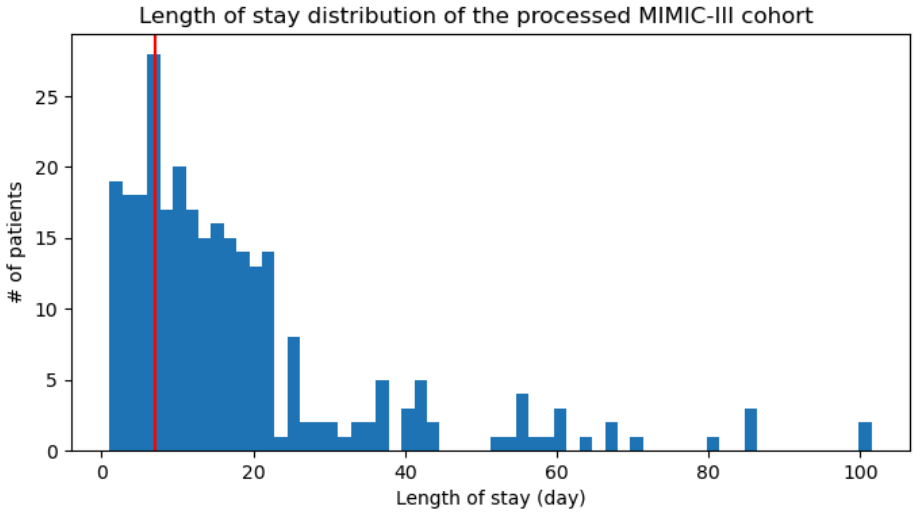
Table 2B: Descriptive statistics of temporal variables by re-admission status

To construct the dataset for this study, we integrated static and temporal data along with clinical notes, each processed to enhance predictive modeling. After organizing data into dedicated directories for streamlined access, we standardized time-related data into uniform intervals and employed zero imputation for missing values. Clinical texts from the first 24 hours of each patient's stay were transformed using Doc2Vec to generate semantic vector representations, crucial for capturing the nuanced dynamics of patient trajectories. This meticulous preprocessing, including data validation and normalization, has allowed for the development of a comprehensive dataset, detailed in splits.json, facilitating robust predictive analysis of mortality and readmission.

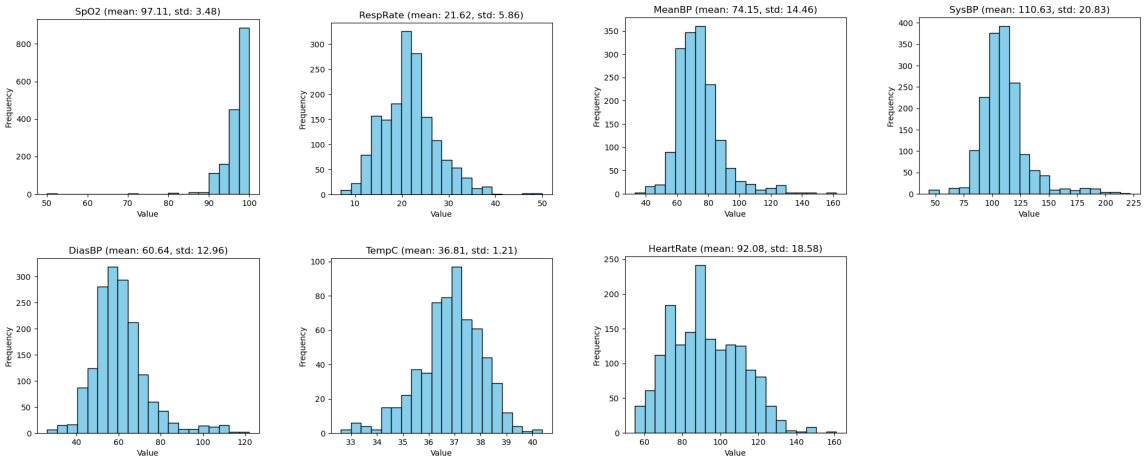
Visualization - EDA

Exploratory data analysis was conducted on the final dataset to assess the distribution of key variables and identify any potential outliers. As illustrated in Figure 1, the distribution of the total length of stay for unique patients was examined. This analysis revealed the presence of outliers, which were subsequently removed to enhance the robustness of the subsequent model building process.

Figure 2 presents the distribution of seven vital signs included in the study dataset. The analysis of these variables indicated that most exhibited a normal distribution, confirming the appropriateness of subsequent analytical assumptions. The only exception was the oxygen saturation (SpO2) levels. This left-skewness is expected and justifiable, given that SpO2 values are capped at 100%, thus constraining the upper range of possible values. Other visualization plots (such as the distribution plots for 19 lab tests) could be found in the github repository under the image folder.



Picture 1. Distribution of LOS



Picture 2. Distribution of 7 vital signs

Predictive Modeling for Readmissions

Interpretation of Results:

When developing a mortality prediction model for patients with sepsis. For Baseline Models, using Logistic Regression and Random Forest Classifier as baseline models, we observed that

their predictive power was relatively limited. The Logistic Regression model achieved an AUC of 0.639, while the Random Forest Classifier produced an AUC of 0.461. These scores indicate that relying solely on one data type and simple classification models is insufficient for accurate mortality prediction in this setting.

For fusion Models, our data integration strategy, as implemented in the Fusion LSTM and Fusion CNN models, demonstrated substantial improvement over the baseline models. The Fusion LSTM model, designed to handle the sequential nature of clinical notes and time-series data, achieved an ROC-AUC of 0.77. Similarly, the Fusion CNN model, which also integrates multiple data types, achieved an ROC-AUC of 0.75. Both models showed significant improvement over traditional single-data-type models.

The notable performance improvements of the Fusion models validate our hypothesis that a multi-faceted approach is essential for accurate outcome prediction in the complex environment of the ICU. Integrating different data types provides a richer, more comprehensive understanding of patient health. Additionally, the superior performance of the Fusion LSTM model (ROC-AUC of 0.77) highlights the importance of capturing how a patient's condition evolves over time. LSTM networks excel in this aspect due to their ability to process sequential data effectively.

Model	**AUC**	**F1 Score**	**AUPR**	**Training Time (sec)**
Logistic Regression (LR)	0.639	0.286	0.341	0.01
Random Forest (RF)	0.461	0.167	0.156	0.13
Long Short-Term Memory (LSTM)	0.77	-	-	-
Convolutional Neural Network (CNN)	0.75	-	-	-

Table 3: Predictive Model Performance for Mortality Rate

These results emphasize the value of advanced machine learning techniques and multi-modal data integration in clinical settings. Moving forward, we aim to continue refining our models to enhance their predictive power. Our goal is to provide clinicians with robust tools that facilitate informed decision-making, ultimately improving patient care and outcomes.

Building on our findings from mortality prediction, we evaluated the performance of similar predictive models for hospital readmissions. Initially, Logistic Regression provided a baseline AUC of 0.50 with an F1 Score of 0.36, achieved rapidly in just 0.02 seconds. While this underscores the challenges in predicting readmissions using simple models, the addition of the Random Forest model modestly improved the AUC to 0.51. However, this model also highlighted the trade-offs in precision and recall, necessitating more sophisticated approaches for better accuracy.

Further advancing our methodology, we implemented LSTM and CNN models, which significantly enhanced our predictive capabilities. The LSTM model elevated the AUC to 0.64, effectively capturing the temporal dynamics crucial for understanding patient readmissions. The CNN model further improved the AUC to 0.69, excelling in integrating multidimensional clinical data, which is vital for comprehending the complex patterns associated with readmissions.

The enhanced performance of these advanced models corroborates the effectiveness of our multi-modal data integration strategy, previously validated in our mortality analysis. Specifically, the superior results of the LSTM and CNN highlight their potential to refine patient management strategies and improve readmission outcomes. This aligns with our broader goal of harnessing sophisticated machine learning techniques to support clinical decision-making and improve healthcare delivery.

Model	**AUC**	**F1 Score**	**AUPR**	**Training Time (sec)**
Logistic Regression (LR)	0.496	0.357	0.476	0.02
Random Forest (RF)	0.508	0.286	0.342	0.18
Long Short-Term Memory (LSTM)	0.64	-	-	-
Convolutional Neural Network (CNN)	0.69	-	-	-

Table 4: Predictive Model Performance for Hospital Readmission

Moving forward, we will continue refining these models to further enhance their accuracy and applicability in clinical settings, with an ongoing focus on optimizing interventions and treatment plans based on robust predictive insights.

CONCLUSION

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.

Our use of advanced LSTM and CNN models has enhanced understanding of factors influencing hospital readmissions. By integrating diverse data types, these models provide deeper insights into patient trajectories, aiding in the development of targeted interventions to reduce readmission rates. This knowledge will inform future research and the creation of more effective discharge and follow-up care strategies.

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.

Implementing these predictive models can lead to personalized patient care post-discharge, reducing readmission risks and improving health outcomes. Integrating these tools into hospital information systems will support clinical decision-making and optimize resource use, potentially leading to shorter hospital stays and lower healthcare costs. Continuous model refinement and provider training will be crucial for effective integration into clinical workflows.

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

Zhang, D., Yin, C., Zeng, J. et al.(2020). Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 20, 280. https://doi.org/10.1186/s12911-020-01297-6