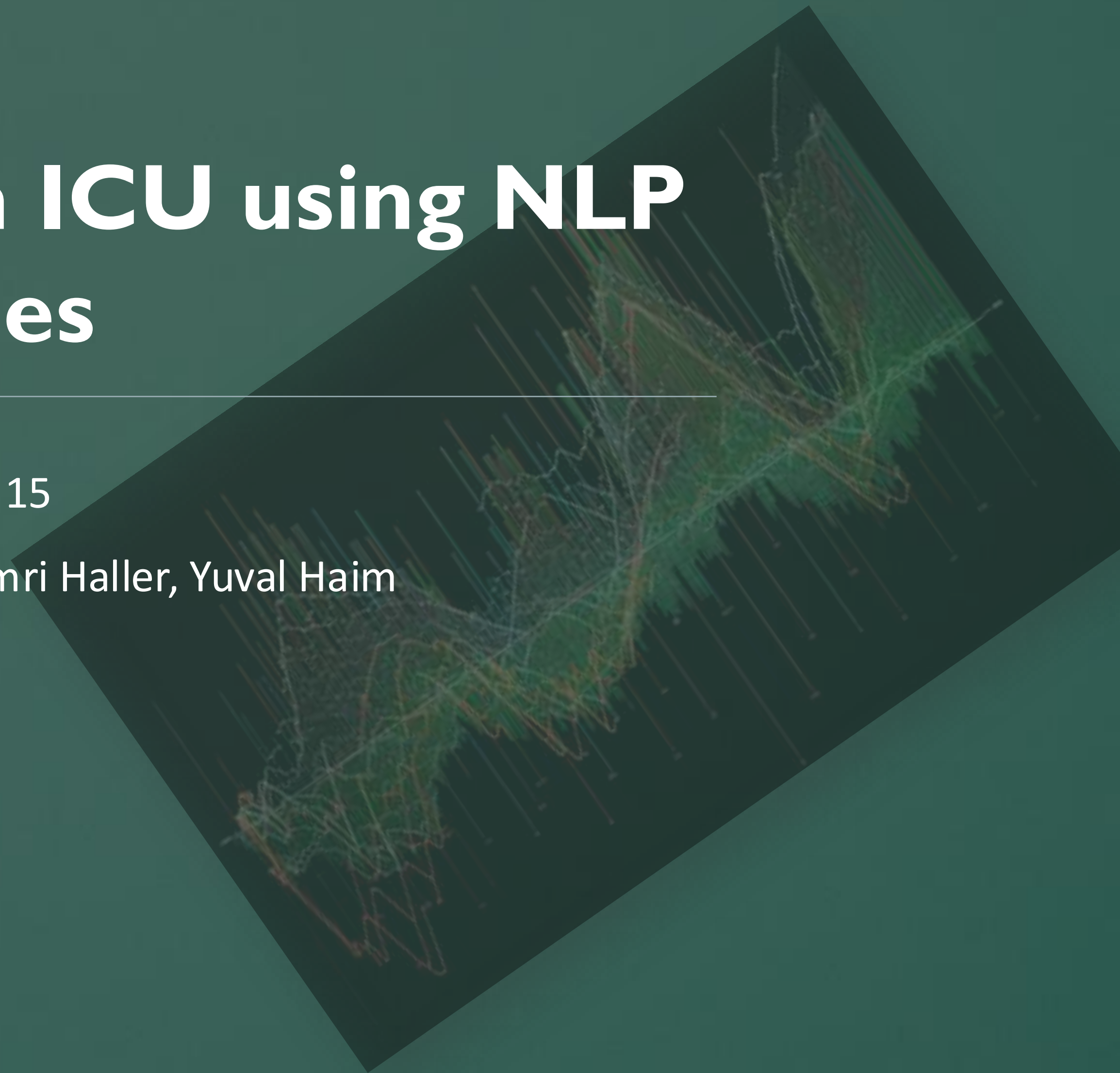
Mortality prediction in ICU using NLP techniques

Project Group 15

Students: Eldar Zosmanovich, Omri Haller, Yuval Haim



BUSINESS PROBLEM

Huge operational burden

5.7 million ICU admissions only in the US **every year.**

Mortality is not trivial

Mean annual of ICU Mortality rate = 10%

Cost pressure

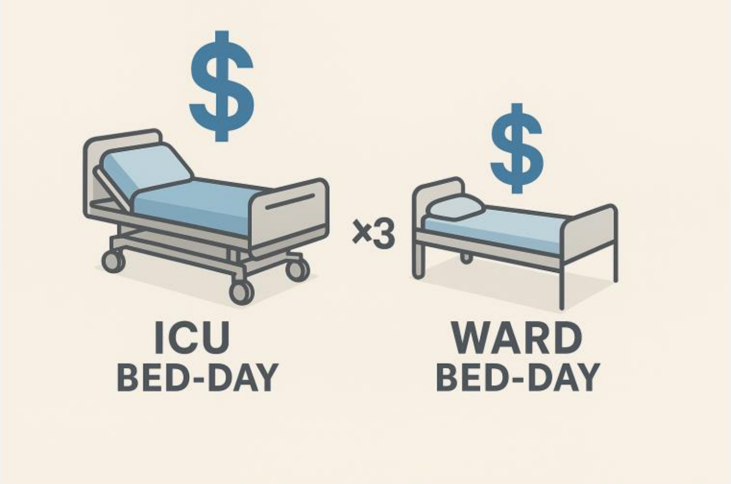
ICUs consume 13% of total hospital spend

Business Goal – Utilize ML Techniques to improve survival prediction and improve allocation of scarce expensive beds while maintaining outcomes

Understand the Problem: Why ICU-Survival Prediction Is important?

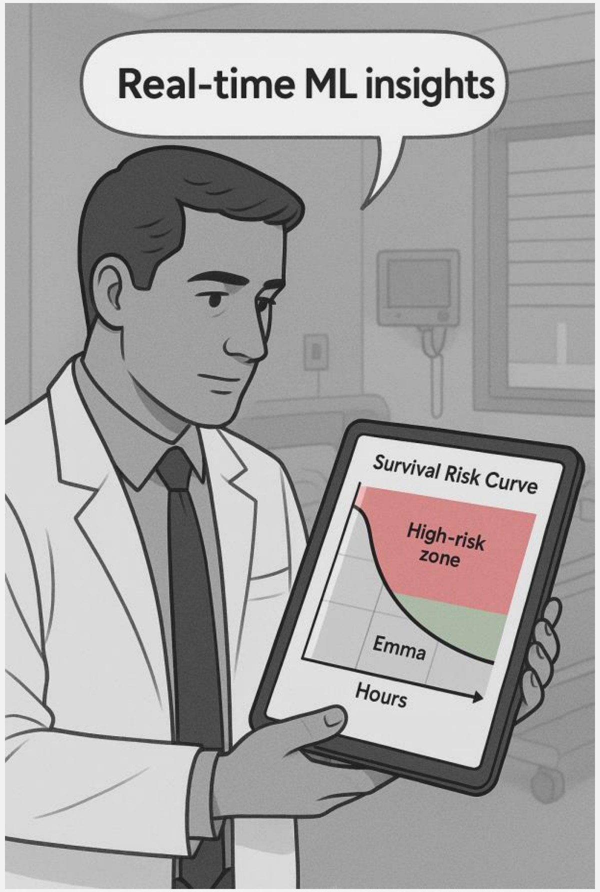
- **High-stakes decisions under pressure**
- **ICU environments are data-rich but overwhelming**
- **Clinician overload leads to missed opportunities**
- **Every decision has economic and human consequences**:

  ICU transfer costs can range from **$2,500 to $25,000 per move.**
- **ICU day costs 3x more than ward (7-15K/day)**
- **Time-critical insights are needed**



Understand the Problem: Example Scenario in ICU

The Scene :"The patient Emma, 72, septic shock. Only one ICU bed left…"



Time-Based Survival probability: could help for the decision process

Traditional scoring systems

Classic score	Snapshot time	Inputs	Key limits
SAPS II	first 24 h	17 vitals/labs	Static; hand-picked features
OASIS	first 24 h	10 features	Static; requires curation
APACHE II	first 24 h	12 variables	Manual abstraction

All rely on a ***manual variable-selection stage*** *that is slow, brittle across hospitals, and ignores the 99 % of EHR data points .*

Our Project Goal

Full Electronic Health Records (EHR) Utilization

The model will use **every chart, lab, and output event** from the EHR without **manual variable selection** or expert curation.

Dynamic Risk Tracking

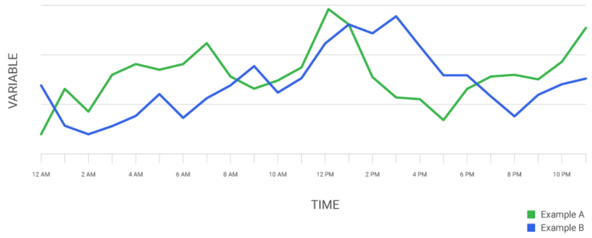
For every patient, the system continuously **updates survival probability hour-by-hour** throughout the ICU stay

Transparent Predictions – Explainable

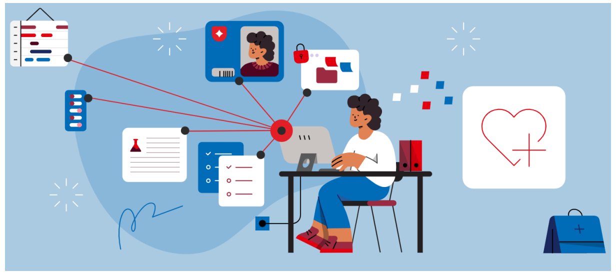
Alongside risk scores, the system provides **interpretable output** by presenting the token variables names that help for support the prediction

Essential Background

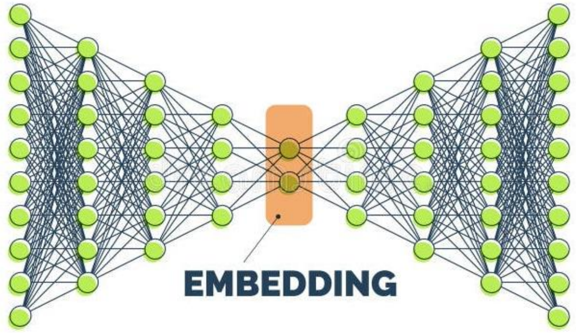
Time Series Data



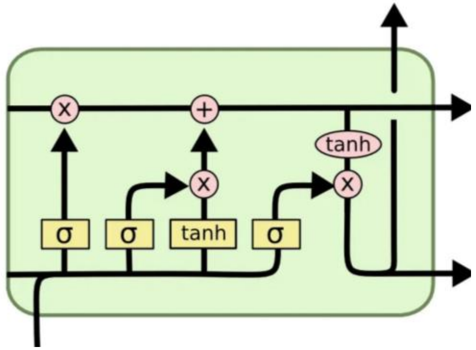
Electronic Health Records



Embedding Layer



LSTM Architecture



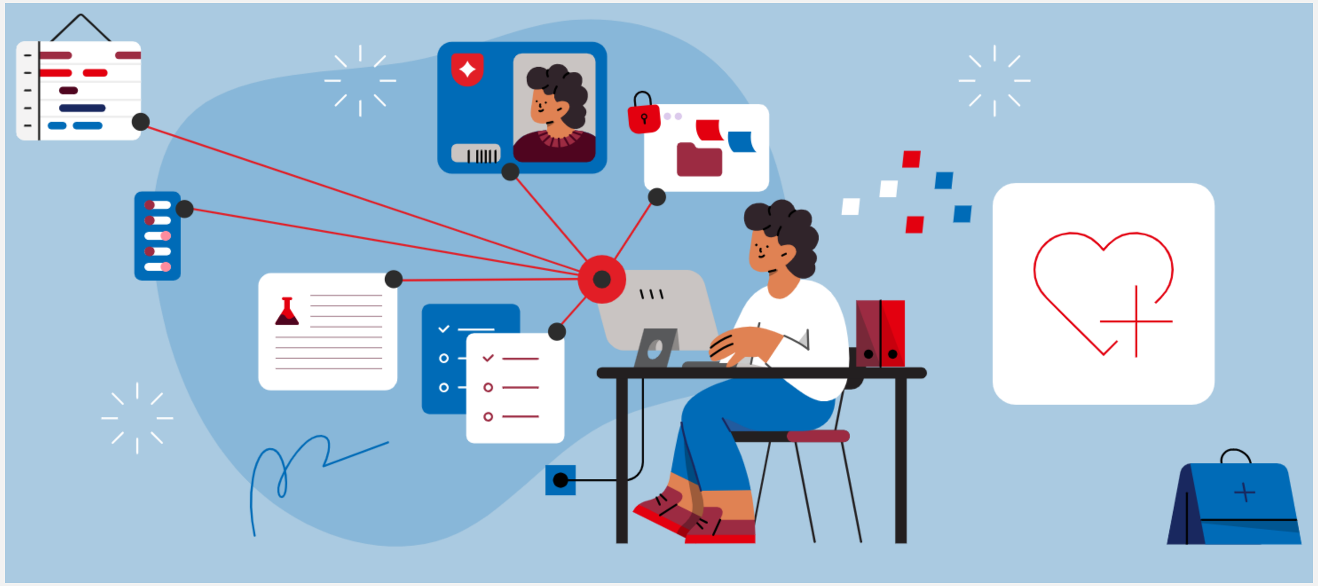
Electronic Health Records (EHR)

- EHR is **Digital patient files** replacing traditional paper charts in hospitals and clinics
- **Comprehensive medical history** stored electronically - diagnoses, treatments, medications, test results
- **Real-time data collection** from multiple sources
  - Nurses entering vital signs and observations
  - Laboratory systems uploading test results automatically
  - Medical devices streaming continuous monitoring data
  - Doctors documenting procedures and assessments
- **Structured and unstructured data** including
  - Numerical values (blood pressure, temperature, lab values)
  - Text notes (physician observations, nursing notes)
  - Coded information (diagnosis codes, medication orders)
  - Time-stamped events (when treatments were given)

Main Challenges

Massive data volume - especially in ICU settings where patients are monitored continuously

Generalization - Hard to train ML model on all types of data without feature selection



Multi-variate Time-Series Data

- **Multiple variables over time**:

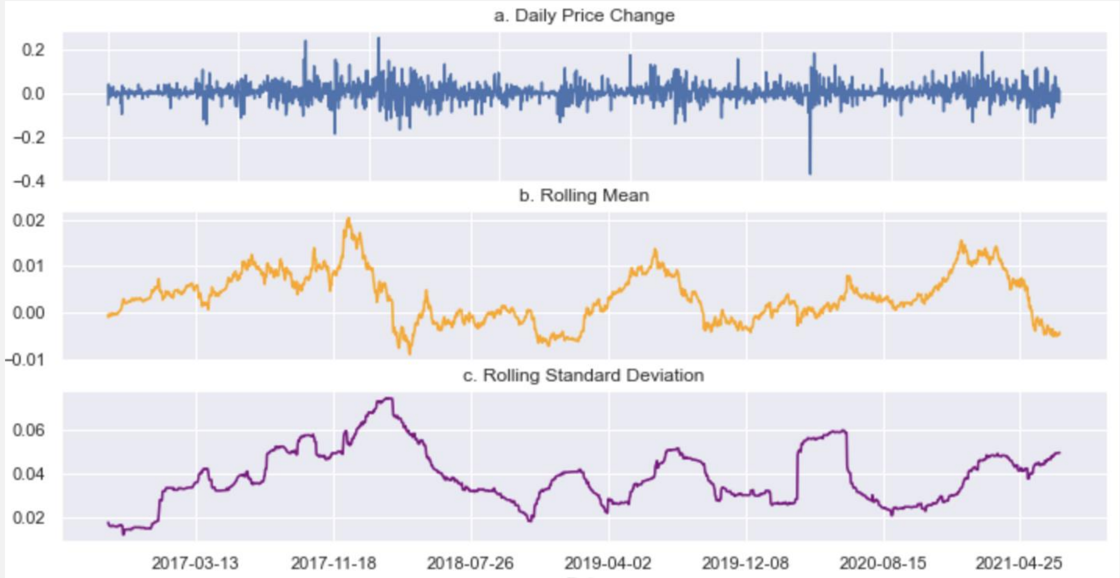
  Multivariate time series data tracks several different features or measurements (e.g., heart rate, blood pressure, oxygen level) recorded over time for a given entity

- **Captures interactions between variables:** help to analyze how variables influence each other over time

- **The Challenge -** Heterogeneous Time-Series Data

  - ICU Generates massive amounts of mixed data types
  - Continuous values (heart rate: 85 bpm, blood pressure: 120/80)
  - Discrete events (medication administered, procedures performed)
  - Different sampling frequencies (hourly vitals vs. daily lab results)

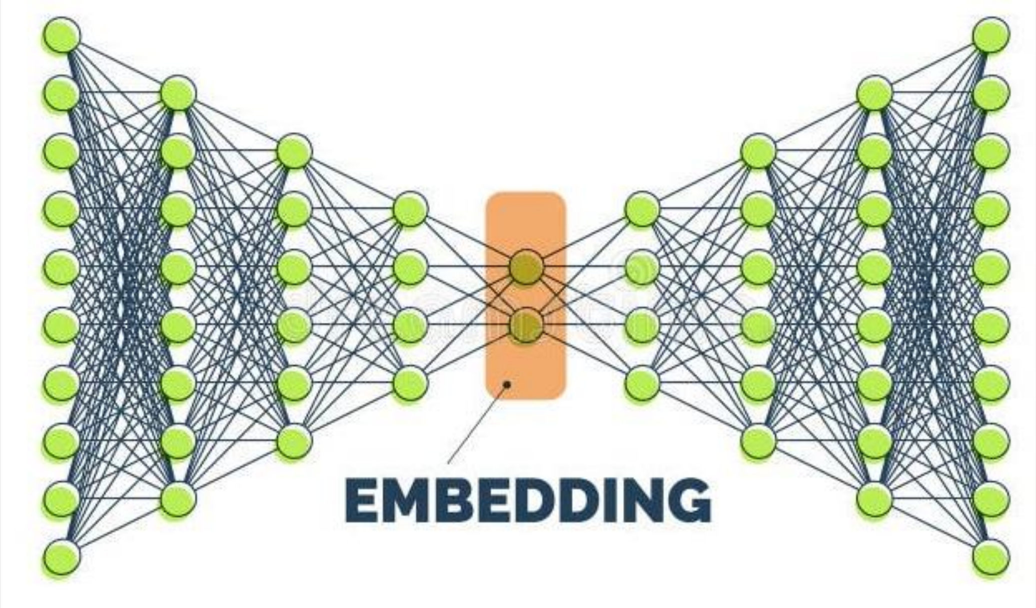
- We model each patient ICU stay by multi-variate time-series data



Embedding Layer

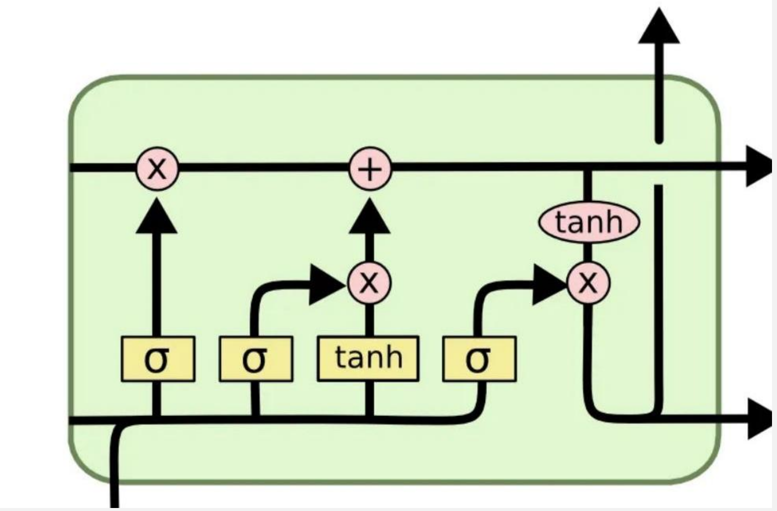
- **The NLP Analogy**:
  - Just like words in a sentence, medical events need mathematical representation
  - "Heart attack" and "myocardial infarction“ should be close in meaning
- **Challenge**: High number of medical variables from the EHR (31,913) – **traditional one-hot encoding is impossible**

How we utilized word embedding for our medical data?



LSTM (Long Short-Term Memory) Networks

- **Memory mechanism for sequences**: Remembers and forgets selectively
- **Solves the vanishing gradient problem**: Prevents information loss over time
- **Hourly updates in this study**: Processes one hour of patient data at a time
- **Dynamic predictions**: Capable to generates new mortality probability every hour
- **Temporal pattern recognition**: Learns complex time-based relationships that are too subtle for humans to systematically track.



MIMIC-III Database

- The project utilizes data from the MIMIC-III (Medical Information Mart for Intensive Care III) v1.4 clinical database, accessed via a PostgreSQL server.
- The initial data collection involves accessing and loading core tables that contain:
  - Patient demographic information
  - Hospital admission details
  - ICU stay specific information
  - Time-series clinical event data from event tables:

Chart Events (e.g., vital signs, charted observations)

Lab Events (laboratory test results)

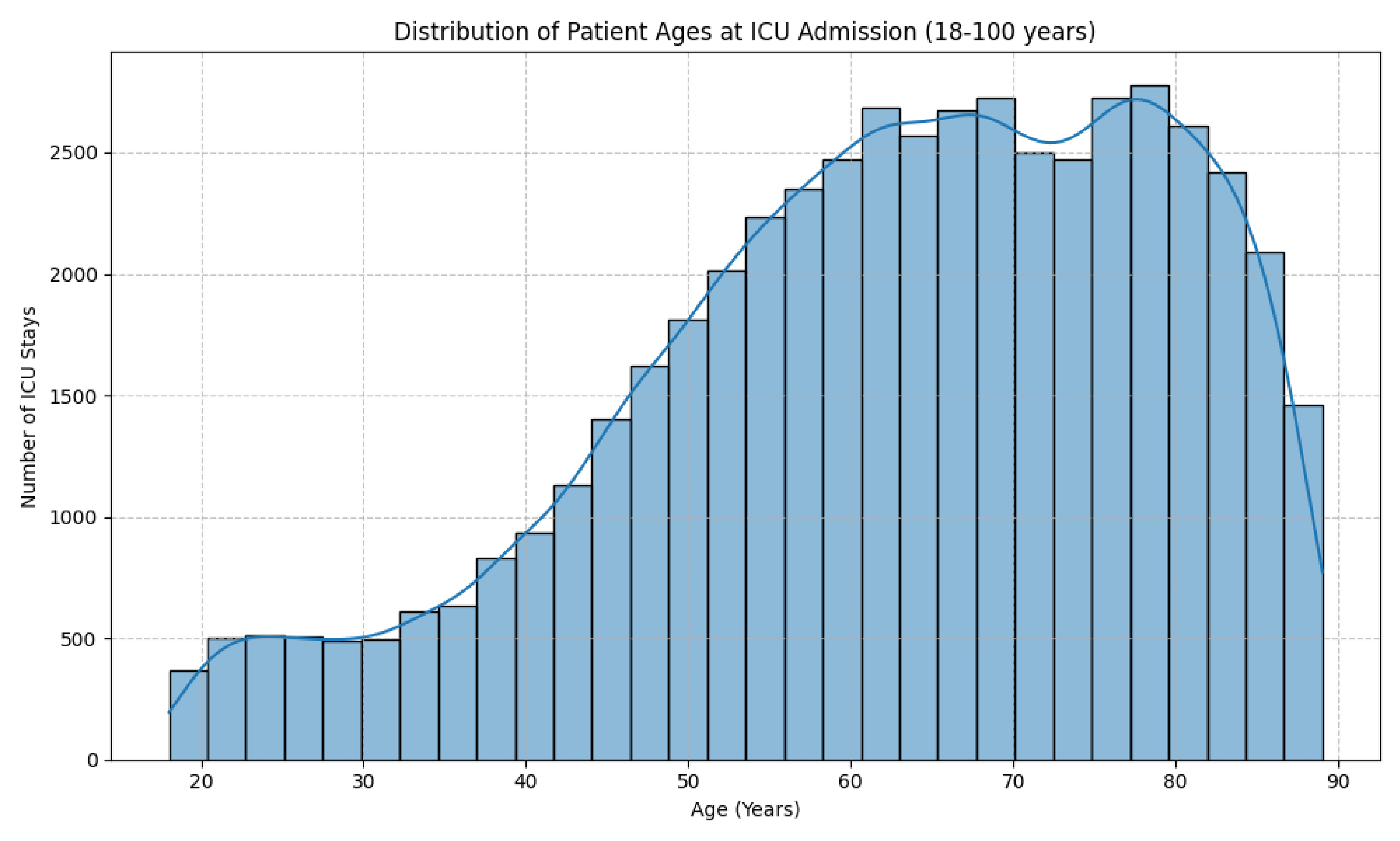
Output Events (e.g., fluid outputs, urine output)

MIMIC-III - Quantity Of Data

Table Name	Number Of Records	Number Of Feilds
PATIENTS	46,520	8
ADMISSIONS	58,976	19
ICUSTAYS	61,532	12
CHARTEVENTS	330,712,483	15
LABEVENTS	27,854,055	9
OUTPUTEVENTS	4,349,218	13

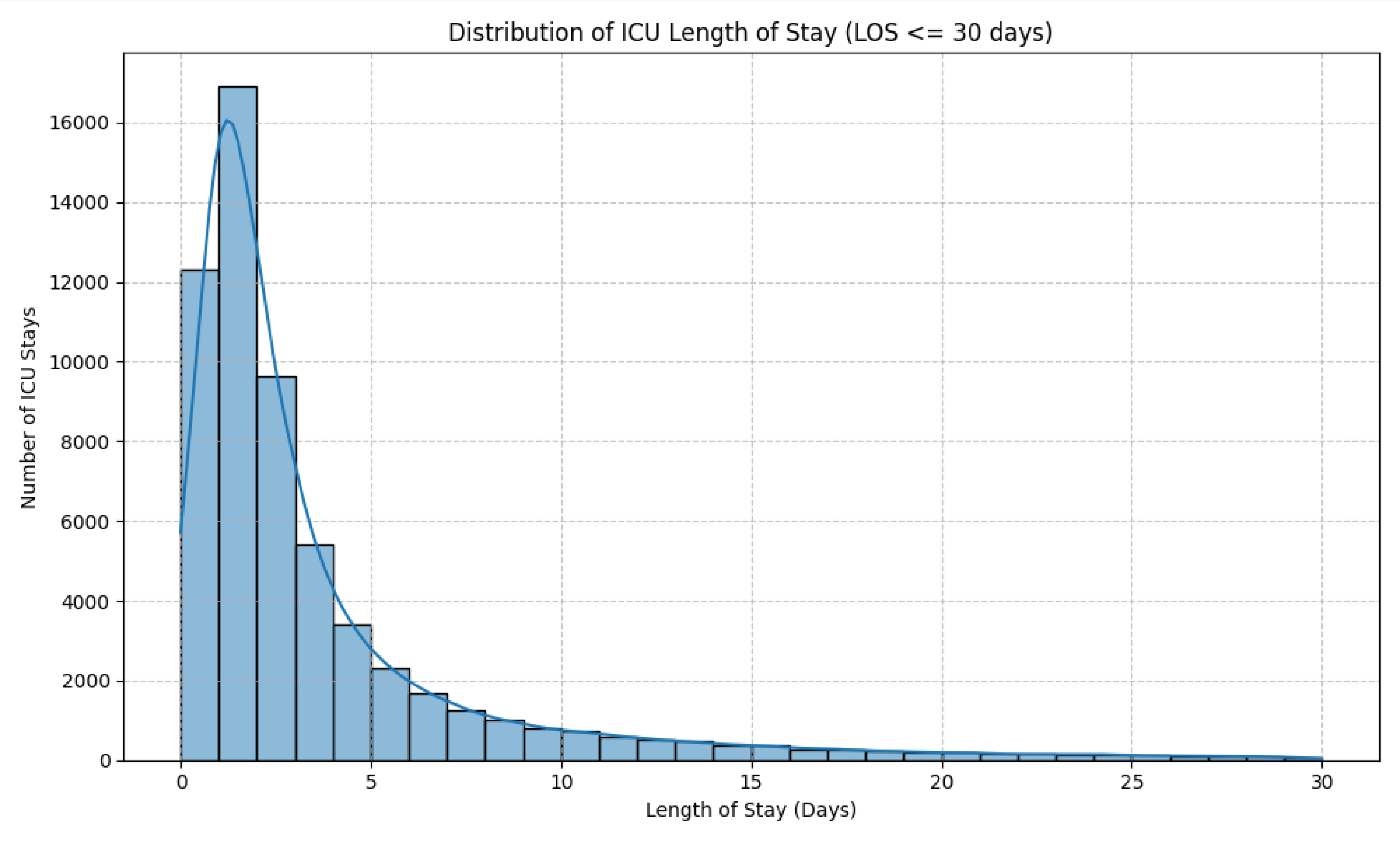
MIMIC-III - EDA

- **Distribution of Patient Ages at ICU Admission:**
  - Majority of patients are between 60 and 80 years old



MIMIC-III - EDA

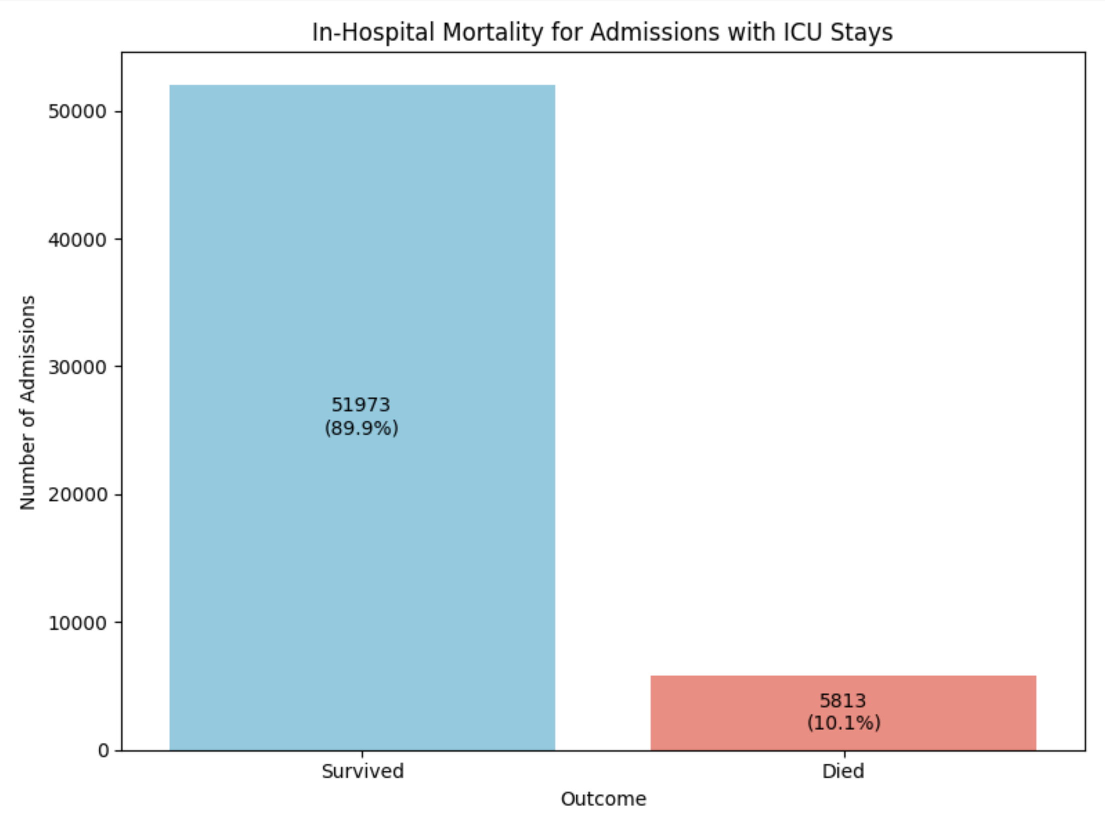
- **Distribution of ICU Length of Stay (LOS):**
  - Skewed right, most ICU stays are under 5 days, long tails exist.



MIMIC-III - EDA

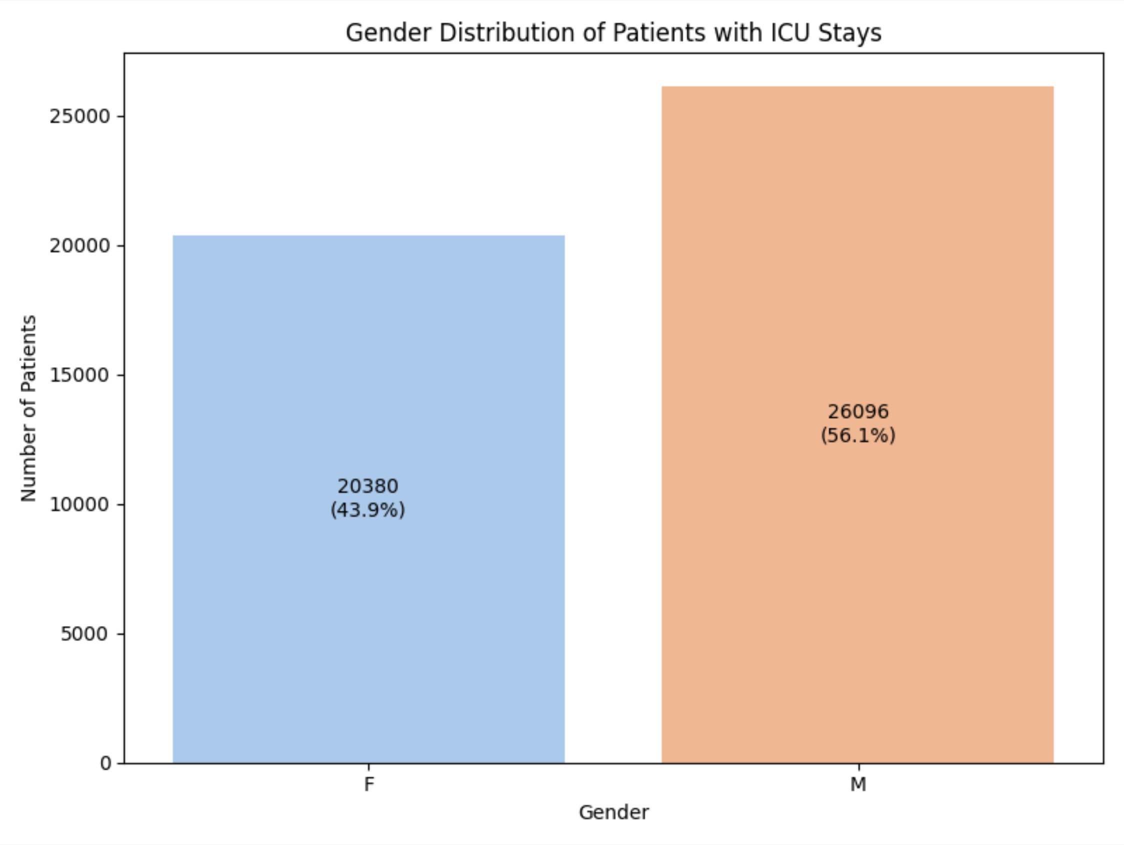
- **In-Hospital Mortality Rate for ICU Cohort:**

  10–15% mortality in ICU cohort (confirms class imbalance).



- **Gender Distribution of ICU Patients:**

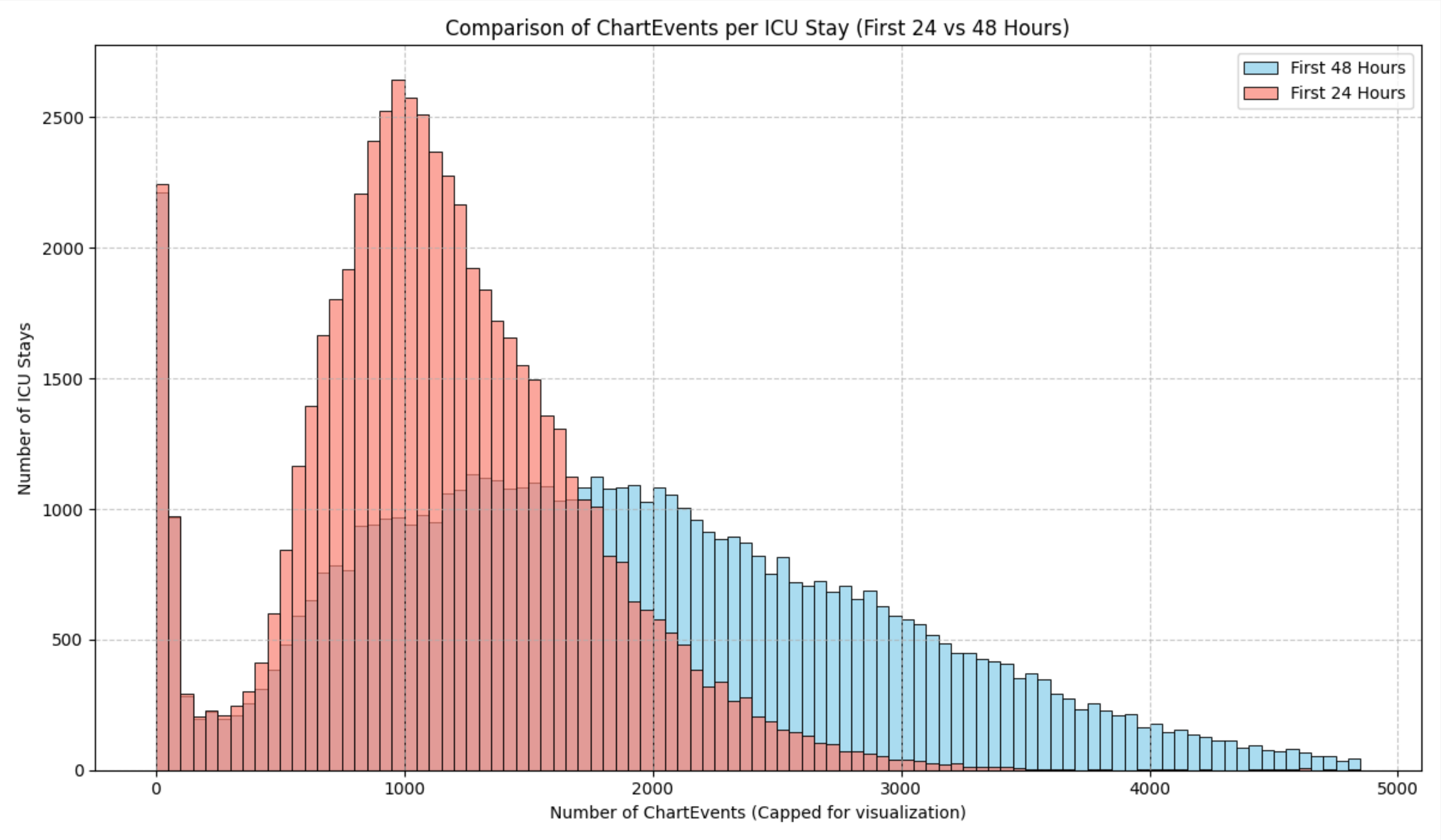
  Slight male majority (56.1% Male, 43.9% Female)



MIMIC-III - EDA

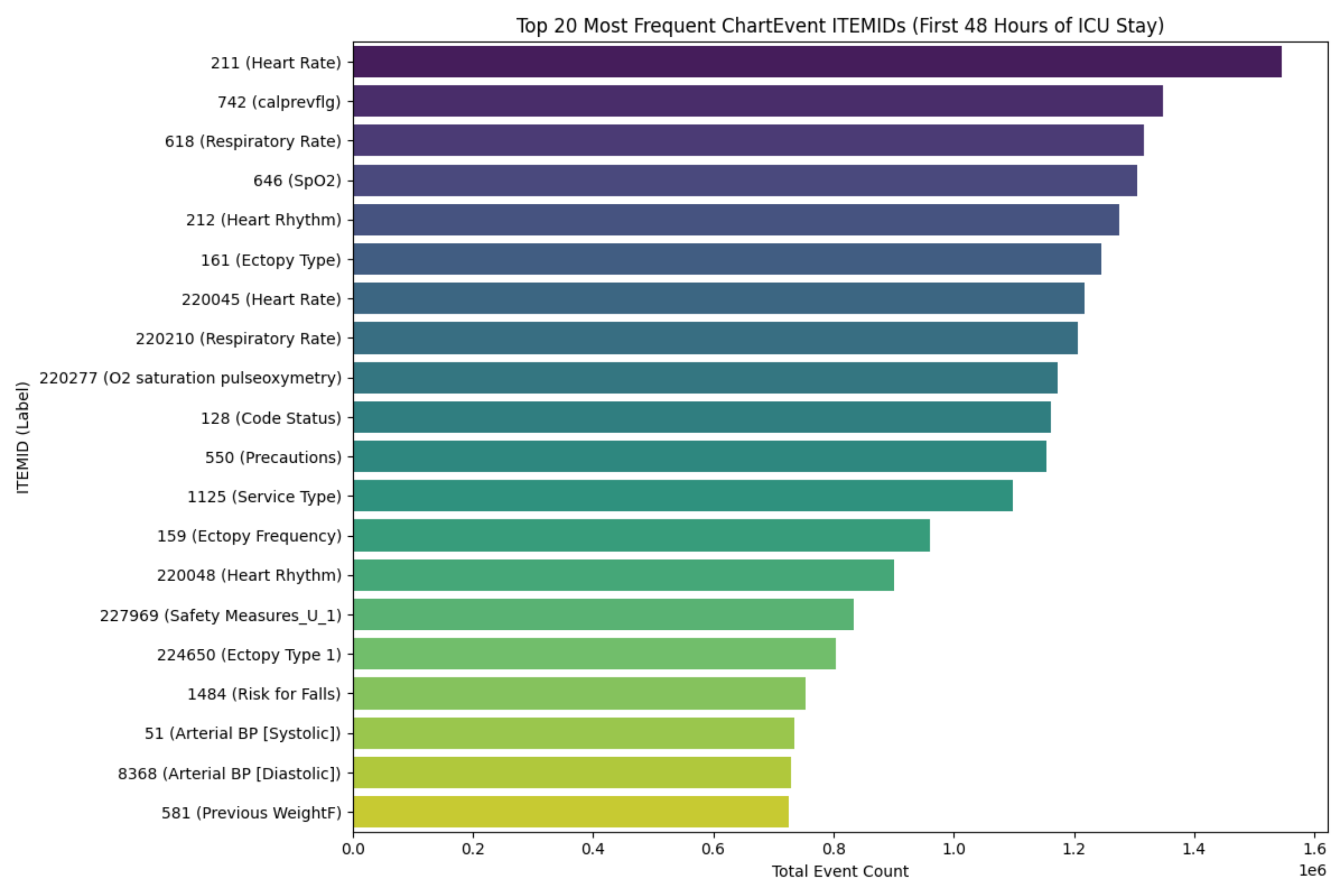
- **Comparison of ChartEvents per ICU Stay (First 24 vs. 48 Hours):**

  Sharp increase from 24h to 48h; supports value in longer observation windows

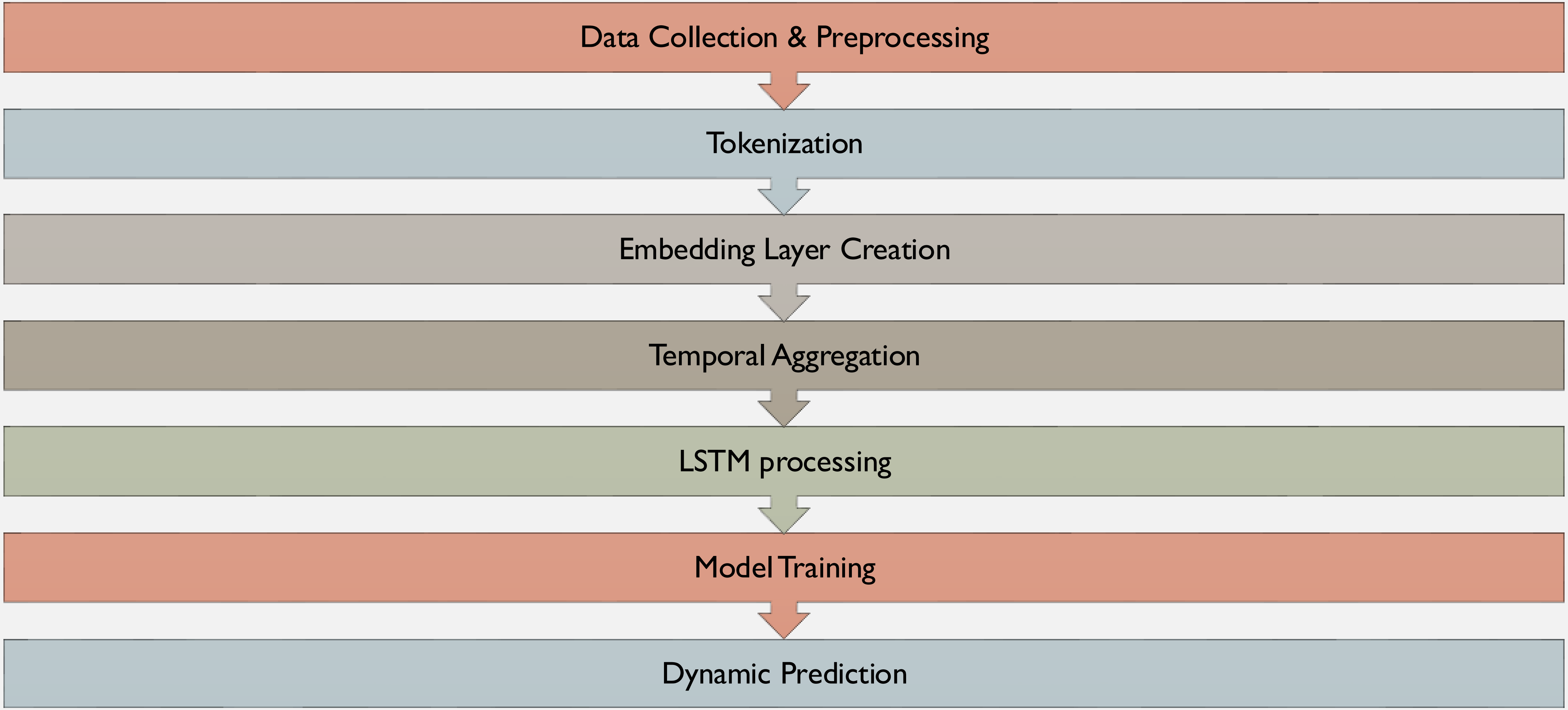


MIMIC-III - EDA

- **Most Frequent ChartEvent ITEMIDs (First 48 Hours):**
  - Features like heart rate, respiratory rate, SpO_2, rhythm dominate early ICU data



Model Pipeline - Overview



Step 1 – Data Collection & Preprocessing

Data Collection

- Retain ALL vital signs, lab, and output events from MIMIC-III db (without cleaning/filtering)
- Assign patient ID, stay ID, and timestamps to each event

Data Preprocessing

- For each patient, we observe up to the first 48 hours of data after ICU admission
- Within each hour, we allow up to 5000 events to be recorded

Step 2 – Tokenization Process

Variable Type Classification

Check if value is continuous or discreate

Example

Continuous: Heart rate (84) / Blood pH (7.4)

Discrete: "Eye Opening 4 Spontaneously"

Continuous Variable Processing

Percentile-based binning: Divide each variable's distribution into 10 bins

Example

Heart Rate: All heart rate values sorted: 60,65,70,75,80,85,90,95,100

8th percentile → becomes "Heart Rate_8_BPM"

Uniform Token Creation

Create Token for each variable in unifrom format:

`[Variable Name]_[Value/Bin]`

Example

Continuous: "Heart Rate_8_BPM" (8th percentile bin)

Discrete: "Eye Opening 4 Spontaneously"

Step 3 – Embedding Layer Creation

After creating tokens for each value in our data each unique token in medical vocabulary gets assigned a learnable vector:

- **Vocabulary size**: 31,913+ unique medical tokens
- **Embedding dimensions**: 16-64 dimensional vectors (optimized via grid search)
- **Learning process**: These vectors start random and are learned during training through backpropagation

$$\overbrace{\begin{bmatrix} \text{Eye Opening 4 Spontaneously} \\ \text{Heart Rate\_8} \\ \text{Code Status Full Code} \\ \vdots \end{bmatrix}}^{\text{Discrete event tokens in 1 h}} \rightarrow \overbrace{\begin{bmatrix} -0.22, 0.34, \ldots, 0.83 \\ 0.52, 0.03, \ldots, -0.28 \\ -0.21, 0.17, \ldots, 0.89 \\ \vdots \end{bmatrix}}^{\text{Corresponding learnt vectors}}.$$

Step 4 – Temporal Aggregation

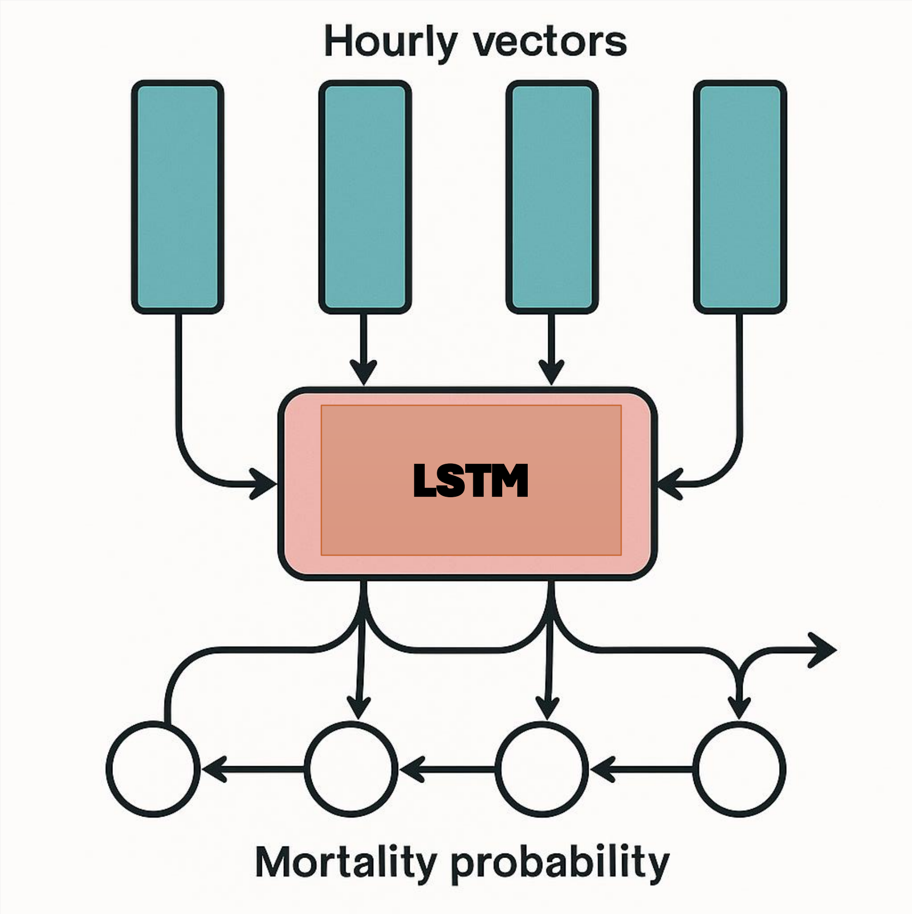
For each hour, multiple medical events are combined:

- Each token's embedding vector gets multiplied by a learned weight
- All weighted embeddings for that hour are summed into a single hourly representation
- Formula: `Aggregated hourly vector = Σ (learned_weight × embedding_vector)`

$$\overbrace{[w_0, w_1, w_3, \ldots]}^{\text{Learnt weights}} \overbrace{\begin{bmatrix} -0.22,\ 0.34, \ldots, \ 0.83 \\ 0.52,\ 0.03, \ldots, \ -0.28 \\ -0.21,\ 0.17, \ldots, \ 0.89 \\ \vdots \end{bmatrix}}^{\text{Learnt vectors}} = \overbrace{[0.04, -0.52, \ldots, -0.72]}^{\text{Aggregated hourly vector}},$$

Step 5 – LSTM Processing

- Feed hourly vectors sequentially to LSTM network
- Update internal memory state each hour
- Generate mortality probability at each timestep



Step 6 – Model Training

- Finally, we use a **densely connected layer** with a sigmoid activation function to output the probability of in-patient mortality given a patient timeseries
- Cross-Entropy Loss:

$$\mathscr{L}(y, \tilde{y}) = -\sum_{i=1}^{N}\sum_{t=0}^{T} \overbrace{\tilde{y}_{it}\log y_i}^{\text{Misclassified death loss}} + \overbrace{(1-\tilde{y}_{it})\log(1-y_i)}^{\text{Misclassified survival loss}},$$

- **Early Stopping:** Training automatically stops when validation(1000 entities) AUROC plateaus for more than 5 epochs
- **Hyperparameter optimization via grid search:**

Embedding dimensions	Hidden LSTM neurons	Dropout probability	Batch Size	Learning Rate
16	32	10%	32	0.005
32	64	20%	64	0.001
48	128	30%	128	0.0005
64	256	0%	256	0.0001

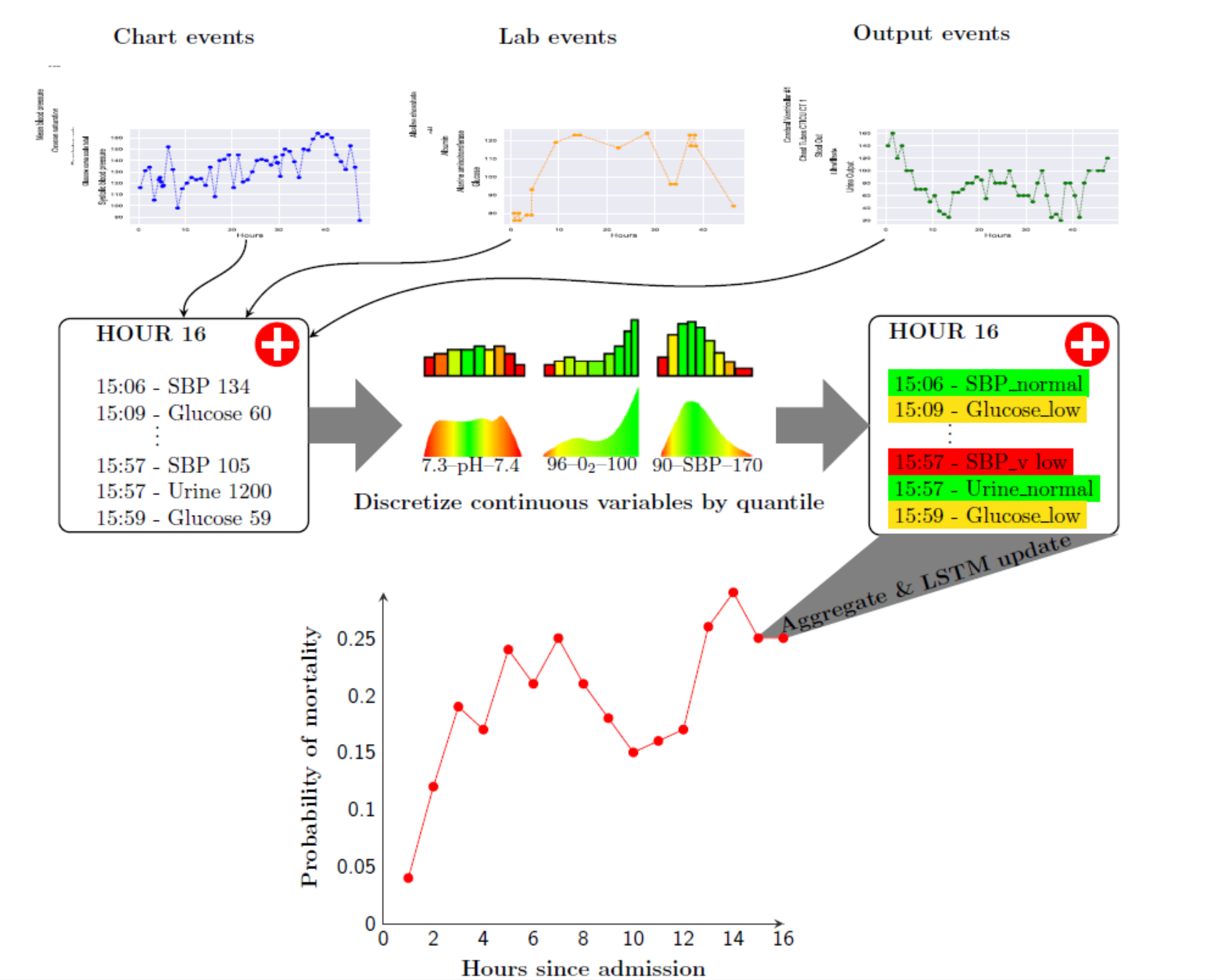
Step 7 – Dynamic Prediction

- Continuous probability updates with confidence interval: Model outputs mortality probability (0-1 scale) every hour for 48 hours.
- Interpretable variable importance rankings: The learned weight system automatically ranks which medical events contributed most to each hourly prediction, providing clinicians with transparent reasoning like "high blood pressure (rank 1), low temperature (rank 2)" for clinical decision support.

```
Hour 36-37:
  Events:
    - Token: 220277_%:1                                      | Meaning: O2 saturation pulseoxymetry
    - Token: 220052_mmHg:3                                   | Meaning: Arterial Blood Pressure mean
    - Token: 220210_insp/min:7                               | Meaning: Respiratory Rate
    - Token: 220051_mmHg:5                                   | Meaning: Arterial Blood Pressure diastolic
    - Token: 220050_mmHg:1                                   | Meaning: Arterial Blood Pressure systolic
    - Token: 220045_bpm:17                                   | Meaning: Heart Rate
```

Predicted Probability : 0.4312

Full Pipeline



Results

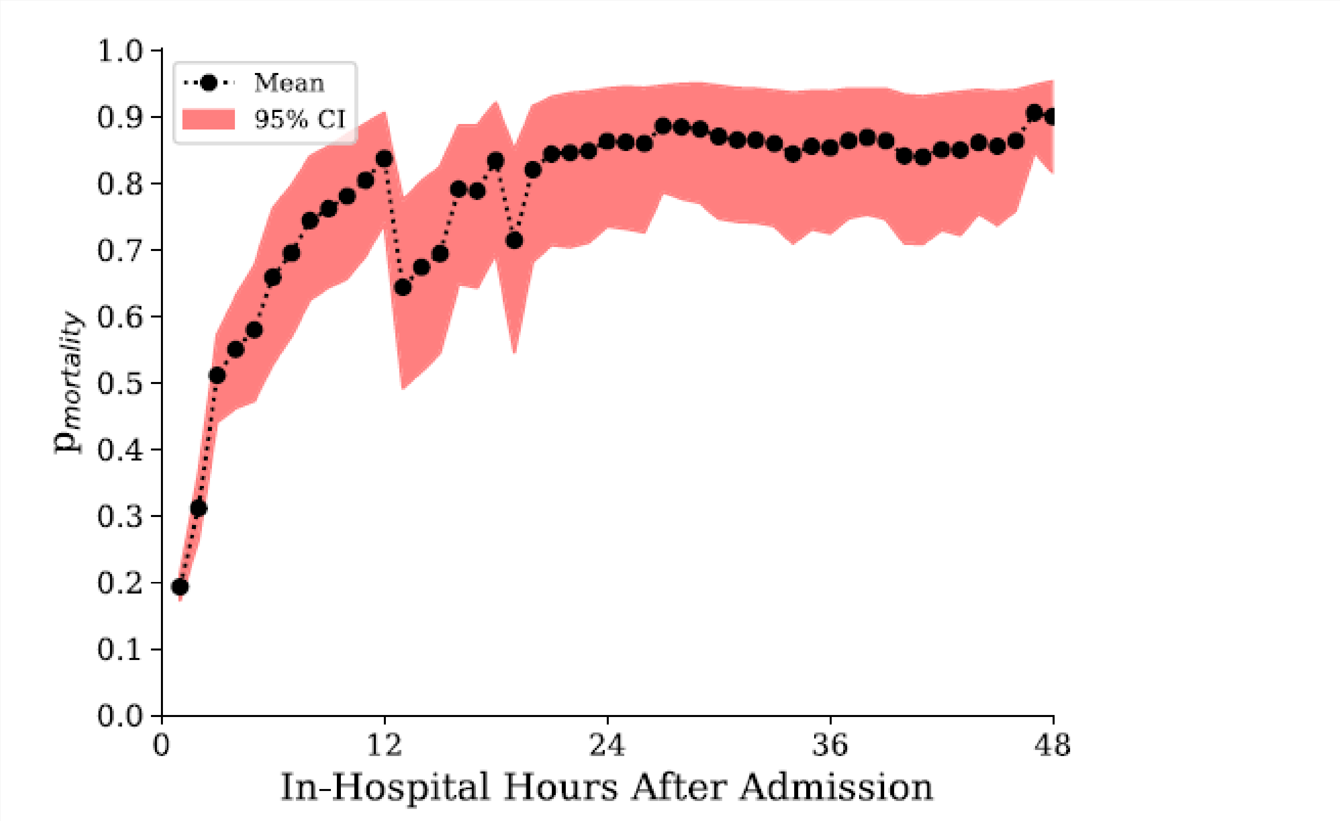
- Best Model Results: **0.8922 Test AUCROC**
- Hyperparameters config:

Batch_size	Hidden_dim	lr	P_dropout	Latent_dim	Number_of_ bins
128	256	0.0005	0.1	64	20

Mortality Prediction Case Study

Dynamic probability of mortality after ICU admission for a patient who subsequently died.

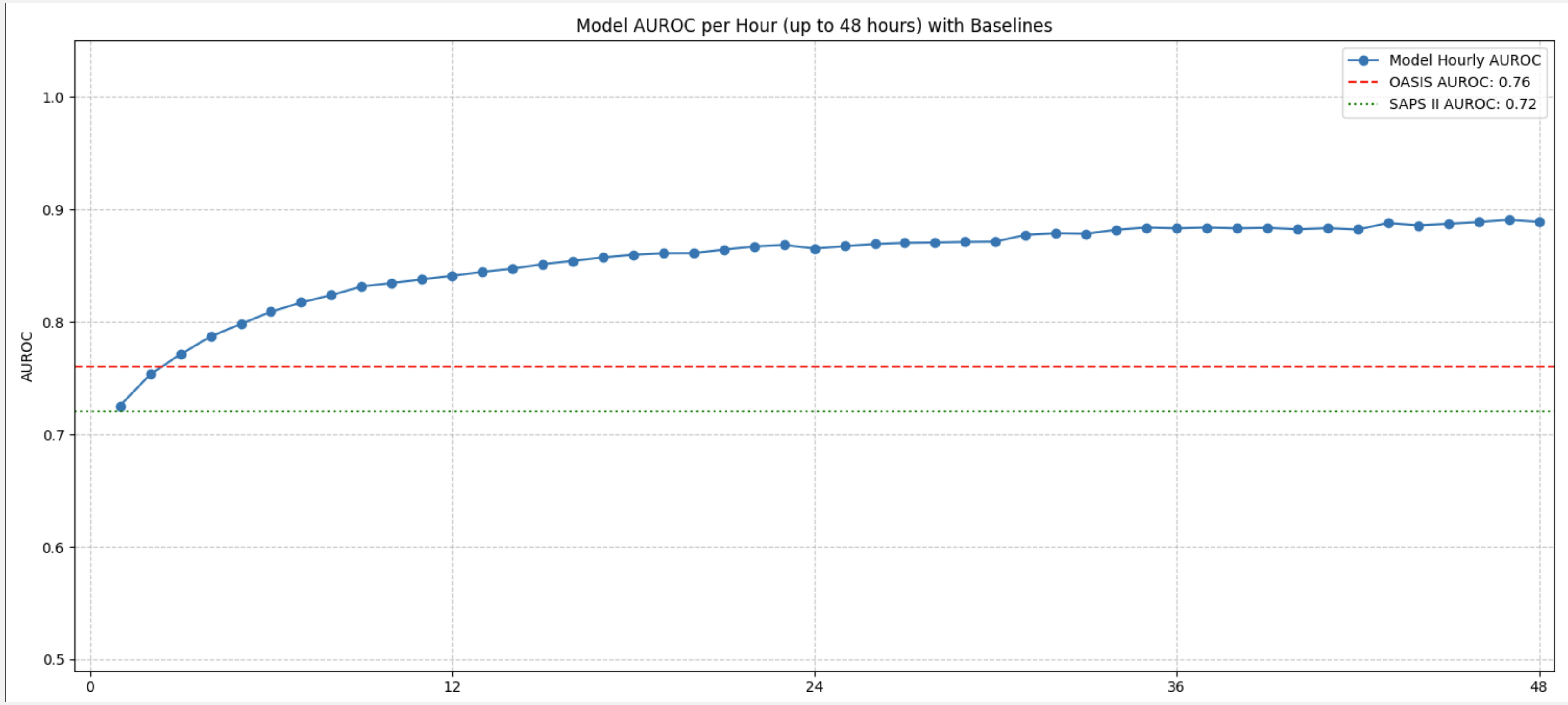
- **Hour 1**: 19% risk (concerning but not critical)
- **Hour 3**: 50% risk (major deterioration detected)
- **Hour 12**: 83% risk (very high confidence of poor outcome)
- **Hours 24-48**: Maintained 85-90% (consistent high-risk prediction)
- **Actual outcome**: Patient died, confirming model accuracy



Comparison of AUC-ROC

•**Dynamic vs. Static:** Our model provides a continuously improving AUC-ROC over time (black line), unlike static traditional scores (OASIS, SAPS II).

•**Superior Performance:** The model outperforms OASIS and SAPS II at ~4 hour mark, offering higher predictive accuracy as more data is collected.



Code Demo

[Colab Demo]