ICU Time Series Analysis: Predictive Modeling and Representation Learning

Bondi Francesco
ETH Zurich
fbondi@ethz.ch
24-942-872

Fontana Saverio
ETH Zurich
sfontana@ethz.ch
24-942-971

Tilman Otto
ETH Zurich
tiotto@ethz.ch
24-963-308

Abstract

ICUs generate extensive time-series data, crucial for patient monitoring and clinical decisions. This project applies statistical and machine learning techniques to analyze vital signs, lab results, and physiological data, identifying trends, anomalies, and predicting deterioration. Using feature engineering, forecasting models, and performance evaluation, we demonstrate the potential of data-driven approaches to enhance patient care.

Q1 Data Processing and Exploration

Q1.1 Data Transformation

We first standardized the time series data (41 features) by aligning them to a 48-hour grid (49 hourly steps, including 00:00). To maintain temporal causality, timestamps were rounded down (not up) to the nearest preceding hour, preventing future information leakage during alignment.

Q1.2 Exploratory Data Analysis

Age distribution The youngest patient was 15, the median age was 67, and the oldest was 90, with no older patients recorded. This sharp cut-off at 90 may reflect data truncation or ICU admission limits. The compressed upper percentiles (80th: 80 years, 90th: 85 years) further suggest fewer very elderly patients. [1]

Single-observation patients Three patients (IDs: 140501, 140936, 141264) had only one time observation. Since the train/validation/test sets each contained 4000 patients, we decided not to remove them. However, future studies should consider their removal, as suggested by Horn et al.[2]

ICU Type and Mechanical Ventilation A Pearson's $\chi^2(3) = 473.49$ ($p < 0.001$) revealed a significant association between ICU type and mechanical ventilation, with moderate effect size (Cramér's $V = 0.34$, 95% CI [0.31, 0.37]). Standardized residuals ($|z| > 2.0$) in Appendix 2 indicated higher ventilation in **Cardiac Surgery (Type 2)** (post-op needs), lower rates in **CCU/Medical (Types 1/3)** (medical management), and moderate use in **Surgical ICU (Type 4)** (post-procedural needs).

Heart Rate by Gender Though statistically significant ($p = 0.008$), the mean difference was clinically negligible (86.90 vs 86.67 bpm, $\Delta = 0.23$ bpm; Appendix 9).[3]

[1] See Appendix Table 1 for the quantiles and Appendix Figure 8 for the complete distribution

[2] Horn, Max, et al. "Set functions for time series." *International Conference on Machine Learning*. PMLR, 2020.

[3] Note this is a preliminary analysis; confounder adjustment is needed for proper causal inference.

Q1.3 Preprocess data for Machine Learning

Data Preprocessing Our preprocessing pipeline addressed missing values and feature scaling while strictly preserving temporal causality to ensure that future information is not used to estimate current values. For missing values, we followed Physionet guidelines where a negative value indicates missingness (see Appendix 3 for details), applying KNN imputation with 10 neighbors for static variables (Weight, Height, Age, Gender) and forward-fill imputation for time-series data, supplemented by time based interpolation for initial missing values (backward filling). Feature scaling was based on an empirical analysis of the data distribution (see Appendix 4): StandardScaler for normally-distributed features like blood pressure, RobustScaler for skewed variables like potassium levels, while MinMaxScaler was used for LSTM inputs. We maintained the original non-regular observations (e.g. 186,416 for Set A) without adding synthetic time points to avoid generating excessive artificial data.

Q2 Supervised Learning

Q2.1 Classic Machine Learning Methods

We aggregated each variable to obtain a single observation per patient. The aggregation functions were carefully selected for each variable based on domain knowledge and exploratory data analysis (see Table 5). Then, we applied two feature engineering strategies. **Feature Lagging**: Shifted time-series variables backward to capture temporal patterns and trends. The result is an augmented dataframe with added context from previous time steps. **Signal Processing Features**: Extracted 215 features (e.g., statistics, spectral trends) via `tsfresh`, then reduced dimensionality using L1 regularization.

Figure 1: Performance of Classic ML Methods

Model	Features	AuROC	AuPRC
Logistic	Basic	0.770	0.392
Logistic	Lagged	0.837	0.491
Logistic (L1)	Signal	0.816	0.479
Random Forest	Basic	0.778	0.410
Random Forest	Lagged	0.831	0.497
Random Forest	Signal	0.810	0.465

Logistic Regression is simple, interpretable and computationally efficient. L1 regularization enables feature selection, beneficial for high-dimensional data. Its main limitation is assuming linear decision boundaries, restricting performance when underlying relationships are non-linear. **Random Forest** handles nonlinear relationships and feature interactions well, resisting overfitting. It works with diverse features (basic, lagged, signal-processed) but is less interpretable. It may underperform simpler models for linear relationships (e.g., logistic regression with lagged features had higher AuPRC). Basic features yielded modest performance (AuROC $\sim$0.77–0.78), whereas lagged features improved results by capturing temporal dynamics, highlighting the importance of time-dependent patterns. Logistic regression with L1 regularization excelled with signal features due to its ability to filter noise and retain discriminative features.

Q2.2 Recurrent Neural Networks

Using MinMax-scaled data, we implemented a 2-layer LSTM (64 hidden units, dropout $p = 0.3$). Max pooling is employed, as it emphasizes extreme values in physiological time-series data—such as peaks in heart rate or drops in blood pressure—which are often critical indicators of patient deterioration. Unlike mean pooling, which averages signals and may dilute infrequent but clinically significant events, max pooling preserves the most abnormal measurements.

We further employ a bidirectional architecture, which provides two key clinical advantages. First, it improves contextual understanding: physiological patterns often have both preceding and succeeding indicators. For instance, a blood pressure crisis may be better recognized when considering gradual changes across the full 48-hour window in both temporal directions. Second, it helps compensate for irregular sampling. Since medical measurements are taken at uneven intervals, bidirectional processing allows the model to interpolate missing information by analyzing the overall trajectory from both past and future perspectives.

Q2.3 Transformers and Tokenizing Time-Series Data

We implemented a transformer model using PyTorch with the following structure: input projection from 41D to 128D, 2 transformer layers with 4 attention heads, using mean pooling over time, with dropout ($p = 0.3$) and layer normalization.

Key advantages over RNNs include the ability of transformers to perform parallel computation, unlike RNNs which process data sequentially. Transformers are also better at capturing long-term dependencies, whereas RNNs may suffer from vanishing gradients. Additionally, transformers are more scalable, as RNNs often require additional techniques to train effectively. The transformer used more memory but gave slightly better results while being easier to train.

Figure 2: ML Model Performance Comparison

Model Variant	AuROC	AuPRC
Unidirectional LSTM	0.779	0.379
Bidirectional LSTM	0.841	0.506
Transformer	0.841	0.510
Tokenized Transformer	0.523	0.164

Tokenizing Time-Series Data We also reported the results using Horn et al.'s triplet encoding (t, z, v) with a Transformer. However, the results were not up to the expectations.

Q3 Representation Learning

Q3.1 Pretraining and Linear Probes

We pretrained an LSTM autoencoder using the time grid representation from Q1.3, maintaining architectural consistency with our supervised LSTM from Q2.2. The pretraining progress was tracked through the mean squared error (MSE) reconstruction loss.
With frozen pretrained embeddings, logistic regression achieved an AuROC of 0.569 and an AuPRC of 0.214. The method did indeed perform poorly with respect to previous ones.

Q3.2 Simulate label scarcity

We compared two approaches under limited labeled data: training logistic regression on 100/500/1000 samples and training logistic regression on 100/500/1000 frozen pretrained embeddings (linear probe).

Figure 3: Performance Under Label Scarcity

Method	Samples	AuROC	AuPRC
Logistic	100	0.638	0.267
Linear Probe	100	0.562	0.178
Logistic	500	0.726	0.341
Linear Probe	500	0.632	0.244
Logistic	1000	0.757	0.354
Linear Probe	1000	0.635	0.258

The results demonstrate that logistic regression trained directly on the data consistently outperforms linear probing on frozen pretrained embeddings across all sample sizes (100, 500, and 1000), achieving higher AuROC and AuPRC values. The findings indicate that pretraining, in this case, does not justify the additional computational cost, as supervised training alone yields better results even with limited labeled data.

Q3.3 Visualising Learned Representations

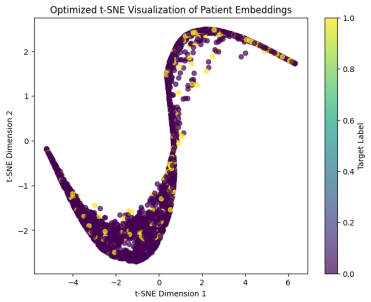

Figure 4: t-SNE plot of patient embeddings colored by mortality

t-SNE separation plot: We projected the 128D patient embeddings to 2D using t-SNE for visualization. Figure 4 shows the 2D projection where data points with different labels are not distributed identically (yellow = deceased, purple = survived). Some class separation is visible, but there remains significant overlap. The quality of the clustering was evaluated using the Silhouette Score; the obtained score of 0.068 confirms the visual observation of weak separation.

Q4 Foundation Models

Q4.1 Prompting an LLM to solve a time-series problem

Prompting & LLM Choice: The best-performing prompting strategy incorporated the most important features identified in Q2.1, providing the LLM with their value range, average, last observed value, and an overall trend. This approach maintained conciseness by minimizing feature count and token usage. Trends were derived by comparing averages across three time segments (0–16h, 16–32h, and 32–48h) and classified as increasing, decreasing, stable, or fluctuating—introducing a temporal dimension to help the LLM infer clinical progression. (See Appendix for an example prompt.)

We augmented this with a system prompt outlining the general task. To evaluate robustness, we compared zero-shot predictions to few-shot predictions (using three training-set examples) and tested binary scoring against a granular 0–10 scale. All experiments were conducted using Ollama's Gemma2:9b model, which outperformed DeepSeek-R1:7b and Llama3.2:3b (See Appendix).

Prediction results: Gemma2:9b achieved the best performance among the LLMs with few-shot prediction, reaching an AuROC of 0.6921 and an AuPRC of 0.2555 when predicting a score from 0 to 10. This supports the idea that few-shot examples improve performance, and that using a continuous scoring scale is more informative than binary outputs. In contrast, binary predictions often lead to less meaningful and overly simplified confidence scores. Despite this, LLM-based prediction still underperforms compared to most other methods used in the project.

Figure 5: LLM Performance Comparison

Model Variant	AuROC	AuPRC
Zero-Shot	0.683	0.285
Few-Shot	0.692	0.256

Q4.2 Using LLMs to retrieve embeddings

LLM-Embedding based Logistic Regression: The logistic regression using LLM-generated embeddings performed worse **(AuROC: 0.707, AuPRC: 0.312)** than both the supervised and domain-specific models. This likely stems from several issues: the embeddings may capture general language patterns rather than clinical signals; the input text is often noisy or loosely related to the target; and the high dimensionality of the embeddings can lead to overfitting, especially with limited training data. Moreover, this approach does not capture temporal or structured information from the original time-series, which is crucial for the task.

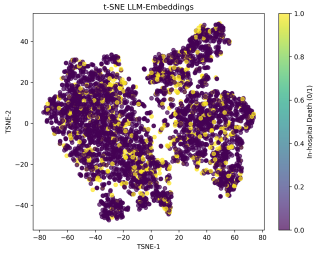

Figure 6: t-SNE plot of patient embeddings

t-SNE separation plot: We projected the 3584D patient embeddings to 2D using t-SNE for visualization. Figure 6 shows the 2D projection where data points with different labels are not distributed identically (yellow=deceased, purple=survived). Compared to Q3.3 we have a lot more original dimensions. The quality of the clustering was evaluated using the Silhouette Score; the obtained score of -0.019 confirms the visual observation of weak separation similar to Q3.3.

Q4.3 Using time-series foundation models

Chronos Usage: Using the pretrained amazon/chronos-t5-small model for the simple aggregation we computed for each patient an embedding for every variable and then averaged these embeddings across variables. A standard logistic regression model is then trained on these patient-level embeddings to predict in-hospital mortality.

Figure 7: Results time-series foundation model

Model Variant	AuROC	AuPRC
Mean-Aggregation	0.636	0.226
MLP Aggregation	0.772	0.384

For a smarter aggregation, we introduced a lightweight TemporalChannelAggregator network, which replaces the averaging step. Instead, each variable v is embedded into a sequence of vectors $\mathbf{e}_{i,v}$ that preserves its temporal structure, and these sequences are processed by a GRU to extract dynamic summaries. We then stack the resulting summaries into a matrix $\mathbf{E}_i \in \mathbb{R}^{|V_i| \times h}$ and feed it into an MLP-based aggregator, specifically two fully-connected layers with ReLU activation.

While the mean aggregation-based logistic regression performs poorly, using the GRU-enhanced aggregation enables the model to capture meaningful temporal patterns in the data. However, there remain alternative approaches that may further improve predictive performance.

Q5 General Questions

Q5.1 Classic methods vs Deep Learning

Classic methods like logistic regression or random forests remained competitive with deep learning models. This is likely because the dataset is not extremely large, and much of the temporal complexity can be captured through feature engineering. Classic models also benefit from being easier to interpret, faster to train, and less prone to overfitting with limited data. Deep learning may only outperform when enough data is available and raw sequences are used effectively without heavy preprocessing.

Q5.2 Attention bottlenecks for long time series

Attention's O(T^2) complexity becomes a bottleneck for very long time series. For such cases, recurrent architectures (RNNs, GRUs, LSTMs) with O(T) complexity or transformer variants with efficient attention (e.g., sparse, linear, or memory-compressed attention) would be more suitable.

Q5.3 Challenges in self-supervised representation learning

Self-supervised learning for medical time series is challenging due to the need for meaningful pretext tasks and clinically valid augmentations. A major difficulty in this project was designing augmentations that preserved important patterns and that embeddings improved downstream tasks. Monitoring progress without labels and tuning hyperparameters without clear validation metrics also added complexity. Additionally, learned embeddings may lack interpretability, making it hard to align them with medical knowledge.

Appendix

Table 1: Patient Age Quantiles

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
15	40	49	56	61	67	72	77	80	85	90

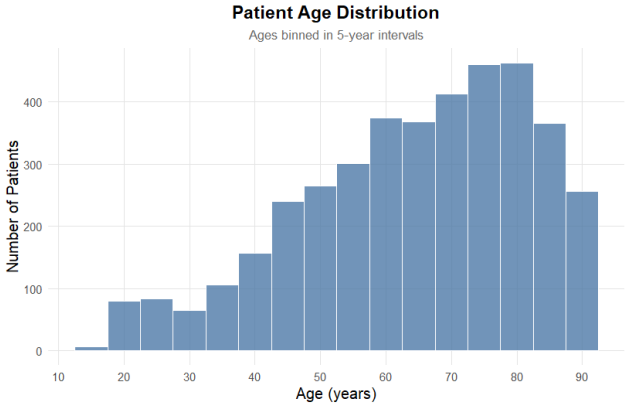

Figure 8: Distribution of patient ages

Table 2: Standardized residuals for mechanical ventilation by ICU type

ICU Type	No Ventilation (0)	Ventilation (1)
Coronary Care (Type 1)	+9.50	-9.50
Cardiac Surgery (Type 2)	-19.32	+19.32
Medical ICU (Type 3)	+12.45	-12.45
Surgical ICU (Type 4)	-3.10	+3.10

Figure 9: Violin plot of Heart Rate by Gender

Table 3: Data Preprocessing Steps for Missing Values and Outliers

Variable	Handling Rule
Age, Gender, Height, ICUType, Weight	Set value '-1' to 'NA' (missing).
Height	Set to 'NA' if < 100 cm or ≥ 300 cm.
Weight	Set to 'NA' if < 20 kg or ≥ 300 kg.
PaO2	• Set to 'NA' if equal to '0'. • If value is '7.47', correct to '74.7'.
pH	• If between '65' and '80', divide by '10'. • If between '650' and '800', divide by '100'.
Temperature	Set to 'NA' if < 20 °C.
MechVent	• Binary indicator: '1' if mechanical ventilation ('1') in any timestep. • '0' otherwise (no mechanical ventilation recorded).

Table 4: List of Normal Features

Height	Na	NIMAP	RespRate
Weight	NIDiasABP	NISysABP	Temp
Age	DiasABP	SysABP	PaCO2
Albumin	Cholesterol	MAP	PaO2
HCO3	Platelets	HR	HCT
Mg			

Table 5: Aggregation functions for clinical variables

Variable	Aggregation	Category
Albumin, ALP, ALT, AST, Bilirubin, BUN	Last measured	Slow-changing labs
Creatinine, HCO3, HCT, K, Mg, Na	Last measured	Metabolic markers
PaCO2, pH, Platelets, WBC, Weight	Last measured	Status indicators
MechVent	Last measured	Intervention status
DiasABP, FiO2, HR, MAP, NIDiasABP	Mean	Hemodynamic
NIMAP, NISysABP, PaO2, RespRate	Mean	Vital signs
SaO2, SysABP, Glucose	Mean	Fluctuating measures
Lactate, TropI, TropT, Temp	Max	Critical peaks
GCS	Min	Worst neurological state
Urine	Sum	Cumulative output

Key: Slow-changing labs use last measurement; hemodynamic/vital signs use mean; critical conditions use max/min; urine output uses sum.

Q4.1 Example Patient Summary: Patient is a 54-year-old female, weights 87.79 kg. Over the first 48 hours:
- Serum bicarbonate in mmol/L ranged from 9.00 to 30.00, shows a trend of ↑. avg: 28.00, last: 28.00
- Hematocrit in % ranged from 25.65 to 43.30, shows a trend of ↓. avg: 30.30, last: 30.30
- Heart rate in bpm ranged from 58.00 to 101.00, shows a trend of ↑. avg: 86.00, last: 86.00
- Serum magnesium in mmol/L ranged from 1.20 to 2.20, shows a trend of ~. avg: 1.90, last: 1.90
- Serum sodium in mEq/L ranged from 131.00 to 141.00, shows a trend of ↓. avg: 136.00, last: 136.00
- Temperature in °C ranged from 35.30 to 38.20, shows a trend of ↑. avg: 37.80, last: 37.80
- Blood urea nitrogen in mg/dL ranged from 8.00 to 65.00, shows a trend of ↓. avg: 8.00, last: 8.00
- Fractional inspired O_2 (0-1) ranged from 0.40 to 1.00, shows a trend of →. avg: 0.60, last: 0.60
- Glasgow Coma Score (3-15) ranged from 14.00 to 15.00, shows a trend of ↓. avg: 15.00, last: 15.00
- Serum potassium in mEq/L ranged from 2.40 to 4.60, shows a trend of ~. avg: 4.00, last: 4.00

Table 6: Ollama Model comparisons - 100 samples, Zero-Shot, Binary Score

Model Variant	AuROC	AuPRC
llama3.2:3b	0.483	0.117
deepseek-r1:7b	0.574	0.252
gemma2:9b	0.641	0.282

Table 7: Binary Score vs. 0-10 Score - 100 samples, Zero-Shot, gemma2:9b

Model Variant	AuROC	AuPRC
Binary Score	0.641	0.282
0-10 Score	0.678	0.437