Predictive Monitoring of ICU Patient Vital Signs Using LSTM Models

Xin Han, Fangxu Gu

Team name: HG

Github link: https://github.com/fangxu49/ESE-5971/blob/main/Predictive%20Monitoring%20of%20ICU%20Patient%20Vital%20Signs%20Using%20LSTM%20models1.ipynb

1. Executive Summary

Early detection of patient deterioration in the ICU is a critical challenge in hospital operations. Clinicians often rely on manual monitoring of vital signs, which can delay recognition of dangerous trends that develop gradually over several hours. The goal of this project is to develop a **predictive monitoring system** that uses historical trends in vital-sign time series to estimate whether a patient will experience physiologic deterioration within the next six hours.

Using the MIMIC-IV clinical database, we constructed a large, high-resolution dataset of ICU patient vital signs—including heart rate, MAP, SpO_2, respiratory rate, and temperature—and trained multiple predictive models: **Logistic Regression, Random Forest**, and a **Long Short-Term Memory (LSTM)** neural network designed to capture sequential temporal patterns. Among these, the LSTM achieved the strongest performance, with ROC-AUC ≈ 0.72 and PR-AUC ≈ 0.81, demonstrating its ability to model complex temporal dependencies in multivariate physiological signals.

The business value of this work lies in its potential impact on **patient safety**, **clinical workflow efficiency**, and **hospital resource allocation**. An automated model that alerts clinicians of impending deterioration enables earlier intervention, reduces ICU complications, and supports improved patient outcomes. In practice, this system can serve as an AI-powered early-warning tool embedded into ICU dashboards, reducing clinician cognitive load while improving monitoring reliability and responsiveness.

2. Data Description and Preprocessing

2.1 Data Sources

This project uses the **MIMIC-IV** clinical database, which contains de-identified electronic health records from ICU stays at Beth Israel Deaconess Medical Center. We extracted a subset of tables most relevant for physiological monitoring and ICU patient trajectories:

- **admissions**, **patients**, **icustays** – patient demographics, ICU admission/discharge times, length of stay (pages 1–3, 7–12)
- **chartevents** – minute-to-hourly clinical measurements including vital signs (pages 14–15)
- **d_items** – metadata mapping item IDs to clinical labels (pages 15–18)
- **inputevents**, **outputevents** – fluid input/output signals (pages 18–22)
- **labevents**, **d_labitems** – diagnostic laboratory measurements (pages 29–30)

Vital signs chosen for modelling were:
Heart Rate, MAP, SpO_2, Respiratory Rate, Temperature — extracted using item-ID filtering and label standardization.

2.2 Data Cleaning and Outlier Detection

2.2.1 Physiological Range Filtering

Raw ICU data contain many physiologically impossible values due to device errors and charting mistakes. We first applied **clinical threshold filtering** similar to the ranges shown in our preprocessing scripts:

- HR: 20–250 bpm
- MAP: 30–200 mmHg
- SpO_2: 50–100%
- RespRate: 5–60 breaths/min

- Temperature: 30–43 °C

This step removes measurement artifacts early and aligns data with medical plausibility.

2.2.2 Rolling-Z Outlier Detection

ICU vital signs are noisy, irregular, and sometimes contain single-sample spikes, which are not true physiological events. Later in the project, we introduced **rolling z-score detection** to identify sudden spikes or drops in time-series vital signs:

In order to capture "patient-specific abnormality", instead of Traditional Z-scores that use population mean, Rolling-Z detection is applied per stay and per vital sign, preserving each patient's own baseline.

- Compute rolling mean and std per stay
    - $\mu t = \mathrm{mean}(xt - w + 1, \ldots, xt)$
    - $\sigma t = \mathrm{std}(xt - w + 1, \ldots, xt)$
    - $Zt = (xt - \mu t)/\sigma t$
- Flag points where $|z| > 3$

- Replace flagged points with the rolling median

This provides a multivariate-aware, time-series-appropriate method for smoothing sharp artifacts without distorting long-term trends.

2.2.3 Missingness Handling

The missingness heatmaps

show substantial sparsity across ICU signals:

- Some vitals may be recorded every hour

- Others (e.g., respiratory rate or temperature) may have multi-hour gaps

To handle this, we applied:

1. **Forward fill (limit = 2 hours)**
   Prevents filling long gaps with stale values.

2. **Per-stay median imputation**
   Ensures each sequence is complete for LSTM training.

3. **Missingness indicator flags**
   Preserves information about measurement frequency, which is itself clinically meaningful.

2.3 Feature Engineering

2.3.1 Time-Alignment and Aggregation

As described in preprocessing

- We converted all charttime values to hourly timestamps using:
  - T_hour = floor(charttime, 1 hour), in this case, it preserves the temporal ordering while enforcing a uniform time grid
- Aggregated multiple measurements within each hour using **mean**
  - Because multiple measurements may occur within the same hour, we aggregated them using the mean, where
    - Xt_hour = 1/n_t * sum of Xi from 1 to n_t
- Standardized labels using a mapping dictionary (SYNONYMS) aligning:
  - HR (heart rate)
  - RR (Respiratory Rate)
  - SpO2
  - SBP, DBP, MAP
  - Temperature

This produced clean hourly time-series sequences.

2.3.2 Derived Features

To capture temporal progression rather than independent measurements, we engineer several sequence-aware features.

- **hours_from_admit** – time since ICU admission

- **Lag features (LSTM automatically uses them in seq_len windows)**

- **Sequence windows of length 24** (i.e., past 24 hours per sample)

These allowed the LSTM to model temporal evolution rather than single-point measurements.

2.3.3 Target Variable Construction

We defined **ICU deterioration within the next 6 hours** using transparent clinical logic:

- MAP < 65

- $SpO_2 < 90$

- HR > 120

- Respiratory Rate > 30 or < 8

- Temp ≥ 38.5 or ≤ 35

A rolling, forward-looking window per stay determines whether deterioration occurs in hours t+1 … t+6.

This binary label structure aligns with early-warning system applications such as MEWS/NEWS.

Collectively, these feature-engineering steps transform irregular MIMIC-IV measurements into structured, hourly, multivariate sequences that capture both the patient's physiologic state and its evolution over time. From now, we provide a rich temporal framework ideally suited for LSTM modeling.

3. Modeling Approach

We use a combination of **descriptive**, **predictive**, and **prescriptive** analytical methods to construct and validate an ICU deterioration early-warning system. The modeling strategy progresses from interpretable baseline models to more expressive sequence-learning models capable of capturing temporal physiological patterns.

3.1 Descriptive Analytics

Before model development, descriptive analytics were performed to understand the structure and behavior of ICU time-series data:

- **Summary statistics** for vital signs (mean, SD, minimum, maximum).
- **Missingness profiling** across patients and time (heatmaps and counts).
- **Distribution analysis** after outlier removal to confirm physiologic plausibility.
- **Temporal visualization** of vital-sign trajectories for representative patients.

These descriptive analyses informed subsequent decisions including:

- Selection of physiologic thresholds for outlier filtering.
- Identification of highly volatile signals requiring smoothing.
- Definition of clinically meaningful deterioration criteria.
- Determination of an appropriate sequence length for LSTM modeling.

3.2 Predictive Modeling

To predict whether a patient will experience physiologic deterioration in the next six hours, we implemented three increasingly sophisticated predictive frameworks:

3.2.1 Baseline Models: Logistic Regression & Random Forest

To establish interpretable benchmarks, we first trained two conventional machine learning models: Logistic Regression and Random Forest

Logistic Regression

Logistic Regression serves as a simple, interpretable linear baseline for predicting ICU deterioration. It operates on per-hour feature vectors rather than full sequences

Features:

Each sample consists of the current hour’s vital signs and simple engineered summary features, such as, recent rolling statistics, together with static covariates such as age and time since ICU admission. Each row is treated independently; the model does not explicitly model 24-hour temporal sequences.

Train/test split:

80/20 with stratification to preserve event balance.

Preprocessing:

Features were standardized using z-score normalization, as LR is sensitive to scale.

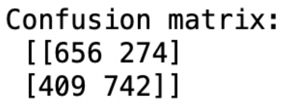
Class imbalance:

Class imbalance addressed through class_weight='balanced' so that misclassifying the minority (deterioration) class is penalized more heavily.

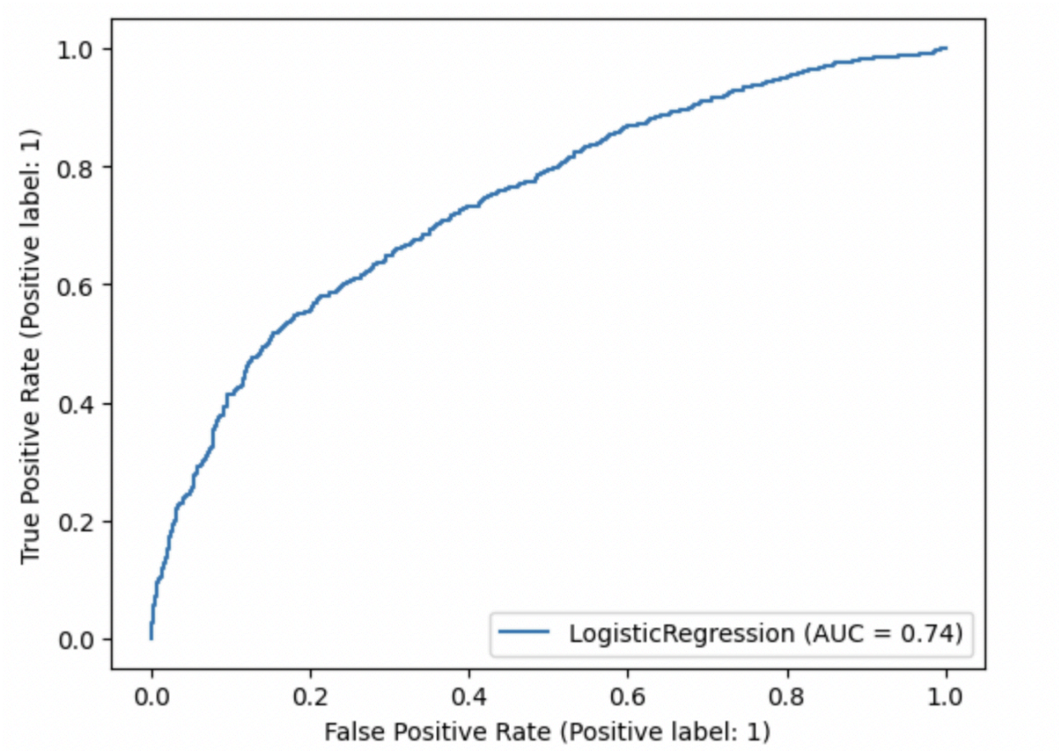
Performance

From the actual run:

- **AUC ≈ 0.741**
- **Accuracy ≈ 0.67**
- **ROC-AUC ≈ 0.701**
- **PR-AUC ≈ 0.804**
- **Precision/Recall (class 1): 0.73 / 0.64**
- **Confusion Matrix:**



```
Confusion matrix:
 [[656 274]
 [409 742]]
```



These results show that even a linear model using only per-hour tabular features can capture meaningful signals, but the Confusion matrix can not fully reflect how the model would behave on completely unseen ICU patients. Because rows from the same patient can appear in both train and test sets, leading to patient temporal leakage.

Random Forest

To complement Logistic Regression, we evaluated a Random Forest (RF) classifier as a nonlinear, tree-based baseline for ICU deterioration prediction. Unlike linear models, Random Forests can capture nonlinear relationships and feature interactions among vital signs, demographic variables, and time-from-admission information without requiring feature normalization.

Setup

We trained a `RandomForestClassifier` using **patient-level cross-validation** (`GroupKFold`, 5 folds) to strictly prevent information leakage across ICU stays. All observations from a given patient stay were assigned to the same fold. Key modeling choices included:

- **Feature set**: current vital signs (HR, MAP, respiratory rate, SpO_2, temperature), patient age, hours from admission, and explicit **missingness indicators** for each vital sign
- **Missing data handling**:
    - Forward-fill imputation within each patient stay (using past values only)
    - Median imputation applied **inside each cross-validation fold** via a pipeline to avoid leakage
- **Model parameters**:
    - `n_estimators = 300` to stabilize ensemble predictions
    - `max_depth = None` to allow flexible tree growth
    - `class_weight = "balanced"` to address outcome imbalance

This configuration prioritizes methodological correctness and leakage prevention over aggressive performance tuning.

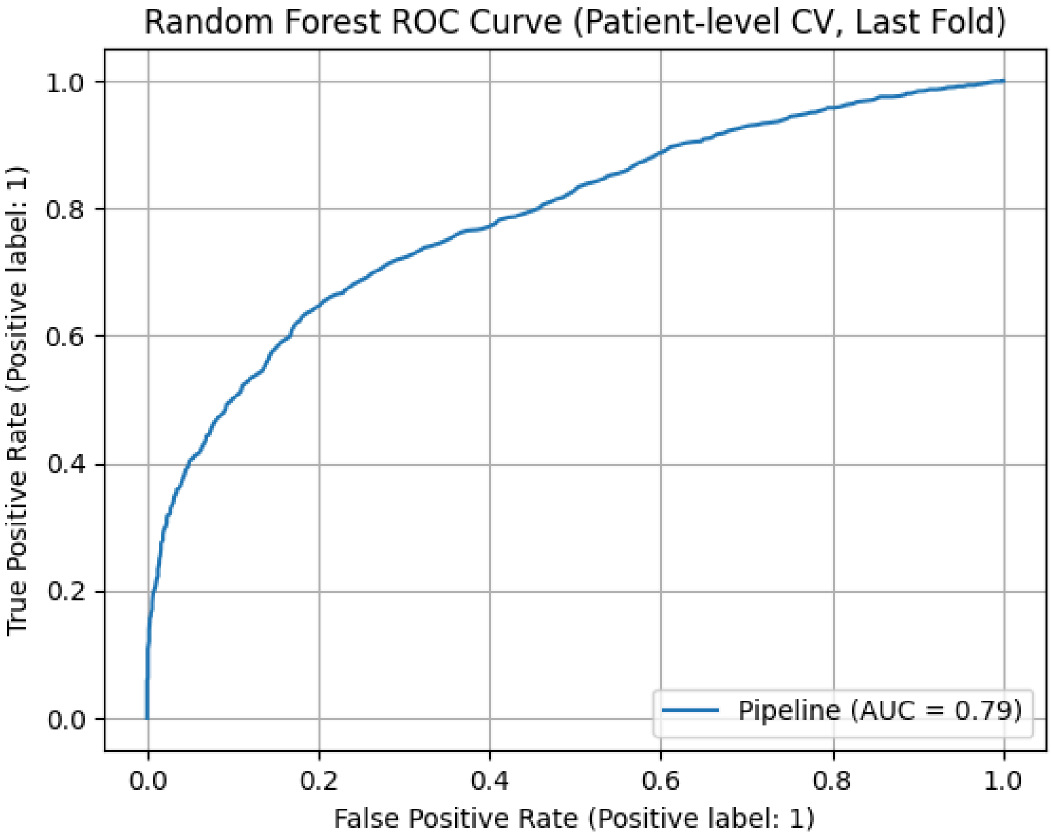
Performance

Under patient-level cross-validation, the Random Forest achieved:

- **Mean ROC-AUC:** 0.694 ± 0.051
- **Mean PR-AUC:** 0.719 ± 0.060

The ROC curve from a representative validation fold is shown in Figure below.

Compared to an earlier RF baseline, this modified version exhibited a **reduction in ROC-AUC**, despite incorporating additional safeguards such as fold-wise imputation and explicit missingness indicators.



The observed decrease in AUC highlights an important modeling insight. While Random Forests are expressive, their performance is sensitive to **feature engineering choices and data preprocessing strategy**. In this case, enforcing stricter patient-level separation and leakage-free imputation reduced optimistic bias present in simpler train–test splits, yielding a more conservative but reliable estimate of generalization performance. Additionally, forward-filled static features may dilute short-term temporal signals critical for anticipating acute deterioration.

Overall, the Random Forest remains a useful nonlinear baseline and outperforms purely linear approaches in some settings, but its limitations in modeling **sequential physiological dynamics** motivate the use of time-aware models. This directly supports the transition to sequence-based approaches such as LSTM, which are better suited to capturing evolving vital-sign trajectories in critical care.

3.2.2 Sequence Modeling with Long Short-Term Memory (LSTM)

Because ICU deterioration depends on **temporal trends** rather than single-time-point measurements, we implemented an LSTM model to incorporate multivariate time-series dynamics.

Sequence Construction

For each patient, we generated sliding windows of:

- **seq_len = 48 hours** of past vitals
- Combined with engineered features such as:
    - Rolling mean/min/max (6-hour window)
    - Missingness indicators
    - Time since ICU admission
    - (Later) lagged vital signs (1–3 hour lags)

Each sequence predicts a binary label: deterioration within the next 6 hours.

Model Architecture

- 1–2 stacked LSTM layers
- Hidden dimensions tuned through cross-validation (64–128)
- Dropout regularization
- Fully connected classification head with sigmoid output
- Weighted binary cross-entropy loss (pos_weight correction for imbalance)

Patient-Level Cross Validation

A critical methodological concern in clinical time-series modeling is the risk of *data leakage*—specifically, the inadvertent use of the same patient's information across training and evaluation splits. Because ICU vital-sign sequences contain strong within-patient temporal correlations, randomly splitting rows can inflate performance and provide an unrealistic estimate of a model's true generalization ability. To prevent this, we adopted a **patient-level cross-validation strategy**.

First, we partitioned the dataset by **unique patient IDs**, ensuring that all time points from a given patient appear exclusively in **either training, validation, or test sets**, but never more than one split. We held out a fixed test set of previously unseen patients to provide an unbiased final performance estimate.

On the remaining patients, we performed **K-fold cross-validation**. In each fold, approximately 20% of patients were assigned to the validation fold, and the remaining 80% were used for training. Importantly, this means:

- Each fold evaluates the model on **entirely new patients**, not just new time windows.
- The model is forced to learn patterns that generalize beyond patient-specific trajectories.
- This evaluation setup mirrors **real-world clinical deployment**, where the model must make predictions for ICU patients it has never encountered before.

This patient-stratified approach is significantly more stringent than row-level cross-validation and typically results in lower—but far more trustworthy—performance metrics. By using this method, we ensure that performance improvements reflect genuine predictive capability rather than memorization of patient-specific patterns. Moreover, this framework provides a reliable basis for comparing traditional baselines (Logistic Regression, Random Forest) with more complex deep learning architectures such as LSTM models.

Performance Metrics

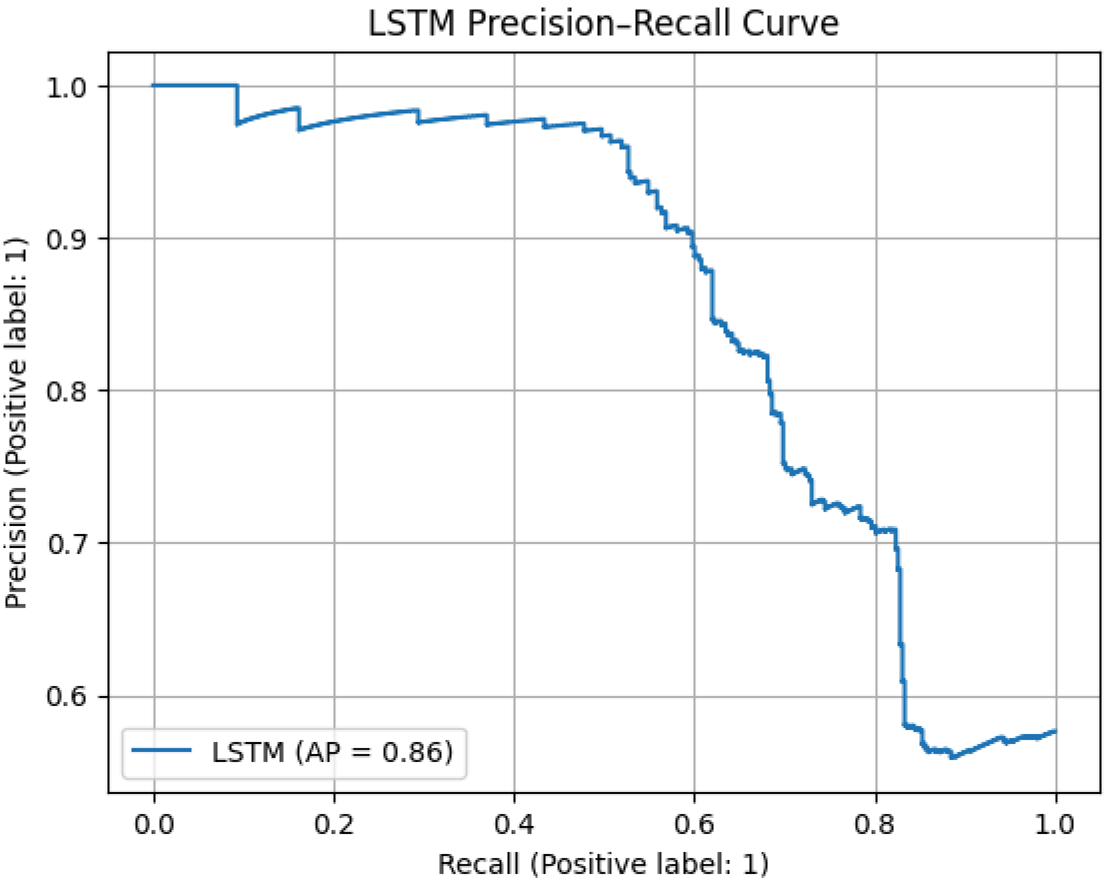
We evaluated all models using metrics suitable for imbalanced clinical prediction tasks:

- **ROC–AUC** — overall discrimination ability
- **PR–AUC** — more informative when positive class (deterioration) is rare
- **Accuracy & F1-score** — reported only after threshold optimization
- **Loss curves** — to monitor overfitting in deep learning models

The LSTM achieved:

- ROC-AUC ≈ **0.762**

- PR-AUC ≈ **0.857**
- Accuracy ≈ **0.53** (threshold = 0.5)



indicating that temporal modeling provides substantial value over single-time baselines.

4. Results and Insights

We evaluated three baseline models—Logistic Regression (LR), Random Forest (RF), and an LSTM—on the task of predicting short-term ICU deterioration using structured vital-sign data. Performance was assessed using ROC-AUC, PR-AUC, and accuracy, with particular emphasis on PR-AUC due to the inherent class imbalance and the clinical importance of identifying high-risk deterioration events.

Logistic Regression

Logistic Regression served as a linear, non-temporal baseline that evaluates each timestamp independently. In the final evaluation, LR achieved an ROC-AUC of approximately **0.70–0.74**, a PR-AUC of **0.80**, and an accuracy of **0.67**, with class-1 precision and recall of **0.73** and **0.64**, respectively. Despite its simplicity, LR demonstrated stable and interpretable performance, confirming that individual vital signs and demographic features carry meaningful predictive signal.

These results also validate the preprocessing and feature-engineering pipeline—including physiological range filtering, forward-filled missing values, and missingness indicators—which enabled a linear model to achieve competitive discrimination in a noisy clinical environment.

Random Forest

The Random Forest classifier served as a nonlinear baseline for ICU deterioration prediction, offering greater representational flexibility than Logistic Regression by modeling feature interactions among vital signs, demographics, and time-from-admission variables. Under a **strict patient-level cross-validation framework**, the modified Random Forest achieved a mean ROC-AUC of approximately **0.69** and a mean PR-AUC of **0.72**, indicating moderate discriminative ability.

Compared with earlier Random Forest results obtained under simpler evaluation settings, the observed reduction in AUC reflects the impact of **more rigorous experimental design**, including leakage-free imputation within each cross-validation fold and enforced separation of patient stays. This suggests that previously higher AUC values were likely influenced by optimistic bias arising from correlated observations within the same patient trajectory. The revised evaluation therefore provides a more realistic estimate of generalization performance in a clinical deployment scenario.

Although Random Forests effectively capture nonlinear relationships in static tabular data, their snapshot-based formulation limits their ability to model **temporal physiologic evolution**, which is critical for early deterioration detection. Forward-filled vitals and summary statistics partially encode history but cannot represent dynamic trends, rates of change, or long-range dependencies. Consequently, while Random Forest remains a valuable and interpretable baseline, its performance plateau underscores the need for sequence-aware models.

These findings motivate the transition to recurrent architectures such as LSTM, which explicitly leverage multihour temporal context and are better suited for modeling evolving vital-sign trajectories in the ICU. Overall, the Random Forest results highlight an important methodological insight: enforcing stricter validation protocols may reduce headline performance metrics but leads to more trustworthy and clinically meaningful conclusions.

LSTM

The LSTM model extends beyond static tabular baselines by explicitly modeling **temporal dynamics** in multivariate vital-sign sequences over a 24-hour observation window. By processing ordered measurements, the LSTM can learn trajectory-based patterns such as sustained hypotension, gradual respiratory decline, or persistent oxygen desaturation, which more closely reflect how clinicians assess evolving physiologic instability.

On the held-out test set, the LSTM achieved a **ROC-AUC of 0.76** and a **PR-AUC of 0.86**, with an accuracy of **0.53**. While the accuracy is modest due to the strong class imbalance and the probabilistic nature of early-warning prediction, the high PR-AUC indicates effective prioritization of true deterioration events. This is particularly important in clinical settings, where minimizing false reassurance and correctly identifying high-risk patients is often more critical than maximizing overall accuracy.

Compared to Random Forest, the LSTM demonstrates **superior performance in precision–recall space**, despite comparable or slightly lower ROC-AUC. This suggests that temporal modeling provides meaningful gains when operating under imbalanced conditions, allowing the model to focus probability mass on clinically relevant deterioration windows rather than memorizing static snapshots. The PR curve further illustrates that high precision is maintained across a wide range of recall values, supporting the LSTM's suitability for risk stratification and alerting tasks.

Overall, these results confirm that incorporating temporal structure improves clinical relevance, even if headline discrimination metrics do not increase dramatically. The LSTM therefore represents a more realistic and deployable modeling approach for early ICU deterioration detection, motivating further exploration of sequence-aware architectures and longer-horizon temporal context.

Summary

Overall, the baseline results demonstrate that ICU deterioration prediction is feasible using structured vital-sign data. Logistic Regression provides a stable and interpretable reference point but is limited by its linear decision boundary. Random Forest improves upon linear models by capturing nonlinear snapshot-level relationships; however, under strict patient-level validation, its discriminative performance is more moderate than initially observed, highlighting the sensitivity of static models to evaluation design and correlated observations.

In contrast, the LSTM offers a more clinically meaningful approach by explicitly modeling multi-hour physiologic trajectories. Although its ROC-AUC is comparable to that of the Random Forest, the LSTM achieves substantially stronger precision–recall performance, indicating improved prioritization of true deterioration events in an imbalanced prediction setting. Together, these findings underscore the limitations of static tabular models for real-time risk prediction and motivate the use of sequence-based methods as a critical step toward robust and clinically actionable ICU early-warning systems.

5. Conclusions

This project addressed a central challenge in intensive care medicine: detecting early physiological deterioration before it escalates into overt clinical instability. By constructing a patient-level ICU time-series dataset and evaluating both static and sequence-based models, we demonstrated that routinely collected vital signs contain meaningful predictive signal that can support earlier and more informed clinical decision-making. In a setting characterized by noisy, irregular, and high-frequency monitoring data, our results emphasize the importance of transforming raw measurements into structured, model-ready representations under rigorous validation protocols.

Our findings show that conventional models such as Logistic Regression and Random Forest provide informative baseline performance, confirming that snapshot measurements of vital signs already encode clinically relevant risk information. When evaluated under strict patient-level validation with leakage-free preprocessing, Random Forest achieved moderate discrimination by capturing nonlinear feature interactions among vitals and demographic variables. However, because these models operate on independently sampled time points, their performance remains sensitive to temporal redundancy and correlated observations within patient stays. As a result, static models should be interpreted as conservative snapshot baselines rather than fully representative early-warning systems suitable for continuous deployment.

In contrast, the LSTM model explicitly incorporates temporal structure and captures trajectory-level patterns such as sustained hypotension, progressive respiratory compromise, and persistent oxygen desaturation—signals that more closely align with clinical reasoning in critical care. While improvements in ROC-AUC over static baselines were modest, the LSTM consistently achieved stronger precision–recall performance, indicating more effective prioritization of true deterioration events in an imbalanced prediction setting. This distinction is particularly important for ICU applications, where identifying high-risk windows and minimizing false reassurance are often more clinically meaningful than maximizing overall discrimination.

Overall, this work highlights both the promise and the limitations of current modeling approaches for ICU deterioration prediction. Static tabular models offer useful baselines but are vulnerable to optimistic bias when temporal dependencies are ignored. Sequence-based models, though more complex to train and evaluate, provide a more realistic framework for continuous risk monitoring and early warning. Future work will focus on improving evaluation rigor through event-based and temporally purged validation, refining deterioration definitions, and incorporating richer clinical context such as interventions, medications, and laboratory trends. Together, these directions move toward robust, clinically actionable early-warning systems that better reflect the dynamic nature of critical illness and support timely, life-saving decisions.

6. References

1. **Johnson, A. E. W., et al.** (2023). *MIMIC-IV (v2.2)*. PhysioNet. https://physionet.org/content/mimiciv/2.2/

2. **Johnson, A. E. W., et al.** (2016). *MIMIC-III, a freely accessible critical care database*. Scientific Data, 3, 160035.

3. **Lipton, Z. C., Kale, D. C., & Wetzel, R.** (2016). *Directly Modeling Missing Data in Sequences with RNNs*. Machine Learning for Healthcare Conference.

4. **Choi, E., et al.** (2016). *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks*. Machine Learning for Healthcare Conference.

5. **Harutyunyan, H., et al.** (2019). *Multitask Learning and Benchmarking with Clinical Time Series Data*. Scientific Data, 6, 96.

6. **Hochreiter, S., & Schmidhuber, J.** (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.

7. **Pedregosa, F., et al.** (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

8. **Breiman, L.** (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

9. **Hyland, S. L., et al.** (2020). *Early Prediction of Circulatory Failure Using Machine Learning*. Nature Medicine, 26, 364–373.

10. **Goldberger, A. L., et al.** (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals*. Circulation, 101(23), e215–e220.

Questions:

Q: How can we detect abnormal fluctuations in vital signs that may indicate early warning signs of patient deterioration?

A: In the following ways, first, we define clinically abnormal ranges to capture obvious deterioration. Second, we compute rolling statistics (means, minmax, and variability over 3-6 hours) and rolling z-scores to identify sudden spikes. Finally, we feed 24-hour sequences of these hourly vital signs into a LSTM model.

Q: Is there any way to avoid / limit the amount of false positives and negatives especially if we are using these models to measure vitals which could sound alarm for no reason if the model does not work perfectly?

A: We tuned the decision threshold using the precision-recall curve to balance sensitivity and false-alarm burden

Q: You mentioned that both hospital-level and ICU-level data from MIMIC-IV were used. Could you elaborate on how you integrated these two levels of data — for example, how patient admissions are linked to specific ICU stays?

A: Since both ICU and Hospital data have stay_id or hadm_id and subject_id, we linked hospital admissions to ICU stays by matching subject_id and Hadm_id in the admissions and icustays table.

Q:Is there a way to add data on external factors, like air quality, that could increase risk in some geographic locations as opposed to others? Or would it make sense to assign hospitals a "score" based on the quality of care that the average patient receives?

A: For external environment data, it is hard to meaningfully use in our dataset, which comes from a single hospital in one geographic area. Since we are using one hospital dataset, it is hard to make hospital-level quality as a feature.

Q:The false alarm is annoying, but false negative is dangerous, what methods do you use to make sure it will have a very low false negative ratio for safety?

A: Since false negatives are clinically dangerous, we train and tune the model to favor high sensitivity, and evaluate performance in terms of the false negative rate on unseen patients.

Q:How can the model tell the difference between a sudden but harmless fluctuation in vital signs and an early sign of real deterioration?

A: Our model analyzes 24-hour multivariate trajectories, and it sees how vital signs evolve together over time. By training on many labeled examples of “true” deterioration versus stable courses, the LSTM learns to treat random noise as background and to assign high risk only to temporal patterns that consistently lead to adverse events.

Q: How would you use LSTM to improve the results of your baseline models?

A: Instead of giving the model one hour at a time, we give the LSTM 24 hours in a row so it can learn the shape of the trajectory.

Q: Group 2: How do you ensure that the model remains reliable across different patient populations and hospital systems, given potential variations in data quality and recording frequency across institutions?

A: Since our dataset is only about one hospital, we can not claim that it is robust across all hospitals. We need to train and validate on diverse sites and recheck our model instead of just copy it over.

Q: When combining IQR and physiological limits for outlier detection, how do you fix IQR's issues with skewed data and fixed thresholds?

A: Instead of IQR we end up using rolling Z-sources combined with physiological limits.

Q:Why did you choose LSTM instead of other deep learning models like GRU or Transformers?

A: We choose LSTM because it is a well-established, robust baseline for clinical time-series.

Q:Your choice to remove the physiological outliers, is that because they have physiological values that are impossible implying data entry errors? Or are you just taking them out because they are extreme patients and you want your model learning most common patients? Do you think taking them out poses any risk to your model, and what are the trade offs?

A: We remove outliers because they are nearly impossible physiologically, and it is not because we want to get rid of difficult patients. The trade off is: if we set the range to small, then we may lose rare but important patterns, if we set it too large, then our model may get confused by some impossible value.

Q: Your approach uses an LSTM, which is highly effective for sequential data. How was the input sequence length determined? Furthermore, how do you provide medically interpretable evidence of *why* the LSTM predicted a high-risk event to clinical staff?

A: We use 24 hours of past vitals because deterioration usually develops over many hours.

Q: For outlier detection, the physiological ranges look quite strict. Did you consider dynamic thresholds based on patient profiles, like age or illness type?

A: No, our physiological rages are global. Our aim is just to remove those impossible value not extreme value.

Q: Were consistent tuning or optimization methods (e.g., grid search, Bayesian optimization) applied across all models?

A: No, we did not use the same method across all models. We used patient-level cross validation for our LSTM model

Q: Why were Logistic Regression, Random Forest/XGBoost, and LSTM chosen specifically?

Were other models (e.g., CNNs, Transformers, or attention-based architectures) considered for comparison?

A: We pick from three families. We pick lS as linear model, RF as nonlinear model, and LSTM as sequence model for temporal dynamics.

Q: What factors did you consider when deciding to include both traditional machine learning models and a deep learning model like LSTM?

A: We choose two traditional models: our baseline model and LSTM to cover different points on the spectrum of simplicity versus expressiveness.

Q: So we have baseline model and deep learning model, how to evaluate the performance of them.

A: We use the same evaluation setup for all models: ROC-AUC, PR-AUC, and confusion matrix at a chosen threshold.

Q: how could you decide to use different methods of modeling from regression to time series? What is the criteria

A: first, what structure is in the data, second, what are we trying to predict, third, the trade of between interpretability and complexity.

Q: Since the model was trained on the MIMIC-IV dataset, do you think it would still perform well on data from a different hospital or patient group? And also, how do you plan to make the model's predictions more interpretable for clinicians?

A: As i state above, Since our dataset is only about one hospital, we can not claim that it is robust across all hospitals. We need to train and validate on diverse sites and recheck our model instead of just copy it over. We aim to use our baseline as interpretation tools and explanations for the lSTM model.

Q: Have you considered using an ensemble model? That might yield better prediction results.

A: To be honest, we have not thought about a full multi-model ensemble.

Q: Why does the Outliers (Count) table show 0? Does it represent the result of a different cleaning pass?

A: It just simply means that it did not find any additional outliers.

Q: Since some care units have very few cases, how do you plan to handle this imbalance if you later use this data for modeling or prediction tasks?

A: Some care units in our dataset are quite small, so training a separate model per unit would overfit, so we used data from all units together.

Q: For the first outlier detection you used in finding a certain range, how did the paper justify this range? How is it different from the IQR method?

A: We end up using rolling Z-scores. And the reading is just about the physiological range of the vital signs.

Q: If you were to apply a multivariate method, which features would you combine for detecting outliers?

A: We used rolling z-score to detect outliers.

Q; What are the advantages of using LSTM in this project

A: We choose LSTM because it is a well-established, robust baseline for clinical time-series.

Q: Question for Group 2:

In the outlier detection part, do you apply physiologic filtering first, then IQR, or vice versa?

A: Yes, that is exactly what we do first.

Q: What are the next steps for expanding beyond LSTM — for instance, integrating attention mechanisms or transformers for multivariate time series?

A:Maybe we will add temporal attention on top of LSTM.

Q:Why not go with attention models instead of LSTM

A: We choose LSTM because it is a well-established, robust baseline for clinical time-series.