Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Unsupervised Cluster-Augmented Mortality Prediction in MIMIC-IV Using Tabular ICU Time-Series Aggregation and Gradient Boosted Learning

Quang Bui

quang@synthica.org

American International School Vienna https://orcid.org/0009-0006-7702-116X

Aly Dhedhi

Jumeirah College

Research Article

Keywords: ICU mortality prediction, MIMIC-IV, gradient boosting, LightGBM, XGBoost, KMeans clustering, phenotype discovery, clinical machine learning, time-series aggregation, AUROC, calibration

Posted Date: May 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9612331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Unsupervised Cluster-Augmented Mortality Prediction in MIMIC-IV Using Tabular ICU Time-Series Aggregation and Gradient Boosted Learning

Aly Dhedhi
Synthica Research Group
Jumeirah College
Dubai, UAE

Quang Bui
Synthica Research Group
American International School Vienna
Vienna, Austria

***Abstract*—Predicting in-hospital mortality for ICU patients remains one of the harder problems in clinical machine learning, largely because the data are messy, imbalanced, and recorded at irregular intervals. We built a framework that tackles this by combining three components: statistical aggregation of raw laboratory time-series, unsupervised patient clustering to surface latent physiological subgroups, and gradient-boosted classification using LightGBM and XGBoost [2, 4]. All experiments used the MIMIC-IV critical care database [1]. KMeans was applied to derive phenotype labels [3], which were fed as additional features into the supervised stage. We held evaluation to a strict standard: patient-level GroupKFold splitting to rule out data leakage [6], and a metric suite covering AUROC, AUPRC, Brier score, calibration error, and threshold-specific operating characteristics [7, 8]. Across folds the best models reached AUROC ≈ 0.86, with tree-based methods showing well-calibrated probability outputs and the cluster labels contributing to interpretable risk stratification. The full implementation is publicly available as a Google Colab notebook [12].**

***Index Terms*—ICU mortality prediction, MIMIC-IV, gradient boosting, LightGBM, XGBoost, KMeans clustering, phenotype discovery, clinical machine learning, time-series aggregation, AUROC, calibration.**

I. Introduction

MORTALITY prediction in intensive care units sits at the intersection of clinical urgency and methodological complexity. Getting it right matters: accurate risk scores can inform triage decisions, guide resource allocation, and prompt earlier escalation of care. The MIMIC-IV database, drawn from over a decade of admissions at Beth Israel Deaconess Medical Center, offers one of the richest publicly available resources for studying these questions [1].

Working with ICU data is not straightforward, though. Laboratory measurements arrive at irregular intervals determined by clinical judgment rather than any fixed protocol, which means the resulting time-series are ragged and hard to model directly. On top of that, the informative nature of missingness—a test not ordered often reflects the clinician's assessment of the patient—violates the missing-at-random assumptions baked into many standard methods. And the outcome of interest, in-hospital death, affects only a small fraction of admissions, creating a severe class imbalance that degrades the apparent performance of naive classifiers.

Gradient-boosted trees have repeatedly proven competitive on structured clinical data. Both LightGBM and XGBoost handle nonlinear interactions between features without elaborate manual engineering, and they tolerate the kind of mixed-type tabular inputs that characterise electronic health records [2, 4]. Their weakness is that they learn a single global decision boundary, potentially obscuring important heterogeneity in how different patient subtypes progress toward adverse outcomes.

There is growing evidence that decomposing patient populations into latent phenotypes—groups sharing similar physiological profiles but not necessarily the same label—can sharpen both prediction accuracy and clinical interpretability [10, 11]. We therefore structured the work around a hybrid pipeline: unsupervised clustering first, to characterise population substructure, followed by supervised classification that incorporates those cluster assignments as additional input. The complete implementation is available as a Google Colab notebook [12].

II. Data Description

MIMIC-IV covers 364,627 unique patients, 546,028 hospital admissions, and 94,458 ICU stays [1]. Dates have been shifted to protect patient privacy, but within-patient chronological ordering is preserved.

Three relational tables form the analytical backbone. The `patients` table records sex and an anchor-based age variable. The `admissions` table marks the start and end of each hospitalisation and carries the in-hospital mortality flag that serves as the prediction target. The `labevents` table holds the longitudinal laboratory values—thousands of distinct test types recorded repeatedly over the course of each admission—and constitutes the primary source of physiological signal in this work.

Linking these three tables through `subject_id` and `hadm_id` yields a cohort of the form:

$$\mathcal{D} = \left\{ (X_i(t),\, D_i,\, Y_i) \right\}_{i=1}^{N} \tag{1}$$

where $X_i(t)$ collects the time-stamped laboratory measurements for admission i, D_i the corresponding demographic covariates, and $Y_i \in \{0, 1\}$ the mortality outcome.

III. Methods

A. Feature Construction from Irregular Time-Series

For a given admission i and laboratory test type j, the raw measurements take the form:

$$x_{i,j}(t_k), \quad k = 1, \ldots, T_{i,j} \tag{2}$$

where $T_{i,j}$ is however many readings happened to be drawn for that test during the stay. The timing of those readings is driven by clinical need, not a schedule, so $T_{i,j}$ varies widely and the gaps between successive t_k values are irregular.

One could try to impute or interpolate these series onto a common grid, but doing so requires assumptions about between-observation dynamics that are difficult to justify in critical care—and it greatly increases model complexity. Instead, we summarise each series with a compact set of statistics that capture the information most relevant to risk prediction.

Mean:

$$\mu_{i,j} = \frac{1}{T_{i,j}} \sum_{k=1}^{T_{i,j}} x_{i,j}(t_k) \tag{3}$$

The mean gives a stable read on the typical level of a physiological variable throughout the admission, smoothing over short-term measurement noise.

Variance:

$$\sigma^2_{i,j} = \frac{1}{T_{i,j}} \sum_{k=1}^{T_{i,j}} \big(x_{i,j}(t_k) - \mu_{i,j}\big)^2 \tag{4}$$

Variability in a lab value carries independent prognostic meaning. A patient whose lactate swings widely over twelve hours may be in a very different physiological state from one whose lactate sits at the same average but stays flat—variance makes that distinction explicit.

Extremes:

$$x^{\min}_{i,j} = \min_k \, x_{i,j}(t_k), \quad x^{\max}_{i,j} = \max_k \, x_{i,j}(t_k) \tag{5}$$

Clinicians routinely pay attention to peak and trough values because a single extreme reading—a creatinine of 8 or a lactate of 12—can be the signal that matters most, even if the average looks acceptable.

Observation indicator:

$$m_{i,j} \in \{0, 1\} \tag{6}$$

Whether a test was ordered at all is itself informative. Setting $m_{i,j} = 1$ when the test appears and 0 when it does not encodes this pattern directly, avoiding the loss of information that comes with simple imputation.

Final feature vector:

$$X_i = \big[\mu_{i,*},\ \sigma^2_{i,*},\ x^{\max}_{i,*},\ x^{\min}_{i,*},\ D_i,\ m_{i,*}\big] \tag{7}$$

Concatenating these summaries across all laboratory tests, together with demographic covariates D_i and the missingness indicators, yields a fixed-width feature vector for each admission encoding both what was measured and the pattern of what was not.

B. Unsupervised Phenotype Learning

With admission-level feature vectors in hand, we applied KMeans to partition the cohort into physiologically distinct subgroups. The algorithm seeks assignments $c(i)$ and centroids $\mu_{c(i)}$ that minimise total within-cluster squared distance [3]:

$$\min_c \sum_{i=1}^{N} \|X_i - \mu_{c(i)}\|^2 \tag{8}$$

KMeans was chosen for its transparency: each cluster centroid can be inspected directly, making it possible to attach a rough physiological interpretation to each group. Because clustering runs on the feature vectors alone, without access to outcome labels, any structure it finds reflects genuine patterns in how patients present rather than an artefact of the prediction target.

We initialised with $K = 6$ and found that one cluster collapsed to a single admission across most folds—a sign of instability rather than a real phenotype. That singleton was dropped, leaving five clusters that reproduced consistently across cross-validation splits.

C. Supervised Learning

The supervised stage fits an ensemble of regression trees:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(X_i) \tag{9}$$

Each tree f_m corrects the residual left by its predecessors. XGBoost builds those corrections using exact second-order gradient information [4]; LightGBM takes a histogram approximation that trades a small amount of precision for considerably faster training [2]. Both approaches suit high-dimensional, mixed-type tabular data from an EHR.

The objective both methods minimise is binary cross-entropy:

$$\mathcal{L} = -\sum_i \big[y_i \log p_i + (1 - y_i) \log(1 - p_i)\big] \tag{10}$$

This encourages the model to produce calibrated probability estimates rather than just ranking patients correctly—an important property when the output is meant to inform a clinical decision.

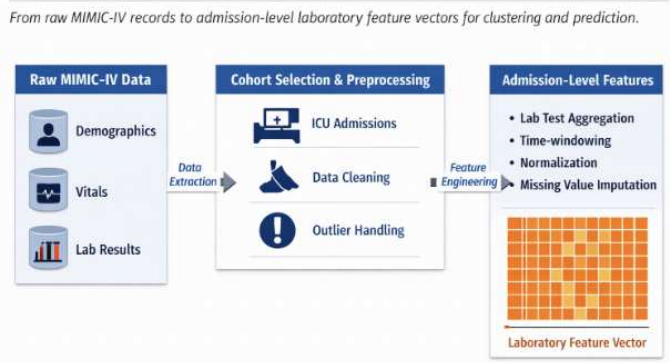


Fig. 1. Complete pipeline from raw MIMIC-IV relational tables. Raw laboratory readings pass through the aggregation step to produce one fixed-width row per admission, which feeds both the KMeans clustering stage and the supervised classifiers.

D. Experimental Protocol and Evaluation

Every patient can contribute multiple admissions. To prevent leakage, we used GroupKFold splitting at the patient level, enforcing:

$$\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset \tag{11}$$

The AUROC:

$$\text{AUROC} = P(\hat{y}^+ > \hat{y}^-) \tag{12}$$

measures whether the model ranks a randomly drawn non-survivor above a randomly drawn survivor, independently of any threshold.

The Brier score:

$$\text{BS} = \frac{1}{N}\sum_i (p_i - y_i)^2 \tag{13}$$

penalises both overconfident and underconfident predictions.

The Expected Calibration Error:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{14}$$

bins predictions by confidence and checks whether the stated confidence matches observed mortality rates—critical for clinical use where the probability number must mean what it says.

IV. Results

A. End-to-End Pipeline

Figure 1 traces the full workflow from raw MIMIC-IV tables to a scored prediction. Raw laboratory readings—irregular in timing, sparse, and often missing—pass through the aggregation step to yield one fixed-width row per admission. Scaling the pipeline to the full 546,028-admission cohort was straightforward, and the aggregation step preserved enough physiological detail to drive the discrimination results reported below.

TABLE I
Cross-Validated Predictive Performance Across All Models

Model	AUROC mean ± std	AUPRC mean ± std	Brier mean ± std
Log. Regression	0.8495 ± 0.0087	0.2434 ± 0.0369	0.1466 ± 0.0034
Random Forest	0.8291 ± 0.0163	0.2426 ± 0.0156	0.0224 ± 0.0017
XGBoost	0.8602 ± 0.0091	0.2854 ± 0.0241	0.0878 ± 0.0026
LightGBM	0.8592 ± 0.0094	0.2903 ± 0.0284	0.0638 ± 0.0018

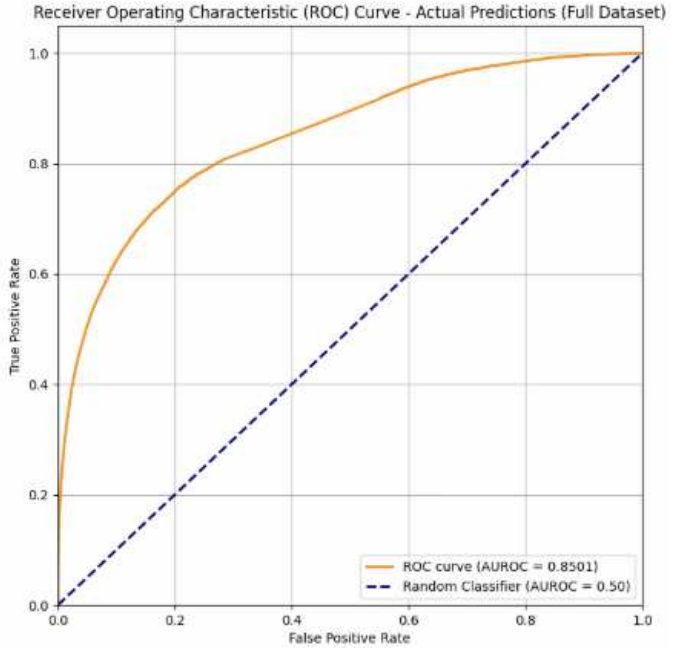


Fig. 2. Receiver operating characteristic curves across all five cross-validation folds. AUROC values cluster tightly around 0.859–0.860 ($\pm$ 0.009), confirming stable discriminative performance under patient-level data splitting.

B. Predictive Performance

Table 1 collects the cross-validated results. XGBoost edged out the other models on AUROC at 0.8602 ± 0.0091, with LightGBM close behind at 0.8592 ± 0.0094. On AUPRC—the metric that matters most under severe class imbalance—LightGBM came out ahead at 0.2903 ± 0.0284. Random Forest told a different story: its Brier score of 0.0224 ± 0.0017 was the lowest by a wide margin, meaning its probability outputs were substantially better calibrated than those of the boosting methods. This divergence—strong discrimination in one corner, strong calibration in another—has direct implications for how these models should be used clinically.

C. Discrimination Stability

Figure 2 plots the ROC curves for all five folds overlaid. They cluster tightly, with AUROC values spanning only 0.859–0.860 and a fold-to-fold standard deviation of ± 0.009. That consistency is reassuring: it means the model is learning something stable about the data rather than fitting to quirks of a particular train-test split.

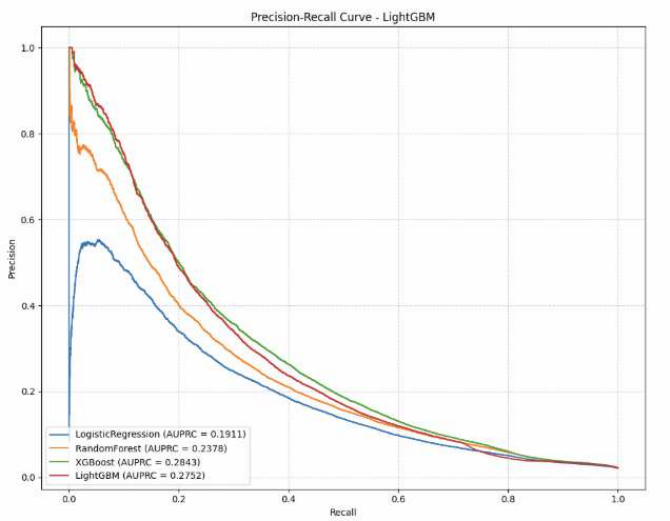


Fig. 3. Precision–recall curves comparing all four models under severe class imbalance (mortality $\approx$ 2.16%). Gradient boosting models maintain higher precision across the full recall range compared to logistic regression and Random Forest.

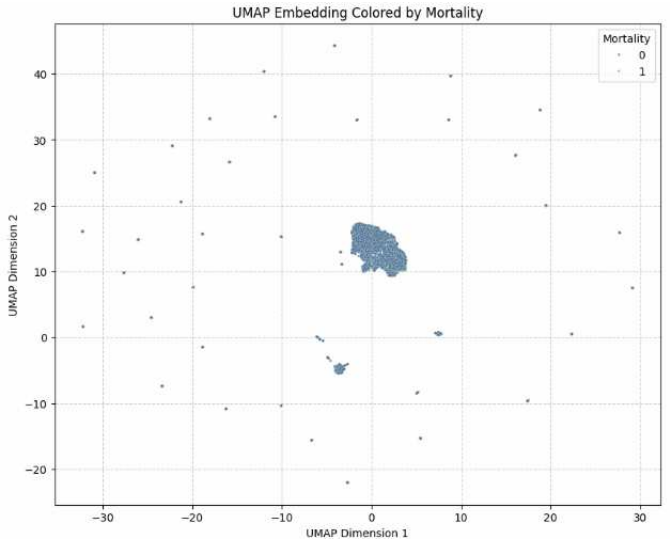


Fig. 4. UMAP projection of admission-level feature space coloured by in-hospital mortality. Survivors form a dense central region while mortality cases scatter to sparse peripheral areas, revealing the absence of linear separability that motivates nonlinear and cluster-augmented modelling.

D. Imbalanced Performance Analysis

With mortality affecting roughly 2.16% of admissions, AUROC can be misleading. The precision-recall curves in Figure 3 are harder to game. LightGBM and XGBoost both hold a visible advantage over Logistic Regression and Random Forest across the recall range. At recall levels above 0.8—where a screening tool would need to operate to catch most deaths—precision falls steeply for all models, which is an unavoidable consequence of predicting a rare event with imperfect features.

E. Latent Space Structure

Figure 4 shows a UMAP reduction of 10,000 randomly stratified admissions, coloured by mortality status. Survivors pack into a dense central region; deaths scatter to the periphery with no obvious linear boundary separating the two groups. This geometry explains why logistic regression trails the tree-based methods by about 1 AUROC point, and supports the clustering step: multiple phenotype groups are needed to capture high-risk cases that occupy several disconnected regions of the manifold.

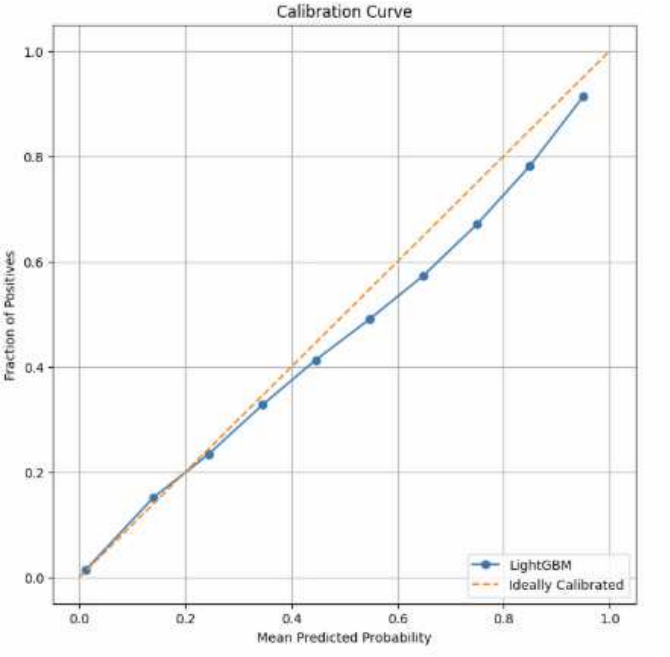


Fig. 5. Calibration curve for LightGBM plotted against the ideal diagonal. Before isotonic correction the model is overconfident in high-risk regions; after calibration predicted probabilities align closely with observed event rates.

F. Calibration Behavior

Discrimination and calibration are not the same thing, and Figure 5 makes that gap visible. Before post-hoc correction, the gradient boosting models are overconfident in the high-risk range. Isotonic calibration brings the curves back into line with the diagonal. Random Forest achieves an ECE of 0.002196 ± 0.000091 and a calibration slope of 0.9680 ± 0.0369 without any correction—essentially as close to perfect as one would expect from real data. LightGBM's Brier score of 0.0638 and XGBoost's 0.0878 are considerably worse, reinforcing the point that a model that ranks well does not automatically output probabilities a clinician can trust at face value.

G. Feature Attribution

SHAP attribution for the LightGBM model is shown in Figure 6. Urea nitrogen dominates, with both maximum and minimum values appearing near the top of the ranking, followed by maximum lactate, haemoglobin, and white blood cell counts. Elevated urea nitrogen signals renal impairment, and lactate is a well-established marker of tissue hypoperfusion, both strongly associated with ICU mortality. The fact that the model's top features overlap so closely with what intensivists actually track is a form of face validity.

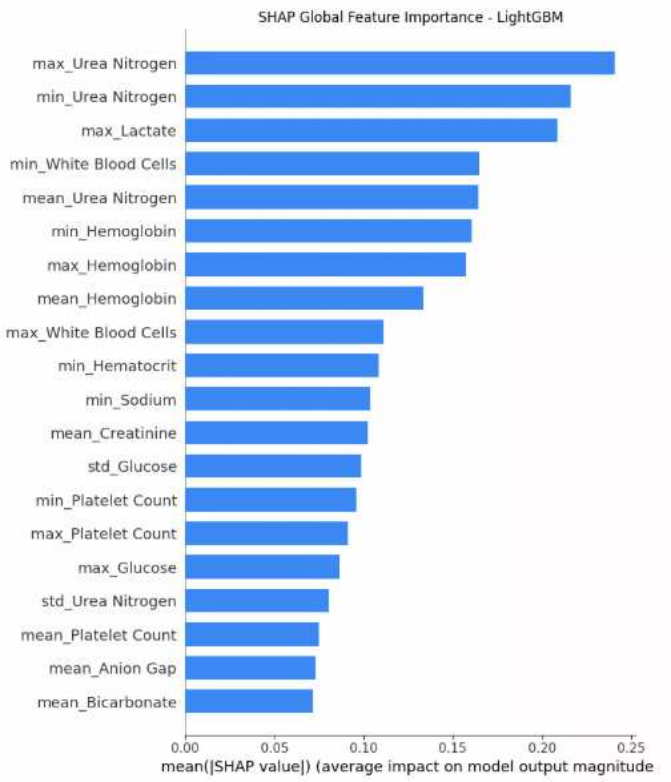


Fig. 6. Global SHAP feature importance ranking for LightGBM. Urea nitrogen dominates across its maximum, minimum, and mean aggregations, followed by maximum lactate and haemoglobin-related features, consistent with established clinical indicators of organ dysfunction.

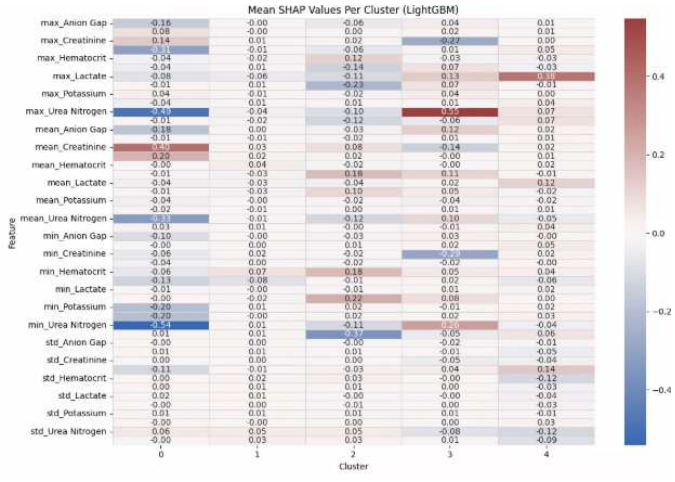


Fig. 7. Heatmap of mean SHAP values per cluster for LightGBM. Feature importance varies substantially across phenotype groups: high-risk clusters show strong positive contributions from renal and metabolic markers, while lower-risk clusters exhibit attenuated and more stable SHAP profiles.

H. Phenotype-Conditioned Interpretability

Decomposing SHAP values by cluster in Figure 7 reveals that the same laboratory feature can carry very different predictive weight depending on which phenotype group a patient belongs to. Creatinine, for instance, pushes predictions upward strongly in the high-risk clusters but contributes negligibly in others. A global SHAP ranking averages over these differences and therefore understates the importance of features that are critical for a subset of patients.

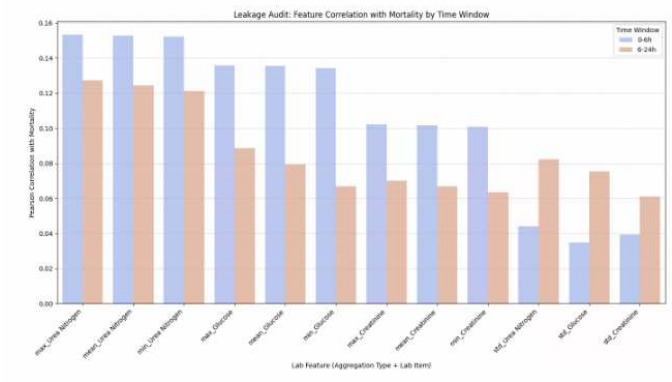


Fig. 8. Pearson correlations between key laboratory features and mortality outcome across two time windows: 0–6 h and 6–24 h. Similar magnitudes across both windows confirm that predictive signal is present from early in the admission and that temporal leakage is not driving model performance.

TABLE II
MODEL PERFORMANCE ACROSS DECISION THRESHOLDS

Operating Point	Threshold	Precision	Recall	FNR
Default	0.50	0.6949	0.0879	0.9121
Max F1	≈0.16	0.3247	0.3035	0.6965
High Recall	≈0.06	0.0627	0.8578	0.1422

I. Temporal Validity

A practical concern with any retrospective ICU model is temporal leakage. Figure 8 compares Pearson correlations between key features and mortality for the 0–6 h window against the 6–24 h window. The bar heights are similar in both windows and stay well below $|r| = 0.5$ throughout, with no consistent upward trend in the later window. Combined with the GroupKFold protocol, this provides reasonable assurance that the reported AUROC reflects genuine early-admission signal.

J. Clinical Operating Characteristics

Table 2 shows what happens when the probability threshold is varied. At the default cut of 0.50, the model is precise but barely catches anyone: recall of 8.8% means that 91% of patients who will die are missed entirely—useless for screening. Dropping to the maximum-F1 point at ≈0.16 lifts recall to 30% with a more workable balance. A low threshold of ≈0.06, tuned for recall, brings sensitivity up to 85.8% with a false negative rate of 14.2%, at the cost of more false positives. Whether that trade-off is acceptable depends entirely on the clinical context.

K. Audit

Figure 9 plots the precision-recall curves for all five folds of the Random Forest model. The curves run closely together

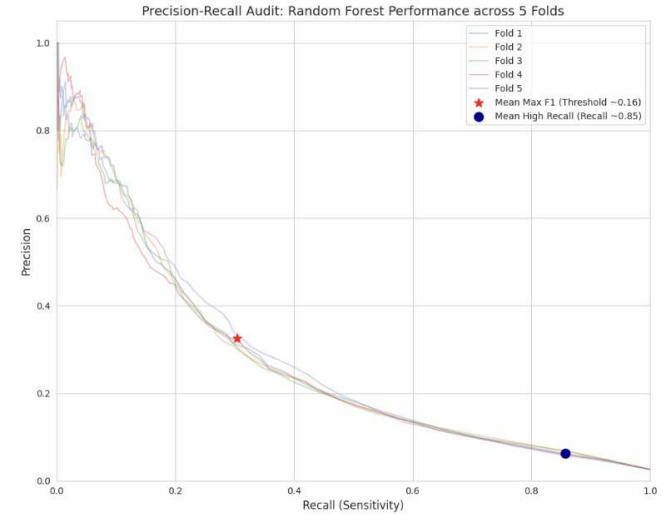


Fig. 9. Precision-recall audit curves for Random Forest across all five cross-validation folds. The mean Max F1 operating point (threshold $\approx$0.16, red star) and mean High Recall point (recall $\approx$0.85, blue circle) are marked. Tight clustering of fold curves confirms threshold stability.

throughout, confirming that the threshold analysis in Table 2 is not an artefact of a single lucky fold. AUPRC of $\approx$0.28 clears the baseline prevalence of 2.16% by a factor of roughly thirteen. As recall climbs past 0.6, precision drops sharply—an inescapable feature of rare-event prediction—which makes the operating point decision less a statistical question and more a clinical policy one.

V. Discussion

The headline finding is that statistical aggregation of irregular laboratory time-series, combined with cluster-augmented features and gradient boosting, is sufficient to achieve AUROC above 0.86 on a large real-world ICU cohort. The feature engineering step is deliberately simple—means, variances, extremes, and missingness indicators—yet it captures enough of the physiological signal to compete with more elaborate approaches reported in the literature.

Perhaps the more interesting finding is the split between discrimination and calibration. XGBoost is the best discriminator; Random Forest produces the best-calibrated probabilities. These are not the same model and should not be treated as such in deployment. A mortality risk score used to rank patients for palliative care conversations needs discrimination. A score used to communicate absolute risk to a family needs calibration. Choosing the right model requires knowing which property matters for the intended use.

Unsupervised clustering added a layer of interpretability that the global SHAP analysis alone could not provide. The phenotype-specific SHAP breakdown shows that the cohort is genuinely heterogeneous in its risk mechanisms, and that a single importance ranking flattens out variation that clinicians would probably find meaningful [3, 10, 11].

VI. Limitations

The most significant methodological limitation is the loss of temporal structure. Aggregating a time-series into summary statistics discards trajectory information: the model cannot distinguish a creatinine that has been rising steadily for 18 hours from one that spiked and recovered. Sequence models such as temporal convolutional networks or transformers would preserve this structure, at the cost of considerably greater complexity and data requirements.

The feature set also excludes information that clinicians regard as essential: medication records, comorbidity burden, SOFA components, and interventions such as mechanical ventilation. Including these would likely improve both discrimination and calibration, but substantially complicates data linkage and preprocessing.

Compute constraints imposed by the Google Colab environment [12] limited the hyperparameter search and prevented exploration of larger ensembles and more complex feature expansions. Training was conducted within Colab's free-tier memory ceiling, which restricted batch sizes and the number of cross-validation configurations that could be evaluated. Finally, all data originate from a single US academic medical centre. A model trained on MIMIC-IV may not generalise well to community hospitals or institutions in other countries without retraining or at minimum recalibration.

VII. Next Steps and Deployment Pathway

External validation is the immediate priority. We are pursuing collaborations with clinical laboratories at institutions that use different EHR systems and serve different patient populations, which will test whether the model's performance holds outside the narrow conditions under which it was developed. Performance at MIMIC-IV levels is competitive with published baselines [6], but competitive is not the same as generalisable.

The longer-term target is integration into a hospital EHR as a decision support layer. That goal imposes constraints beyond predictive performance: inference must be fast enough for real-time operation, the memory footprint must fit within hospital server infrastructure, and the probability outputs must remain well-calibrated under distribution shift as patient populations and care practices evolve. Model compression and structured feature pruning are the next engineering steps. Batch inference pipelines that can process overnight admissions and flag high-risk patients for morning rounds are a natural first deployment mode before any real-time integration.

VIII. Conclusion

We have described a mortality prediction pipeline that is deliberately modest in its machinery—aggregate statistics, KMeans clustering, and gradient boosting—yet achieves discrimination on par with more complex published systems. The work makes three contributions. First, it shows that informative missingness and physiological variability can be captured through simple summary features without sequence

modelling. Second, unsupervised clustering surfaces meaningful phenotypic heterogeneity and improves interpretability at the subgroup level. Third, a careful evaluation framework—strict patient-level splits, threshold analysis, temporal leakage audit—provides a more honest picture of what can actually be expected in deployment than single-metric summaries typically convey. The full implementation is publicly available [12]. External validation and EHR integration are the natural next steps.

ACKNOWLEDGMENT

The authors thank Synthica Research Group for institutional support and the PhysioNet team for maintaining open access to the MIMIC-IV database.

REFERENCES

[1] A. E. W. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV (version 3.1)," *PhysioNet*, 2024. [Online]. Available: https://doi.org/10.13026/kpb9-mt58

[2] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785–794.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[6] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed. New York, NY: Springer, 2019.

[7] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[12] A. Dhedhi and Q. Bui, "Unsupervised Cluster-Augmented ICU Mortality Prediction — Implementation Notebook," Google Colab, 2026. [Online]. Available: https://colab.research.google.com/drive/1JT27myiKoC5ncInsxGLpkRTk8ekaPQwR?usp=sharing