Predicting Mortality in ICU Patients with Hypertension Using Machine Learning on EHR Data

Stephanie Chen∗
Department of Biomedical Informatics, Harvard Medical School
stephanie_chen@hms.harvard.edu

Rishabh Goel∗
Department of Biomedical Informatics, Harvard Medical School
rishabh_goel@hms.harvard.edu

Andrew B. Gumbert∗
Department of Biomedical Informatics, Harvard Medical School
andrew_gumbert@hms.harvard.edu

∗All authors contributed equally to this work.

Abstract

Hypertension is a leading preventable cause of mortality, with over 120 million patients affected in the US alone. Yet current predictive models are often limited to specific subtypes of the disease and only predict short-term outcomes. This study introduces a unified, time-aware, ensemble for mortality risk prediction that generalizes across all hypertension subtypes and multiple horizons while addressing selection effects from longer observation windows. Using MIMIC-IV EHR data, we trained models at 24, 48, and 72 hours and combined their outputs with learned convex weights. Evaluated with stratified 5-fold cross-validation for in-hospital, 30-, 60-, 90-day, and 1-year mortality, the ensemble achieved AUCs of 0.854, 0.847, 0.841, and 0.801, with F1 > 0.50 for all horizons except in-hospital. Learned temporal weights placed minimal weight on later windows, indicating limited incremental signal once selection bias is controlled.

1 Introduction

Hypertension, a medical condition characterized by high blood pressure of at least 140/90 mmHg, affects over one billion people and is a leading cause of premature mortality in the global population [1]. Mortality can occur in hypertension because the condition can lead to heart disease, kidney disease, stroke, and other life-threatening complications [1-3]. Worldwide, only around 42% of hypertension cases are diagnosed and treated, and only around 21% of cases are well-controlled [1].

Given the global health burden of hypertension, accurate early risk stratification in the ICU could target resources to the highest-risk patients. Prior research has investigated predictive models for mortality in hypertension and other serious health conditions based on ICU data [2-5]. This prior research has successfully identified clinical features in ICU data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) that can facilitate mortality predictions [2-5]. For example, one study by Huang et al. found that an XGBoost model provided accurate predictions of 28-day mortality in patients with hypertensive stroke, with an AUC value of 0.739 in the test cohort [2]. Also, Peng et al. created a model with Neural Networks to predict 28-day mortality among patients with

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

hypertensive heart failure, yielding an AUC of 0.764 in the test cohort [3]. In addition, Zhang and Ye used Convolutional Neural Networks on ICU data from critically ill patients with hypertension to predict mortality with an AUC of 0.774 [4]. Despite the success of these studies, this existing research is primarily limited to specific subtypes of hypertension, critical cases only, or short timeframes of less than one month [2-4].

To address these gaps, we develop a unified, time-aware convex ensemble for mortality prediction in hypertensive ICU patients. We include all hypertension phenotypes using established ICD-9/10 codes (401–405; I10–I15), covering essential hypertension, hypertensive heart disease, hypertensive kidney disease, combined heart and kidney disease, and secondary hypertension [6-8]. Using MIMIC-IV EHR data, we train separate models using measurements available by 24, 48, and 72 hours and combine their outputs with learned nonnegative weights that sum to one. Our primary aim is an anytime predictor that produces valid estimates at 24 hours and incorporates later data only when it adds signal. We also quantify the incremental value of later windows after debiasing and evaluate performance for in-hospital, 30-, 60-, 90-day, and 1-year mortality.

2 Methods

2.1 Data source and cohort construction

We used the MIMIC-IV database, a publicly available ICU dataset comprising over 500,000 patient stays at Beth Israel Deaconess Medical Center [9]. Patients were included if they had a recorded hypertension diagnosis (ICD-9/10 codes I10–I15, 401–405) and were experiencing their first ICU admission [6-7]. We constructed three cohorts corresponding to observation windows of 24 hours (N = 24,924), 48 hours (N = 22,732), and 72 hours (N = 20,900).

2.2 Feature engineering and preprocessing

For each cohort, we extracted initial values of demographics, vital signs, laboratory results, and comorbidities within the designated window. Variables with more than 10% missingness were excluded at 24h, while those with more than 20% missingness were excluded at 48h and 72h [10]. Patients missing more than 20% of the retained variables were removed. Documentation fields related to medical device settings and alarms were removed, as they are not direct clinical measurements [11]. Outlier values (>3 standard deviations from the mean) were set to missing, and all remaining missingness was imputed with multivariate chained equations [12]. Highly correlated variables (mean absolute correlation > 0.5) were pruned to reduce redundancy. Composite features were derived from clinical information: age brackets, fall risk scores, chronic kidney disease stage, and total Glasgow Coma Scale [13-16]. See Appendix A.2 for more details on feature selection and data preprocessing.

2.3 Modeling at each time window

At each window $t \in \{24, 48, 72\}$, we trained 5 classifiers including logistic regression (LR), random forest (RF), XGBoost (XGB), LightGBM (LGBM), and a soft-voting ensemble (XGB + LGBM) on the t-hour feature set to produce class-probability predictions. For each time window t and model $m \in \{LR, RF, XGB, LGBM\}$, we learn a function:

$$f_m^t : X^t \to [0, 1].$$

For a patient with feature vector x^t at time t, this function predicts the probability of mortality:

$$\hat{p}_m^t(x^t) = f_m^t(x^t).$$

We also constructed a soft-voting ensemble of gradient boosting models, with $\alpha \geq 0, (1 - \alpha) \geq 0$:

$$\hat{p}_{ens}^t(x^t) = \alpha * \hat{p}_{XGB}^t(x^t) + (1 - \alpha)\hat{p}_{LGBM}^t(x^t),$$

2.4 Convex ensemble across time windows

For a patient with features available up to time t, we define the final predicted probability of mortality as a convex combination of the individual time-window predictions:

$$\hat{p}(x) = \sum_{t \in \{24,48,72\}} \lambda_t \hat{p}^t(x^t),$$

with non-negative weights λ_t assigned to each time window, where $\sum_t \lambda_t = 1$.

The weights λ_t are calculated using the weighted binary cross-entropy loss:

$$L(\lambda) = -\sum_i \sum_t w^t(x_i^t) \left[y_i * log(\hat{p}(x_i)) + (1 - y_i) log(1 - \hat{p}(x_i)) \right].$$

2.5 Evaluation

Data were split into train and test sets with an 80:20 ratio, stratified by outcome prevalence (fixed random seed). Within the training set, five-fold cross-validation was applied. We report median and standard deviation across folds for AUROC, accuracy, precision, recall, and F1-score. Statistical comparisons across time windows and model types were conducted to identify the optimal modeling strategy.

3 Results

Table 1: Model type, time window, and performance metrics of the highest performing models in each mortality horizon.

Mortality Horizon	Best Model + Time Window	AUC (±SD)	Accuracy (±SD)
In-Hospital	Ensemble (LGBM+XGB), 24h	0.861 ± 0.006	0.821 ± 0.005
30-Day	Ensemble (LGBM+XGB), 24h	0.854 ± 0.003	0.801 ± 0.002
60-Day	Ensemble (LGBM+XGB), 24h	0.847 ± 0.003	0.784 ± 0.003
90-Day	Ensemble (LGBM+XGB), 24h	0.841 ± 0.008	0.778 ± 0.004
1-Year	Ensemble (LGBM+XGB), 24h	0.801 ± 0.006	0.713 ± 0.005

3.1 Cohort characteristics and sample attrition

Requiring longer observation windows reduced sample size and skewed population characteristics. The number of eligible patients declined from 24,924 (24h) to 22,732 (48h) and 20,900 (72h). Crude 30-day mortality dropped from 13.8% at 24 hours to 11.1% at 72 hours. 1-year mortality remained consistently high (>28%) across all cohorts.

3.2 Model performance across mortality horizons

Across all five mortality targets, in-hospital, 30-day, 60-day, 90-day, and 1-year mortality, the ensemble model combining LightGBM and XGBoost trained on 24-hour data consistently achieved the highest discriminative performance. Specifically, it yielded AUCs of 0.861 (in-hospital), 0.854 (30-day), 0.847 (60-day), 0.841 (90-day), and 0.801 (1-year), with corresponding accuracies all exceeding 0.71 (Table 1). F1-scores exceeded 0.50 for all targets except in-hospital death (F1 = 0.45).

Notably, adding additional data from 48 or 72 hours did not improve performance and sometimes reduced it. For example, the AUC for 30-day mortality dropped from 0.854 (24h) to 0.843 (48h) and 0.826 (72h). A similar pattern was observed across other horizons. This trend was mirrored in F1-score declines, driven largely by reduced precision at longer windows. These results suggest that while more extended data windows include more information, they also introduce selection bias (due to early discharge or death) and add limited incremental predictive signal after 24 hours.

3.3 Model comparison

Figure 1 shows that the 24h ensemble outperformed all other models in both AUC and F1 across most targets. LightGBM performed competitively but slightly lagged on recall. Random forests had inflated accuracy but poor F1 due to low sensitivity. Logistic regression maintained decent AUCs but lower F1 across the board. Learned weights in the convex temporal ensemble placed minimal emphasis on later windows, suggesting that early data capture most of the useful signal for prediction.

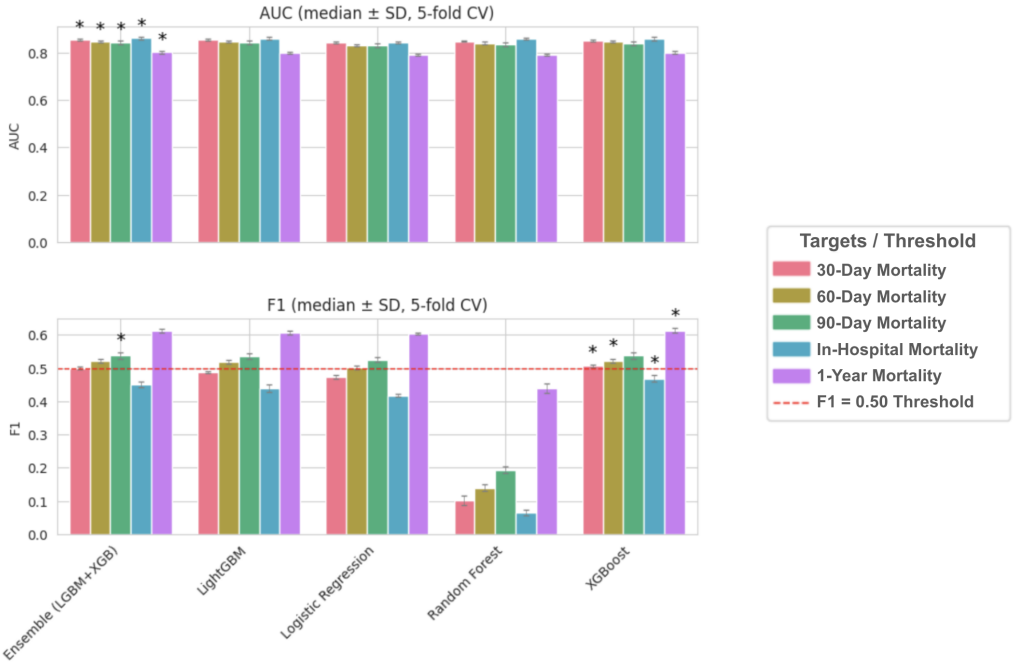

Figure 1: Discriminative performance of five classifiers trained on the 24-hour feature set (Day-1 cohort). See Appendix A.1 for other feature sets. Upper panel: median area under the ROC curve (AUC) across five stratified cross-validation folds; lower panel: corresponding F1 scores. Bars represent the mean of the fold medians and whiskers denote the standard deviation. Asterisks mark the best-performing model for each outcome (tied values all receive a star). The dashed red line highlights an F1 threshold of 0.50, above which precision–recall balance is considered clinically actionable.

4 Discussion and conclusion

Our study demonstrates that routinely collected data from the first 24 hours of an ICU stay can stratify mortality risk in cases of hypertension with promising performance metric results. By including all hypertension subtypes and a one-year horizon, our work broadens clinical applicability and confirms that early physiologic profiles contain sufficient signal for long-term prognostication.

Contrary to our initial hypothesis, adding longitudinal measurements from the second and third ICU days did not benefit model performance. Two mechanisms likely explain this observation. First, requiring 48- or 72-hour completeness selectively removes early discharges and deaths, reducing sample size and event prevalence. Second, many key laboratory and vital sign abnormalities are captured during the first day, so additional values from beyond the first day contribute limited information.

Overall, extending the observation window did not improve discrimination, and an alert system based on day-1 data could be simpler to implement and less vulnerable to missingness than models that wait for later data. Future research should validate the framework in additional external datasets and explore further improvements in interpretability and multi-center applicability. See Appendix A.5 for

more detailed limitations and possible extensions. From a clinical perspective, the ability to identify high-risk hypertensive patients within 24 hours of ICU admission offers a practical window for more intensive monitoring and treatment to ultimately improve health outcomes.

Acknowledgments and Disclosure of Funding

The authors would like to express sincere gratitude to Dr. Marinka Zitnik (Department of Biomedical Informatics, Harvard Medical School), for providing mentorship and feedback during the planning and development of this project. The authors designed an earlier version of the project as part of Dr. Zitnik's BMI 702 course at Harvard Medical School. This work did not receive external funding. The authors have no conflict of interest to declare.

5 References

[1] WHO. Hypertension. World Health Organization. Published March 16, 2023. https://www.who.int/news-room/fact-sheets/detail/hypertension

[2] Huang J, Chen H, Deng J, et al. Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the ICU: a multi-center retrospective cohort study with internal and external cross-validation. Frontiers in Neurology. 2023;14. doi:https://doi.org/10.3389/fneur.2023.1185447

[3] Peng S, Huang J, Liu X, et al. Interpretable machine learning for 28-day all-cause in-hospital mortality prediction in critically ill patients with heart failure combined with hypertension: A retrospective cohort study based on medical information mart for intensive care database-IV and eICU databases. Frontiers in Cardiovascular Medicine. 2022;9. doi:https://doi.org/10.3389/fcvm.2022.994359

[4] Zhang Z, Ye J. Predicting mortality in critically ill patients with hypertension using machine learning and deep learning models. Published online August 22, 2024. doi:https://doi.org/10.1101/2024.08.21.24312399

[5] Yong L, Zhenzhou L. Deep learning-based prediction of in-hospital mortality for sepsis. Scientific Reports. 2024;14(1):372. doi:https://doi.org/10.1038/s41598-023-49890-9

[6] Massen GM, Stone PW, Kwok HHY, et al. Review of codelists used to define hypertension in electronic health records and development of a codelist for research. Open Heart. 2024;11(1):e002640. Published 2024 Apr 15. doi:10.1136/openhrt-2024-002640

[7] Saladini F, Dorigatti F, Santonastaso M, et al. Natural History of Hypertension Subtypes in Young and Middle-Age Adults. American Journal of Hypertension. 2009;22(5):531-537. doi:https://doi.org/10.1038/ajh.2009.21

[8] Parminder Singh Reel, Reel S, Josie, et al. Machine learning for classification of hypertension subtypes using multi-omics: A multi-centre, retrospective, data-driven study. 2022;84:104276-104276. doi:https://doi.org/10.1016/j.ebiom.2022.104276

[9] Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV. Physionet.org. Published October 11, 2024. https://physionet.org/content/mimiciv/3.1/

[10] Dong Y, Peng CYJ. Principled Missing Data Methods for Researchers. SpringerPlus. 2013;2(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/

[11] Carencotte R, Oliver M, Allou N, Ferdynus C, Allyn J. Exploring Clinical Practices of Critical Alarm Settings in Intensive Care Units: A Retrospective Study of 60,000 Patient Stays from the MIMIC-IV Database. Journal of Medical Systems. 2024;48(1). doi:https://doi.org/10.1007/s10916-024-02107-6

[12] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research. 2011;20(1):40-49. doi:https://doi.org/10.1002/mpr.329

[13] Jewell VD, Capistran K, Flecky K, Qi Y, Fellman S. Prediction of Falls in Acute Care Using The Morse Fall Risk Scale. Occupational Therapy In Health Care. 2020;34(4):1-13. doi:https://doi.org/10.1080/07380577.2020.1815928

[14] Zsom L, Zsom M, Salim SA, Fülöp T. Estimated Glomerular Filtration Rate in Chronic Kidney Disease: A Critical Review of Estimate-Based Predictions of Individual Outcomes in Kidney Disease. Toxins. 2022;14(2):127. doi:https://doi.org/10.3390/toxins14020127

[15] Fu J, Kosaka J, Morimatsu H. Impact of Different KDIGO Criteria on Clinical Outcomes for Early Identification of Acute Kidney Injury after Non-Cardiac Surgery. Journal of Clinical Medicine. 2022;11(19):5589. doi:https://doi.org/10.3390/jcm11195589

[16] Jain S, Iverson LM. Glasgow Coma Scale. Nih.gov. Published June 12, 2023. https://www.ncbi.nlm.nih.gov/sites/books/NBK513298/

A Appendix

A.1 Performance of classifier models on all feature sets

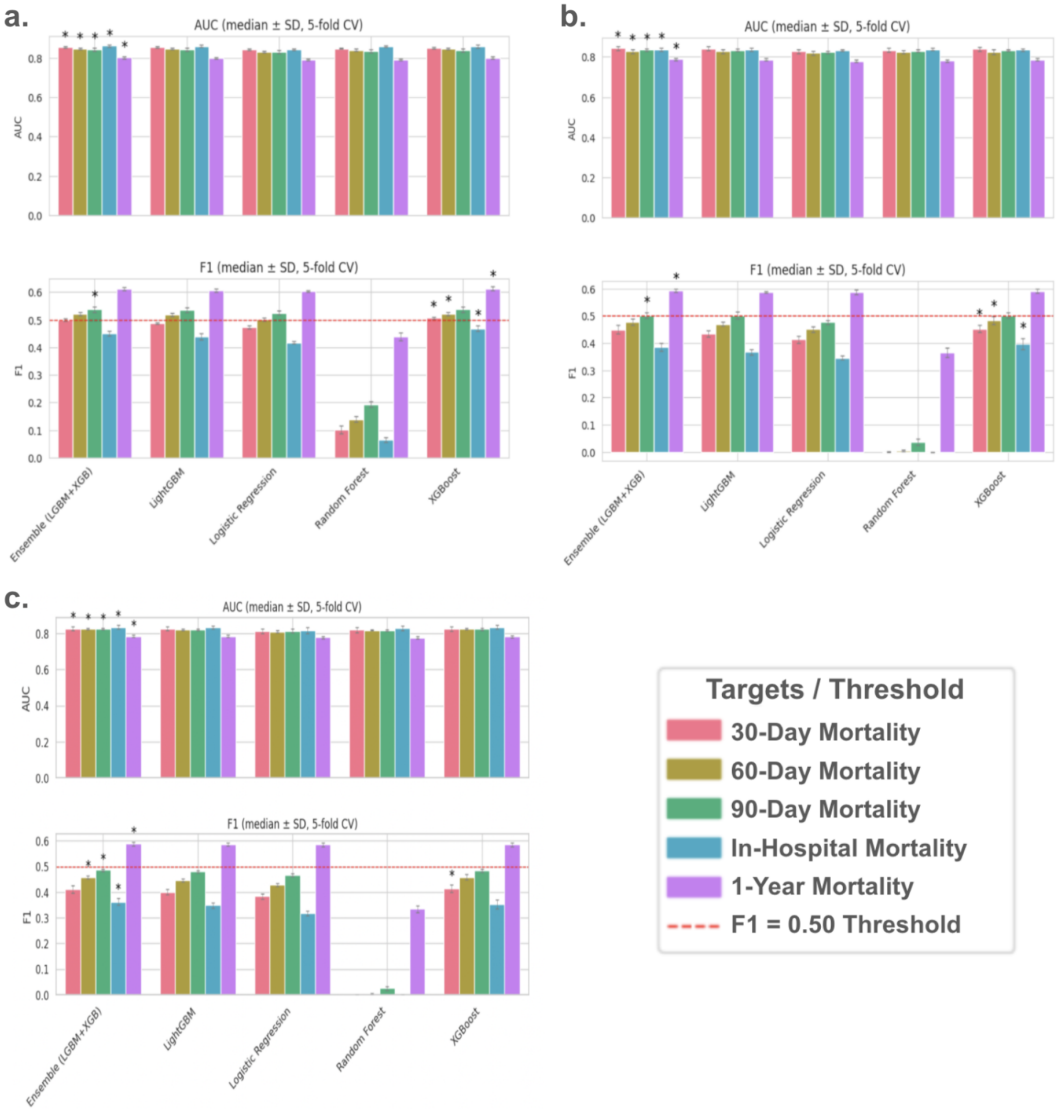

Figure 2: Discriminative performance of five classifiers trained on (a.) the 24-hour feature set (Day-1 cohort), (b.) the 48-hour feature set (Day-2 cohort), and (c.) the 72-hour feature set (Day-3 cohort). Upper panel: median area under the ROC curve (AUC) across five stratified cross-validation folds; lower panel: corresponding F1 scores. Bars represent the mean of the fold medians and whiskers denote the standard deviation. Asterisks mark the best-performing model for each outcome (tied values all receive a star). The dashed red line highlights an F1 threshold of 0.50, above which precision–recall balance is considered clinically actionable.

A.2 Feature selection and data preprocessing

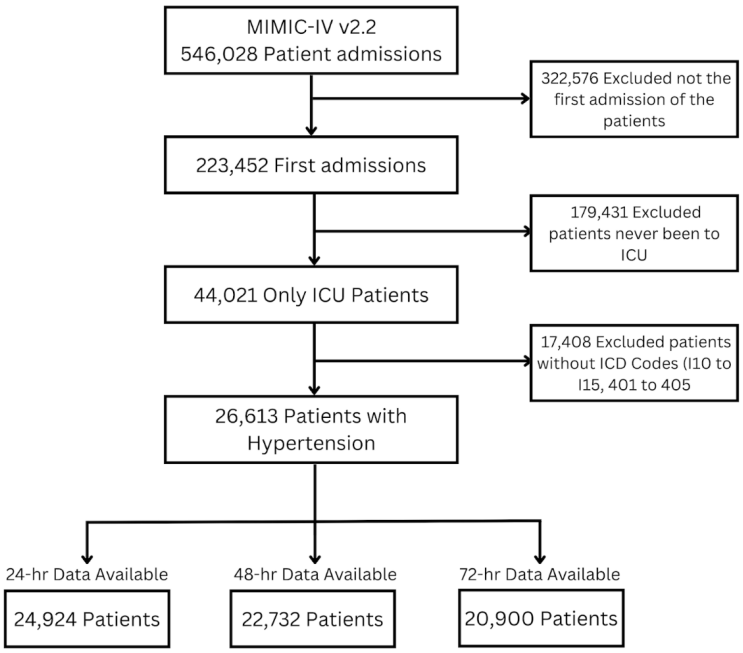

Figure 3: Patient selection workflow to identify patients with hypertension on first admission and with measurements from the first 24, 48, and 72 hours.

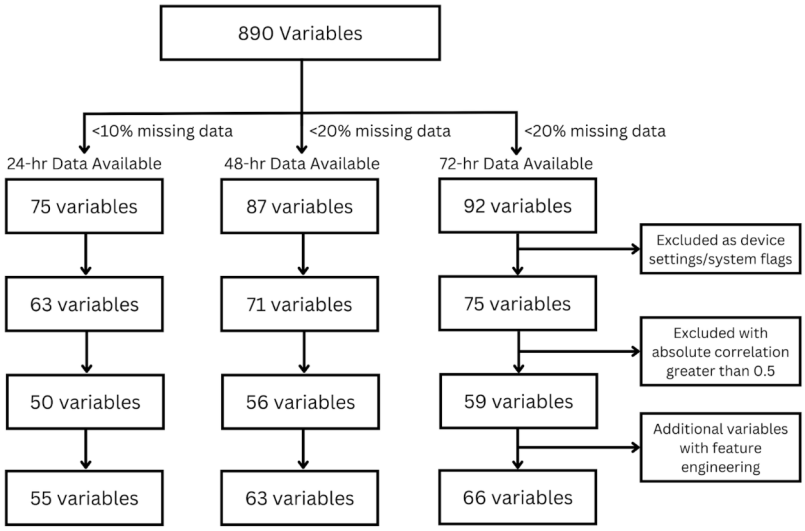

Figure 4: Variable selection pipeline to arrive at the final predictors for each cohort.

Feature selection. For feature selection, we first assessed the missingness of each variable by calculating the proportion of missing values across all observations. For day 1, we chose a 10% missingness threshold, resulting in 75 clinical variables [10]. Given that fewer data points were collected on later days, we adjusted the missingness thresholds to 20% for day 2 and day 3, yielding 87 and 92 variables, respectively. Without raising these thresholds, we would not have had enough meaningful variables

to support robust model development. Documentation fields related to medical device settings and alarms were removed, as they are not direct clinical measurements [11]. Additionally, we excluded patients with less than 80% observed values across all clinical variables to ensure data completeness (Figure 5).

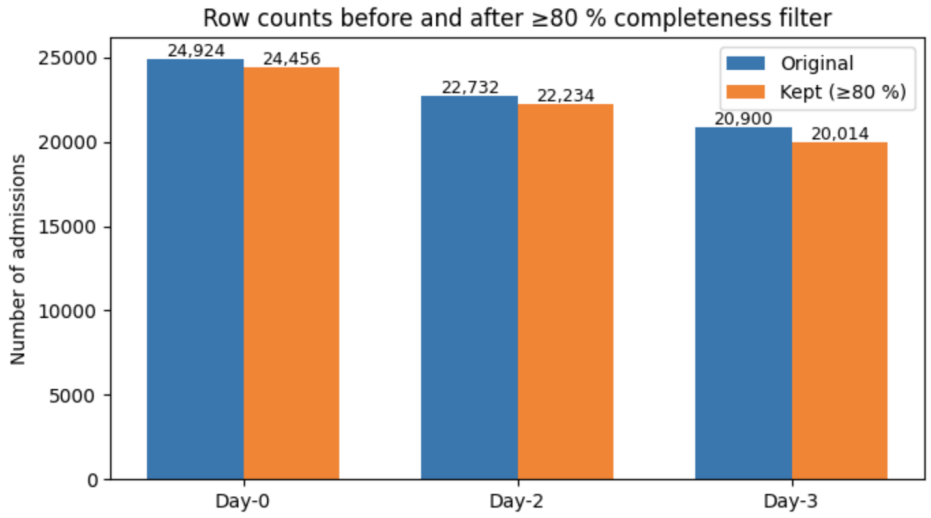

Figure 5: Admissions retained after applying an $\geq$80% completeness filter at each time point. Bar heights show the total number of ICU admissions before (blue) and after (orange) filtering for at least 80% non-missing measurements at day 0, day 2, and day 3.

Outlier handling. We addressed outliers within continuous variables using the empirical rule, defining outliers as observations beyond three standard deviations from the mean. These outliers were replaced with NA values to prevent distortion during imputation, while categorical variables were excluded from this step.

Imputation. Remaining missingness was handled using Multivariate Imputation by Chained Equations (MICE), implemented via the IterativeImputer function from the sklearn.impute package (https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer) [12]. Before imputation, there were missing values in each dataset; following imputation, no missingness remained.

Multicollinearity handling. In order to mitigate multicollinearity, which can destabilize machine learning parameter estimates (especially in linear models) and make feature-importance measures unreliable, we conducted a correlation analysis. We quantified the strength of association between every pair of our candidate predictors, using absolute Pearson's r for continuous–continuous pairs, Cramér's V (from a chi-squared test) for categorical–categorical pairs, and the square-root of eta-squared (or point-biserial r for binary factors) for mixed pairs, and assembled these into a symmetric "association matrix." We then computed each variable's mean absolute association (its average correlation with all others) and flagged any variable pair whose mutual association exceeded 0.5. Within each such highly correlated pair, we dropped the variable with the higher mean absolute association (i.e. the more globally redundant feature). This pruning removed several variables, yielding a final set of 50, 56, and 59 variables whose pairwise associations all fall below the 0.5 threshold (for days 1, 2, and 3, respectively).

Feature engineering. Finally, we derived several clinically-informed features to augment our raw measurements and aid downstream modeling. First, we binned age into five clinically meaningful brackets (< 50, 50–64, 65–74, 75–84, 85+) to accommodate non-linear age effects. Next, we collapsed six binary safety/fall-risk indicators (e.g. history of falls, mental status, use of ambulatory aid) into a single fall_risk_score (sum of the six components) and further discretized this into "No," "Low," and "High" risk groups [13]. We then summarized renal function with an estimated glomerular filtration rate (eGFR) calculated from day-1 serum creatinine, age and sex using the IDMS-aligned CKD-EPI-2021 equation (with a small ε to avoid log-zero) [14]. eGFR was then mapped to KDIGO

stages G1-G5 ($\geq$90, 60-89, 45-59, 30-44, 15-29, <15 mL min^{-1} 1.73 m^{-2}) to yield an ordinal CKD stage [15]. We then created an additive fall-risk score by coercing six Morse/Braden components (recent fall, mental status, ambulatory aid, gait/transferring, IV-line, secondary diagnoses) to 0/1 and summing them (range 0–6); the score was further binned into "No", "Low" (25-49) and "High" ($\geq$50) risk categories to mirror nursing documentation [13]. For days 2 and 3, we also incorporated neurological status via the Glasgow Coma Scale (GCS) [16]. We summed the three sub-scores (Eye, Verbal, Motor) to obtain a total GCS (range 3–15) and categorized severity as Severe (3–8), Moderate (9–12), or Mild (13–15) [16]. These new composite features (age_group, fall_risk_score & group, eGFR & CKD_stage, GCS_total & severity) were then combined with our remaining predictors for each cohort.

A.3 Event prevalence across cohorts

Table 2: Baseline event prevalence across progressively stricter cohorts. Each row shows the number of hypertensive ICU admissions that met the specified minimum length-of-stay requirement (24 h, 48 h, or 72 h) and the corresponding crude rates of in-hospital expiry, 30-, 60-, and 90-day mortality, and one-year death.

Cohort Minimum Stay	Patients (n)	In-hospital Expiry	30-day Mortality	60-day Mortality	90-day Mortality	One-year Death
≥ 24hr (Day 1 Model)	24,456	10.80%	13.80%	16.20%	17.80%	30.20%
≥ 48hr (Day 1+2 Model)	22,234	8.70%	11.60%	14.20%	15.90%	28.70%
≥ 72hr (Day 1+2+3 Model)	20,014	8.20%	11.10%	13.80%	15.60%	28.50%

A.4 Detailed performance metrics of classifiers at each time range

Table 3: Discriminative performance of five machine-learning classifiers trained on variables recorded within the first 24 hours of ICU admission (Day-1 cohort). Values represent the median ± standard deviation across five held-out cross-validation folds for each mortality endpoint. Bold type indicates the highest (or tied-highest) value within each metric–endpoint combination that has an F1>0.5.

Target	Model	AUC	Accuracy	F1	Precision	Recall
In-Hospital Mortality	Ensemble (LGBM+XGB)	0.861 ± 0.006	0.821 ± 0.005	0.451 ± 0.008	0.338 ± 0.008	0.681 ± 0.015
	LightGBM	0.860 ± 0.006	0.799 ± 0.006	0.440 ± 0.012	0.315 ± 0.009	0.730 ± 0.025
	Logistic Regression	0.844 ± 0.004	0.773 ± 0.003	0.418 ± 0.004	0.289 ± 0.002	0.758 ± 0.017
	Random Forest	0.858 ± 0.003	0.895 ± 0.001	0.065 ± 0.008	0.750 ± 0.070	0.034 ± 0.005
	XGBoost	0.859 ± 0.007	0.843 ± 0.006	0.469 ± 0.006	0.518 ± 0.006	0.643 ± 0.013
30-Day Mortality	**Ensemble (LGBM+XGB)**	**0.854 ± 0.003**	**0.801 ± 0.002**	**0.500 ± 0.004**	**0.384 ± 0.003**	**0.716 ± 0.008**
	LightGBM	0.854 ± 0.003	0.778 ± 0.003	0.487 ± 0.003	0.358 ± 0.003	0.751 ± 0.007
	Logistic Regression	0.842 ± 0.005	0.769 ± 0.006	0.473 ± 0.005	0.345 ± 0.006	0.751 ± 0.009
	Random Forest	0.848 ± 0.002	0.867 ± 0.001	0.102 ± 0.014	0.787 ± 0.030	0.055 ± 0.008
	XGBoost	0.852 ± 0.004	0.815 ± 0.003	0.507 ± 0.004	0.401 ± 0.004	0.684 ± 0.004
60-Day Mortality	**Ensemble (LGBM+XGB)**	**0.847 ± 0.003**	**0.784 ± 0.003**	**0.522 ± 0.005**	**0.407 ± 0.005**	**0.730 ± 0.006**
	LightGBM	0.846 ± 0.003	0.772 ± 0.002	0.519 ± 0.005	0.395 ± 0.004	0.755 ± 0.009
	Logistic Regression	0.831 ± 0.004	0.758 ± 0.005	0.502 ± 0.005	0.377 ± 0.006	0.752 ± 0.007
	Random Forest	0.840 ± 0.005	0.847 ± 0.001	0.140 ± 0.009	0.747 ± 0.066	0.076 ± 0.006
	XGBoost	0.846 ± 0.003	0.793 ± 0.004	0.523 ± 0.006	0.418 ± 0.006	0.698 ± 0.007
90-Day Mortality	**Ensemble (LGBM+XGB)**	**0.841 ± 0.008**	**0.778 ± 0.004**	**0.538 ± 0.010**	**0.428 ± 0.007**	**0.724 ± 0.020**
	LightGBM	0.841 ± 0.008	0.762 ± 0.004	0.537 ± 0.009	0.412 ± 0.008	0.753 ± 0.019
	Logistic Regression	0.830 ± 0.008	0.760 ± 0.004	0.526 ± 0.008	0.406 ± 0.006	0.747 ± 0.017
	Random Forest	0.836 ± 0.000	0.835 ± 0.001	0.194 ± 0.009	0.752 ± 0.026	0.111 ± 0.006
	XGBoost	0.838 ± 0.000	0.786 ± 0.003	0.538 ± 0.010	0.437 ± 0.006	0.701 ± 0.024
1-Year Mortality	**Ensemble (LGBM+XGB)**	**0.801 ± 0.006**	**0.713 ± 0.005**	**0.613 ± 0.007**	**0.526 ± 0.005**	**0.749 ± 0.016**
	LightGBM	0.799 ± 0.005	0.720 ± 0.004	0.607 ± 0.007	0.512 ± 0.005	0.729 ± 0.017
	Logistic Regression	0.791 ± 0.003	0.719 ± 0.005	0.587 ± 0.008	0.501 ± 0.007	0.713 ± 0.010
	Random Forest	0.791 ± 0.005	0.748 ± 0.003	0.439 ± 0.015	0.668 ± 0.004	0.328 ± 0.011
	XGBoost	0.800 ± 0.007	0.702 ± 0.005	0.614 ± 0.006	0.505 ± 0.005	0.773 ± 0.016

Table 4: Discriminative performance of five machine-learning classifiers trained on variables recorded within the first 48 hours of ICU admission (Day-2 cohort). Values represent the median ± standard deviation across five held-out cross-validation folds for each mortality endpoint. Bold type indicates the highest (or tied-highest) value within each metric–endpoint combination that has an F1>0.5.

Target	Model	AUC	Accuracy	F1	Precision	Recall
In-Hospital Mortality	Ensemble (LGBM+XGB)	0.837 ± 0.007	0.828 ± 0.005	0.385 ± 0.015	0.278 ± 0.011	0.626 ± 0.021
	LightGBM	0.837 ± 0.007	0.790 ± 0.000	0.368 ± 0.012	0.248 ± 0.007	0.702 ± 0.020
	Logistic Regression	0.832 ± 0.005	0.758 ± 0.008	0.346 ± 0.008	0.226 ± 0.006	0.741 ± 0.017
	Random Forest	0.837 ± 0.006	0.913 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	XGBoost	0.835 ± 0.007	0.860 ± 0.006	0.397 ± 0.020	0.316 ± 0.017	0.532 ± 0.026
30-Day Mortality	Ensemble (LGBM+XGB)	0.843 ± 0.010	0.806 ± 0.006	0.450 ± 0.015	0.336 ± 0.011	0.683 ± 0.028
	LightGBM	0.842 ± 0.010	0.776 ± 0.005	0.434 ± 0.012	0.308 ± 0.009	0.739 ± 0.024
	Logistic Regression	0.827 ± 0.011	0.754 ± 0.008	0.413 ± 0.013	0.286 ± 0.010	0.747 ± 0.024
	Random Forest	0.833 ± 0.010	0.884 ± 0.000	0.000 ± 0.002	0.000 ± 0.200	0.000 ± 0.001
	XGBoost	0.840 ± 0.010	0.827 ± 0.004	0.453 ± 0.014	0.359 ± 0.010	0.613 ± 0.024
60-Day Mortality	Ensemble (LGBM+XGB)	0.827 ± 0.009	0.782 ± 0.004	0.478 ± 0.011	0.366 ± 0.007	0.683 ± 0.025
	LightGBM	0.827 ± 0.009	0.760 ± 0.004	0.470 ± 0.007	0.341 ± 0.005	0.735 ± 0.023
	Logistic Regression	0.820 ± 0.009	0.746 ± 0.009	0.451 ± 0.010	0.325 ± 0.008	0.737 ± 0.027
	Random Forest	0.826 ± 0.008	0.858 ± 0.000	0.006 ± 0.003	0.600 ± 0.346	0.003 ± 0.002
	XGBoost	0.825 ± 0.010	0.800 ± 0.006	0.483 ± 0.014	0.385 ± 0.011	0.635 ± 0.027
90-Day Mortality	**Ensemble (LGBM+XGB)**	**0.835 ± 0.005**	**0.778 ± 0.005**	**0.502 ± 0.010**	**0.389 ± 0.008**	**0.714 ± 0.012**
	LightGBM	0.834 ± 0.006	0.761 ± 0.007	0.502 ± 0.013	0.375 ± 0.010	0.748 ± 0.017
	Logistic Regression	0.823 ± 0.006	0.746 ± 0.005	0.477 ± 0.007	0.356 ± 0.006	0.735 ± 0.010
	Random Forest	0.829 ± 0.006	0.843 ± 0.001	0.036 ± 0.010	0.692 ± 0.110	0.018 ± 0.006
	XGBoost	0.834 ± 0.004	0.790 ± 0.005	0.502 ± 0.011	0.402 ± 0.010	0.665 ± 0.014
1-Year Mortality	**Ensemble (LGBM+XGB)**	**0.788 ± 0.006**	**0.707 ± 0.004**	**0.592 ± 0.008**	**0.493 ± 0.005**	**0.742 ± 0.012**
	LightGBM	0.786 ± 0.006	0.709 ± 0.004	0.587 ± 0.004	0.478 ± 0.009	0.727 ± 0.010
	Logistic Regression	0.779 ± 0.005	0.714 ± 0.005	0.587 ± 0.008	0.501 ± 0.007	0.709 ± 0.013
	Random Forest	0.780 ± 0.005	0.750 ± 0.005	0.365 ± 0.017	0.670 ± 0.221	0.249 ± 0.013
	XGBoost	0.787 ± 0.006	0.699 ± 0.006	0.591 ± 0.008	0.485 ± 0.006	0.762 ± 0.015

Table 5: Discriminative performance of five machine-learning classifiers trained on variables recorded within the first 72 hours of ICU admission (Day-3 cohort). Values represent the median ± standard deviation across five held-out cross-validation folds for each mortality endpoint. Bold type indicates the highest (or tied-highest) value within each metric–endpoint combination that has an F1>0.5.

Target	Model	AUC	Accuracy	F1	Precision	Recall
In-Hospital Mortality	Ensemble (LGBM+XGB)	0.834 ± 0.007	0.831 ± 0.002	0.363 ± 0.012	0.262 ± 0.007	0.678 ± 0.027
	LightGBM	0.833 ± 0.009	0.793 ± 0.002	0.349 ± 0.012	0.236 ± 0.007	0.678 ± 0.027
	Logistic Regression	0.816 ± 0.016	0.745 ± 0.006	0.317 ± 0.009	0.204 ± 0.006	0.736 ± 0.025
	Random Forest	0.828 ± 0.012	0.918 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	XGBoost	0.833 ± 0.011	0.857 ± 0.003	0.353 ± 0.017	0.278 ± 0.011	0.483 ± 0.031
30-Day Mortality	Ensemble (LGBM+XGB)	0.826 ± 0.012	0.795 ± 0.007	0.412 ± 0.015	0.300 ± 0.012	0.656 ± 0.023
	LightGBM	0.826 ± 0.011	0.765 ± 0.006	0.401 ± 0.009	0.279 ± 0.008	0.719 ± 0.012
	Logistic Regression	0.811 ± 0.013	0.737 ± 0.005	0.384 ± 0.010	0.261 ± 0.007	0.728 ± 0.020
	Random Forest	0.819 ± 0.014	0.889 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	XGBoost	0.823 ± 0.013	0.819 ± 0.006	0.414 ± 0.014	0.321 ± 0.012	0.589 ± 0.015
60-Day Mortality	Ensemble (LGBM+XGB)	0.824 ± 0.004	0.777 ± 0.004	0.459 ± 0.006	0.346 ± 0.005	0.688 ± 0.009
	LightGBM	0.821 ± 0.005	0.748 ± 0.005	0.447 ± 0.006	0.322 ± 0.006	0.727 ± 0.009
	Logistic Regression	0.809 ± 0.005	0.733 ± 0.005	0.430 ± 0.006	0.305 ± 0.006	0.742 ± 0.008
	Random Forest	0.817 ± 0.003	0.862 ± 0.000	0.004 ± 0.003	0.500 ± 0.332	0.002 ± 0.001
	XGBoost	0.823 ± 0.004	0.795 ± 0.005	0.459 ± 0.010	0.361 ± 0.008	0.641 ± 0.012
90-Day Mortality	Ensemble (LGBM+XGB)	0.823 ± 0.005	0.769 ± 0.003	0.486 ± 0.005	0.371 ± 0.004	0.700 ± 0.013
	LightGBM	0.821 ± 0.005	0.751 ± 0.001	0.480 ± 0.004	0.357 ± 0.002	0.740 ± 0.014
	Logistic Regression	0.813 ± 0.009	0.736 ± 0.002	0.467 ± 0.006	0.342 ± 0.003	0.737 ± 0.016
	Random Forest	0.816 ± 0.004	0.845 ± 0.000	0.025 ± 0.007	0.667 ± 0.055	0.013 ± 0.004
	XGBoost	0.822 ± 0.005	0.782 ± 0.003	0.484 ± 0.005	0.385 ± 0.005	0.655 ± 0.010
1-Year Mortality	**Ensemble (LGBM+XGB)**	**0.784 ± 0.007**	**0.696 ± 0.003**	**0.586 ± 0.005**	**0.488 ± 0.006**	**0.732 ± 0.005**
	LightGBM	0.784 ± 0.007	0.705 ± 0.005	0.586 ± 0.006	0.479 ± 0.009	0.729 ± 0.017
	Logistic Regression	0.777 ± 0.006	0.713 ± 0.009	0.585 ± 0.008	0.498 ± 0.010	0.707 ± 0.006
	Random Forest	0.776 ± 0.007	0.746 ± 0.003	0.334 ± 0.012	0.650 ± 0.016	0.223 ± 0.011
	XGBoost	0.781 ± 0.006	0.689 ± 0.007	0.585 ± 0.007	0.473 ± 0.007	0.761 ± 0.010

A.5 Limitations and potential extensions

Several limitations merit consideration. First, co-morbidity burden was not explicitly modeled. Although every subject carried a hypertension diagnosis, many ICU admissions were precipitated by other acute or chronic conditions (e.g., sepsis, heart failure) that independently affect survival. Because MIMIC-IV is optimized for acute-care documentation, the granularity of outpatient comorbidity coding is limited; consequently, our models may partly reflect risk signatures of co-existing disease rather than hypertension per se.

Second, external validity remains untested. Performance metrics were obtained with five-fold cross-validation on a single centre's data. External evaluation on geographically and demographically distinct EHRs, such as eICU, HiRID or multi-hospital Cerner networks, is required before clinical deployment.

Third, right-censoring in this cohort differs from classical survival settings. Patients were most often censored at hospital discharge rather than death. We addressed this by training separate models on 24-, 48- and 72-hour feature windows, thereby conditioning on length of stay and partially accounting for informative censoring. Nonetheless, longer follow-up windows and formal competing-risk methods could yield more robust longitudinal insight. It is worth noting, however, that our comparison of 24-, 48-, and 72-hour observation windows inherently mixes two effects that are difficult to disentangle. First, later windows offer more longitudinal data, which could in principle improve predictive performance, but, requiring patients to remain in the ICU for 48 or 72 hours also changes the underlying case-mix: early deaths and early discharges are removed, which narrows the distribution of illness severity and eliminates many of the most and least predictable cases. This case-mix shift makes later-window cohorts intrinsically harder to classify, while the marginal information added by day-2 and day-3 measurements is often limited. As a result, lower performance at 48 and 72 hours should not be interpreted solely as evidence that additional ICU data are unhelpful; rather, it reflects the combined influence of selection bias and diminishing incremental signal in later-day measurements.

Fourth, interpretability is constrained. Gradient-boosted ensembles capture complex non-linearities but operate as "black boxes." We mitigated this with post-hoc SHAP analyses, yet prospective adoption may still benefit from inherently explainable architectures (for eg, generalized additive or rule-based models) to support bedside acceptance.

Finally, temporal context is coarse. The dataset does not record when hypertension was first diagnosed; nor does it contain outpatient blood-pressure trajectories or medication adherence. Hence, our predictions are anchored to a single ICU admission and cannot distinguish long-standing, treatment-resistant hypertension from newly diagnosed disease. Linking longitudinal outpatient records or claims data to the index admission would enable finer-grained risk stratification.