frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 11 December 2025
DOI 10.3389/frai.2025.1686378

OPEN ACCESS

EDITED BY
Christos A. Frantzidis,
University of Lincoln, United Kingdom

REVIEWED BY
De Rong Loh,
Duke-NUS Medical School, Singapore
Ioanna Maria Spyrou,
Berry College, United States

*CORRESPONDENCE
Jiancheng Ye
jiancheng.ye@u.northwestern.edu

RECEIVED 15 August 2025
REVISED 02 November 2025
ACCEPTED 21 November 2025
PUBLISHED 11 December 2025

CITATION
Zhang S, Ding S, Xu Z and Ye J (2025) Machine learning-based mortality prediction in critically ill patients with hypertension: comparative analysis, fairness, and interpretability.
Front. Artif. Intell. 8:1686378.
doi: 10.3389/frai.2025.1686378

Machine learning-based mortality prediction in critically ill patients with hypertension: comparative analysis, fairness, and interpretability

Shenghan Zhang[1], Sirui Ding[2], Zidu Xu[3] and Jiancheng Ye[4]*

[1]Department of Biomedical Informatics, Harvard University, Boston, MA, United States, [2]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, United States, [3]School of Nursing, Columbia University, New York, NY, United States, [4]Weill Cornell Medicine, Cornell University, New York, NY, United States

Background: Hypertension is a leading global health concern, significantly contributing to cardiovascular, cerebrovascular, and renal diseases. In critically ill patients, hypertension poses increased risks of complications and mortality. Early and accurate mortality prediction in this population is essential for timely intervention and improved outcomes. Machine learning (ML) and deep learning (DL) approaches offer promising solutions by leveraging high-dimensional electronic health record (EHR) data.
Objective: To develop and evaluate ML and DL models for predicting in-hospital mortality in hypertensive patients using the MIMIC-IV critical care dataset, and to assess the fairness and interpretability of the models.
Methods: We developed four ML models—gradient boosting machine (GBM), logistic regression, support vector machine (SVM), and random forest—and two DL models—multilayer perceptron (MLP) and long short-term memory (LSTM). A comprehensive set of features, including demographics, lab values, vital signs, comorbidities, and ICU-specific variables, were extracted or engineered. Models were trained using 5-fold cross-validation and evaluated on a separate test set. Feature importance was analyzed using SHapley Additive exPlanations (SHAP) values, and fairness was assessed using demographic parity difference (DPD) and equalized odds difference (EOD), with and without the application of debiasing techniques.
Results: The GBM model outperformed all other models, with an AUC-ROC score of 96.3%, accuracy of 89.4%, sensitivity of 87.8%, specificity of 90.7%, and F1 score of 89.2%. Key features contributing to mortality prediction included Glasgow Coma Scale (GCS) scores, Braden Scale scores, blood urea nitrogen, age, red cell distribution width (RDW), bicarbonate, and lactate levels. Fairness analysis revealed that models trained on the top 30 most important features demonstrated lower DPD and EOD, suggesting reduced bias. Debiasing methods improved fairness in models trained with all features but had limited effects on models using the top 30 features.
Conclusion: ML models show strong potential for mortality prediction in critically ill hypertensive patients. Feature selection not only enhances interpretability and reduces computational complexity but may also contribute to improved model fairness. These findings support the integration of interpretable and equitable AI tools in critical care settings to assist with clinical decision-making.

KEYWORDS

hypertension, mortality prediction, machine learning, deep learning, intensive care unit, fairness in AI, SHAP

Introduction

Hypertension, commonly known as high blood pressure, is one of the most prevalent chronic conditions globally and a leading contributor to morbidity and mortality. According to the World Health Organization, approximately 1.13 billion people worldwide suffer from hypertension, with fewer than 20% achieving adequate blood pressure control (World Health Organization, 2021). The condition significantly increases the risk of cardiovascular disease, stroke, and chronic kidney disease, often resulting in poor health outcomes and elevated healthcare costs (Nowbar et al., 2019). Despite advances in pharmacological treatments and lifestyle interventions (Ye et al., 2022; Chen et al., 2024), hypertension remains a major public health challenge and a key contributor to premature mortality (Vaduganathan et al., 2022).

In critically ill patients, the presence of hypertension further complicates clinical trajectories, especially in intensive care unit (ICU) settings where rapid deterioration can occur (Ye and Sanchez-Pinto, 2020). Early identification of high-risk individuals among hypertensive patients admitted to the ICU is essential for timely and targeted interventions that could reduce mortality rates and optimize resource allocation (Bress et al., 2024). However, traditional mortality prediction models, such as risk scoring systems, often rely on a narrow set of pre-selected variables and may not adequately capture the complex, nonlinear relationships and dynamic physiological changes present in critically ill populations. Recent advances in machine learning (ML) and deep learning (DL) have opened new avenues for clinical risk prediction by enabling the analysis of large-scale electronic health record (EHR) data with minimal prior assumptions (Ye et al., 2020). These models are capable of learning complex patterns from high-dimensional data and can incorporate a wide range of clinical variables—including demographic data, comorbidities, laboratory test results, vital signs, and treatment information—into predictive frameworks (Zhang and Ye, 2025). Among the rich sources of EHR data available for research, the Medical Information Mart for Intensive Care IV (MIMIC-IV) database presents a comprehensive, publicly available resource containing detailed longitudinal data on ICU patients.

In this study, we utilize the MIMIC-IV dataset to develop and evaluate a suite of ML and DL models for predicting in-hospital mortality among patients diagnosed with hypertension. We explore six modeling approaches, including four machine learning algorithms—gradient boosting machine (GBM), logistic regression, support vector machine (SVM), and random forest—and two deep learning architectures—a multilayer perceptron (MLP) and long short-term memory (LSTM) network. We incorporate a broad set of features and introduce new variables related to ICU complications and patient conditions to enhance predictive performance.

Our work has three objectives: (1) to assess the comparative performance of these models in predicting mortality in critically ill patients with hypertension, (2) to identify key clinical features associated with increased mortality risk using explainable techniques such as SHAP (SHapley Additive exPlanations), and (3) to examine the fairness of model predictions across demographic groups while testing debiasing strategies. By integrating prediction accuracy, interpretability, and fairness evaluation, this study also aims to contribute to the growing body of evidence supporting the use of AI-driven models in clinical decision-making. Ultimately, our goal is to lay the groundwork for developing scalable, equitable, and actionable tools to assist clinicians in managing hypertensive patients in critical care settings.

Methods

Dataset and study population

This study leverages the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset, which comprises detailed clinical data from patients admitted to critical care units at a large tertiary care hospital (Johnson et al., 2023). The dataset includes demographics, vital signs, laboratory results, medications, interventions, and outcomes for over 40,000 patients. MIMIC-IV is structured into modules corresponding to various hospital departments, allowing access to patient data at both hospital and ICU levels. Using unique patient identifiers, we integrated information across these modules to construct a comprehensive cohort. For this study, we focused on predicting mortality among patients diagnosed with hypertension. Figure 1 illustrates the overall study pipeline. We identified all adult patients (aged 18 and older) with hypertension using ICD-9 and ICD-10 diagnosis codes. Patients were included regardless of whether hypertension was the primary admission diagnosis or a comorbidity, reflecting the real-world scenario where hypertensive patients present to the ICU with various acute conditions. Our inclusion criteria included: (1) age $\geq$18 years at admission, (2) at least one documented hypertension diagnosis code during the hospital stay, and (3) complete outcome data (hospital mortality status). Exclusion criteria included: (1) missing information, (2) incomplete vital signs or laboratory data exceeding 80% missingness, and (3) duplicate records. Patients with incomplete records or missing outcome data were excluded from the analysis.

Data preprocessing pipeline

Data preprocessing was essential to ensure the quality and robustness of our machine learning models. The pipeline included four major steps: (1) data extraction, (2) data cleaning, (3) feature engineering, and (4) feature selection. Given the breadth of conditions represented in MIMIC-IV, we first isolated patients with hypertension using relevant ICD-10 codes. ICD-9 codes were mapped to ICD-10 using the General Equivalence Mapping (GEM) tables provided by the Centers for Medicare & Medicaid Services (CMS), ensuring a unified coding system. Mortality status was derived from the hospital_expire_flag variable. Each patient was identified using a unique subject_id to extract

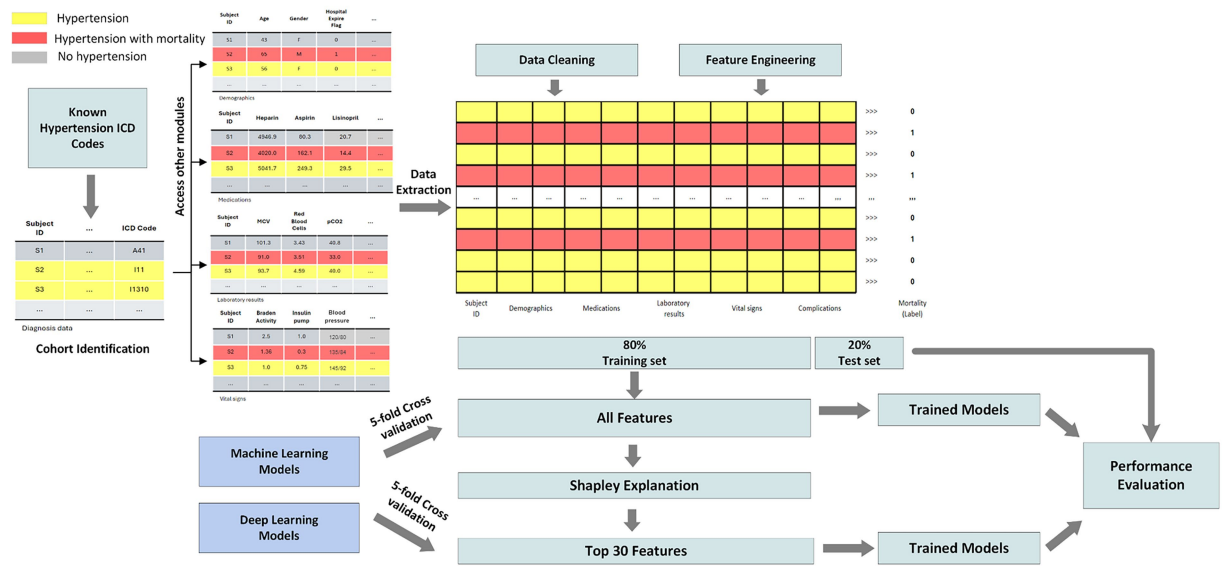

FIGURE 1
Overview of the hypertension mortality prediction pipeline.

information across modules. The features collected included (1) demographics (age, sex, self-reported race), (2) laboratory test results, (3) medication history, (4) vital signs, and (5) clinical complications.

To address data quality, we eliminated duplicate records by retaining the most recent entry per patient. Features with more than 80% missing data were discarded, and remaining missing values were imputed using the median. Patients with ambiguous demographic information (e.g., "unknown" or "other" for race) were excluded to minimize potential biases. Feature engineering involved creating new variables based on hospital and ICU-level data. We calculated the length of hospital stay (in days), and binary indicators were created for ICU admission and clinical complications (based on ICD-10 codes). Continuous variables were normalized to ensure comparability. Race was encoded using binary indicators for each category.

Feature selection and model explainability

Feature selection combined domain expertise with data-driven methods to ensure both causal plausibility and predictive power. Our approach consisted of two stages: (1) clinically-informed feature extraction, and (2) SHAP-based feature ranking and selection. In the first stage, we extracted features based on established clinical knowledge and known causal pathways to mortality in critically ill patients. These included: physiological measures with direct biological mechanisms (vital signs reflecting hemodynamic and respiratory status; laboratory values indicating organ function and metabolic state), validated clinical assessment instruments with established prognostic value (Glasgow Coma Scale for neurological function; Braden Scale for functional status and frailty), documented comorbidities and complications with known mortality associations based on ICD codes, and treatment-related variables (medications, procedures). This clinically-grounded foundation ensured that all candidate features had face validity and plausibility for inclusion in mortality prediction, rather than being selected purely through algorithmic screening.

In the second stage, we applied SHAP (SHapley Additive exPlanations, version 0.47.1) to quantify the marginal contribution of each clinically-justified feature to mortality predictions (Lundberg and Lee, 2017). SHAP values were computed for all six models and averaged across cross-validation folds to ensure robust importance estimates. Importantly, we examined whether the highest-ranked features aligned with known causal mechanisms. The identified top features—including GCS components (neurological compromise, potentially reflecting cerebrovascular complications), Braden scores (functional status and frailty), urea nitrogen (renal function, a key target organ in hypertension), lactate and bicarbonate (tissue perfusion and metabolic dysfunction)—all have well-established causal links to mortality in critically ill populations and biological plausibility in the context of hypertensive complications. Features were ranked by their average SHAP value, and the top 30 were retained for subsequent model development.

This hybrid approach balances prediction-oriented feature selection with causal reasoning: we began with a set of features justified by clinical mechanisms and literature, then identified which of these causally-plausible features were most predictive in our specific population. This methodology guards against including features that may be predictive through spurious correlation or confounding while lacking genuine causal relationships with mortality.

Model development

To comprehensively evaluate model performance, we developed and trained four machine learning (ML) models and two deep learning (DL) models on our hypertension patient cohorts. The ML models included logistic regression, random forest, support vector machines (SVM), and gradient boosting machines (GBM), selected for their ability to handle the complexity of healthcare data and their

varying degrees of interpretability (Seki et al., 2021). Logistic Regression is a linear model widely used in medical research due to its simplicity and high interpretability, making it a common baseline for binary classification tasks. Random Forest is an ensemble learning method that aggregates multiple decision trees to model non-linear relationships and interactions between features, improving prediction robustness. SVM classify outcomes by identifying the optimal hyperplane that separates data points in a high-dimensional space. SVMs are particularly effective when the relationship between predictors and outcomes is non-linear. GBM are another ensemble method that iteratively combines weak learners, often decision trees, to minimize a loss function using gradient descent. GBMs are known for strong predictive performance and flexibility. These four ML models have been widely applied in the literature for mortality prediction using EHR data and are well-regarded for balancing accuracy and interpretability (Chen et al., 2019).

Additionally, we implemented two deep learning models—Multi-Layer Perceptrons (MLP) and Long Short-Term Memory (LSTM) networks—both commonly applied in clinical outcome prediction tasks (Thorsen-Meyer et al., 2020). MLP are feedforward neural networks comprising multiple layers of interconnected neurons. We implemented a standard three-layer MLP, with the input layer matching the number of patient features, a hidden layer containing 128 neurons, and an output layer with two neurons and softmax activation. LSTM networks is a specialized type of recurrent neural network, are designed to capture temporal dependencies in sequential data. This makes them well-suited for modeling longitudinal patient data. We developed an LSTM model with an input dimension based on patient features and a hidden state size of 128, followed by a softmax-activated output layer with two neurons.

All models were developed using Python 3.8.19. ML models (logistic regression, random forest, SVM) were implemented using the scikit-learn library (v1.3.0), and GBM was implemented using the XGBoost library (v2.0.3). DL models were developed using PyTorch (v2.3.0), trained for 500 epochs using the Adam optimizer (learning rate = 1e-3, batch size = 1,024).

Model fairness and bias mitigation

Fairness is a critical consideration when developing machine learning models for clinical use. Ensuring fairness entails minimizing algorithmic bias across demographic groups, such as sex or race (Ye and Ren, 2022). In this study, we focused on identifying and mitigating biases across sex subgroups. Biases can generally be classified into two categories: data-level bias and model-level bias. To address both, we implemented three types of debiasing strategies described in recent literature (Ding et al., 2024): (1) Correlation Removal, a pre-processing technique, eliminates features that are highly correlated with sensitive demographic variables to reduce data-level bias; (2) Reduction, an in-processing method, introduces a fairness-related penalty term in the model's loss function during training to discourage biased predictions; (3) Threshold Optimization, a post-processing method, adjusts the model's decision threshold to improve fairness metrics without retraining.

We used the Fairlearn library (v0.10.0) to implement the reduction and threshold optimization methods for ML models. Due to Fairlearn's limited support for DL models, we manually implemented equivalent debiasing strategies, including a custom reduction term and correlation-based feature filtering.

Model evaluation and statistical analysis

The dataset was randomly partitioned into a training set (80%) and a test set (20%). We applied 5-fold cross-validation on the training set to tune hyperparameters and prevent overfitting. Final evaluations were conducted on the held-out test set to assess generalizability. To evaluate the utility of our feature selection strategy, models were trained using both the original extracted features and a reduced feature set filtered by SHAP values. Performance comparisons on the test set enabled assessment of the impact of feature selection on prediction accuracy. Model performance was assessed using standard metrics for clinical prediction tasks, including accuracy, sensitivity (recall), specificity (precision), F1 score, area under the receiver operating characteristic curve (AUC-ROC), and area under the precision-recall curve (AUPRC). Accuracy and AUC-ROC reflect overall model discriminative ability, while sensitivity, specificity, and F1 score provide insight into performance among individuals with mortality.

To compare patient characteristics between those with and without mortality, Welch's t-tests were used for continuous variables and Chi-square tests for categorical variables. A p-value < 0.05 was considered statistically significant. For performance metrics, we computed 95% confidence intervals (CI) using the following approaches: (1) Normal approximation for cross-validation results; (2) Wald intervals for accuracy, sensitivity, and specificity on the test set; (3).

Bootstrapping (1,000 iterations) for complex metrics, such as AUROC and F1 score. Model fairness was evaluated using two metrics: (1) Demographic Parity Difference (DPD): the difference in positive prediction rates between demographic groups; and (2) Equalized Odds Difference (EOD): the difference in true positive and false positive rates across groups.

Bootstrapping (1,000 iterations) was used to calculate 95% CIs for both fairness metrics.

Study approval

We solely used publicly available MIMIC-IV data for this study.

Results

Patient characteristics

Table 1 demonstrates the characteristics of the patient cohort, including distributions for age, length of hospital stay, sex, race, and clinical complications. After applying the inclusion and exclusion criteria, we identified 88,084 patients diagnosed with hypertension. Among them, 3.9% (3,442 patients) died during their hospital stay. The cohort was composed of 48.6% females and 51.4% males, with a mean age of 48.9 years [standard deviation (SD) = 14.8 years]. The average age of patients in the mortality group was significantly higher at 56.4 years (SD = 12.8), compared to 48.6 years (SD = 14.8) in the

TABLE 1 Characteristics of the patients.

Characteristics	Total (n = 88,084)	Mortality (n = 3,442)	Non-mortality (n = 84,642)	p-value
Sex				0.451
Male	45,241 (51.4%)	1790 (52.0%)	43,451 (51.3%)	
Female	42,843 (48.6%)	1,652 (48.0%)	41,191 (48.7%)	
Age				
Mean (SD)	48.93 (14.79)	56.38 (12.78)	48.63 (14.79)	<0.001
Length of stay, days				
Median (IQR)	6.9 (2.9–16.2)	7.49 (2.7–16.6)	6.9 (2.9–16.2)	<0.001
Race				<0.01
American Indian	175 (0.2%)	2 (0.1%)	173 (0.2%)	
Asian	2,696 (3.1%)	112 (3.3%)	2,584 (3.1%)	
Black	12,299 (14.0%)	257 (7.5%)	12,042 (14.2%)	
Hispanic	3,831 (4.3%)	61 (1.8%)	3,770 (4.5%)	
White	60,425 (68.6%)	2,129 (61.9%)	58,296 (68.9%)	
Others	8,658 (9.8%)	881 (25.6%)	7,777 (9.2%)	
ICU stay				<0.001
Yes	34,868 (39.6%)	2,993 (87.0%)	31,875 (37.7%)	
No	53,216 (60.4%)	449 (13.0%)	52,767 (62.3%)	
Complication				
Diabetes mellitus	51,690 (58.7%)	1,585 (46.0%)	50,105 (59.2%)	0.194
Heart failure	43,593 (49.5%)	2,193 (63.7%)	41,400 (48.9%)	<0.001
Arrhythmias	8,278 (9.4%)	421 (12.2%)	7,857 (9.3%)	<0.001
Myocardial infarction	7,591 (8.6%)	526 (15.3%)	7,065 (8.3%)	<0.001
Chronic obstructive pulmonary disease	14,424 (16.4%)	788 (22.9%)	13,636 (16.1%)	<0.001
Asthma	9,889 (11.2%)	177 (5.1%)	9,712 (11.5%)	<0.001
Pneumonia	13,085 (14.9%)	970 (28.2%)	12,115 (14.3%)	<0.001
Stroke	3,397 (3.9%)	278 (8.1%)	3,119 (3.7%)	<0.001
Dementia	8,427 (9.6%)	379 (11.0%)	8,048 (9.5%)	<0.001

non-mortality group, suggesting that older patients face a greater risk of in-hospital mortality. This age difference was also reflected in the median values: 59 years (IQR: 48–67) in the mortality group versus 49 years (IQR: 39–60) in the non-mortality group. There was a substantially higher ICU stay rate among patients who died (87.0%) compared to those who survived (37.7%).

Model performance using all extracted features

We trained several models on the hypertension cohort and evaluated their performance using 5-fold cross-validation and a separate held-out test set. Performance was assessed using five metrics: Accuracy, Area under the Receiver Operating Characteristic Curve (AUROC), Sensitivity (Recall), Specificity (Precision), and F1 Score. For each metric, we report the mean and the corresponding 95% confidence interval (CI), reflecting the variability across data partitions. Full cross-validation results are available in Supplementary Table 1. Overall, deep learning models outperformed traditional machine learning approaches, indicating their strength in capturing complex patterns. The Long Short-Term Memory (LSTM) network achieved the highest performance across all metrics, with accuracy of 93.9%, sensitivity of 94.3%, specificity of 93.5%, and an F1 score of 93.9%. Among machine learning models, logistic regression (LR) performed the worst, while gradient boosting machines (GBM) showed the best performance, particularly in accuracy, AUROC, and specificity.

To evaluate generalizability, we also assessed performance on the test set (Table 2). Notably, both MLP and LSTM models exhibited performance declines—3.2 and 5.5% decreases in accuracy, respectively—across all metrics, suggesting potential overfitting. In contrast, the machine learning models (LR, RF, SVM, and GBM) maintained stable performance, underscoring better generalizability and lower risk of overfitting. Figure 2 displays the AUC-ROC and Precision-Recall (PR) curves. The ROC curves illustrate the trade-off between true positive rate and false positive rate, with GBM achieving the highest AUC (0.97), closely followed by LR, RF, and SVM (all with

TABLE 2 Model performance for hypertension mortality prediction on the test set using all features.

Models\Metrics	Accuracy, 95% CI	AUROC, 95% CI	Sensitivity, 95% CI	Specificity, 95% CI	F1 Score, 95% CI
Logistic regression	0.881 [0.864, 0.898]	0.956 [0.945, 0.965]	0.864 [0.845, 0.882]	0.895 [0.879, 0.911]	0.879 [0.860, 0.895]
Random forest	0.888 [0.871, 0.904]	0.958 [0.949, 0.967]	0.904 [0.889, 0.920]	0.875 [0.858, 0.892]	0.889 [0.872, 0.905]
Support vector machine	0.885 [0.869, 0.902]	0.956 [0.945, 0.965]	0.872 [0.855, 0.890]	0.896 [0.880, 0.912]	0.884 [0.866, 0.900]
Gradient boost machine	0.901 [0.886, 0.917]	0.968 [0.960, 0.976]	0.893 [0.876, 0.909]	0.908 [0.893, 0.924]	0.900 [0.882, 0.917]
Multi-layer perceptron	0.882 [0.865, 0.899]	0.925 [0.910, 0.938]	0.891 [0.875, 0.908]	0.876 [0.858, 0.893]	0.884 [0.865, 0.901]
Long short-term memory neural network	0.884 [0.867, 0.901]	0.935 [0.921, 0.947]	0.897 [0.881, 0.913]	0.874 [0.857, 0.892]	0.885 [0.867, 0.901]

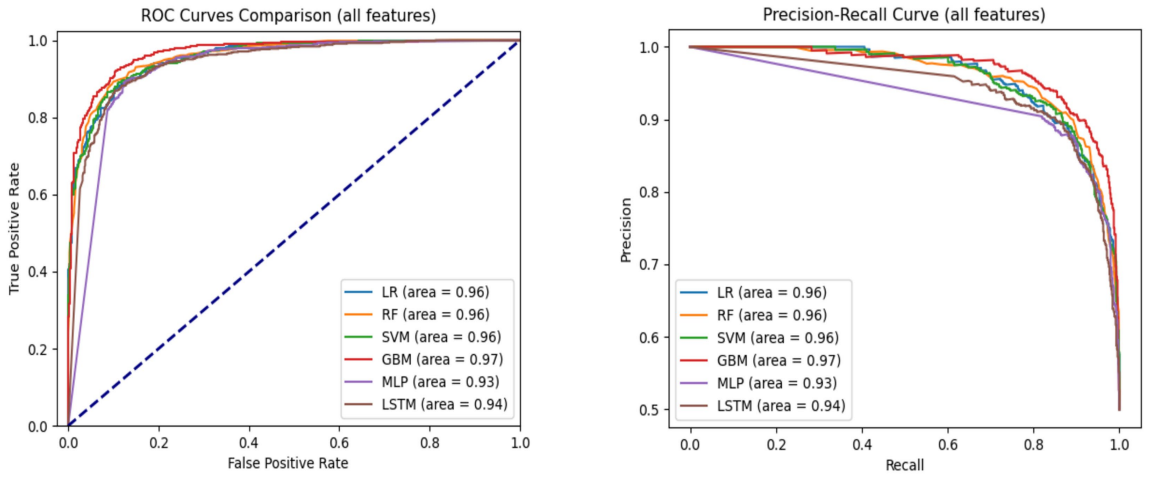

FIGURE 2
AUC-ROC and precision-recall curves comparing model performance for hypertension mortality prediction on the test set using all features.

AUC = 0.96). The PR curves evaluate the balance between precision and recall, where GBM again outperformed other models (AUC = 0.97), followed by LR, RF, and SVM (0.96), LSTM (0.94), and MLP (0.93). These curves indicate that GBM consistently offers the best balance of sensitivity and precision for predicting mortality in hypertensive patients.

Feature importance analysis and feature selection

To address potential issues related to overfitting and model complexity, we conducted a feature importance analysis to guide feature selection. Using SHAP values, we quantified the importance of each feature across all models (LR, RF, SVM, GBM, MLP, and LSTM), based on the 5-fold cross-validation results. Figure 3 presents the top 30 features ranked by average SHAP values. The most influential predictors were Glasgow Coma Scale (GCS) eye-opening response, emergency admission type, anchor age, Braden mobility score, and urea nitrogen level. These features consistently ranked highly across models and are likely key indicators of mortality risk in patients with hypertension. To reduce model complexity and mitigate overfitting, we limited the feature set to these top 30 most important variables for subsequent modeling. To quantify feature importance consistency, we computed Spearman's rank correlation coefficients between feature importance rankings across all model pairs, which ranged from $\rho = 0.76$ to $\rho = 0.91$ (all $p < 0.001$), indicating strong agreement. For the top 10 most important features, the average intersection size across all six models was 8.2 ± 1.3 features, and for the top 30 features, the overlap averaged 25.8 ± 2.4 features. This convergence across diverse algorithmic families—from linear models (LR) to tree-based ensembles (RF, GBM) to neural architectures (MLP, LSTM)—provides strong evidence that these features capture fundamental, algorithm-independent predictive signals associated with mortality risk in hypertensive ICU patients.

Model performance with selected important features

After selecting the top 30 most important features, we retrained and fine-tuned all models using 5-fold cross-validation. These validated models were then evaluated on a held-out 20% test set to assess their generalizability to unseen data. Across models, performance differences between using all features versus the top 30 were minimal—typically under 0.2%—indicating stable model behavior with no signs of overfitting.

Model performance patterns observed with all features were largely preserved when using only the top 30 features during cross-validation. Among all models, the Gradient Boosting Machine (GBM) achieved the highest performance, with an AUROC of 96.0% (95% CI: 95.6, 96.4%) and an F1 score of 89.1% (95% CI: 88.2, 90.0%). Both the

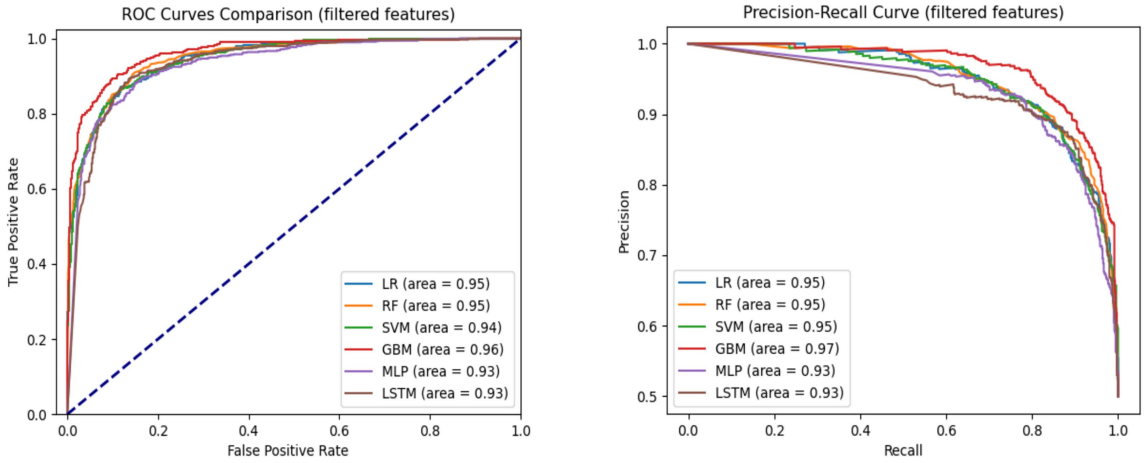

FIGURE 3
AUC-ROC and precision-recall curves comparing model performance for hypertension mortality prediction on the test set using top 30 features.

Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models outperformed most traditional machine learning models, except for GBM. The Logistic Regression (LR) model had the lowest performance across accuracy (86.1%), sensitivity (84.0%), and F1 score (85.8%). The MLP model reported the lowest AUROC (93.2%), while the Random Forest (RF) had the lowest specificity (86.9%). Full cross-validation metrics with 95% confidence intervals are provided in Supplementary Table 2.

Test set evaluation

On the test set, the LR model achieved an accuracy of 86.9%, AUROC of 94.6%, sensitivity of 85.2%, specificity of 88.3%, and F1 score of 86.7%. The RF model slightly outperformed LR, with an accuracy of 87.4%, AUROC of 94.8%, sensitivity of 88.8%, specificity of 86.4%, and F1 score of 87.6%. The Support Vector Machine (SVM) also demonstrated strong performance: accuracy of 87.2%, AUROC of 94.5%, sensitivity of 85.3%, specificity of 88.7%, and F1 score of 87.0%. GBM again delivered the best performance across most metrics, achieving an accuracy of 89.4% and AUROC of 96.3%, with sensitivity and specificity of 87.8 and 90.7%, respectively, and an F1 score of 89.2%. In contrast, the deep learning models saw a slight decline in test performance. The MLP achieved 86.0% accuracy, 93.2% AUROC, sensitivity of 86.4%, specificity of 85.7%, and an F1 score of 86.0%. LSTM performed better than MLP, with 87.4% accuracy, 92.7% AUROC, sensitivity of 87.5%, specificity of 87.3%, and an F1 score of 87.4%. These results suggest that while deep learning models performed competitively in cross-validation, they did not surpass traditional machine learning models on the held-out test set. Test set results are summarized in Table 3.

Figure 3 presents the AUC-ROC and Precision-Recall curves for each model using the top 30 input features. GBM achieved the highest AUROC (0.96), followed by LR and RF (0.95), and SVM (0.94). In terms of Precision-Recall curves, GBM also led with an area under the curve (AUC) of 0.97, while LR, RF, and SVM followed closely at 0.95, and MLP and LSTM at 0.93. These results reinforce GBM's superiority in predicting hypertension-related mortality, even with a reduced feature set.

Comparative analysis

We conducted a comparative analysis between models trained on all available features and those trained on the top 30 selected features. The results demonstrate that reducing the number of input features from over 400 to just 30 did not significantly degrade performance for most models. This finding underscores the effectiveness of our feature selection process.

To statistically validate the consistency of feature selection across models, we computed Kendall's tau correlation coefficients between the feature importance rankings of all model pairs (Bolboaca and Jäntschi, 2006). The correlations ranged from $\tau = 0.72$ to $\tau = 0.89$ (all $p < 0.001$), indicating strong agreement in feature prioritization across diverse algorithms. Additionally, we calculated the intersection of top-k features ($k = 10, 20, 30$) across models. For the top 10 features, there was an average overlap of 8.3 ± 1.2 features across all six models, and for the top 30 features, the average overlap was 24.7 ± 2.1 features. This high degree of convergence demonstrates that our SHAP-based feature selection captured robust, model-agnostic predictive signals rather than algorithm-specific artifacts. Furthermore, to assess whether performance differences between using all features versus the top 30 were statistically significant, we conducted paired t-tests on the cross-validation AUROC scores. None of the six models showed significant performance degradation when using the reduced feature set (all $p > 0.05$), with the largest mean AUROC difference being only 0.015 for the LSTM model. This statistical evidence confirms that the feature selection process effectively eliminated redundant or weakly predictive variables without compromising model performance.

Across all evaluation metrics—accuracy, sensitivity, and F1 score—every model outperformed the logistic regression (LR) baseline. However, regarding AUROC, the two deep learning models showed a slight decline in performance relative to the others. Notably, performance trends emerged when grouped by model type. Tree-based models such as Random Forest (RF) and Gradient Boosting Machine (GBM) exhibited a stronger ability to capture complex patterns associated with hypertension-related mortality while simultaneously mitigating overfitting. This advantage is likely attributable to their hierarchical structure and ensemble learning

TABLE 3 Summary of model performance for hypertension mortality prediction on the test set using the top 30 selected features.

Models\Metrics	Accuracy, 95% CI	AUROC, 95% CI	Sensitivity, 95% CI	Specificity, 95% CI	F1 Score, 95% CI
Logistic regression	0.869 [0.852, 0.887]	0.946 [0.935, 0.958]	0.852 [0.833, 0.871]	0.883 [0.866, 0.900]	0.867 [0.849, 0.886]
Random forest	0.874 [0.857, 0.892]	0.948 [0.938, 0.959]	0.888 [0.872, 0.905]	0.864 [0.846, 0.882]	0.876 [0.858, 0.894]
Support vector machine	0.872 [0.855, 0.890]	0.945 [0.934, 0.956]	0.853 [0.835, 0.872]	0.887 [0.870, 0.904]	0.870 [0.850, 0.888]
Gradient boost machine	0.894 [0.878, 0.910]	0.963 [0.955, 0.972]	0.878 [0.861, 0.895]	0.907 [0.892, 0.922]	0.892 [0.875, 0.910]
Multi-layer perceptron	0.860 [0.842, 0.878]	0.932 [0.919, 0.945]	0.864 [0.845, 0.882]	0.857 [0.839, 0.876]	0.860 [0.840, 0.880]
Long short-term memory neural network	0.874 [0.856, 0.891]	0.927 [0.913, 0.941]	0.875 [0.858, 0.893]	0.873 [0.855, 0.89]	0.874 [0.855, 0.892]

strategy. These trends were consistent across both five-fold cross-validation and test set evaluations. Among all models, GBM achieved the highest performance across most metrics, surpassing even its tree-based counterpart, RF. This highlights GBM's robust capability to distinguish between patients who survived and those who did not. RF also demonstrated substantial improvements over the baseline, ranking second among the six models evaluated.

Model explainability

To interpret model predictions, we calculated the mean absolute SHAP values across all six trained models, as shown in Figure 4. This analysis identified the top 30 features most predictive of hypertension-related mortality. Among these, 21 were laboratory variables (e.g., GCS - Eye Opening, Braden Mobility, Urea Nitrogen), 4 were related to admission type (e.g., Emergency, Urgent), 3 were demographic (e.g., anchor age, ICU stay, hospital stay length), 1 was medication-related (5% Dextrose), and 1 was a complication feature (Pneumonia). Interestingly, several top-ranked features belong to specific clinical assessment tools. For example, the Glasgow Coma Scale (GCS), a standard measure of consciousness, was represented by all three components—eye opening, motor response, and verbal response—within the top 30 features, emphasizing its prognostic relevance. Another key group of features were derived from the Braden Scale (BS), used to assess pressure ulcer risk. Five distinct BS sub-scores appeared among the top features: Braden Mobility, Friction/Shear, Nutrition, Sensory Perception, and Moisture. Additional important laboratory markers included RDW, lymphocyte count, anion gap, white blood cell count, and albumin.

For local interpretability, we selected two representative cases from the hypertension cohort and applied SHAP to explain individual predictions using the GBM model—the best-performing model. One case resulted in a positive prediction (mortality), and the other negative (non-mortality). Only features with SHAP contribution scores above the default 0.05 threshold are visualized. In the positive case (Supplementary Figure 1), key contributors to the high predicted mortality risk included two GCS components, three Braden scores, and bicarbonate level. In the negative case (Supplementary Figure 2), despite some mortality-risk indicators (e.g., lymphocyte count and urea nitrogen), protective features—such as non-urgent admission type, Braden Mobility score, and RDW—contributed more strongly to a negative prediction. These case studies reinforce the overall significance of our top 30 features in predicting hypertension-related mortality.

Fairness evaluation

We further assessed model fairness using several debiasing techniques and compared results across models trained on all features versus the top 30. Table 4 presents the Disparate Parity Difference (DPD) and Equal Opportunity Difference (EOD) metrics with 95% confidence intervals for each model under different conditions. The Long Short-Term Memory (LSTM) model consistently demonstrated the best fairness across all scenarios. Without any debiasing, the LSTM achieved a DPD of 0.015 (95% CI: 0.001, 0.068) using all features, and 0.004 (95% CI: 0.001, 0.061) with the top 30 features. Its EOD values were similarly low: 0.008 (95% CI: 0.005, 0.066) for all features and 0.004 (95% CI: 0.005, 0.065) for the reduced feature set.

In contrast, the GBM model exhibited the poorest fairness performance, with DPD values of 0.037 (all features) and 0.029 (top 30), and EOD values of 0.058 (all features) and 0.036 (top 30). Debiasing strategies yielded measurable improvements. For example, correlation removal reduced LR's DPD from 0.032 to 0.028 and SVM's from 0.032 to 0.026. The reweighting method improved both DPD and EOD for SVM and GBM, while threshold optimization proved effective for most models, except RF. These findings highlight the potential of targeted mitigation techniques to enhance algorithmic fairness without sacrificing predictive performance.

Discussion

Interpretation of findings

In this study, we developed and evaluated four machine learning models (Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting Machine) and two deep learning models (Multilayer Perceptron and Long Short-Term Memory) to predict mortality among patients with hypertension using the MIMIC-IV dataset. By extracting a comprehensive set of features—including clinical, demographic, laboratory, medication, and complication-related variables—we aimed to enhance model performance and support clinical decision-making. Our findings demonstrate that the Gradient Boosting Machine (GBM) outperformed other models in terms of accuracy, sensitivity, F1 score, and AUROC. This suggests GBM's superior capability to model the complex, non-linear relationships inherent in clinical data. Notably, tree-based models like GBM and Random Forest also showed resilience against overfitting and were more effective in capturing patterns related to hypertension mortality. These models' ensemble

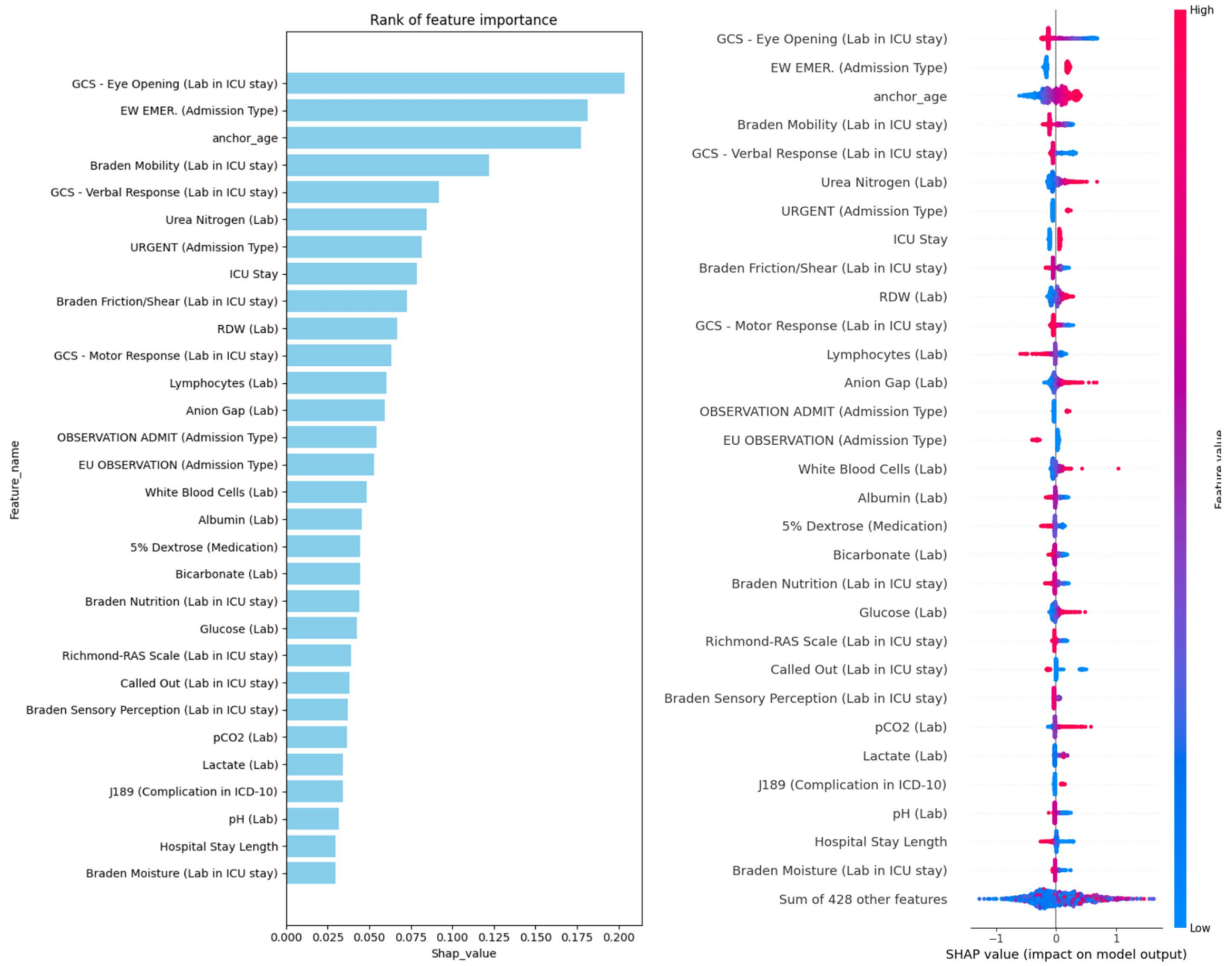

FIGURE 4
Ranking of the top 30 features based on average SHAP values, with information sources specified in the feature names.

learning strategies and hierarchical structure may explain their stronger performance relative to traditional methods and even deep learning models in this context (Zhang et al., 2019).

Importantly, we found that reducing the feature set from over 400 variables to the top 30 features, based on SHAP importance scores, did not significantly degrade model performance. This reinforces the value of our feature selection process and highlights the potential for streamlined, interpretable, and computationally efficient models that could be implemented in real-world clinical settings using EHR data similar to MIMIC-IV. The most predictive features identified, such as GCS scores, Braden Scale scores, blood urea nitrogen, age, and complications like pneumonia, align well with clinical intuition and prior studies (Lan et al., 2021; Wang et al., 2020). For instance, elevated blood urea nitrogen has been shown to predict adverse cardiovascular outcomes, and age is a well-established risk factor for hypertension-related mortality (Ye et al., 2018). Red Cell Distribution Width (RDW) has emerged in recent literature as a marker for mortality in cardiovascular diseases, while variables like lymphocyte count, bicarbonate, and lactate have also been linked to in-hospital mortality in hypertensive or critically ill populations (Perlstein et al., 2009; Fava et al., 2019). Our findings further emphasize the prognostic importance of neurological and functional assessments such as the GCS and Braden Scale. Low GCS scores—particularly in eye, motor, and verbal responses—were strongly associated with higher mortality, consistent with literature on traumatic brain injury and critical illness (Barmparas et al., 2014). Similarly, five Braden sub-scores were negatively associated with mortality risk, echoing findings from COVID-19 mortality studies that associated lower Braden scores with worse outcomes (Lovicu et al., 2021).

Synthesis

Our study specifically focused on mortality prediction in critically ill patients with hypertension rather than the general ICU population. This targeted approach was chosen for several clinical and practical reasons. First, hypertension is one of the most prevalent chronic conditions globally, affecting a substantial proportion of ICU admissions either as a contributing factor to critical illness (e.g., hypertensive emergencies, stroke, acute coronary syndrome) or as an important comorbidity requiring careful management. Second, hypertensive patients in the ICU face unique clinical challenges, including blood pressure management complexities, heightened risk of end-organ damage, and specific medication interactions that may

10.3389/frai.2025.1686378

TABLE 4 Fairness evaluation of models based on demographic parity difference (DPD) and equalized odds difference (EOD).

Metric	All features								Top 30 features							
	DPD				EOD				DPD				EOD			
Models\ Methods	No debiasing, 95% CI	Correlation Removal, 95% CI	Reduction, 95% CI	Threshold Optimization, 95% CI	No debiasing, 95% CI	Correlation Removal, 95% CI	Reduction, 95% CI	Threshold Optimization, 95% CI	No debiasing, 95% CI	Correlation Removal, 95% CI	Reduction, 95% CI	Threshold Optimization, 95% CI	No debiasing, 95% CI	Correlation Removal, 95% CI	Reduction, 95% CI	Threshold Optimization, 95% CI
Logistic regression	0.032 [0.002, 0.084]	0.028 [0.002, 0.081]	0.032 [0.002, 0.083]	0.021 [0.001, 0.073]	0.044 [0.011, 0.096]	0.038 [0.010, 0.086]	0.044 [0.011, 0.097]	0.028 [0.007, 0.080]	0.018 [0.001, 0.071]	0.018 [0.001, 0.073]	0.018 [0.001, 0.071]	0.008 [0.001, 0.062]	0.028 [0.007, 0.079]	0.028 [0.007, 0.082]	0.028 [0.007, 0.084]	0.006 [0.005, 0.068]
Random forest	0.022 [0.001, 0.076]	0.022 [0.001, 0.071]	0.022 [0.001, 0.072]	0.066 [0.015, 0.119]	0.016 [0.006, 0.067]	0.029 [0.007, 0.082]	0.016 [0.005, 0.069]	0.114 [0.055, 0.174]	0.005 [0.001, 0.065]	0.016 [0.001, 0.068]	0.005 [0.001, 0.061]	0.067 [0.013, 0.117]	0.014 [0.005, 0.067]	0.020 [0.006, 0.067]	0.014 [0.006, 0.069]	0.077 [0.048, 0.131]
Support vector machine	0.032 [0.002, 0.080]	0.026 [0.002, 0.081]	0.002 [0.001, 0.060]	0.014 [0.001, 0.066]	0.055 [0.012, 0.104]	0.049 [0.013, 0.100]	0.036 [0.010, 0.086]	0.032 [0.008, 0.080]	0.012 [0.001, 0.066]	0.012 [0.001, 0.07]	0.015 [0.001, 0.069]	0.006 [0.001, 0.063]	0.019 [0.006, 0.077]	0.019 [0.006, 0.074]	0.025 [0.007, 0.078]	0.022 [0.008, 0.079]
Gradient boost machine	0.037 [0.002, 0.091]	0.037 [0.002, 0.093]	0.029 [0.002, 0.081]	0.031 [0.002, 0.083]	0.058 [0.019, 0.105]	0.058 [0.018, 0.108]	0.034 [0.008, 0.080]	0.048 [0.012, 0.097]	0.029 [0.001, 0.082]	0.029 [0.002, 0.084]	0.024 [0.001, 0.077]	0.025 [0.001, 0.077]	0.036 [0.008, 0.085]	0.036 [0.008, 0.091]	0.033 [0.008, 0.086]	0.034 [0.006, 0.085]
Multi-layer perceptron	0.024 [0.002, 0.078]	0.021 [0.001, 0.073]	0.024 [0.002, 0.077]	0.024 [0.002, 0.077]	0.026 [0.007, 0.072]	0.026 [0.008, 0.075]	0.026 [0.007, 0.075]	0.026 [0.007, 0.075]	0.020 [0.001, 0.070]	0.017 [0.001, 0.068]	0.020 [0.001, 0.075]	0.020 [0.001, 0.071]	0.044 [0.013, 0.097]	0.031 [0.008, 0.083]	0.044 [0.013, 0.096]	0.027 [0.007, 0.076]
Long short-term memory neural network	0.015 [0.001, 0.068]	0.014 [0.001, 0.067]	0.015 [0.001, 0.068]	0.012 [0.001, 0.062]	0.008 [0.005, 0.066]	0.016 [0.006, 0.069]	0.008 [0.005, 0.063]	0.006 [0.006, 0.064]	0.004 [0.001, 0.061]	0.006 [0.001, 0.061]	0.004 [0.001, 0.059]	0.013 [0.001, 0.065]	0.004 [0.005, 0.065]	0.017 [0.006, 0.075]	0.004 [0.005, 0.063]	0.006 [0.006, 0.068]

influence mortality risk. By developing a model specifically for this population, we enable more precise risk stratification tailored to the clinical context of managing critically ill hypertensive patients (Zhang and Ye, 2025).

While many of our identified features, such as GCS scores and laboratory markers, are associated with mortality across various patient populations, their relative importance, threshold effects, and interaction patterns may differ in hypertensive versus non-hypertensive cohorts. For instance, the combination of neurological compromise (reflected in GCS) with cardiovascular and renal dysfunction markers may have distinct prognostic significance in hypertensive patients who already face elevated baseline risks for cerebrovascular and cardiovascular events (Wang et al., 2025). Our hypertension-focused model captures these population-specific patterns and provides clinically actionable predictions for a well-defined patient group, facilitating more targeted implementation in clinical workflows where hypertensive patients represent a substantial and readily identifiable subpopulation.

Fairness evaluation

We also examined model fairness through the lens of two standard metrics: Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD). To mitigate potential biases, we applied three debiasing strategies—correlation removal, reduction, and threshold optimization. Overall, the Long Short-Term Memory (LSTM) model demonstrated the best fairness across scenarios, achieving the lowest DPD and EOD scores, especially when trained on the top 30 features. Interestingly, models trained on the reduced feature set generally exhibited better fairness metrics compared to those using all features, suggesting that the feature selection process may inherently reduce biases. While the debiasing methods improved fairness for most models trained on the full feature set, they were less impactful when applied to models using the top 30 features. This pattern suggests that reducing the feature space to the most important variables may inherently support more equitable predictions—potentially by minimizing noise or spurious correlations linked to sensitive attributes.

Comparison with existing models

Compared to prior studies in hypertensive patient risk stratification, our work offers several advancements. First, we leveraged a broader and more comprehensive set of features extracted from the MIMIC-IV dataset, including novel variables such as ICU length of stay and specific complications. Second, we systematically compared multiple machine learning and deep learning models, rather than focusing on a single approach. Third, we incorporated a robust model interpretability framework using SHAP values and conducted an in-depth fairness evaluation—an area often neglected in earlier studies. Moreover, our comparison of models trained on all features versus only the top 30 features revealed minimal performance trade-offs. This not only validates the importance of the identified top predictors but also offers a path toward more efficient and interpretable clinical models. Reducing the feature set can ease the burden of data collection and processing in clinical environments, making deployment more feasible across various EHR systems.

Limitations

Despite the promising results, this study has several limitations that should be acknowledged. First, our analysis was based solely on data from the MIMIC-IV database, which, while comprehensive, represents a single healthcare system and may not generalize to other populations or clinical settings. External validation using data from different institutions, regions, or healthcare systems is necessary to confirm the robustness and generalizability of our models. Additionally, the exclusion of patients with incomplete records could introduce bias, as these patients may have distinct characteristics compared to those included in the study. Second, the definition of hypertension-related mortality was inferred from the available data and may not fully capture the complexity of clinical decision-making or cause of death in critically ill patients. Further refinement in outcome definitions, including cause-specific mortality, could enhance model accuracy and relevance. Third, while we assessed and mitigated bias using fairness metrics and debiasing strategies, our analysis was limited to demographic parity and equalized odds. Other aspects of algorithmic fairness, such as calibration across subgroups or individual-level fairness, were not explored and warrant future investigation (Ascher et al., 2024). Finally, while we identified top predictive features and demonstrated their utility in building more efficient models, clinical validation of these features and how they might be used in practice to guide treatment decisions remain an area for future research (Safiri et al., 2022). Integration into clinical workflows will also require collaboration with clinicians, careful usability testing, and assessments of real-world impact (Ezzati et al., 2015; Ye, 2021).

Future directions

Our findings suggest that machine learning models, particularly GBM, can play a valuable role in predicting hypertension-related mortality and potentially guiding clinical decision-making in intensive care settings (Niu et al., 2021). The identification of interpretable and clinically relevant predictors, combined with fairness-aware modeling strategies, strengthens the potential for safe and equitable integration into clinical workflows (Zhang et al., 2024). Future work should explore the external validation of our models in different hospital systems and populations, as well as real-time deployment feasibility. This study did not account for hypertension severity or degree of blood pressure control, which may influence mortality risk (Brunström and Carlberg, 2018). Hypertension severity can range from mild cases managed with lifestyle modifications to severe, refractory hypertension requiring multiple medications and associated with end-organ complications such as chronic kidney disease (Agarwal et al., 2024). Future studies should incorporate measures of hypertension severity, duration, and control status to better understand how these factors modify mortality risk and whether prediction models should be stratified by disease severity (Echouffo-Tcheugui et al., 2013).

In this study, we assessed and mitigated bias using fairness metrics and debiasing strategies, focusing specifically on group fairness measures—namely, DPD and EOD. While these metrics offer valuable insights, other critical dimensions of algorithmic fairness, such as calibration across subgroups and individual-level fairness, warrant further exploration (Wang et al., 2020). Individual fairness, which ensures that similar individuals receive similar predictions regardless of sensitive attributes, provides a vital complement to group fairness. Incorporating individual fairness constraints can be approached in

several ways: for tree-based models like Random Forest and GBM, fair splitting criteria could be introduced to penalize candidate splits that separate individuals with similar clinical profiles but differing sensitive attributes; for neural networks, similarity-preserving loss functions could be integrated during training. Future research should investigate these individual fairness techniques in conjunction with group fairness metrics to enable a more holistic evaluation of model fairness. Moreover, applying fairness assessments through intersectional lenses (e.g., combinations of race, sex, and age) could help uncover vulnerable subpopulations that may face compounded unfairness. Further studies may also investigate causal relationships between identified features and outcomes, and assess how these models could support or augment clinician judgment at the bedside.

In addition, incorporating clinical notes and unstructured text data through natural language processing could capture rich contextual information unavailable in structured fields, such as medication adherence patterns, symptom descriptions, and clinical reasoning documented in admission and progress notes (Ye et al., 2024). Multimodal prediction models integrating diverse data types, including continuous waveform data (e.g., ECG, arterial blood pressure tracings), medical imaging (e.g., echocardiograms, brain CT scans), time-series physiological signals, and clinical text—could substantially improve risk stratification (Ye et al., 2024; Wang et al., 2024). Hypertensive complications often manifest across multiple modalities: imaging may reveal end-organ damage, continuous monitoring captures blood pressure variability patterns, and notes document clinical trajectories (Zhang et al., 2025). Deep learning architectures combining convolutional networks for images, recurrent networks for time-series, and transformers for text could learn complementary representations that outperform single-modality approaches (Ye et al., 2025). These extensions would leverage the full spectrum of ICU data, moving toward comprehensive decision support tools for managing critically ill hypertensive patients.

Conclusion

This study developed and evaluated multiple machine learning and deep learning models to predict in-hospital mortality among patients with hypertension using the MIMIC-IV dataset. Our findings highlight the strong performance of the gradient boosting machine (GBM) model, which outperformed other approaches in capturing the complex, non-linear relationships inherent in clinical data. By incorporating a broad set of features, including novel variables related to patient complications and ICU stays, and by identifying a reduced set of the most important predictors, we demonstrated that accurate and efficient mortality prediction is achievable using routinely collected EHRs data. Additionally, we evaluated the fairness of our models across demographic subgroups and applied debiasing strategies to mitigate potential unfairness in performance. Our results indicate that feature selection may play a role in enhancing model fairness, offering a promising direction for building more equitable clinical prediction tools. These findings support the potential of interpretable, fair, and efficient machine learning models to assist in clinical decision-making for critically ill patients with hypertension. Future work should focus on external validation, clinical integration, and real-world impact evaluation to ensure these models can meaningfully improve patient outcomes in diverse care settings.

Data availability statement

All data are available at Medical Information Mart for Intensive Care: https://mimic.physionet.org/. The relevant code and analyses are available at: https://github.com/ShenghanZhang1123/hypertension_mortality_pred.

Author contributions

SZ: Formal analysis, Writing – original draft. SD: Methodology, Writing – review & editing. ZX: Writing – review & editing. JY: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1686378/full#supplementary-material

References

Agarwal, A., Mehta, P. M., Jacobson, T., Shah, N. S., Ye, J., Zhu, J. J., et al. (2024). Fixed-dose combination therapy for the prevention of atherosclerotic cardiovascular disease. *Nat. Med.* 30:2371. doi: 10.1038/s41591-024-03128-x

Ascher, S. B., Kravitz, R. L., Scherzer, R., Berry, J. D., de Lemos, J. A., Estrella, M. M., et al. (2024). Incorporating individual-level treatment effects and outcome preferences into personalized blood pressure target recommendations. *J. Am. Heart Assoc.* 13:e033995. doi: 10.1161/JAHA.124.033995

Barmparas, G., Liou, D. Z., Lamb, A. W., Gangi, A., Chin, M., Ley, E. J., et al. (2014). Prehospital hypertension is predictive of traumatic brain injury and is associated with higher mortality. *J. Trauma Acute Care Surg.* 77, 592–598. doi: 10.1097/TA.0000000000000382

Bolboaca, S.-D., and Jäntschi, L. (2006). Pearson versus spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo J. Sci.* 5, 179–200.

Bress, A. P., Anderson, T. S., Flack, J. M., Ghazi, L., Hall, M. E., Laffer, C. L., et al. (2024). The management of elevated blood pressure in the acute care setting: a scientific statement from the American Heart Association. *Hypertension* 81, e94–e106. doi: 10.1161/HYP.0000000000000238

Brunström, M., and Carlberg, B. (2018). Association of blood pressure lowering with mortality and cardiovascular disease across blood pressure levels: a systematic review and meta-analysis. *JAMA Intern. Med.* 178, 28–36. doi: 10.1001/jamainternmed.2017.6015

Chen, H., Simmons, W., Hashish, M., and Ye, J. (2024). Telehealth utilization and patient experiences: the role of social determinants of health among individuals with hypertension and diabetes. *medRxiv* 2024:24311392. doi: 10.1101/2024.08.01.24311392

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., et al. (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access* 7, 150960–150968. doi: 10.1109/ACCESS.2019.2946980

Ding, S., Zhang, S., Hu, X., and Zou, N. (2024). Identify and mitigate bias in electronic phenotyping: a comprehensive study from computational perspective. *J. Biomed. Inform.* 156:104671. doi: 10.1016/j.jbi.2024.104671

Echouffo-Tcheugui, J. B., Batty, G. D., Kivimäki, M., and Kengne, A. P. (2013). Risk models to predict hypertension: a systematic review. *PLoS One* 8:e67370. doi: 10.1371/journal.pone.0067370

Ezzati, M., Obermeyer, Z., Tzoulaki, I., Mayosi, B. M., Elliott, P., and Leon, D. A. (2015). Contributions of risk factors and medical care to cardiovascular mortality trends. *Nat. Rev. Cardiol.* 12, 508–530. doi: 10.1038/nrcardio.2015.82

Fava, C., Cattazzo, F., Hu, Z. D., Lippi, G., and Montagnana, M. (2019). The role of red blood cell distribution width (RDW) in cardiovascular risk assessment: useful or hype? *Ann. Transl. Med.* 7:581. doi: 10.21037/atm.2019.09.58

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* 10:1. doi: 10.1038/s41597-022-01899-x

Lan, Q., Zheng, L., Zhou, X., Wu, H., Buys, N., Liu, Z., et al. (2021). The value of blood urea nitrogen in the prediction of risks of cardiovascular disease in an older population. *Front. Cardiovasc. Med.* 8:614117. doi: 10.3389/fcvm.2021.614117

Lovicu, E., Faraone, A., and Fortini, A. (2021). Admission Braden scale score as an early independent predictor of in-hospital mortality among inpatients with COVID-19: a retrospective cohort study. *Worldviews Evid.-Based Nurs.* 18, 247–253. doi: 10.1111/wvn.12526

Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* 2017:7874. doi: 10.48550/arXiv.1705.07874

Niu, M., Wang, Y., Zhang, L., Tu, R., Liu, X., Hou, J., et al. (2021). Identifying the predictive effectiveness of a genetic risk score for incident hypertension using machine learning methods among populations in rural China. *Hypertens. Res.* 44, 1483–1491. doi: 10.1038/s41440-021-00738-7

Nowbar, A. N., Gitto, M., Howard, J. P., Francis, D. P., and Al-Lamee, R. (2019). Mortality from ischemic heart disease: analysis of data from the World Health Organization and coronary artery disease risk factors from NCD risk factor collaboration. *Circ. Cardiovasc. Qual. Outcomes* 12:e005375. doi: 10.1161/CIRCOUTCOMES.118.005375

Perlstein, T. S., Weuve, J., Pfeffer, M. A., and Beckman, J. A. (2009). Red blood cell distribution width and mortality risk in a community-based prospective cohort. *Arch. Intern. Med.* 169, 588–594. doi: 10.1001/archinternmed.2009.55

Safiri, S., Karamzad, N., Singh, K., Carson-Chahhoud, K., Adams, C., Nejadghaderi, S. A., et al. (2022). Burden of ischemic heart disease and its attributable risk factors in 204 countries and territories, 1990–2019. *Eur. J. Prev. Cardiol.* 29, 420–431. doi: 10.1093/eurjpc/zwab213

Seki, T., Kawazoe, Y., and Ohe, K. (2021). Machine learning-based prediction of in-hospital mortality using admission laboratory data: a retrospective, single-site study using electronic health record data. *PLoS One* 16:e0246640. doi: 10.1371/journal.pone.0246640

Thorsen-Meyer, H.-C., Nielsen, A. B., Nielsen, A. P., Kaas-Hansen, B. S., Toft, P., Schierbeck, J., et al. (2020). Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Health* 2, e179–e191. doi: 10.1016/S2589-7500(20)30018-2

Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., and Roth, G. A. (2022). The global burden of cardiovascular diseases and risk: a compass for future health. *J. Am. Coll. Cardiol.* 80, 2361–2371. doi: 10.1016/j.jacc.2022.11.005

Wang, R., Harper, F. M., and Zhu, H. (2020). *Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences*. in Proceedings of the 2020 CHI conference on human factors in computing systems.

Wang, X., Ren, Z., and Ye, J. (2025). *Predicting survival time for critically ill patients with heart failure using conformalized survival analysis*. AMIA summits on translational science proceedings, p. 576.

Wang, Y., Yin, C., and Zhang, P. (2024). Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon* 10:e26772. doi: 10.1016/j.heliyon.2024.e26772

Wang, C., Yuan, Y., Zheng, M., Pan, A., Wang, M., Zhao, M., et al. (2020). Association of age of onset of hypertension with cardiovascular diseases and mortality. *J. Am. Coll. Cardiol.* 75, 2921–2930. doi: 10.1016/j.jacc.2020.04.038

World Health Organization (2021). Guideline for the pharmacological treatment of hypertension in adults. Geneva: World Health Organization.

Ye, J. (2021). The impact of electronic health record–integrated patient-generated health data on clinician burnout. *J. Am. Med. Inform. Assoc.* 28, 1051–1056. doi: 10.1093/jamia/ocab017

Ye, J., Bronstein, S., Hai, J., and Hashish, M. A. (2025). DeepSeek in healthcare: a survey of capabilities, risks, and clinical applications of open-source large language models. *arXiv* 2025:01257. doi: 10.48550/arXiv.2506.01257

Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., et al. (2018). Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J. Med. Internet Res.* 20:e22. doi: 10.2196/jmir.9268

Ye, J., Hai, J., Song, J., and Wang, Z. (2024). Multimodal data hybrid fusion and natural language processing for clinical prediction models. *AMIA Jt Summits Transl. Sci. Proc.* 2024:191

Ye, J., He, L., Hai, J., Xu, C., Ding, S., and Beestrum, M. (2024). Development and application of natural language processing on unstructured data in hypertension: a scoping review. *medRxiv* 2024:24303468. doi: 10.1101/2024.02.27.24303468

Ye, J., Orji, I. A., Baldridge, A. S., Ojo, T. M., Shedul, G., Ugwuneji, E. N., et al. (2022). Characteristics and patterns of retention in hypertension Care in Primary Care Settings from the hypertension treatment in Nigeria program. *JAMA Netw. Open* 5, e2230025–e2230025. doi: 10.1001/jamanetworkopen.2022.30025

Ye, J., and Ren, Z. (2022). Examining the impact of sex differences and the COVID-19 pandemic on health and health care: findings from a national cross-sectional study. *JAMIA Open* 5:76. doi: 10.1093/jamiaopen/ooac076

Ye, J., and Sanchez-Pinto, L. N. (2020). *Three data-driven phenotypes of multiple organ dysfunction syndrome preserved from early childhood to middle adulthood*. In AMIA annual symposium proceedings American Medical Informatics Association.

Ye, J., Yao, L., Shen, J., Janarthanam, R., and Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med. Inform. Decis. Mak.* 20, 1–7. doi: 10.1186/s12911-020-01318-4

Zhang, B., Ren, J., Cheng, Y., Wang, B., and Wei, Z. (2019). Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm. *IEEE Access* 7, 32423–32433. doi: 10.1109/ACCESS.2019.2902217

Zhang, Z., Wu, Q., Ding, S., Wang, X., and Ye, J. (2024). Echo-vision-FM: a pre-training and fine-tuning framework for echocardiogram videos vision foundation model. *medRxiv* 2024:24315195. doi: 10.1101/2024.10.09.24315195

Zhang, Z., and Ye, J. (2025). Predicting mortality in critically ill patients with hypertension using machine learning and deep learning models. *Front Cardiovasc Med* 12:1568907. doi: 10.3389/fcvm.2025.1568907

Zhang, W., Zhang, Z., He, M., and Ye, J. (2025). Organ-aware multi-scale medical image segmentation using text prompt engineering. *arXiv* 2025:13806. doi: 10.48550/arXiv.2503.13806