Early versus late mortality among fatal trauma cases: a comparative analysis using statistical methods and machine learning algorithms

M. Fevzi Esen,[1] Tutku Tuncalı Yaman[2]

[1]Department of Health Information Systems, University of Health Sciences, İstanbul;
[2]Department of Management Information Systems, Marmara University, İstanbul, Türkiye

Correspondence: M. Fevzi Esen, Department of Health Information Systems, University of Health Sciences, İstanbul, Türkiye.
E-mail: fevzi.esen@sbu.edu.tr

Key words: emergency medicine, mortality, risk assessment, machine learning.

Abstract

Early mortality after trauma is a major concern in emergency care. This study aimed to compare deaths occurring within the first 24 hours with those occurring later using statistical analysis and supervised machine learning, and to assess the effect of class imbalance on model performances. This retrospective study included all consecutive trauma deaths recorded between May and August 2021 in the hospital information system of a tertiary care training and research hospital. The cohort was restricted to patients who died during hospitalization because the aim was to distinguish mortality timing among fatal trauma cases rather than to predict mortality versus survival. Mortality time was classified as death within 24 hours or death after 24 hours from emergency department admission. Clinical, laboratory, and injury-related variables were analyzed using univariate tests. Multiple supervised machine learning models were trained with a stratified 80:20 train/test split and evaluated by AUC. Five resampling techniques were employed to address class imbalance. Hemoglobin and hematocrit differed significantly between early and late mortality groups ($p<0.05$), and injury patterns were associated with mortality timing. In imbalanced data, Support Vector Machine achieved the highest AUC (0.800). After resampling, Gaussian Naive Bayes

combined with ADASYN showed the best performance (AUC=0.847). These findings suggest that mortality timing among fatal trauma cases can be explored using supervised machine learning; however, the retrospective single-center design, modest sample size, class imbalance, and exclusion of survivors limit direct applicability to real-time emergency department triage. The model should therefore be interpreted as an exploratory mortality-timing classifier rather than as a general trauma mortality prediction tool.

Introduction

Trauma remains a leading cause of death and disability worldwide, and a substantial proportion of preventable deaths occur early after injury, during the period when rapid assessment, hemorrhage control, airway management, and prioritization of resources have the greatest impact. Time to death is not uniform: "early" deaths often reflect physiologic collapse and hemorrhage or devastating injury, whereas deaths occurring later may relate to evolving secondary insults, complications, and the downstream consequences of initial injury severity. From an emergency care perspective, mortality timing is clinically relevant because the first 24 hours represent a critical window for triage, hemorrhage control, airway management, operative prioritization, and intensive monitoring.

Several clinical studies have supported the use of the first 24 hours of hospitalization as a clinically meaningful threshold to separate early from late mortality in trauma populations. For example, in pelvic trauma, investigators have highlighted that many deaths occur within the first 24 hours and that predictors of early versus late mortality may differ, reinforcing the clinical value of a 24-hour discrimination point.[1] Similar reasoning has also been applied in wider trauma populations, where early mortality is treated as a separate endpoint. This reflects variations in underlying mechanisms and necessity of early interventions.[2]

Traditional trauma severity scores and admission physiology are fundamental to emergency risk assessment. However, their performance can differ across various settings, and their accuracy may decrease when the outcome is time specific rather than overall in hospital mortality. In parallel, Machine Learning (ML) has been increasingly used to develop prediction models in emergency and trauma care. ML approaches can achieve significant discrimination but also demonstrate variability in development practices, validation strategies,

and reporting quality.[3,4] This has contributed to calls for clearer reporting and more reproducible analytic pipelines in clinical prediction modeling.[5]

A significant challenge in trauma prediction modeling, particularly for time-specific outcomes, is class imbalance. In this situation, one outcome class substantially outnumbers the other. Imbalance can inflate apparent accuracy while masking poor performance for the less frequent outcome, which is often clinically important. Resampling strategies such as synthetic minority oversampling methods have been proposed to mitigate biased learning and improve discrimination for minority outcomes; however, the comparative impact of different balancing methods and classifier families remains context dependent and merits empirical evaluation in real ED trauma datasets.

In this study, we analyzed a cohort of fatal trauma cases using admission demographics, physiological parameters, laboratory values, and injury-related features to distinguish deaths occurring within the first 24 hours from those occurring after 24 hours. The study was not designed to predict mortality versus survival among all trauma admissions; rather, it focused specifically on mortality timing among patients who died during hospitalization. We first compared early and late mortality groups using conventional statistical methods. We then trained and evaluated multiple supervised machine learning classifiers under both imbalanced and balanced training conditions, applying several resampling strategies to assess the influence of class imbalance on model performance. The primary objective was to evaluate whether mortality timing among fatal trauma cases could be differentiated using a reproducible exploratory modeling pipeline.

Materials and Methods

This retrospective single-center study was conducted in the emergency department of a tertiary care training and research hospital after approval was obtained from the Clinical Research Ethics Committee of the University of Health Sciences Kocaeli Training and Research Hospital (approval number: 2021/138; dated December 23, 2021). The dataset was derived from the hospital information system and included all consecutive trauma patients who died during hospitalization between May and August 2021. No eligible trauma death was excluded. The final dataset consisted of 276 observations and 37 features.

Mortality timing was defined using the interval between the recorded time of emergency department admission and the recorded time of in-hospital death. The primary outcome was defined as a binary variable classifying patients as early mortality (death within 24 hours) or late mortality (death after 24 hours). The cohort was restricted to fatal trauma cases because

the objective was not to predict mortality versus survival among all trauma admissions, but to compare and model the timing of death among patients with fatal outcomes. The dataset was imbalanced, with a higher proportion of early deaths compared with late deaths. All analyses were implemented in Python version 3.12.7. Data preprocessing was conducted using pandas version 2.2.3 and NumPy version 1.26.4. Statistical analyses were performed using SciPy version 1.11.4, machine learning models were developed using scikit-learn version 1.3.2, and resampling procedures were implemented using imbalanced-learn version 0.11.0. Visualizations were generated using Matplotlib version 3.8.2.

For numerical variables, missing entries were imputed via K-Nearest Neighbors (KNN) imputation (KNNImputer) with k=5. This method replaces each missing value by the average of the k closest instances in feature space, preserving multivariate relationships. For categorical variables, missing entries were imputed using the mode for each feature. Only three missing values were observed in INR (1.1%), and no other substantial missingness was identified.

Study population and eligibility criteria

The study population consisted of all consecutive trauma patients who died during hospitalization after admission to the emergency department within the study period. Since the dataset was assembled specifically from consecutive trauma deaths, survivors were not included in the analytic cohort. No additional exclusion criteria were applied after identifying eligible fatal trauma cases in the hospital information system. This design allowed a focused comparison of early and late mortality trajectories among fatal trauma cases, but it also means that the resulting models should not be interpreted as tools for predicting death among all trauma patients.

Data source and variables

Data were retrospectively extracted from the hospital information system. The extracted variables included demographic characteristics, trauma mechanism, anatomical injury variables, admission vital signs, laboratory parameters, and trauma severity scores. The recorded time of emergency department admission and the recorded in-hospital time of death were used to calculate mortality timing. The available variables were analyzed as admission-level predictors; longitudinal changes during hospitalization were not included.

Discretization for Bayesian networks

For Bayesian network-based analysis, 20 numerical features were discretized into five bins using quantile-based binning (qcut). This procedure ensures that each bin contains an approximately equal number of observations and facilitates structure learning in the networks. After preprocessing, the dataset was stratified and split into an 80% training set (n=221) and a 20% test set (n=55), maintaining the class proportions of early and late deaths in both subsets. In the training set, 151 deaths occurred within the first 24 hours, while 70 deaths were recorded after 24 hours. In the test set, 32 patients died within 24 hours and 23 patients died after this period. These partitions were consistently used for model training and independent test-set evaluation.

Bayesian Network models were included as probabilistic graphical classifiers to provide an additional comparison with conventional machine learning algorithms. Two structure-learning approaches were evaluated: Hill Climbing Bayesian Network and PC Algorithm Bayesian Network. Because Bayesian Network structure learning requires discrete input states, numerical variables were discretized before model estimation. These models were trained and evaluated using the same training and test partitions as the other classifiers.

Analysis

Descriptive statistics were computed for all variables. To compare categorical variables between patients who died within the first 24 hours and those who died after 24 hours, the chi-square test was used. For numerical variables, distributional assumptions were assessed; as normality assumptions were not satisfied for most variables, between group comparisons were performed using the Mann-Whitney U test. A $p<0.05$ was considered statistically significant.

Supervised classification models were developed to distinguish between death within 24 hours and death after 24 hours. Ten supervised classification models were evaluated: Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), a simple feed-forward Neural Network (NN), Hill Climbing Bayesian Network, and PC Algorithm Bayesian Network. All models were trained on the training set and evaluated on the independent test set. Model performance was quantified using accuracy and the area under the receiver operating characteristic curve (Area Under the Curve, AUC). ROC curves were produced for chosen models to show the trade-off between sensitivity and specificity at various classification thresholds. The 95% confidence interval for the AUC of the best-performing model was estimated using DeLong's method.

Given the single-center design and modest sample size, the machine learning analysis was considered exploratory and comparative rather than intended to produce a definitive clinical prediction model.

Handling class ımbalance

Imbalance in the outcome variable can lead traditional classification algorithms to biased learning, often favoring the dominant class and thereby underperforming in identifying less frequent, yet clinically critical, outcomes. To address this issue, multiple data balancing techniques were employed prior to model training, including SMOTE, ADASYN, Random Oversampling, SMOTE combined with Tomek Links, and SMOTE combined with Edited Nearest Neighbors (ENN). These methods were applied to the training dataset to reduce class imbalance. Subsequently, the same set of classification algorithms was retrained on each balanced dataset to observe how resampling affected the model's performance. Implementation of these strategies expanded the training set to 348 observations, achieving a balanced class distribution (Death after 24 h: 175; Death within first 24 h: 173). Eight classifiers were retrained on each resampled dataset and then evaluated using the original test set. Bayesian Network models were evaluated on the original imbalanced dataset after discretization of numerical variables and were included as probabilistic graphical model comparators. Supplementary materials, Table 1 represents comparative results across these strategies and algorithms. Resampling was applied only to the training dataset after the train-test split. The original test set was not resampled and was used only for independent model evaluation. This approach was used to reduce the risk of information leakage and to preserve the natural class distribution during final performance assessment.

Results

Descriptive findings for categorical trauma-related features and their distribution according to mortality are presented in Table 1.

Trauma mechanisms

The most frequent mechanism of injury was pedestrian traffic accidents, accounting for 39.4% (n=109) of the cases. The analysis revealed a statistically significant relationship between the mechanism of trauma and the time of death. Specifically, pedestrian traffic accident cases demonstrated a distinct mortality pattern compared to other mechanisms

(p=0.012), with a relatively higher proportion of patients (44.0%) remaining alive beyond the first 24 hours before in-hospital death compared to high-velocity injuries. In contrast, in-vehicle traffic accidents and motorcycle accidents were associated with significantly higher rates of rapid mortality. For in-vehicle traffic accidents (n=40), 88.0% of deaths occurred within the first 24 hours (p=0.003). Similarly, motorcycle accidents (n=26) showed an 86.7% early mortality rate (p=0.008). Falls represented 21.73% of the sample and showed a significant association with mortality time (p=0.032), where 68.0% of deaths occurred within the first day. The "Other causes" category did not show a statistically significant difference in mortality time (p=0.078).

Type of trauma and injury severity

The distribution of mortality timing varied significantly across different anatomical injury types. While Head-Neck trauma did not reach statistical significance (p=0.058), severe head trauma (AIS>3) was a significant predictor of mortality timing (p=0.024). Patients with severe head trauma showed a reverse trend compared to other injuries; 62.0% of these patients died after the first 24 hours, suggesting a more prolonged clinical course compared to other trauma types.

Conversely, injuries involving fractures and internal organ damage were strongly associated with early mortality (within 24 hours). Significant associations were observed for Maxillofacial injuries (p=0.013) and Vascular nerve injuries (p=0.009), with early mortality rates of 96.0% and 88.0%, respectively. Orthopedic injuries also consistently pointed toward high early mortality. Vertebral fractures (p=0.022), extremity fractures (p=0.011), and pelvic fractures (p=0.045) were all statistically significant, with early death rates exceeding 87% in these groups. Femur fractures also followed this pattern (p=0.029). Regarding internal trauma, both thoracic and abdominal injuries were significant factors. Thoracic trauma (n=104) was associated with mortality timing (p=0.046), a relationship that became stronger when isolated to severe thoracic trauma (AIS>3; p=0.007). Similarly, abdominal trauma (p=0.048) and severe abdominal trauma (AIS>3; p=0.031) were significantly linked to death within the first 24 hours.

The comparison of clinical characteristics and trauma severity scores between patients who died within the first 24 hours and those who died after 24 hours is presented in Table 2. The analysis revealed a statistically significant difference in hematological markers between the two groups. HGB levels were found to be significantly higher in patients who died within the first 24 hours compared to those who died after 24 hours (p=0.006). HCT values were

significantly elevated in the early mortality group relative to the late mortality group ($p=0.021$). Conversely, no statistically significant differences were observed regarding demographic data or injury severity. Mean age of patients dying within the first 24 hours did not differ significantly from those dying after 24 hours ($p=0.184$). Trauma severity scores, including ISS ($p=0.740$), GCS ($p=0.744$), and RTS ($p=0.647$), were statistically similar between two groups. Furthermore, hemodynamic and metabolic parameters including SBP, MAP, Heart Rate, pH, LAC, and BE showed no significant variance between the groups ($p>0.05$ for all). The length of LHS and SI also did not differ significantly based on the timing of death.

Exploratory Pearson ccorrelation analyses were conducted to examine the associations between mortality timing and selected admission-level physiological and laboratory indicators, using pairwise complete observations. Early mortality was coded as death within the first 24 hours. Most numerical indicators did not show statistically significant correlations with early mortality. APACHE II score ($r=-0.028$; $p=0.768$), Shock Index ($r=-0.031$; $p=0.745$), INR ($r=0.111$; $p=0.253$), RTS ($r=-0.127$; $p=0.186$), GCS ($r=-0.162$; $p=0.091$), MAP ($r=0.031$; $p=0.746$), diastolic blood pressure ($r=0.037$; $p=0.704$), systolic blood pressure ($r=0.023$; $p=0.810$), base excess ($r=-0.149$; $p=0.120$), and bicarbonate ($r=-0.135$; $p=0.159$) were not significantly associated with mortality timing. Only pH showed a weak but statistically significant negative correlation with early mortality ($r=-0.196$; $p=0.040$), indicating that lower pH values were modestly associated with death within the first 24 hours. Inter-variable correlations supported the physiological consistency of the dataset. MAP was strongly correlated with systolic blood pressure ($r=0.976$; $p<0.001$) and diastolic blood pressure ($r=0.985$; $p<0.001$). Shock Index showed a strong negative correlation with MAP ($r=-0.792$; $p<0.001$) and a strong positive correlation with heart rate ($r=0.777$; $p<0.001$). APACHE II score was strongly and inversely correlated with RTS ($r=-0.708$; $p<0.001$) and moderately inversely correlated with MAP ($r=-0.564$; $p<0.001$). In addition, APACHE II score showed a moderate positive correlation with lactate ($r=0.381$; $p<0.001$), consistent with the expected relationship between physiological deterioration and trauma severity.

Figure 1 shows the performance of the classification algorithms, including conventional classifiers and Bayesian Network models trained on the original imbalanced dataset. Among the models, trained without resampling, Multinomial Naive Bayes and Random Forest achieved the highest accuracy (76.4%), whereas Support Vector Machine demonstrated the highest AUC (0.800). These findings indicate that accuracy and AUC did not rank models in the same way, underscoring the importance of using threshold-independent metrics when

evaluating performance under class imbalance. The results also support the need to examine resampling strategies to improve minority class prediction, particularly for deaths occurring after 24 hours. For this reason, ROC curves were generated for further investigation, shown as Figure 2.

Figure 2 presents ROC curves for the models on imbalanced test dataset. SVM achieved the highest AUC value (0.800), indicating strong discriminatory capacity in distinguishing between early and late deaths. LR and GNB also demonstrated AUC scores (0.776 and 0.788, respectively), suggesting reliable performance despite the class imbalance. Some models with lower accuracy still demonstrated relatively high AUCs, highlighting the importance of using threshold-independent metrics such as AUC when evaluating performance on imbalanced datasets. Although the baseline models demonstrated acceptable discriminatory performance, the imbalanced distribution of the outcome variable and the modest sample size required a more cautious evaluation. Therefore, before the final comparative assessment, resampling techniques were applied only to the training set, and the same classifiers were retrained on the balanced training datasets while preserving the original test set for independent evaluation.

Classification results with different balancing techniques

Table 3 highlights the impact of different balancing strategies on predictive accuracy and class separability, illustrating that balancing the dataset significantly improves the performance of most models, particularly in terms of AUC values. ADASYN produced the most favorable outcomes, particularly when combined with the GNB model. The model achieved the highest AUC of 0.847 (95% CI: 0.664–1.000, DeLong method), indicating the best overall discrimination among the evaluated model-resampling combinations.

Figure 3 visually summarizes the top 15 model–resampling combinations ranked by AUC. The GNB model trained with ADASYN showed the highest AUC (0.847) and an accuracy of 81.82% on the independent test set, suggesting comparatively better discrimination within this exploratory analysis. This visualization reinforces that the choice of resampling method critically affected model performance, and that ADASYN, which targets hard-to-classify minority instances, can be particularly effective in imbalanced healthcare datasets.

Results of Gaussian Naive Bayes model

Classification metrics for GNB model trained on the ADASYN-balanced dataset are presented in Table 3. The model demonstrated an overall accuracy of 81.82% on the test set. However, a disparity in class-wise performance is evident: the model achieved strong recall

(0.94) and F1-score (0.86) for predicting early deaths (majority class), but performance was lower for late deaths (minority class), with recall of 0.65 and F1-score of 0.75. These results reflect the ongoing challenge of achieving balanced sensitivity across both classes, even after applying a sophisticated resampling method such as ADASYN.

The confusion matrix showed that the ADASYN–GNB model correctly classified 93.8% of early deaths and 65.2% of late deaths. Conversely, 6.3% of early deaths were misclassified as late deaths, while 34.8% of late deaths were misclassified as early deaths. Accordingly, the model demonstrated higher sensitivity for early mortality than for late mortality, indicating that prediction of later mortality remained comparatively more limited despite favorable overall accuracy and AUC. Figure 4 shows the ROC curve for the GNB model trained with ADASYN. The AUC value of 0.8471 indicates good discrimination between early and late death outcomes. The curve indicates that the model significantly outperforms random chance and exhibits good trade-off attributes between sensitivity and specificity. However, this finding should be interpreted cautiously because the analysis was restricted to fatal trauma cases and class-wise performance remained lower for deaths occurring after 24 hours.

Discussion

This study evaluated early versus later mortality among trauma patients admitted to the emergency department by integrating conventional statistical analyses with supervised machine learning approaches. By focusing on mortality timing within and beyond first 24 hours, the analysis addressed a clinically meaningful endpoint corresponding to the early decision window in trauma care, during which rapid assessment and intervention are most likely to influence survival.[6-7]

It is important to emphasize that this study focused exclusively on fatal trauma cases. Therefore, the analysis does not address the broader clinical question of whether a trauma patient will survive or die. Instead, it evaluates whether the timing of death among patients with fatal outcomes can be differentiated using admission-level variables. This distinction is central to the interpretation of the findings and limits direct translation of the model into a general emergency department triage tool.

Clinical findings and time of mortality

Univariate analyses demonstrated that several clinical and injury-related characteristics differed according to mortality timing. Among numerical variables, hemoglobin and hematocrit levels were significantly higher in patients who died within the first 24 hours,

whereas demographic characteristics and most global severity scores did not differ significantly between early and late death groups. While low hemoglobin has been linked to hemorrhage and negative outcomes in trauma, admission laboratory values may indicate complex physiological conditions shaped by injury mechanism, timing of blood loss, and prehospital interventions, which can differ among populations.[8]

Categorical analyses showed that mortality timing varied according to trauma mechanism and injury pattern. Several injury types, including thoracic, abdominal, and orthopedic injuries, were predominantly associated with deaths occurring within the first 24 hours. In contrast, severe head trauma demonstrated a relatively higher proportion of deaths after 24 hours, suggesting a more prolonged clinical course in this subgroup. Similar distinctions between early and late trauma mortality have been reported previously, supporting the concept that early deaths are often related to immediate physiological collapse, whereas later deaths may reflect secondary injury complications.[6-7]

Correlation analysis provided a more cautious interpretation of admission-level physiological indicators. Most numerical variables, including APACHE II score, Shock Index, INR, RTS, GCS, blood pressure parameters, base excess, and bicarbonate, were not significantly correlated with early mortality. Only pH showed a weak negative association with death within the first 24 hours. However, inter-variable correlations among MAP, systolic and diastolic blood pressure, Shock Index, heart rate, APACHE II score, RTS, and lactate supported the physiological consistency of the dataset. These findings suggest that although several variables are clinically meaningful and internally coherent, their isolated admission-level correlations with mortality timing were limited in this fatal-case-only cohort. These findings are consistent with prior trauma literature identifying hemodynamic instability and metabolic dysregulation as indicators of early adverse outcomes.[9] Importantly, these associations represent statistical relationships and should not be interpreted as causal or as evidence of predictor importance within ML models.

Machine learning findings in context

In this study, machine learning was used as an analytical tool to explore a clinically defined emergency care problem rather than as a purely computational exercise. The machine learning results of this study align with existing trauma and emergency medicine literature demonstrating that predictive performance is strongly influenced by algorithm selection, outcome definition, and class distribution. Systematic reviews of ML in trauma outcomes have highlighted substantial heterogeneity in model development and evaluation practices,

with class imbalance repeatedly identified as a key methodological challenge, particularly for time-defined endpoints such as early mortality.[4]

From a clinical prediction modeling perspective, model performance should be interpreted in relation to sample size, predictor dimensionality, outcome frequency, and validation strategy. Prediction models developed in modest-sized datasets may show unstable performance estimates, particularly when multiple algorithms are compared or when flexible modeling approaches are used. Therefore, internal validation and cautious interpretation are essential when evaluating apparent differences between candidate models.[10]

In the baseline analysis using the imbalanced dataset, SVM achieved the highest AUC, while LR showed comparatively strong accuracy. Similar patterns have been reported in emergency and trauma prediction studies, where classical linear or margin-based models often perform comparably to more complex machine learning algorithms, particularly in datasets with modest sample sizes and limited predictor sets.[11-12] Studies have shown that differences between algorithms are frequently modest and highly dependent on study context rather than algorithmic complexity alone.[12]

A defining feature of the present analysis was the imbalance between early and late deaths, a pattern commonly observed in trauma cohorts where early mortality constitutes the majority of fatal outcomes. This imbalance is particularly relevant for a ≤24 hour mortality endpoint, as deaths occurring after the first day represent a smaller but clinically important subgroup. Methodological studies have shown that class imbalance can inflate overall accuracy while masking reduced sensitivity for the minority outcome if not explicitly addressed.[13-14]

After applying resampling techniques to the training data, discrimination improved across several model resampling combinations. The highest AUC was observed with ADASYN combined with GNB. Although trauma-specific head-to-head comparisons of ADASYN with alternative oversampling methods remain limited, prior machine learning studies in clinical and trauma-related datasets have demonstrated that synthetic oversampling approaches, including SMOTE and ADASYN variants, can improve discrimination in imbalanced settings.[15-17] Despite improved AUC, class-wise performance remained uneven, with lower sensitivity for deaths occurring after 24 hours. This observation is consistent with reports indicating that later mortality may depend on evolving physiological processes and in-hospital events not fully captured by admission-level variables.[18]

Effect of sample size on machine learning performance

The modest sample size is an important limitation of this study and may have influenced model performance. Although several machine learning algorithms were evaluated, more flexible models such as neural networks, random forests, and support vector machines can be sensitive to small sample sizes and may be prone to overfitting or unstable performance estimates. In this context, the observed differences between models should not be interpreted as definitive evidence of algorithmic superiority. Rather, they should be viewed as exploratory comparisons within a single-center dataset. The relatively strong performance of the ADASYN–GNB combination may reflect the interaction between this specific dataset, the class distribution, and the resampling method. Larger multicenter datasets are required to determine whether this performance pattern is reproducible.

Methodological and clinical implications

A key strength of this study is the use of a structured and reproducible analytic framework that integrates standard statistical comparisons with systematic machine learning evaluation under both imbalanced and balanced conditions. By separating univariate statistical associations from predictive modeling, the analysis avoids conflating statistical significance with model-derived importance and provides complementary perspectives on mortality timing.

However, the absence of explicit model interpretability analyses limits conclusions regarding the relative contribution of individual predictors within the machine learning models. As emphasized in recent reporting guidelines for prediction modeling, including TRIPOD-AI, interpretability and transparency are essential for clinical adoption but were beyond the scope of the present work.[5] Consequently, the findings should be regarded as an indication of predictive potential rather than mechanistic explanation. Future studies should validate these findings in larger, multicenter trauma cohorts. Incorporation of longitudinal physiological and laboratory data may improve prediction of later mortality. In addition, applying explainable artificial intelligence approaches, such as SHAP-based feature attribution, may enhance transparency and facilitate clinical interpretation of model predictions.[19]

Clinical applicability of the fatal-case-only design

Restricting the dataset to patients who died during hospitalization limits the clinical applicability of the models for real-time emergency department decision-making. Since survivors were not included, the models cannot estimate the risk of death among all trauma admissions. Instead, they distinguish early versus late death among fatal trauma cases. This

design was chosen because the available dataset consisted of consecutive trauma deaths and because the study objective was to explore mortality timing rather than mortality occurrence. Nevertheless, future studies should include both survivors and non-survivors to develop clinically deployable prediction models for emergency triage. Existing trauma survival prediction studies using machine learning provide a relevant methodological basis for this broader mortality-versus-survival prediction task.[20] A clinically oriented future model may require a two-stage framework: first predicting mortality versus survival among all trauma patients, and then distinguishing early versus late mortality among patients at high risk of death.

This study has several limitations. First, it was retrospective and single-center, which limits generalizability. Second, the sample size was modest for machine learning analysis, particularly for more complex algorithms. Third, the cohort included only patients who died during hospitalization; therefore, the models should not be interpreted as mortality prediction tools for all trauma patients. Fourth, although resampling improved overall discrimination, prediction of deaths occurring after 24 hours remained limited. Finally, the analysis relied on admission-level variables and did not include longitudinal physiological or laboratory trends, treatment variables, or in-hospital complications, which may be especially relevant for later mortality.

Conclusions

This study provides exploratory evidence that early and late mortality timing among fatal trauma cases may be differentiated using admission-level clinical variables, statistical analysis, and supervised machine learning. The findings highlight differences in injury patterns and physiological indicators across mortality timing and demonstrate the influence of class imbalance on model performance. Adaptive oversampling improved overall discrimination; however, prediction of deaths occurring after 24 hours remained limited. Because the cohort was restricted to patients who died during hospitalization, the findings should not be interpreted as a general trauma mortality prediction model. Larger multicenter studies including both survivors and non-survivors are needed before such approaches can be considered for real-time emergency department decision support.

These findings should be interpreted as exploratory and require external validation in larger multicenter cohorts that include both survivors and non-survivors.

References

[1] Wang H, Robinson RD, Moore B, et al. Predictors of early versus late mortality in pelvic trauma patients. Scand J Trauma Resusc Emerg Med 2016;24:27.

[2] Jin WYY, Jeong JH, Kim DH, et al. Factors predicting the early mortality of trauma patients. Ulus Travma Acil Cerrahi Derg 2018;24:532-8.

[3] Naemi A, Schmidt T, Mansourvar M, et al. Machine learning techniques for mortality prediction in emergency departments: a systematic review. BMJ Open 2021;11:e052663.

[4] Zhang T, Nikouline A, Lightfoot D, Nolan B. Machine learning in the prediction of trauma outcomes: a systematic review. Ann Emerg Med 2022;80:440-55.

[5] Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:q902.

[6] Sobrino J, Shafi S. Timing and causes of death after injuries. Proc (Bayl Univ Med Cent) 2013;26:120-3.

[7] Pfeifer R, Tarkin IS, Rocos B, Pape HC. Patterns of mortality and causes of death in polytrauma patients-has anything changed? Injury 2009;40:907-911.

[8] Kauvar DS, Lefering R, Wade CE. Impact of hemorrhage on trauma outcome. J Trauma 2006;60:S3-S11.

[9] Rady MY, Smithline HA, Blake H, Nowak R, Rivers E. Shock index vs conventional vital signs. Ann Emerg Med 1994;24:685-90.

[10] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer; 2019.

[11] Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning vs regression in TBI prognostication. J Clin Epidemiol 2020;122:95-107.

[12] Christodoulou E, Ma J, Collins GS, et al. No performance benefit of ML over logistic regression. BMJ 2019;366:l5398.

[13] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263-84.

[14] Branco P, Torgo L, Ribeiro RP. Predictive modeling on imbalanced domains. ACM Comput Surv 2016;49:31.

[15] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE. J Artif Intell Res 2002;16:321-57.

[16] He H, Bai Y, Garcia EA, Li S. ADASYN. In: Proc Int Joint Conf Neural Netw 2008:1322-1328.

[17] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-Learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Machine Learning Res 2017;18:1–5.
[18] Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure using ML. Nat Med 2020;26:364-73.
[19] Lundberg SM, Lee, SI. A unified approach to interpreting model predictions. Adv Neural Informat Processing Systems 2017;30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
[20] Rau CS, Wu SC, Chuang JF, et al. ML survival prediction in trauma. J Clin Med 2019;8:799.

Table 1. Descriptives of categorical features.

Feature	**n (%)**	**Death after 24h (%)**	**Death within first 24h (%)**	**p**
Trauma Mechanism				
Pedestrian traffic accident	109 (39.40)	44.00	56.00	0.012
In vehicle traffic accident	40 (14.50)	12.00	88.00	0.003
Motorcycle accident	26 (9.42)	13.30	86.70	0.008
Fall	60 (21.73)	32.00	68.00	0.032
Other causes	41 (14.95)	14.70	85.30	0.078
Type of Trauma				
Head-neck trauma	166 (60.14)	52.00	48.00	0.058
Head trauma AIS>3	157 (56.88)	62.00	38.00	0.024
Maxillofacial injury	21 (7.60)	4.00	96.00	0.013
Vascular-nerve injury	25 (9.05)	12.00	88.00	0.009
Vertebral fracture	26 (9.42)	8.00	92.00	0.022
Extremity fracture	43 (15.57)	10.70	89.30	0.011
Pelvic fracture	33 (11.95)	12.40	87.60	0.045
Femur fracture	30 (10.86)	18.00	82.00	0.029
Thoracic trauma	104 (37.68)	36.00	64.00	0.046

Thoracic trauma AIS>3	62 (22.46)	22.70	77.30	0.007
Abdominal trauma	91 (32.97)	29.30	70.70	0.048
Abdominal trauma AIS>3	65 (23.55)	14.70	85.30	0.031

Trauma mechanisms are treated as mutually exclusive categories. Conversely, injury types are non-exclusive to reflect the reality of polytrauma; thus, a patient with multiple injuries (e.g., combined vertebral and extremity fractures) is counted in each relevant category. P-values refer to row-wise exploratory comparisons of the presence versus absence of each categorical feature between mortality timing groups; they were not adjusted for multiple comparisons.

Table 2. Descriptive statistics and group comparisons of numerical clinical features by mortality timing.

Feature	**Mean (SD)**	**Death after 24h (Mean)**	**Death within first 24h (Mean)**	**p**
Age	45.01 (18.22)	47.53	43.88	0.184
Glasgow Coma Scale (GCS)	6.81 (4.82)	6.98	6.75	0.744
Injury Severity Score (ISS)	46.34 (18.55)	45.42	46.69	0.740
Length of Hospital Stay (days)	10.23 (23.11)	8.26	10.26	0.712
Hemoglobin (HGB)	10.89 (2.73)	10.17	11.16	0.006
Hematocrit (HCT)	32.75 (8.22)	31.11	33.36	0.021
pH	7.22 (0.18)	7.24	7.21	0.502
Partial Pressure of CO_2 (PCO_2)	42.81 (12.26)	43.05	42.73	0.719
Bicarbonate (HCO_3)	18.03 (5.47)	18.65	17.80	0.114
Lactate (LAC)	5.27 (3.03)	4.78	5.45	0.123
Base excess (BE)	-9.03 (7.42)	-8.21	-9.33	0.235
Systolic blood pressure	86.02 (39.21)	82.41	87.88	0.368
Diastolic blood pressure	49.52 (25.77)	47.20	50.04	0.555
Mean arterial pressure (MAP)	60.73 (29.20)	58.39	61.60	0.483
Heart rate (Pulse)	117.6 (29.57)	117.05	116.09	0.827
Shock Index	1.78 (1.12)	1.87	1.74	0.381
APACHE II	23.08 (7.88)	23.33	23.52	0.591
Mortality (%)	49.31 (23.35)	50.24	48.96	0.763
Revised Trauma Score (RTS)	4.32 (1.93)	4.39	4.29	0.647
INR	1.21 (0.27)	1.19	1.22	0.697

Values are presented as mean and standard deviation for the overall sample and as group means for patients who died after 24 hours and within the first 24 hours. Between-group comparisons were performed using the Mann–Whitney U test because normality assumptions were not satisfied for most numerical variables. A p-value <0.05 was considered statistically significant. GCS: Glasgow Coma Scale; ISS: Injury Severity Score; LHS: length of hospital stay; HGB: hemoglobin; HCT: hematocrit; PCO_2: partial pressure of carbon dioxide; HCO_3: bicarbonate; LAC: lactate; BE: base excess; MAP: mean arterial pressure; SI: shock index; APACHE II: Acute Physiology and Chronic Health Evaluation II; RTS: Revised Trauma Score; INR: international normalized ratio.

Table 3. Performance Metrics of the Gaussian Naïve Bayes Model

Metric	**Death after 24h**	**Death within 24h**	**Accuracy**	**Macro Avg.**	**Weighted Avg.**
Precision	0.88	0.79	-	0.84	0.83
Recall / Sensitivity	0.65	0.94	-	0.79	0.82
Specificity	0.94	0.65	-	-	-
F1-score	0.75	0.86	-	0.80	0.81
Balanced accuracy	-	-	0.79	-	-
Support	23	32	55	55	55

Values were calculated on the independent test set. Death within 24 hours represents early mortality, whereas death after 24 hours represents late mortality. Sensitivity corresponds to recall for each class. Specificity was calculated by treating each outcome category as the reference class in a one-versus-rest manner. Macro averages represent the unweighted mean of class-wise metrics, while weighted averages account for the number of patients in each class. Resampling was applied only to the training set; the original test set was not resampled.

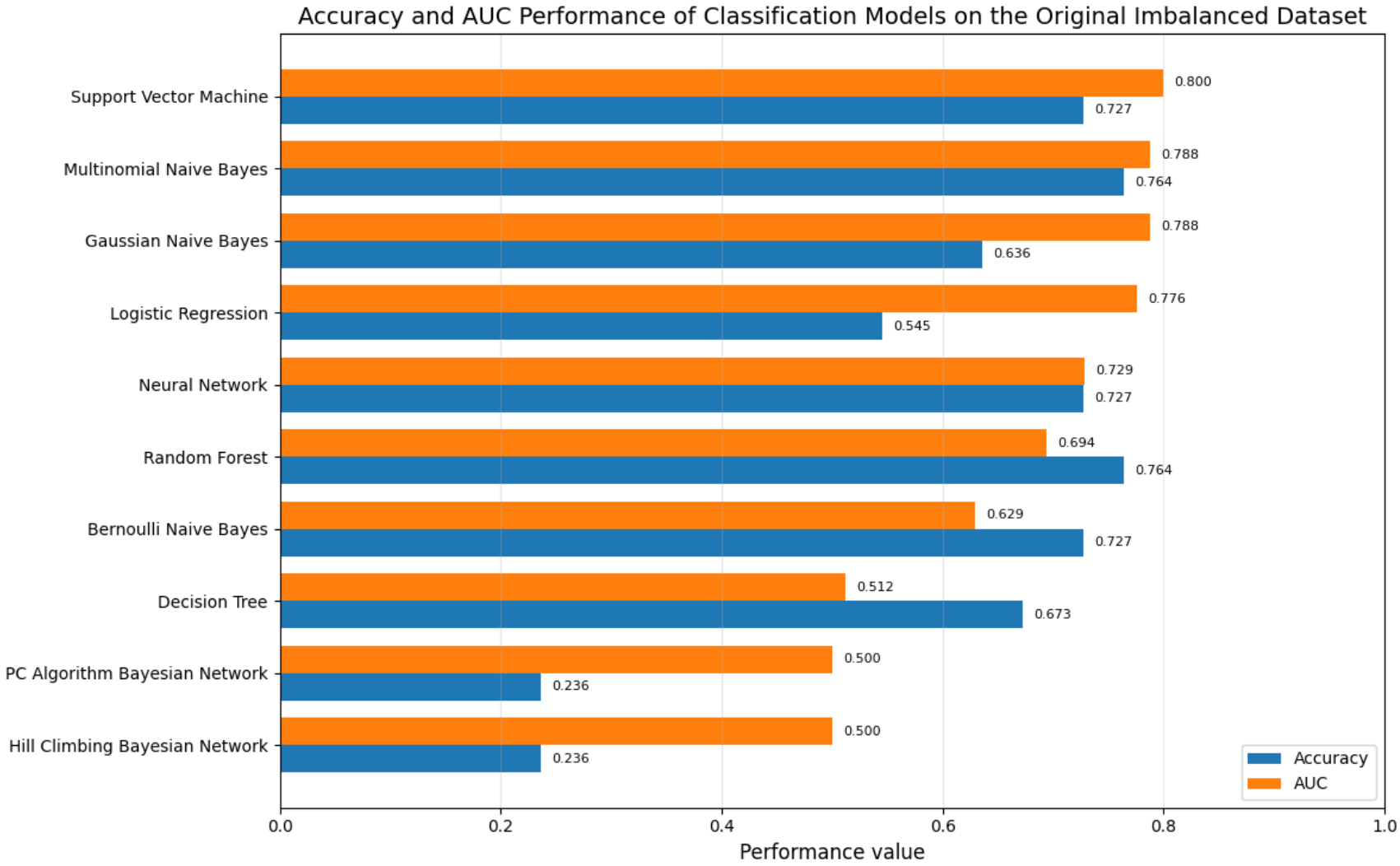

Figure 1. Accuracy and AUC Performance of Classification Models on the Original Imbalanced Dataset

Note. The figure compares the accuracy and AUC values of classification models trained on the original imbalanced training dataset and evaluated on the independent test set. Values correspond to the “None” condition reported in Table 3. Accuracy reflects the proportion of correctly classified cases, whereas AUC reflects threshold-independent discrimination between death within 24 hours and death after 24 hours.

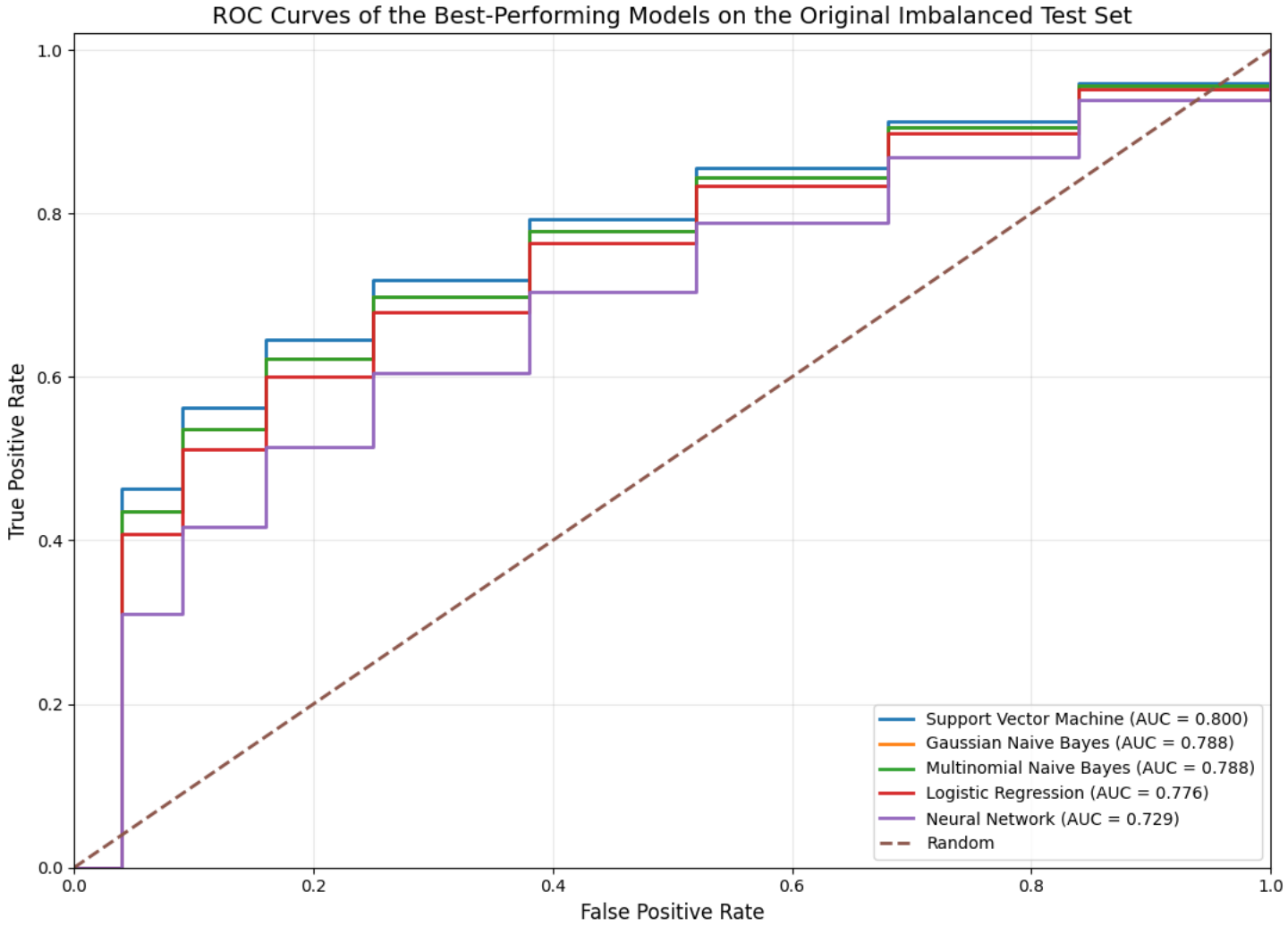

Figure 2. ROC Curves of the Best-Performing Models on the Original Imbalanced Test Set

Note. Receiver operating characteristic curves are presented for the models with comparatively higher discriminatory performance in the baseline imbalanced analysis. AUC values indicate the ability of each model to distinguish between early mortality, defined as death within 24 hours, and late mortality, defined as death after 24 hours.

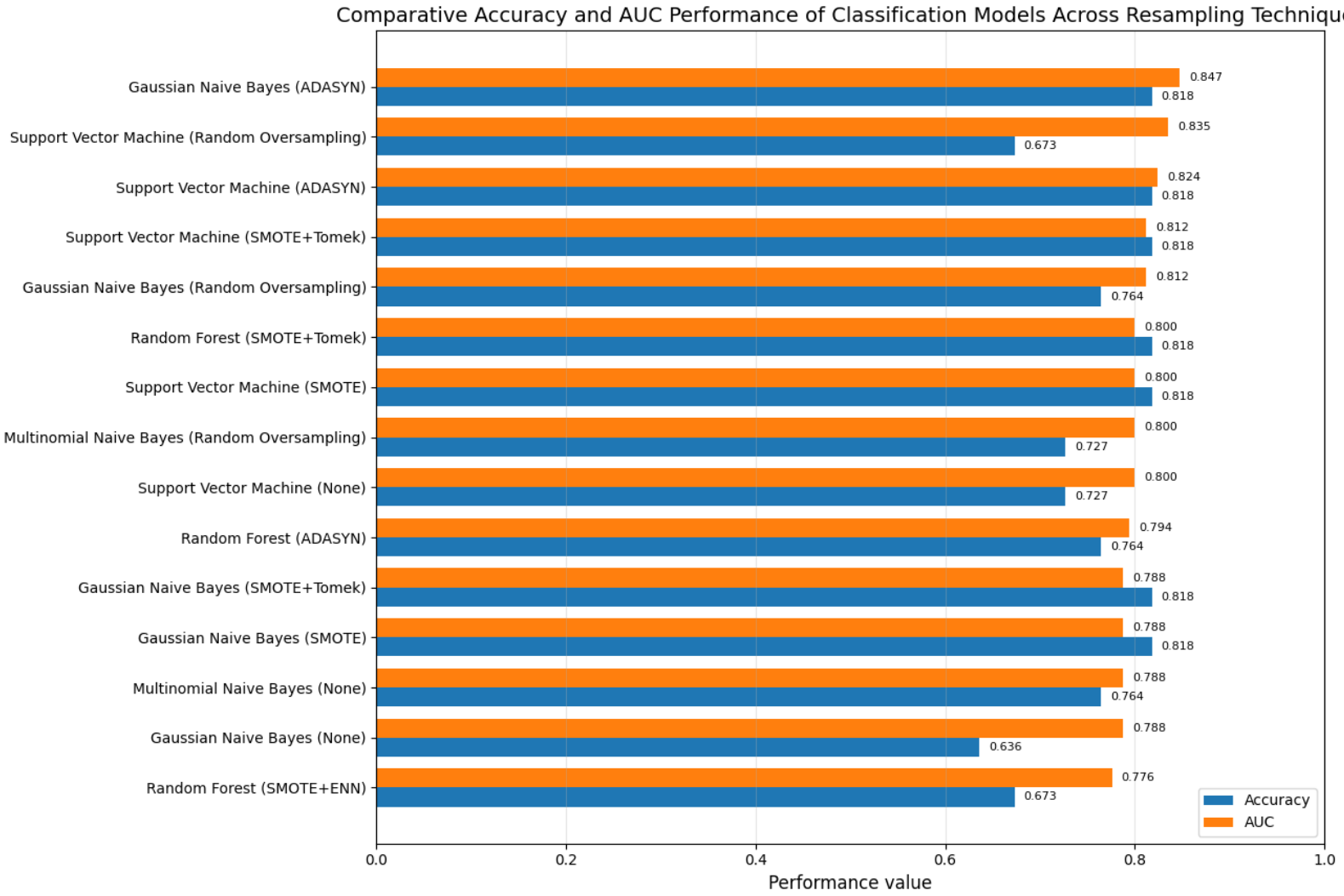

Figure 3. Comparative Accuracy and AUC Performance of Classification Models Across Resampling Techniques

Note. The figure presents the top 15 model–resampling combinations ranked according to AUC values reported in Table 3. Accuracy and AUC were calculated on the independent test set. Resampling techniques were applied only to the training dataset, while the original test set was preserved for model evaluation. The highest AUC was observed for the ADASYN-balanced Gaussian Naive Bayes model.

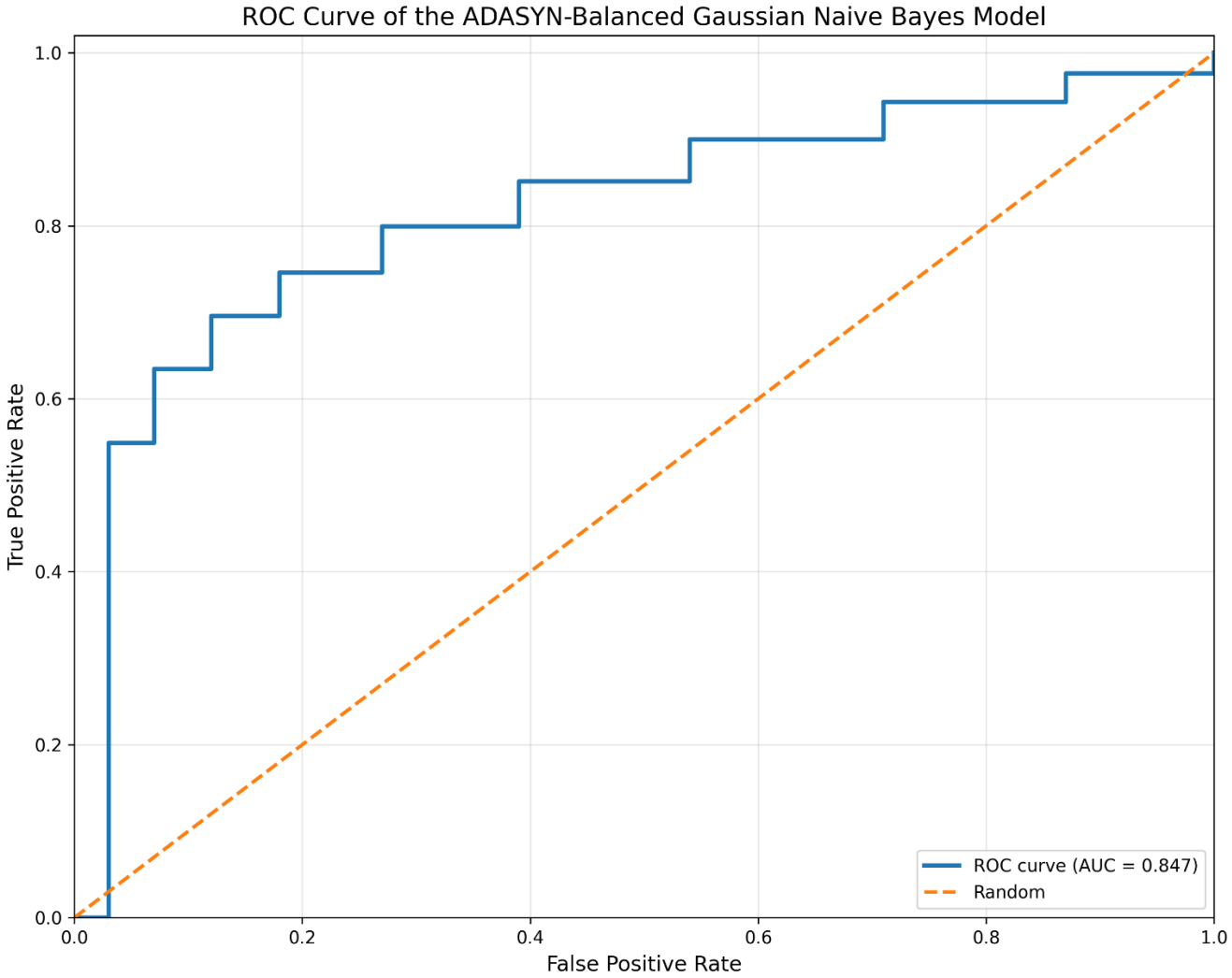

Figure 4. ROC Curve of the ADASYN-Balanced Gaussian Naive Bayes Model

Note. The ROC curve shows the discriminatory performance of the Gaussian Naive Bayes model trained on the ADASYN-balanced training dataset and evaluated on the original independent test set. The AUC value reflects the model's overall ability to distinguish between early and late mortality; however, it should be interpreted together with class-wise sensitivity, specificity, and F1-score because class-wise performance remained uneven.

Contributions: M. Fevzi Esen and Tutku Tuncalı Yaman, conceptualization, formal analysis, investigation, resources, supervision; Tutku Tuncalı Yaman methodology, software, validation; M. Fevzi Esen, data curation, visualization, writing-review & editing; M. Fevzi Esen and Tutku Tuncalı Yaman, writing-original draft preparation; Tutku Tuncalı Yaman, project administration.

Conflicts of interest: The authors declared no conflicts of interest.

Ethics approval: Ethical approval was obtained from the Clinical Research Ethics Committee of the University of Health Sciences Kocaeli Training and Research Hospital (approval number: 2021/138; dated December 23, 2021).

Data availability: Due to privacy and confidentiality concerns, the data are not publicly available. However, they can be made available upon reasonable request to the corresponding author, subject to the completion of a signed data access agreement.

Funding: This research did not receive any funding.

Online supplementary materials

Table 1. Comparison of model performances according to different balancing techniques.