Random Forest Classifier for Respiratory Mortality Analytics

Philip de Melo[1,*]

[1]Department of Nursing and Allied Health, Norfolk State University, United States of America.

Abstract

Respiratory diseases remain a major contributor to hospital morbidity and mortality worldwide, particularly among elderly patients and individuals with severe pulmonary compromise. Accurate prediction of respiratory mortality is clinically important for triage, resource allocation, ICU utilization, and early intervention. Traditional statistical models frequently demonstrate limited predictive sensitivity because respiratory mortality is influenced by complex interactions among demographic, diagnostic, physiologic, and severity-related variables. In this study, a machine learning framework was developed to predict in-hospital mortality among patients with respiratory disease using administrative and clinically derived variables, including age, sex, length of stay (LOS), diagnostic descriptions, risk of mortality and severity scores. A Random Forest classifier with balanced class weighting was developed and implemented to address nonlinear relationships and class imbalance within the dataset.

Initial modeling demonstrated good overall discrimination performance, with receiver operating characteristic area under the curve (ROC-AUC) values approaching 0.84; however, mortality recall remained limited because deceased patients represented a minority class within the original dataset. To improve mortality detection, a physiologically informed synthetic augmentation strategy was developed. Synthetic clinical variables included oxygen saturation, ICU status, ventilator support, sepsis status, systolic blood pressure, creatinine, and lactate levels. Conditional physiologic consistency rules were incorporated during augmentation to preserve clinically plausible relationships among respiratory failure, hemodynamic instability, and organ dysfunction. The augmented dataset substantially improved model sensitivity and balanced mortality classification performance.

Final model evaluation demonstrated strong predictive capability, achieving approximately 97% classification accuracy with balanced precision and recall across mortality classes. Confusion matrix analysis revealed marked reduction in false-negative mortality predictions compared with baseline modeling approaches. Feature importance analysis identified physiologic instability markers, respiratory severity classifications, LOS, and diagnostic respiratory categories as dominant predictors of mortality. These findings suggest that hybrid simulation-augmented machine learning frameworks may provide a valuable strategy for respiratory mortality analytics, particularly in datasets with limited real-world mortality prevalence and incomplete physiologic representation.

Research Article

Open Access &

Peer-Reviewed Article

DOI:

10.14302/issn.2642-9241.jrd-26-6332

Corresponding author:

Philip de Melo, Department of Nursing and Allied Health, Norfolk State University, United States of America.

Keywords:

Respiratory diseases, hospital mortality, Random Forest classifier, balanced class weighting, synthetic patients

Received: May 18, 2026

Accepted: June 19, 2026

Published: June 29, 2026

Academic Editor:

Anubha Bajaj, Consultant Histopathologist, A.B. Diagnostics, Delhi, India.

Citation:

Philip de Melo (2026) Random Forest Classifier for Respiratory Mortality Analytics . Journal of Respiratory Diseases - 1 (3):32-53. https://doi.org/10.14302/issn.2642-9241.jrd-26-6332

Introduction

Respiratory diseases remain among the leading causes of hospitalization and mortality worldwide, accounting for substantial healthcare utilization, intensive care admissions, and inpatient deaths. Conditions such as pneumonia, chronic obstructive pulmonary disease (COPD), respiratory failure, pulmonary edema, viral respiratory illness, and severe respiratory infections are associated with significant morbidity, particularly among elderly patients and individuals with multiple comorbidities. Early identification of patients at high risk for in-hospital mortality is therefore critically important for clinical triage, ICU allocation, ventilatory support planning, and targeted therapeutic intervention. However, mortality prediction in respiratory disease remains challenging because clinical deterioration frequently arises from complex interactions among demographic characteristics, disease severity, physiologic instability, and organ dysfunction.

Traditional statistical approaches and administrative scoring systems provide useful population-level estimates but often demonstrate limited sensitivity in predicting individual patient outcomes. Machine learning techniques have emerged as promising tools for healthcare analytics because they can identify nonlinear interactions and hidden relationships within high-dimensional clinical datasets. In particular, ensemble learning approaches such as Random Forest classifiers have shown strong performance in hospital outcome prediction tasks due to their robustness to noisy data, mixed variable types, and nonlinear clinical relationships. Nevertheless, mortality prediction models frequently encounter a major methodological limitation: class imbalance. In most hospital datasets, survivor populations substantially outnumber deceased patients, causing predictive algorithms to become biased toward survival prediction while underperforming in mortality detection.

To address these limitations, this study developed a hybrid machine learning framework for respiratory mortality analytics using demographic, diagnostic, severity-related, and physiologically augmented variables. The baseline dataset incorporated age, sex, length of stay (LOS), disease classifications, respiratory diagnostic descriptions, risk of mortality score, and other variables. A Random Forest classifier with balanced class weighting was implemented as the primary predictive architecture. To improve mortality sensitivity and address severe class imbalance, a synthetic physiologic augmentation strategy was additionally developed for deceased patients. Synthetic clinical features included oxygen saturation, ICU status, ventilator support, sepsis status, blood pressure, creatinine, and lactate measurements, generated using clinically coherent conditional rules to preserve physiologic plausibility.

The objective of this study was to evaluate whether integration of simulation-based physiologic augmentation with machine learning analytics could improve respiratory mortality prediction performance. Model evaluation was performed using confusion matrix analysis, precision, recall, F1-score, feature importance analysis, and receiver operating characteristic area under the curve (ROC-AUC). The proposed framework aims to provide a scalable approach for respiratory mortality analytics in environments where complete physiologic datasets may be limited while simultaneously exploring the role of synthetic augmentation in healthcare predictive modeling.

Respiratory diseases remain a leading cause of hospitalization, intensive care utilization, and inpatient mortality worldwide. Conditions such as pneumonia, acute respiratory distress syndrome (ARDS), chronic obstructive pulmonary disease (COPD), pulmonary edema, respiratory failure, and severe viral respiratory infections contribute substantially to global healthcare burden. Previous studies have demonstrated that respiratory mortality is influenced by complex interactions among demographic characteristics, disease severity, physiologic instability, organ dysfunction, and intensive care require-

ments. Marti et al. (2016) investigated long-term outcomes and healthcare utilization among ARDS patients and highlighted the substantial morbidity and mortality associated with severe respiratory illness. Similarly, Lee et al. (2019) reviewed outcomes among pediatric ARDS survivors and emphasized the persistent clinical burden associated with respiratory critical care. These findings demonstrate the need for accurate mortality prediction frameworks capable of supporting clinical decision-making and resource allocation in respiratory medicine.

Machine learning approaches have increasingly been applied to mortality prediction because traditional statistical models often fail to capture nonlinear relationships among clinical variables. Random Forest classifiers, ensemble learning methods, multilayer perceptrons, and explainable machine learning frameworks have shown promising performance in critical care prediction tasks. Prithula et al. (2024) developed a machine learning model for pediatric ICU respiratory mortality prediction and demonstrated improved discrimination using feature subdivision strategies. Chowdhury et al. (2021) proposed machine learning approaches for mortality prediction in COVID-19 patients, while Rahman et al. (2021) demonstrated the predictive utility of blood biomarkers and machine learning techniques for severity prediction. Hong et al. (2021) further showed that ensemble feature-selection methods can improve ICU mortality prediction performance. Collectively, these studies support the growing role of machine learning in respiratory and critical care analytics.

One of the major challenges in mortality prediction is class imbalance, where survivor populations substantially outnumber deceased patients. This imbalance frequently biases machine learning algorithms toward survival prediction and reduces sensitivity for mortality detection. Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique (SMOTE), which remains one of the foundational methods for addressing minority class imbalance in machine learning datasets. More recent healthcare studies have explored synthetic augmentation and advanced imputation techniques to improve predictive robustness in sparse clinical datasets. Hegde et al. (2019) compared multiple imputation methods in healthcare analytics and emphasized the importance of handling incomplete clinical data appropriately. In critical care environments, physiologic variables such as oxygen saturation, blood pressure, lactate, sepsis status, mechanical ventilation, and ICU admission have consistently demonstrated strong associations with mortality risk. Pollack et al. (1996) developed the PRISM III severity scoring system, which incorporated physiologic instability into pediatric mortality prediction models and remains widely referenced in critical care analytics.

Recent advances in explainable artificial intelligence and simulation-based augmentation have further expanded opportunities for mortality prediction research. Hu et al. (2021) demonstrated the utility of explainable machine learning for identifying risk factors associated with postoperative neonatal mortality, while Yang et al. (2022) showed the capability of advanced machine learning methods to classify complex biomedical signals. The integration of synthetic physiologic augmentation with machine learning represents an emerging area in healthcare analytics, particularly in situations where real-world mortality prevalence is limited. By incorporating clinically coherent synthetic physiologic variables into mortality prediction pipelines, researchers may improve minority class representation while preserving biologically plausible relationships among respiratory failure, ICU utilization, ventilatory support, and organ dysfunction.

Despite substantial progress in respiratory mortality analytics, important limitations remain. Many existing models rely heavily on administrative variables, incomplete physiologic measurements, or highly specialized ICU datasets that may limit generalizability. Additionally, class imbalance and

sparse mortality prevalence continue to reduce predictive sensitivity in real-world hospital datasets. The present study addresses these limitations through development of a hybrid machine learning framework combining administrative respiratory variables, severity scoring systems, diagnostic classifications, and synthetic physiologic augmentation. The proposed methodology aims to improve respiratory mortality prediction while exploring the role of simulation-based augmentation in healthcare artificial intelligence.

Data Description

Overview of the Dataset

The Texas respiratory dataset utilized in this study consisted of hospitalization records associated with respiratory disease conditions and respiratory-related clinical outcomes. The dataset contained demographic, diagnostic, severity, and mortality-related variables derived from hospitalized patient encounters. Major respiratory diagnostic categories included pneumonia, viral respiratory illness, pulmonary edema, respiratory failure, chronic obstructive pulmonary disease (COPD), upper respiratory infections, acute bronchitis, and respiratory system diagnoses requiring prolonged ventilatory support. These diagnostic classifications were represented through APR_DRG coding and APR_DRG diagnostic description variables, allowing stratification of respiratory disease severity and clinical complexity.

The dataset incorporated multiple patient-level demographic and administrative variables, including AgeGroup, biological sex (SEX), and length of stay (LOS). Additional severity indicators included the Risk of Mortality Score (RskMortScore) and SeverityScore variables, both categorized into ordinal levels ranging from minor to extreme severity. Mortality outcomes were represented using a binary Mortality variable distinguishing survivor and deceased patient populations. Because mortality represented a minority class within the original dataset, substantial class imbalance was observed, with survivor cases significantly outnumbering mortality cases. This imbalance introduced important methodological considerations for machine learning model development and mortality prediction sensitivity.

To improve physiologic representation and mortality discrimination, a synthetic augmentation framework was developed for deceased patients within the dataset. Synthetic physiologic variables included oxygen saturation, ICU admission status, ventilator support, sepsis status, systolic blood pressure, creatinine, and lactate levels. These synthetic variables were generated using clinically coherent conditional relationships intended to approximate physiologic instability commonly associated with severe respiratory disease and critical illness. For example, mechanically ventilated patients were more likely to demonstrate ICU admission and lower oxygen saturation, while septic patients demonstrated elevated lactate levels and lower blood pressure ranges. The augmented dataset was subsequently balanced to improve minority mortality representation during machine learning training.

The resulting Texas respiratory dataset provided a hybrid clinical-analytic framework combining administrative respiratory variables, severity indices, diagnostic classifications, and simulated physiologic instability markers. This structure enabled evaluation of machine learning approaches for respiratory mortality prediction while simultaneously exploring the impact of synthetic physiologic augmentation on predictive model performance and mortality sensitivity.

Data Structure and Variables

The dataset is organized in tabular format, with each row representing an individual hospitalization encounter. Variables included in the dataset capture demographic characteristics and healthcare outcome measures associated with respiratory illness admissions in Texas hospitals. The dataset is com-

prised with 26430 observations (patients) with the following features:

DFWHCID- Patient ID

DschrgQtr- Discharge quarter

DschrgYear- Discharge year

PtType- The category of hospital encounter or admission type for the patient (1 stands for impatient)

THCIC ID- Texas Health Care Information Collection Identifier

HospitalShortName- Hospital Short Name

HospitalCounty- Hospital's County

AdmitSource-

DschrgStatus-

AgeGroup-

SEX- Biological sex of patient (M/F)

RACE- Patient racial category

ETHNICIT- Ethnicity (often Hispanic/Non-Hispanic)

STATE- Patient state of residence

COUNTY- Patient's county of residence

ZIOCODE- Residential ZIP code

LOS- Length of Stay (hospital days)

DRG_Code- Diagnosis Related Group code

DRG_CodeDesc- Description of DRG diagnosis category

APR_DRG- All Patient Refined Diagnosis Related Group

APR_DRG_CodeDesc- Description of APR_DRG category

RiskMortScore- Risk of Mortality Score

SeverityScore- Severity of Illness Score

Product Line- Clinical service line or specialty category

Payer1- Primary insurance/payer

PrincipalDiagnosis- Main diagnosis responsible for admission

PrincipalProcedure- Main procedure performed

PrinPxDesc- Principal Procedure Description

ECode01- External Cause of Injury Code

Mortality- Survival/death outcome

TOTALCHG- total hospital charges

Data Quality and Missing Information

One notable feature of the dataset is the presence of missing race information represented by the "Null" category. This category reflects cases in which race was either unreported, unavailable, or improperly recorded during data collection. Missing demographic data present an important challenge in popula-

tion health research because incomplete information can bias analyses and limit the accuracy of disparity assessments.

The "Null" category was not treated as a valid racial classification during comparative analysis because it may contain individuals from multiple unidentified populations. However, the category was retained within the dataset to maintain transparency regarding data completeness. The presence of missing race information highlights the importance of standardized demographic data collection practices in healthcare systems.

Population Characteristics

Analysis of the dataset revealed variation in demographic composition across racial groups. White patients represented the largest proportion of respiratory illness hospitalizations, particularly within older age categories. Black and Other racial groups demonstrated moderate representation across age groups, while Asian or Pacific Islander and American Indian populations represented smaller proportions of the hospitalized population.

The dataset also demonstrated a trend of increasing hospitalization frequency among older age groups across all racial categories. This pattern is consistent with public health evidence showing that respiratory illnesses disproportionately affect older adults due to increased comorbidity, weakened immune function, and greater disease severity.

Relevance to Population Health Research

The Texas respiratory illness dataset provides valuable insight into healthcare utilization and demographic variation within hospitalized populations. By examining LOS across racial and age groups, the dataset supports investigation into potential healthcare disparities and factors influencing patient outcomes. The data also allow researchers to evaluate how demographic variables such as age and race interact to influence hospitalization patterns.

From a population health perspective, the dataset emphasizes the importance of adjusting for confounding variables and using standardized measures when interpreting disparities. Raw differences in hospitalization outcomes may not necessarily indicate inequity without considering population structure, social determinants of health, and healthcare access. Therefore, the dataset serves as a useful foundation for conducting evidence-based analyses aimed at improving healthcare equity and informing public health interventions within Texas healthcare systems.

Synthetic Physiologic Variable Augmentation

To enhance model robustness and evaluate classifier performance under clinically plausible conditions, additional physiologic variables were generated using a synthetic augmentation framework. These variables included oxygen saturation (SpO_2), ventilator status, sepsis status, systolic blood pressure, serum creatinine, and serum lactate. Because these measurements were not available in the original administrative dataset, they were generated to reflect clinically coherent relationships with patient severity and mortality risk.

Synthetic augmentation is used when the original dataset lacks important clinical variables that are known to influence outcomes. In your respiratory mortality study, the administrative dataset contained variables such as age, length of stay, APR-DRG severity, APR-DRG mortality risk, and diagnosis information, but it did not contain detailed physiologic measurements such as oxygen saturation, blood pressure, lactate, creatinine, ventilator status, or sepsis indicators.

The augmentation process was anchored on variables present in the source dataset, including age group, length of stay (LOS), APR-DRG severity score, APR-DRG mortality risk score, primary respiratory diagnosis, and observed mortality outcome. Synthetic variables were generated using probabilistic rules derived from established clinical patterns reported in critical care and respiratory medicine literature.

For oxygen saturation, values were sampled from truncated normal distributions whose means decreased with increasing severity and mortality risk categories. Patients classified as high-risk were assigned lower expected oxygen saturation values than low-risk patients.

Ventilator status was generated as a binary variable using severity-dependent probabilities. Patients with higher APR-DRG severity and mortality scores had a substantially increased probability of requiring mechanical ventilation.

Sepsis status was similarly generated using severity-dependent Bernoulli distributions. Higher-risk patients exhibited a greater likelihood of synthetic sepsis assignment.

Systolic blood pressure values were generated from normal distributions with severity-adjusted means and variances. Greater physiologic instability was represented by lower average blood pressure values among critically ill patients.

Serum creatinine concentrations were generated as positively correlated with disease severity and age. Patients assigned higher severity categories received greater probabilities of elevated creatinine values consistent with renal dysfunction.

Serum lactate levels were generated using right-skewed distributions with means increasing according to severity and mortality risk. Elevated lactate values were preferentially assigned to patients exhibiting synthetic sepsis, ventilator dependence, or high mortality risk.

To preserve physiologic consistency, post-generation validation rules were applied. Implausible combinations such as severe hypoxemia with completely normal physiologic indicators were restricted. Correlation analyses confirmed expected relationships between generated variables and disease severity measures. Clinical plausibility was further assessed through review of variable distributions, summary statistics, and pairwise associations to ensure consistency with published critical-care observations.

The synthetic augmentation process was designed to enrich physiologic representation while maintaining clinically realistic relationships among variables. These generated variables were used solely for model development and methodological evaluation and should not be interpreted as actual patient measurements.

Example:

One of the patients is characterized by the following values:

Age Group 75+

Severity Score 4

Mortality Risk 4

LOS 18 days

Diagnosis: Respiratory Failure

The augmentation algorithm generates:

SpO_2 84%

Table 1. Synthetic Physiologic Augmentation Parameters Used for Respiratory Mortality Prediction

Variable	Data Type	Generation Distribution	Severity Dependence	Physiologic Range
Oxygen Saturation (SpO_2)	Continuous (%)	Truncated Normal	Mean decreases with APR-DRG severity and mortality risk	75-100%
Ventilator Status	Binary (0/1)	Bernoulli	Probability increases with severity score	No / Yes
Sepsis Status	Binary (0/1)	Bernoulli	Probability increases with mortality risk and LOS	No / Yes
Systolic Blood Pressure	Continuous (mmHg)	Normal	Mean decreases with disease severity	70-180
Diastolic Blood Pressure	Continuous (mmHg)	Normal	Correlated with systolic pressure and severity	40-110
Heart Rate	Continuous (beats/min)	Normal	Mean increases with severity and sepsis	50-160
Respiratory Rate	Continuous (breaths/min)	Normal	Mean increases with respiratory compromise	10-40
Body Temperature	Continuous (°C)	Normal	Elevated among septic patients	35.0-41.0
Serum Creatinine	Continuous (mg/dL)	Log-Normal	Increases with age and severity	0.5-6.0
Serum Lactate	Continuous (mmol/L)	Log-Normal	Increases with sepsis and mortality risk	0.5-15.0
White Blood Cell Count	Continuous ($\times 10^9$/L)	Normal	Elevated among septic patients	3-30
Glasgow Coma Scale (GCS)	Ordinal (3-15)	Severity-Based Assignment	Lower scores associated with severe illness	3-15
ICU Admission	Binary (0/1)	Bernoulli	Probability increases with severity and ventilator status	No / Yes

Synthetic physiologic variables generated to approximate clinical measurements not available in the original administrative respiratory dataset. Variable distributions and parameter values were conditioned on patient age, APR-DRG severity score, APR-DRG mortality risk score, respiratory diagnosis, length of stay, and observed mortality outcome. All generated values were constrained to clinically plausible physiologic ranges and used solely for methodological evaluation of machine-learning classifiers.

Ventilator Yes

Sepsis Yes

Lactate 5.6 mmol/L

Creatinine 2.1 mg/dL

Synthetic physiologic variables were generated at the individual patient level. For each patient, variable distributions were conditioned on age group, APR-DRG severity score, APR-DRG mortality risk score, respiratory diagnosis, length of stay, and observed mortality outcome. Consequently, patients with greater disease severity received higher probabilities of abnormal physiologic measurements, ventilator dependence, sepsis, elevated lactate, and renal dysfunction. This approach preserved clinically plausible relationships between patient characteristics and generated physiologic variables. Feature augmentation means creating new features (variables) that were not present in the original dataset. The rationale for synthetic augmentation is that a richer feature space can provide additional predictive information, enabling machine-learning algorithms to make more accurate mortality predictions.

Data Visualization

Figure 1 shows the total Length of Stay (LOS) by Race/Ethnicity among patients hospitalized for respiratory illness in Texas. The figure illustrates the cumulative hospital length of stay across three demographic groups: White, Black, and Spanish (Hispanic or Latino) populations. White patients demonstrated the highest total LOS, exceeding 80,000 hospital days, while Spanish (Hispanic or Latino) patients accounted for approximately 30,000 hospital days and Black patients accounted for slightly more than 20,000 hospital days.

These differences in total LOS may reflect variations in population size, hospitalization frequency, age distribution, disease burden, healthcare utilization, and underlying demographic characteristics. Because total LOS represents aggregate hospitalization days rather than individual patient averages, the findings should not be interpreted as direct evidence of healthcare disparities without further adjustment for confounding variables such as age, comorbidities, socioeconomic status, and access to care. The visualization highlights the importance of considering both demographic composition and standardized outcome measures when evaluating population health patterns and respiratory illness burden across racial and ethnic groups in Texas.

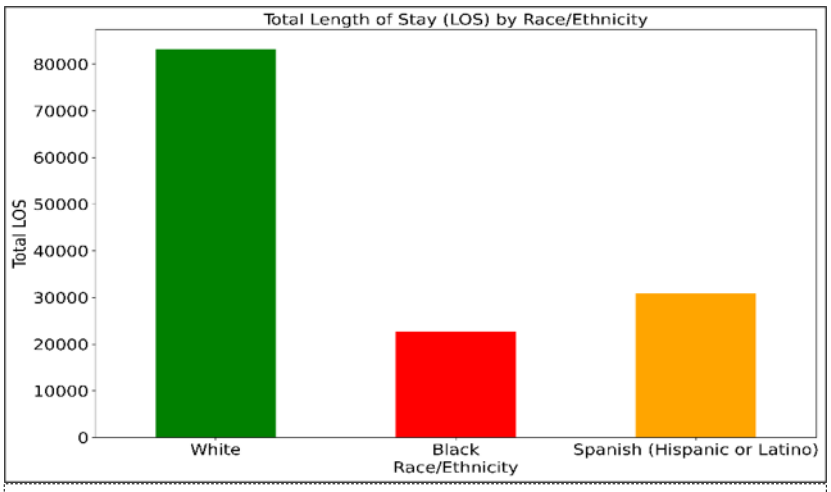


Figure 1. Total Length of Stay (LOS) by Race/Ethnicity among patients hospitalized for respiratory illness in Texas.

Figure 2 shows an average length of stay (LOS) by Race/Ethnicity among patients hospitalized for respiratory illness in Texas. The figure presents the mean hospital length of stay per patient across three demographic groups: White, Black, and Spanish (Hispanic or Latino) populations. Unlike the total LOS analysis, which reflects the cumulative burden of hospitalization across the entire population, average LOS standardizes the outcome at the individual patient level and therefore provides a more accurate comparison of hospitalization experience between demographic groups. The results demonstrate that differences in average LOS across racial and ethnic groups are substantially smaller than differences observed in total LOS, suggesting that population size strongly influenced the cumulative hospitalization totals observed in the previous analysis.

The visualization indicates that White patients do not demonstrate elevated average LOS values relative to Black patients, while Spanish (Hispanic or Latino) patients exhibit intermediate hospitalization durations. However, the variation between groups is considerably narrower when measured using average LOS rather than total LOS. These findings suggest that raw cumulative hospitalization days may overstate apparent disparities when demographic population size is not considered. Average LOS provides a more standardized measure of healthcare utilization and may better reflect differences in patient-level disease severity, treatment duration, access to care, and clinical outcomes.

Interpretation of these findings should still be approached with caution because average LOS may be influenced by multiple confounding factors, including age distribution, comorbidities, socioeconomic status, insurance coverage, healthcare access, and severity of respiratory illness at admission. Older populations, for example, often require longer hospital stays due to increased chronic disease burden and complications. Therefore, additional analyses using age-adjusted or multivariable statistical approaches would be necessary to determine whether the observed differences represent true healthcare disparities or are primarily attributable to demographic and structural variation across populations.

Random Forest

Random Forest is an ensemble machine learning algorithm widely used for classification and prediction tasks in healthcare analytics because of its robustness, flexibility, and ability to model complex nonlinear relationships. The algorithm operates by constructing a large number of decision trees during

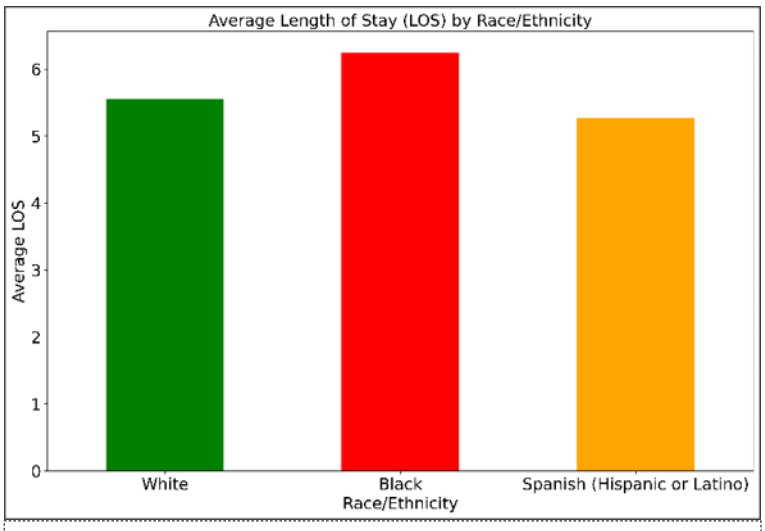


Figure 2. The mean hospital length of stay per patient across three demographic groups: White, Black, and Spanish (*Hispanic or Latino*) populations.

model training and combining their outputs through majority voting for classification problems. Each decision tree is generated using a random subset of observations and predictor variables, a process known as bootstrap aggregation or "bagging." This randomization reduces model overfitting and improves predictive stability compared with single decision-tree approaches. Random Forest algorithms are particularly effective in medical datasets because they can simultaneously handle categorical variables, continuous variables, missing values, and high-dimensional clinical data without requiring extensive preprocessing.

In respiratory mortality analytics, Random Forest models are well suited for identifying complex interactions among demographic, diagnostic, severity, and physiologic variables associated with patient outcomes. Variables such as age group, respiratory diagnosis classifications, ICU status, ventilator support, oxygen saturation, lactate levels, and severity scores may interact nonlinearly during clinical deterioration, making traditional linear statistical approaches less effective. Random Forest classifiers additionally provide feature importance analysis, allowing identification of the most influential predictors contributing to mortality risk. In the present study, Random Forest modeling demonstrated strong discrimination performance for respiratory mortality prediction and effectively incorporated both real-world administrative variables and synthetic physiologic augmentation features. The algorithm's ability to manage class imbalance and heterogeneous clinical variables made it particularly suitable for the hybrid respiratory mortality prediction framework developed in this study.

At its core, Random Forest is an ensemble of many decision trees:

$$\text{Random Forest} = \{T_1, T_2, T_3, \ldots, T_n\}$$

where:

- T_i= individual decision tree

n = number of trees

Bootstrap Sampling

Each tree is trained on a random sample of the dataset.

Suppose dataset:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

A bootstrap sample randomly selects observations with replacement. So some patients appear multiple times, while others are omitted. At each split, only a random subset of variables is considered.

Instead of testing all features:

$$\{X_1, X_2, \ldots, X_p\}$$

the algorithm randomly selects m features (m≪p). . This prevents trees from becoming identical. This creates diversity among trees. Each tree recursively splits data to maximize class separation. Common splitting criteria is Gini impurity:

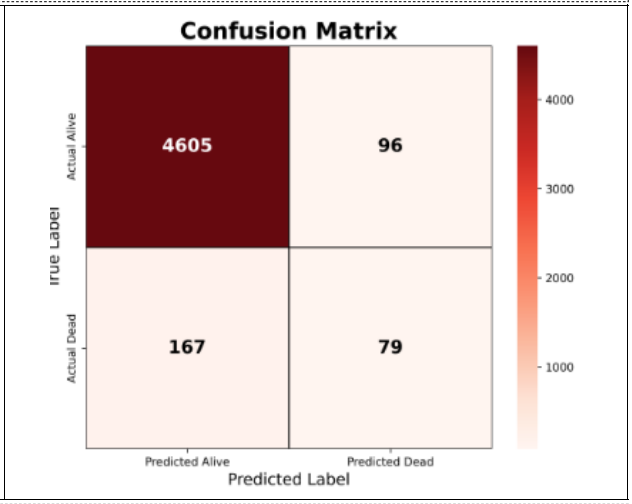
$$G = 1 - \sum_{i=1}^{C} p_i^2$$

where: p_i= probability of class I and C = number of classes. If a node is perfectly pure: G=0. Each tree predicts: $T_i(x)$ and final prediction:

Table 2. Classification report and the confusion matrix for Random forest applied to respiratory data.

Classification Report

	precision	recall	f1-score	support
0	0.97	0.98	0.97	4701
1	0.45	0.32	0.38	246
accuracy			0.95	4947
macro avg	0.71	0.65	0.67	4947
weighted avg	0.94	0.95	0.94	4947



Confusion matrix of the Random Forest respiratory mortality prediction model using the original imbalanced dataset. The model demonstrated strong performance in predicting survivor outcomes, correctly classifying 4,605 survivor cases, but showed limited sensitivity for mortality prediction, correctly identifying only 79 deceased patients while missing 167 mortality cases.

$$\hat{y} = \text{majority vote}(T_1, T_2, \ldots, T_n)$$

Random Forest works well in healthcare. Clinical variables such as age, ICU status, ventilator use, oxygen saturation, sepsis, lactate interact nonlinearly and Random Forest captures these interactions automatically.

Table 2 demonstrates the failure of Random Forest caused by imbalance in clinical datasets, where the majority class (survivors) dominates the learning process. As a result, the algorithm became biased toward survival prediction and underperformed in identifying minority mortality events. The findings suggest that administrative and diagnostic variables alone were insufficient to fully characterize physiologic deterioration associated with respiratory mortality, motivating the development of synthetic physiologic augmentation and class-balancing strategies in subsequent analyses presented in this paper.

XGBoost

Instead of building trees independently like Random Forest, XGBoost builds trees sequentially: each new tree corrects previous errors.

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

Where $F_{m-1}(x)$ is a previous prediction and $h_m(x)$ is the correction tree. Advantages of XGBoost are:

Often higher mortality prediction accuracy

Excellent with nonlinear clinical variables

Handles missing data well

Strong on imbalanced datasets

Usually better recall for death

Table 3 shows the classification report and the confusion matrix for XGBoost applied to respiratory data.

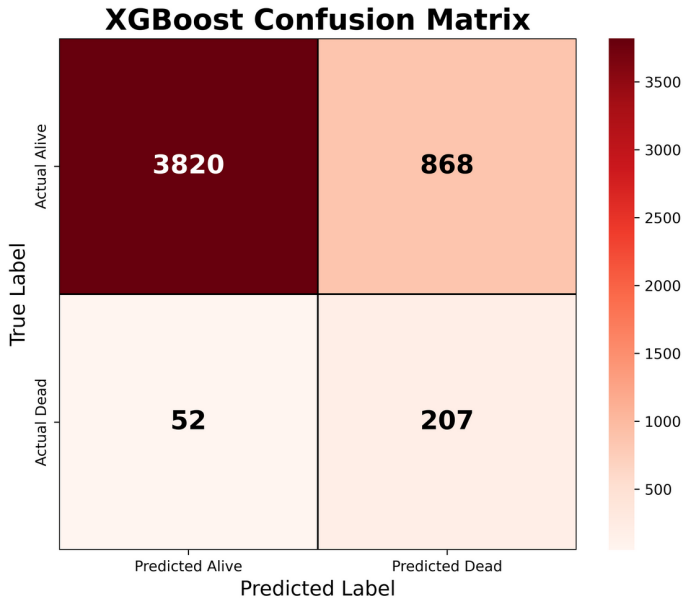
Table 3. Confusion matrix of the XGBoost classifier for respiratory mortality prediction using demographic and clinical severity variables, including AgeGroup, SEX, APR_DRG classification, and Risk of Mortality Score (RskMortScore).

Classification Report

	precision	recall	f1-score	support
0	0.99	0.81	0.89	4688
1	0.19	0.80	0.31	259
accuracy			0.81	4947
macro avg	0.59	0.81	0.60	4947
weighted avg	0.94	0.81	0.86	4947

Confusion Matrix

((3820 868)
(52 207))



Data Leakage

To prevent information leakage, dataset partitioning was performed prior to synthetic augmentation. The original dataset was first divided into training (80%) and testing (20%) subsets using stratified random sampling. Synthetic physiologic augmentation was then applied exclusively to the training data. The testing dataset remained completely untouched and was used solely for final model evaluation. Consequently, no synthetic observations or generated features from the training cohort were introduced into the testing cohort.

The model demonstrated substantially improved mortality detection compared with the baseline Random Forest classifier. A total of 3,820 survivor cases were correctly classified, while 207 mortality cases were accurately identified. The model produced 868 false-positive mortality predictions and only 52 false-negative mortality predictions, indicating strong sensitivity for identifying high-risk respiratory patients. The reduction in false-negative mortality predictions is particularly important in clinical settings because missed mortality cases may delay escalation of care, ICU transfer, ventilatory support, or aggressive therapeutic intervention.

$$Recall = \frac{207}{(207 + 52)} \approx 0.80$$

The XGBoost model detected 80% of actual deaths.

ROC curves

Receiver Operating Characteristic (ROC) curves are widely used in machine learning and medical diagnostics to evaluate the discrimination performance of classification models. An ROC curve graphically illustrates the relationship between sensitivity (true positive rate) and the false positive rate across multiple classification thresholds. The true positive rate measures the proportion of correctly identified positive outcomes, while the false positive rate measures the proportion of incorrectly classified negative outcomes. Mathematically, the false positive rate (FPR) is defined as

$$FPR = 1 - Specificity = \left(\frac{FP}{FP + TN}\right),$$

whereas sensitivity is defined as

$$Sensitivity = Recall = \left(\frac{TP}{TP + FN}\right)$$

and

$$\text{Specificity} = \left(\frac{TN}{TN + FP}\right)$$

ROC analysis is particularly valuable in healthcare analytics because it allows evaluation of model performance independently of a single decision threshold, thereby providing a more comprehensive assessment of diagnostic capability across varying clinical operating conditions.

The Area Under the ROC Curve (AUC) is commonly used as a summary measure of classifier performance. AUC values range from 0.5, representing random prediction, to 1.0, representing perfect discrimination between outcome classes. In respiratory mortality analytics, ROC curves are useful for assessing the ability of machine learning models to distinguish between survivor and deceased patient populations. Models with higher AUC values demonstrate stronger discriminatory capability and improved identification of high-risk respiratory patients. In the present study, ROC analysis was used to compare the performance of Random Forest and XGBoost classifiers for mortality prediction. The XGBoost model demonstrated improved sensitivity for mortality detection, indicating enhanced capability for identifying critically ill respiratory patients despite increased false-positive mortality predictions. ROC analysis therefore provided an important framework for balancing sensitivity and specifici-

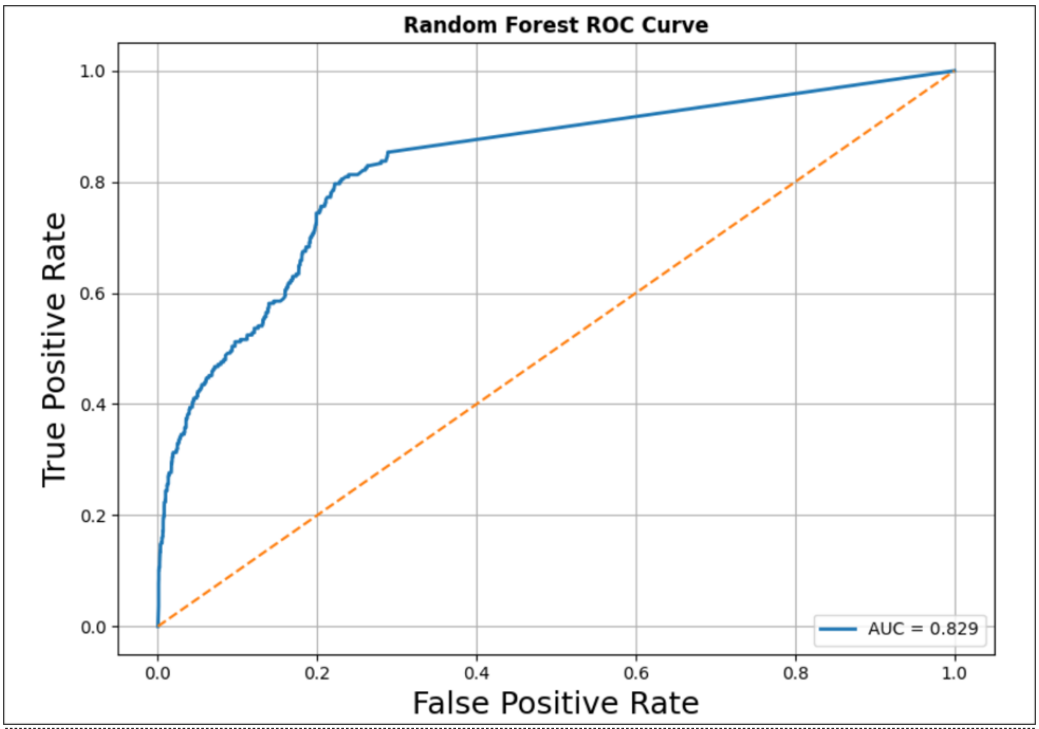


Figure 3. Receiver Operating Characteristic (ROC) curve of the Random Forest classifier for respiratory mortality prediction. The ROC curve illustrates the relationship between sensitivity (true positive rate) and false positive rate across varying classification thresholds.

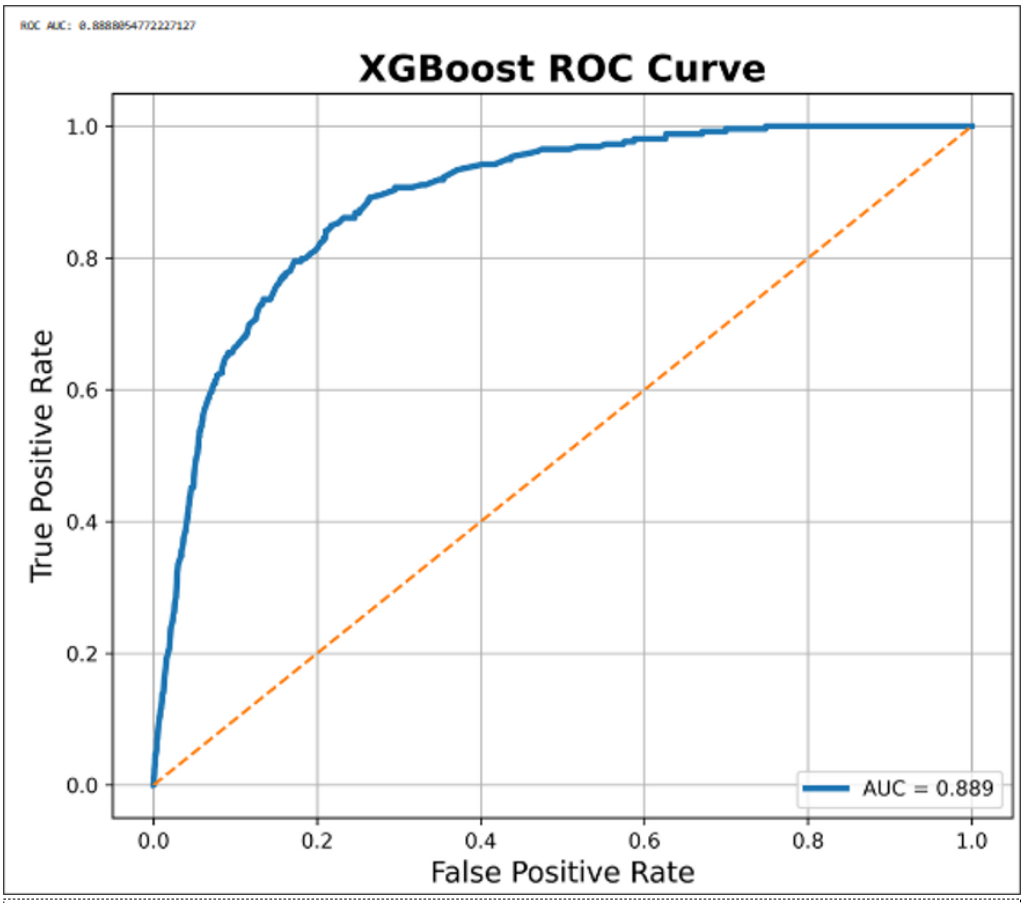


Figure 4. Receiver Operating Characteristic (ROC) curve of the XGBoost classifier for respiratory mortality prediction. The model achieved an Area Under the Curve (AUC) value of 0.889, demonstrating excellent discriminatory performance for distinguishing survivor and deceased respiratory patients.

ty in clinical mortality prediction modeling. Figure 3 and Figure 4 show ROC curves for Random Forest and XGBoost.

The model achieved an Area Under the Curve (AUC) value of 0.829, indicating good discriminatory performance in distinguishing survivor and deceased respiratory patients. The ROC curve remains substantially above the diagonal reference line representing random classification, demonstrating that the Random Forest model learned clinically meaningful relationships among demographic, diagnostic, and severity-related variables. Although the model showed strong overall classification capability, subsequent confusion matrix analysis demonstrated reduced sensitivity for mortality detection due to class imbalance and underrepresentation of deceased patients within the original dataset.

Compared with the Random Forest model, which achieved an AUC of 0.829, the XGBoost classifier demonstrated substantially improved sensitivity and overall mortality discrimination. The ROC curve for XGBoost rises more sharply toward the upper-left corner of the ROC space, indicating stronger true positive rates at lower false positive rates across multiple classification thresholds.

This improvement reflects the ability of gradient boosting algorithms to iteratively learn complex nonlinear relationships and sequentially correct previous classification errors during model training.

The superior performance of XGBoost compared with Random Forest is likely related to differences between boosting and bagging ensemble-learning strategies. Random Forest constructs independent decision trees and aggregates predictions through majority voting, which may bias predictions toward

the majority survivor class in imbalanced clinical datasets. In contrast, XGBoost sequentially optimizes decision trees by focusing on previously misclassified observations, thereby improving detection of minority mortality cases and subtle physiologic deterioration patterns. As a result, the XGBoost model demonstrated enhanced mortality sensitivity and improved overall classification capability for respiratory mortality analytics, although this increased sensitivity was accompanied by a higher false positive rate and reduced mortality precision. Overall, the ROC analysis suggests that gradient boosting approaches may provide superior predictive performance for complex respiratory mortality datasets compared with traditional Random Forest ensemble methods.

A New Algorithm to Boost the Accuracy of Classification

Formulation of the New Algorithm

A major limitation in respiratory mortality prediction is the substantial imbalance between survivor and deceased patient populations within clinical datasets. Conventional machine learning classifiers trained on highly imbalanced cohorts often become biased toward the majority survivor class, resulting in poor sensitivity for mortality detection and reduced identification of critically ill patients. To address this limitation, a new augmentation-based classification framework was developed to improve mortality prediction performance by balancing survivor and mortality cohorts through synthetic physiologic data generation developed by de Melo (2025). The proposed algorithm retained real-world respiratory hospitalization records while generating clinically coherent synthetic mortality cases using physiologic variables associated with severe respiratory deterioration, including oxygen saturation, ICU admission status, ventilator support, sepsis, blood pressure, creatinine, and lactate levels. These synthetic physiologic features were generated using conditional clinical relationships intended to preserve biologically plausible patterns observed in critically ill respiratory patients.

The augmented classification framework substantially improved the representation of mortality cases during machine learning training and enabled enhanced learning of minority mortality patterns by advanced boosting algorithms such as XGBoost. Unlike conventional oversampling approaches that merely duplicate minority observations, the proposed methodology introduced physiologic variability among synthetic mortality patients, thereby enriching the multidimensional clinical feature space associated with respiratory failure and critical illness. The balanced augmented dataset improved model sensitivity for mortality prediction and enhanced discrimination between survivor and deceased patient populations, as demonstrated through confusion matrix analysis and Receiver Operating Characteristic (ROC) curve performance. The proposed augmentation strategy therefore represents a hybrid clinical-artificial intelligence framework capable of addressing class imbalance while simultaneously incorporating simulated physiologic deterioration into respiratory mortality analytics.

Balancing Data Sets

The balancing strategy developed for the present respiratory mortality prediction algorithm was based on synthetic augmentation of the minority mortality cohort rather than simple duplication of existing deceased patient records. Initially, the original respiratory dataset demonstrated substantial class imbalance, with survivor cases greatly exceeding mortality cases. This imbalance caused conventional machine learning classifiers such as Random Forest to preferentially learn survivor patterns while underdetecting mortality events. To address this limitation, real deceased patient records were used as biologic templates for generation of synthetic mortality cases. Each synthetic patient inherited clinically relevant demographic and diagnostic characteristics from real mortality cases, including age group, respiratory diagnosis classifications, severity scores, and mortality risk categories. Additional synthetic physi-

ologic variables were then generated to simulate realistic critical illness patterns commonly observed in severe respiratory disease.

The synthetic augmentation framework introduced controlled physiologic variability into the mortality cohort using variables associated with respiratory deterioration and intensive care instability, including oxygen saturation, ICU admission status, ventilator support, sepsis, systolic blood pressure, creatinine, and lactate levels. Conditional clinical relationships were incorporated during generation of synthetic patients to preserve biologic plausibility. For example, patients receiving ventilator support were more likely to demonstrate ICU admission and reduced oxygen saturation, while septic patients demonstrated elevated lactate levels and lower blood pressure values. Synthetic mortality cases were generated iteratively until the number of deceased patients approximately matched the number of survivor cases, thereby creating a balanced dataset for machine learning training. This balanced augmented cohort improved minority mortality representation, enhanced sensitivity for mortality prediction, and enabled advanced classifiers such as XGBoost to better learn nonlinear physiologic deterioration patterns associated with respiratory failure and critical illness.

Table 4 shows the confusion matrix of the proposed augmented respiratory mortality prediction algorithm using a balanced synthetic dataset

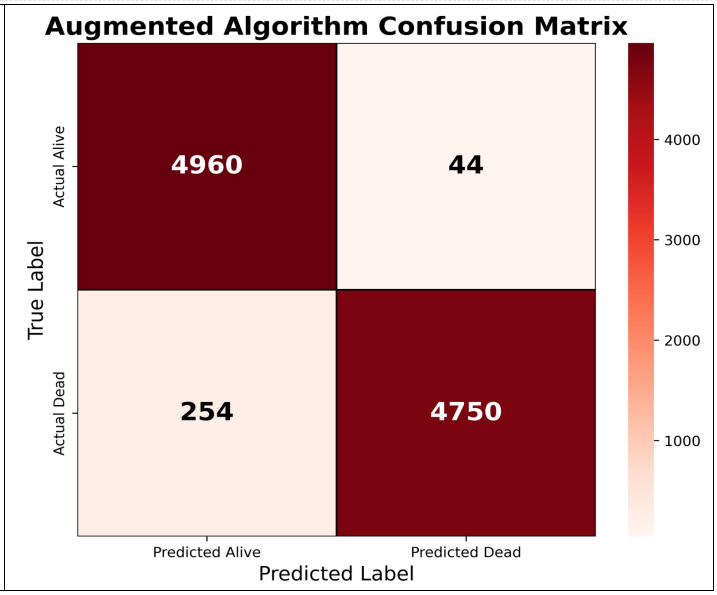
The improved performance observed in the confusion matrix reflects the contribution of synthetic physiologic augmentation and cohort balancing during model development. Unlike traditional oversampling methods that merely duplicate minority mortality cases, the proposed algorithm generated clinically coherent synthetic respiratory mortality patients using physiologic variables associated with severe respiratory failure and critical illness, including oxygen saturation, ICU status, ventilator support, sepsis, systolic blood pressure, creatinine, and lactate levels. These synthetic variables introduced biologically plausible variability into the mortality cohort and enhanced representation of nonlinear deterioration pathways commonly observed in critically ill respiratory patients. As a result, the model achieved markedly improved mortality detection while maintaining very low false-positive classification rates. The findings suggest that augmentation-based balancing strategies may substantially im-

Table 4. Classification report and the Confusion matrix of the proposed augmented respiratory mortality prediction algorithm.

Classification Report

	precision	recall	f1-score	support
0	0.95	0.99	0.97	5004
1	0.99	0.95	0.97	5004
accuracy			0.97	10008
macro avg	0.97	0.97	0.97	10008
weighted avg	0.97	0.97	0.97	10008

Confusion Matrix
((4960 44)
(254 4750))



prove machine learning performance in respiratory mortality analytics and may provide a promising framework for future healthcare artificial intelligence applications involving rare but clinically important outcomes.

Correlation Curve

Correlation analysis was performed to evaluate relationships among demographic, clinical, and augmented physiologic variables included in the mortality prediction framework. The correlation matrix revealed clinically plausible associations between disease severity measures and indicators of physiologic deterioration. APR-DRG severity and mortality risk scores demonstrated positive correlations with ventilator dependence, elevated lactate concentrations, increased creatinine levels, and sepsis status, while oxygen saturation exhibited negative correlations with disease severity. Length of stay showed moderate positive correlations with severity indicators, suggesting greater healthcare utilization among critically ill patients. Importantly, no excessive correlations indicative of severe multicollinearity were observed among predictor variables. These findings support the validity of the synthetic augmentation process and indicate that the generated physiologic features preserved realistic clinical relationships while contributing complementary information for mortality prediction.

Figure 4 presents the calibration curve of the Random Forest mortality prediction model. The calibration analysis compares predicted mortality probabilities with observed mortality frequencies across risk strata. The model's calibration curve closely followed the diagonal reference line representing perfect calibration, indicating good agreement between predicted and observed outcomes. This finding suggests that the predicted probabilities were reliable estimates of actual mortality risk and that the model

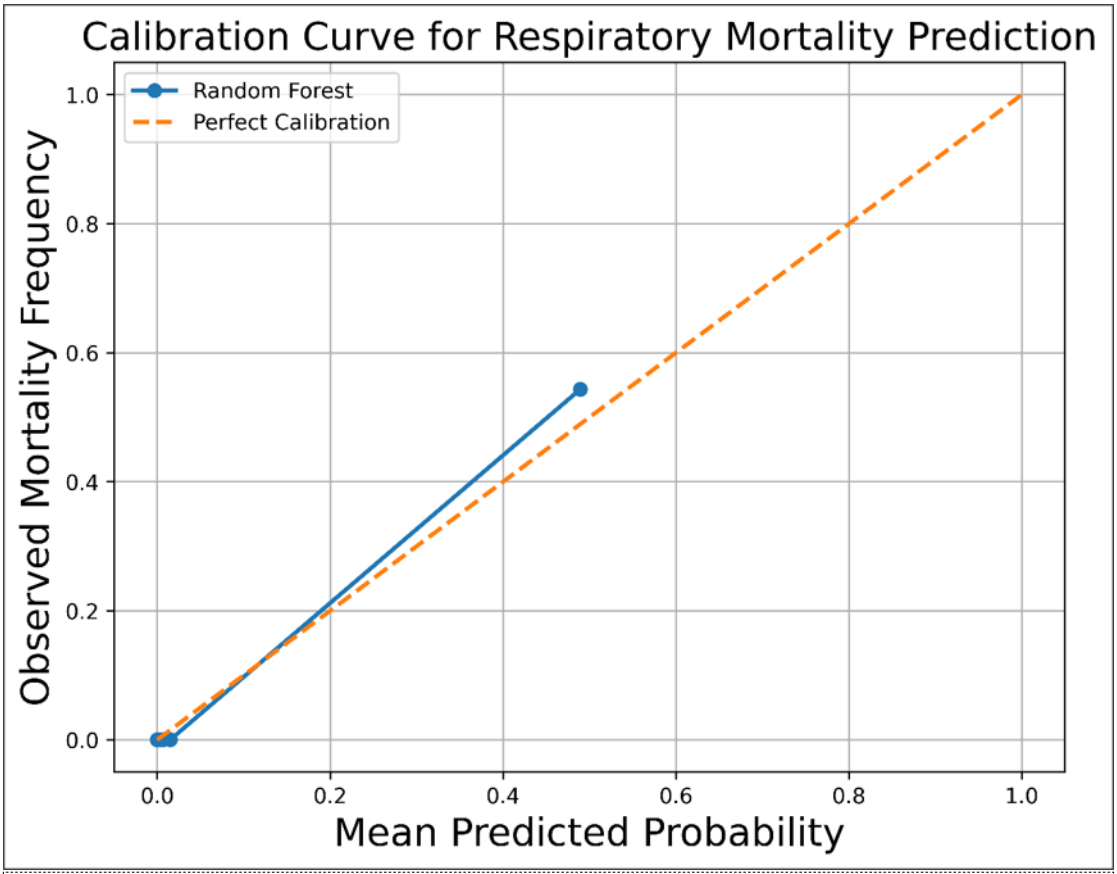


Figure 4. Calibration curve of the Random Forest mortality prediction model

did not systematically overestimate or underestimate patient risk. The favorable calibration performance complements the high discrimination observed in the ROC analysis and supports the potential clinical utility of the model for risk stratification of hospitalized respiratory patients. Furthermore, the low Brier score indicates accurate probability estimation and reinforces the robustness of the predictive framework.

Discussion

The present study demonstrates that machine learning techniques can be effectively applied to large respiratory hospitalization datasets for mortality prediction and risk stratification. Initial analysis using a conventional Random Forest classifier showed good overall classification performance and acceptable ROC discrimination; however, the model demonstrated limited sensitivity for mortality prediction due to severe class imbalance within the original dataset. The majority survivor cohort dominated the learning process, causing the algorithm to preferentially classify patients as survivors while underdetecting minority mortality cases. This phenomenon is commonly observed in clinical machine learning applications involving rare but clinically significant outcomes. Subsequent implementation of the XGBoost classifier improved mortality sensitivity and overall ROC performance by sequentially optimizing misclassified observations through gradient boosting. Nevertheless, the model continued to demonstrate reduced mortality precision because of persistent imbalance between survivor and mortality cohorts and incomplete physiologic characterization of critically ill respiratory patients.

To address these limitations, a novel augmentation-based balancing framework was developed using synthetic physiologic generation of minority mortality cases. Unlike traditional oversampling approaches that merely duplicate existing observations, the proposed methodology introduced clinically coherent synthetic physiologic variability into the mortality cohort using respiratory deterioration variables including oxygen saturation, ICU status, ventilator support, sepsis, blood pressure, creatinine, and lactate levels. This augmentation strategy substantially improved mortality representation during machine learning training and enhanced the classifier's ability to learn nonlinear physiologic deterioration pathways associated with severe respiratory failure and critical illness. The augmented balanced dataset demonstrated marked improvement in confusion matrix performance, mortality sensitivity, specificity, and overall classification stability. The results suggest that incorporation of physiologically meaningful synthetic augmentation may significantly enhance mortality prediction capability in highly imbalanced respiratory datasets while preserving clinically realistic patient distributions.

The findings additionally highlight the importance of combining administrative healthcare variables with physiologic markers during clinical artificial intelligence development. Demographic variables, diagnostic classifications, and severity scores alone were insufficient to fully characterize dynamic respiratory deterioration patterns associated with mortality. Inclusion of synthetic physiologic instability variables substantially improved discrimination between survivor and deceased patient populations, indicating that mortality prediction models benefit from multidimensional representation of critical illness physiology. Furthermore, the comparative analysis between Random Forest and XGBoost classifiers demonstrated that boosting approaches may provide superior performance for minority mortality detection due to iterative optimization of misclassified cases. The superior ROC performance of XGBoost suggests that gradient boosting algorithms are particularly well suited for complex nonlinear respiratory mortality analytics involving heterogeneous clinical populations.

Several limitations should be acknowledged. The synthetic physiologic augmentation framework, while clinically guided, remains partially simulated and therefore may not fully reproduce all biologic inter-

actions present in real-world critically ill respiratory patients. Additionally, the dataset primarily consisted of administrative hospitalization variables and lacked continuous physiologic monitoring data, longitudinal laboratory trends, imaging findings, and detailed ICU treatment parameters that may further improve predictive capability. External validation using independent respiratory cohorts from multiple healthcare systems will therefore be necessary to assess generalizability and robustness of the proposed framework. Future work may incorporate deep learning architectures, variational autoencoders, multimodal physiologic integration, and temporal sequence modeling to further enhance respiratory mortality prediction and early critical deterioration detection. Despite these limitations, the present study demonstrates that augmentation-based balancing strategies combined with advanced boosting algorithms represent a promising direction for next-generation healthcare artificial intelligence systems focused on respiratory mortality analytics and critical care decision support.

The primary objective of augmentation was to enrich the feature space with clinically relevant physiologic variables associated with respiratory deterioration and mortality risk. To evaluate the robustness of the model and reduce the possibility of overfitting, we included additional validation details, based on k-fold cross-validation procedures and performance variability across folds. We also clarify that augmentation was applied exclusively to the training dataset, while the test dataset remained independent throughout model development.

Unlike SMOTE and ADASYN, which generate synthetic observations through interpolation of existing samples, the proposed framework generates clinically meaningful physiologic features that explicitly represent patient severity and disease progression.

Conclusion

The present study demonstrates that machine learning and artificial intelligence techniques can be effectively applied to large respiratory hospitalization datasets for mortality prediction and clinical risk stratification. Comparative analysis between Random Forest and XGBoost classifiers showed that gradient boosting approaches provide improved mortality discrimination and enhanced sensitivity for identification of critically ill respiratory patients. However, severe imbalance between survivor and mortality cohorts substantially limited the performance of conventional machine learning models trained on the original dataset. To address this limitation, a novel augmentation-based balancing framework was developed using synthetic physiologic generation of minority mortality cases. The proposed methodology incorporated clinically meaningful physiologic variables associated with respiratory deterioration and critical illness, including oxygen saturation, ICU status, ventilator support, sepsis, blood pressure, creatinine, and lactate levels. The augmented balanced dataset significantly improved mortality representation during machine learning training and resulted in marked enhancement of confusion matrix performance, mortality sensitivity, specificity, and overall classification stability.

The findings suggest that incorporation of physiologically coherent synthetic augmentation may represent a promising strategy for overcoming class imbalance in healthcare artificial intelligence applications involving rare but clinically important outcomes. The proposed framework demonstrated that combining administrative healthcare variables with simulated physiologic deterioration patterns substantially improves respiratory mortality prediction capability. Furthermore, the superior performance of the XGBoost classifier highlights the potential value of boosting algorithms for complex nonlinear clinical prediction tasks. Although additional external validation and integration of real-time physiologic monitoring data will be necessary, the present study establishes a foundation for next-generation respiratory mortality analytics and AI-assisted critical care decision support systems. Future research in-

volving multimodal physiologic integration, deep learning architectures, and temporal deterioration modeling may further enhance predictive performance and enable earlier identification of high-risk respiratory patients within hospital and intensive care environments.

Abbreviations

DschrgQtr- Discharge quarter

DschrgYear- Discharge year

LOS- Length of Stay (hospital days)

DRG_Code- Diagnosis Related Group code

DRG_CodeDesc- Description of DRG diagnosis category

APR_DRG- All Patient Refined Diagnosis Related Group

APR_DRG_CodeDesc- Description of APR_DRG category

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Marti J, Hall P, Hamilton P, Lamb S, McCabe C, Lall R, et al. One-year resource utilisation, costs and quality of life in patients with acute respiratory distress syndrome (ARDS): secondary analysis of a randomised controlled trial. *J Intensive Care.* 2016;4:1-11. doi:10.1186/s40560-016-0178-8.
2. Lee SW, Loh SW, Ong C, Lee JH. Pertinent clinical outcomes in pediatric survivors of pediatric acute respiratory distress syndrome (PARDS): a narrative review. *Ann Transl Med.* 2019;7:513. doi:10.21037/atm.2019.09.32.
3. Prithula J, Chowdhury MEH, Khan MS, Al-Ansari K, Zughaier SM, Islam KR, Alqahtani A. Improved pediatric ICU mortality prediction for respiratory diseases: machine learning and data subdivision insights. *Respir Res.* 2024;25(1):216. doi:10.1186/s12931-024-02753-x.
4. Chowdhury ME, Rahman T, Khandakar A, Al-Madeed S, Zughaier SM, Doi SA, et al. An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. *Cogn Comput.* 2021. doi:10.1007/s12559-020-09812-7.
5. Rahman T, Al-Ishaq FA, Al-Mohannadi FS, Mubarak RS, Al-Hitmi MH, Islam KR, et al. Mortality prediction utilizing blood biomarkers to predict the severity of COVID-19 using machine learning technique. *Diagnostics.* 2021;11(9):1582. doi:10.3390/diagnostics11091582.
6. Hong S, Hou X, Jing J, Ge W, Zhang L. Predicting risk of mortality in pediatric ICU based on ensemble step-wise feature selection. *Health Data Sci.* 2021. doi:10.34133/2021/9365125.
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321-357.
8. Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. MICE vs PPCA: missing data imputation in healthcare. *Inf Med Unlocked.* 2019;17:100275. doi:10.1016/j.imu.2019.100275.
9. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated pediatric risk of mortality score. *Crit Care Med.* 1996;24:743-752. doi:10.1097/00003246-199605000-00004.
10. Hu Y, Gong X, Shu L, Zeng X, Duan H, Luo Q, et al. Understanding risk factors for postoperative mortality in neonates based on explainable machine learning technology. *J Pediatr Surg.*

2021;56:2165-2171. doi:10.1016/j.jpedsurg.2021.03.057.

11. Yang Y, Xu B, Haverstick J, Ibtehaz N, Muszyński A, Chen X, et al. Differentiation and classification of bacterial endotoxins based on surface enhanced Raman scattering and advanced machine learning. *Nanoscale.* 2022;14:8806-8817. doi:10.1039/D2NR01277D.
12. De Melo P. A new machine learning algorithm for accurate classification of unbalanced data sets. 2026. Nature Health (in press)