

Original Paper

Development and Interpretability Analysis of a Stacking Ensemble Model for Early Prediction of Nutritional Risk in Intensive Care Unit Patients: Retrospective Cohort Study

Xu Zhang¹, BSc; An Fang¹, MD; Pei Lou¹, MSc; Kuanda Yao¹, MSc; Tianci Huang², MSc; Jiahui Hu¹, PhD

¹Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

²State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Corresponding Author:

Jiahui Hu, PhD

Institute of Medical Information/Medical Library

Chinese Academy of Medical Sciences and Peking Union Medical College

No. 3 Yabao Road, Chaoyang District

Beijing 100020

China

Phone: 86 01052328782

Email: hu.jiahui@imicams.ac.cn

Abstract

Background: Malnutrition in critically ill patients is associated with increased morbidity and mortality, yet traditional screening tools such as the modified NUTRIC (mNUTRIC) score often rely on subjective assessments or delayed data, limiting their utility for early intervention in the dynamic intensive care unit (ICU) environment. Real-time, data-driven approaches using electronic health records offer a promising solution for automated and objective risk stratification.

Objective: This study aimed to develop and validate a machine learning model, the E-NUTRIC (Ensemble-NUTRIC), for the early prediction of malnutrition risk within the first 24 hours of ICU admission. By integrating multiple algorithms through stacking ensemble learning, we sought to improve predictive performance over traditional scoring systems and individual machine learning models while maintaining clinical interpretability.

Methods: We conducted a retrospective cohort study using data from the Medical Information Mart for Intensive Care (MIMIC-IV, version 3.1). Adult ICU stays exceeding 24 hours were included, and the primary outcome was malnutrition diagnosis. Variables from the first 24 hours (demographics, vitals, and laboratory tests) were extracted and harmonized. Missingness was addressed with k-nearest neighbors imputation, features were standardized, and class imbalance was mitigated via random undersampling. The proposed E-NUTRIC model used a stacking ensemble with 4 base learners—Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine—and a logistic metalearner. Performance was assessed on a stratified 80/20 holdout test set using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, and calibration curves. The mNUTRIC score served as the clinical benchmark. Model interpretability was derived by applying Shapley Additive Explanations (SHAP) specifically to the highly predictive XGBoost component, while clinical utility was assessed using Platt scaling recalibration.

Results: The final cohort comprised 51,483 patients, of whom 4384 (8.5%) were classified as being at malnutrition risk. The E-NUTRIC model showed superior discrimination with an AUROC of 0.875 (95% CI 0.864-0.885), outperforming the mNUTRIC score (AUROC=0.635, $P<.001$). Relative to individual base learners, E-NUTRIC achieved the best overall performance, exceeding those of the best-performing individual models XGBoost (AUROC=0.871) and Light Gradient Boosting Machine (AUROC=0.866). The area under the precision-recall curve of E-NUTRIC was 0.424, representing approximately a 3.4-fold increase over mNUTRIC (0.126). SHAP analysis highlighted minimum serum albumin, admission weight, early hypokalemia, and specific ICU admission types as key nonlinear predictors of malnutrition risk. Unlike the traditional mNUTRIC score, which compressed predictions into a low-risk tier, the recalibrated E-NUTRIC model effectively spanned the full probability spectrum, thereby isolating high-risk phenotypes.

Conclusions: The E-NUTRIC stacking ensemble provides an interpretable approach for nutritional risk screening in the ICU using routinely available electronic health records data. In this retrospective cohort study, it demonstrated superior discrimination to the mNUTRIC score and offered clinically consistent feature attributions.

Keywords: malnutrition; intensive care units; machine learning; stacking ensemble; model interpretability

Introduction

Malnutrition is a critical prognostic risk factor for patients in the intensive care unit (ICU), significantly influencing clinical outcomes such as length of ICU stay, duration of mechanical ventilation, infection risk, and mortality [1-3]. A meta-analysis of 20 studies reported that the prevalence of malnutrition among ICU patients ranges from 38% to 78%. The physiological stress of critical illness exacerbates this vulnerability, with nutritional risk escalating during prolonged hospitalization [4]. Crucially, research indicates that early prediction and identification of malnourished patients can facilitate timely and appropriate nutritional interventions, thereby mitigating these adverse outcomes. A national cohort study in Switzerland demonstrated that nutritional support significantly reduced patients' all-cause mortality, reduced 30-day rehospitalization rates, and increased the likelihood of discharge to home [5].

Clinical guidelines recommend that critically ill patients undergo nutritional risk screening within 48 hours of ICU admission to guide therapy [6]. However, the widely used nutritional risk screening tools, such as the Nutritional Risk Screening 2002 [7] and the modified NUTRIC score (mNUTRIC) [8], along with biomarkers such as plasma albumin, have recognized limitations in the ICU setting. These methods can be subjective, require manual data collection which is often delayed, and have demonstrated low recommendation levels for accurately assessing and predicting nutritional status in this dynamic population [9,10]. This highlights a clear need for an objective and early detection method for malnutrition risk in the ICU.

To meet this need, researchers have begun to explore machine learning (ML) algorithms. Early studies often used traditional ML models such as Extreme Gradient Boosting (XGBoost) and Logistic Regression (LR), demonstrating the potential of using electronic health records (EHRs) data to predict malnutrition [11,12]. However, single model approaches often struggle with generalization across diverse patient populations and may suffer from overfitting on imbalanced datasets [13]. Furthermore, previous studies frequently used data from the entire hospitalization period, introducing "look-ahead bias" that limits their utility for early prospective screening [14]. Crucially, complex "black-box" models often lack interpretability, hindering clinical adoption as clinicians require transparent reasoning behind risk predictions to make informed decisions [15]. Few studies have rigorously benchmarked advanced ensemble methods against established clinical scoring systems such as mNUTRIC within a strictly prospective early prediction framework.

This study uses the Medical Information Mart for Intensive Care (MIMIC-IV) database (v3.1) [16], a large, publicly

available repository containing comprehensive clinical data. This rich dataset enables the development of robust predictive models based on routinely collected parameters. However, analyzing real-world EHR data presents significant challenges, including high rates of missing data, heterogeneity in variable recording, and pronounced class imbalance, as malnutrition is less common than normal nutritional status [17]. Addressing these challenges through rigorous data preprocessing and advanced modeling strategies is essential for building a reliable clinical decision support tool.

Therefore, the objective of this study is to develop and validate an interpretable stacking ensemble model, named E-NUTRIC (Ensemble-NUTRIC), for the early prediction of malnutrition risk in ICU patients using data exclusively from the first 24 hours of admission. We propose a Stacking Ensemble Learning framework that integrates multiple base learners, including LR, Random Forest (RF), XGBoost, and Light Gradient Boosting Machine (LightGBM), to enhance predictive stability and accuracy compared with individual algorithms. By integrating a well-established ML interpretability framework, Shapley Additive Explanations (SHAP) [18,19], we aim to enhance the transparency and clinical applicability of our model. Ultimately, this work seeks to provide a precise and automated screening tool that outperforms traditional scoring systems, enabling clinicians to intervene earlier and improve patient outcomes.

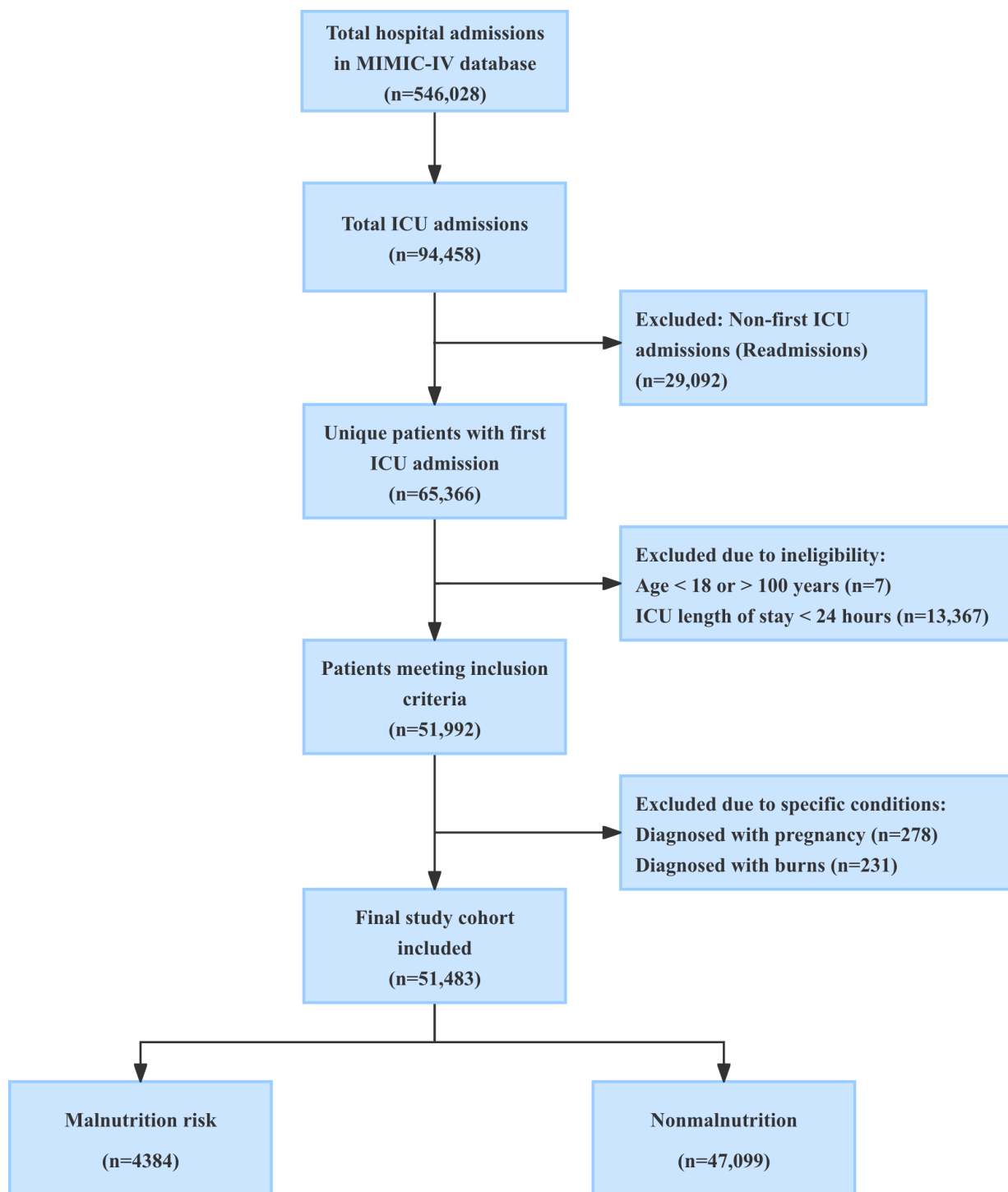
Methods

Data Source and Study Population

This study was a retrospective cohort analysis using the MIMIC-IV database (version 3.1) [16]. MIMIC-IV is a large, single-center, publicly available relational database containing comprehensive clinical data from patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2008 and 2019. The database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center. The requirement for individual patient informed consent was waived because all protected health information was deidentified.

We systematically screened all hospital admissions in the MIMIC-IV database to identify eligible patients. To ensure data consistency and minimize bias from repeated measures, we restricted our analysis to the first ICU admission of the first hospitalization for each patient. Patients younger than 18 years or older than 100 years, as well as those with an ICU stay of <24 hours, were excluded from both the positive and negative cohorts. [Figure 1](#) presents the data flow diagram.

Figure 1. Flowchart of patient selection. The diagram illustrates the inclusion and exclusion process applied to the MIMIC-IV database. Starting from total hospital admissions, we sequentially filtered for ICU patients, first admissions, valid age ranges, and sufficient length of stay. Patients with confounding conditions (pregnancy and severe burns) were excluded. The final cohort (n=51,483) was stratified into the malnutrition risk group (defined by ICD-9 [International Classification of Diseases, Ninth Revision] and ICD-10 [International Statistical Classification of Diseases, Tenth Revision] codes) and the nonmalnutrition group for model development. ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care.



For the positive cohort (patients with malnutrition), we initially identified 23,099 hospital admissions with a diagnosis of malnutrition. After applying these malnutrition diagnosis criteria and excluding non-first hospitalizations and non-first ICU stays, 5081 patients remained. Further applying the age and ICU stay exclusions yielded a final positive cohort of 4384 patients.

For the negative cohort (patients with no malnutrition), we started with 546,028 total hospital admissions, corresponding to 94,458 ICU admissions. After excluding duplicate ICU admissions, we retained 65,366 first-time ICU patients. Applying the same age and ICU stay exclusions left 51,992 eligible patients. To ensure a clear control group, we further excluded patients diagnosed with conditions that could confound nutritional status assessment, namely, pregnancy

and severe burns. Finally, after removing the 4384 patients already identified in the positive cohort, the final negative cohort comprised 47,099 patients.

The final study population consisted of 51,483 adult patients, with malnutrition prevalence in 4384 (8.5%) patients. The primary outcome was the diagnosis of malnutrition, defined using *ICD-9 (International Classification of Diseases, Ninth Revision)* codes 260-263 and *ICD-10 (International Statistical Classification of Diseases, Tenth Revision)* codes E40–E46. These codes encompass a spectrum of nutritional deficiencies, including kwashiorkor, nutritional marasmus, and severe to mild protein-calorie malnutrition. All data partitioning into training (80%) and test (20%) sets were performed strictly at the patient level using unique participant identifiers to prevent data leakage.

Feature Extraction and Data Preprocessing

To facilitate early and accurate risk stratification, data extraction was strictly limited to the first 24 hours of ICU admission. This temporal restriction prevents look-ahead bias and ensures that the model relies solely on information available during the early critical phase. We extracted a comprehensive set of clinical variables based on their clinical relevance and data availability in the MIMIC-IV database [16]. These concepts encompassed demographics; vital signs; laboratory tests; therapeutic interventions such as ventilation, dialysis, and vasopressor usage; and clinical severity scores including the Acute Physiology and Chronic Health Evaluation II and the Sequential Organ Failure Assessment (SOFA) scores.

Specific emphasis was placed on laboratory indicators related to nutritional status, inflammation, and metabolic function. The extracted laboratory panel included serum albumin, total cholesterol, lymphocytes, C-reactive protein, erythrocyte sedimentation rate, glucose, lactate, electrolytes, and renal function markers. To capture the dynamic fluctuations inherent in critical illness, we did not rely on single time-point measurements. Instead, for every time-varying vital sign and laboratory parameter, we computed 6 statistical aggregate metrics: mean, SD, minimum, maximum, first value, and last value within the 24-hour window. These statistics, combined with nominal categorical variables, including gender, race, insurance, language, marital status, and ICU unit type, were represented using one-hot encoding with explicit missing-category indicators. After preprocessing and feature expansion, the final model used 296 predictors (see [Multimedia Appendix 1](#) for details).

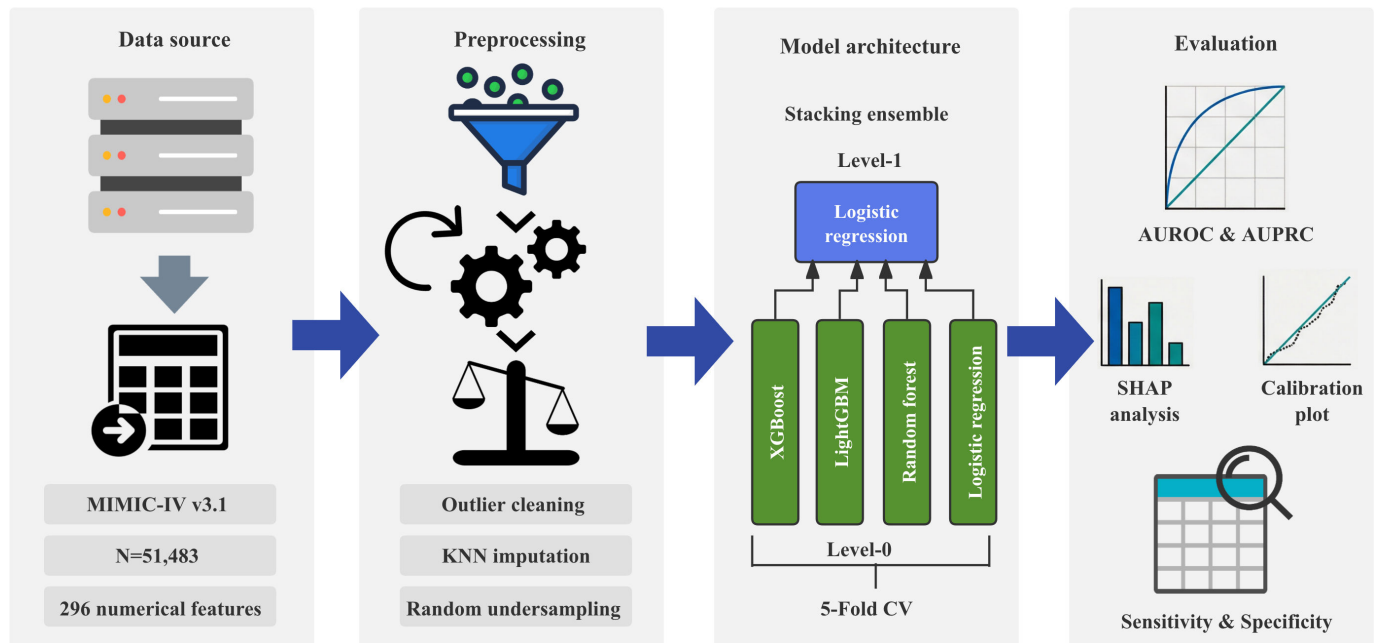
Before modeling, rigorous data cleaning was performed to address physiological artifacts and measurement errors based on expert consensus [20]. We enforced clinically plausible ranges for key variables; for instance, weight was restricted to 30–300 kg and BMI to 10–100 kg/m². Temperature readings required specific handling to harmonize units: values exceeding 50 were interpreted as Fahrenheit and converted to Celsius, after which a valid range of 25 °C to 45 °C was enforced. Values falling outside these predefined thresholds were designated as missing data.

To handle missing data without discarding valuable patient information, we used a *k*-nearest neighbors (KNN) imputation strategy [21]. This method estimates missing entries by identifying the *k* most similar patients in the feature space, thereby preserving the multivariate correlation structure of the data. We validated this choice through preliminary experiments, where KNN imputation demonstrated superior area under the receiver operating characteristic curve (AUROC) in 5-fold cross-validation compared with simple mean imputation or multivariate imputation by chained equations. Following imputation, all continuous features were standardized using *Z*-score normalization to achieve zero mean and unit variance. This step is critical for algorithms sensitive to feature scaling, such as LR [22], which serves as both a base learner and the metalearner in our ensemble framework.

Given the pronounced class imbalance in our cohort (8.5% malnutrition prevalence), standard training procedures would likely bias the model toward the majority class. To mitigate this, we integrated a random undersampling (RUS) strategy into the training pipeline. Comparative analysis of resampling techniques revealed that RUS yielded a higher area under the precision-recall curve (AUPRC) than Synthetic Minority Over-sampling Technique or class-weighting adjustments. Importantly, this resampling was applied exclusively to the training folds within the cross-validation process, ensuring that the validation and test sets maintained the natural clinical prevalence of malnutrition to provide realistic performance estimates. Preliminary comparisons of imputation methods are detailed in [Multimedia Appendix 2](#).

The final dataset was partitioned into a training set (80%) and an independent test set (20%) using stratified sampling. This approach ensured that the proportion of malnutrition cases remained consistent across splits. No data from the test set were used during the feature engineering, imputation fitting, or model training phases to strictly prevent data leakage. [Figure 2](#) presents the methodological framework of this study.

Figure 2. Overall experimental workflow and model development pipeline. Clinical data extracted from the MIMIC-IV v3.1 database yielded a final analytic cohort of 51,483 patients. After preprocessing, the final modeling matrix comprised 296 predictors, including one-hot encoded nominal categorical variables. The data preprocessing phase involved outlier cleaning, KNN imputation, and random undersampling to address class imbalance. Subsequently, a 2-level Stacking Ensemble model was constructed under a 5-fold cross-validation framework, using XGBoost, LightGBM, Random Forest, and Logistic Regression (LR) as base learners (level 0) to feed an LR metalearner (level 1). Finally, the predictive performance and clinical interpretability of the model were systematically evaluated using metrics including AUROC, AUPRC, and SHAP analysis and calibration plots. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CV: cross-validation; KNN: k-nearest neighbors; LightGBM: Light Gradient Boosting Machine; MIMIC-IV: Medical Information Mart for Intensive Care; SHAP: Shapley Additive Explanations; XGBoost: Extreme Gradient Boosting.



Model Development

The development of the predictive model was centered on a Stacking Ensemble Learning framework, termed E-NUTRIC, designed to integrate the strengths of diverse ML algorithms for robust risk stratification. This approach was selected to mitigate the limitations of single-model classifiers, which may suffer from overfitting or fail to capture complex nonlinear interactions within high-dimensional clinical data [23,24]. The ensemble architecture consisted of 2 layers: a base layer comprising 4 distinct algorithms and a metalearner layer that synthesized their outputs into a final calibrated probability.

The base layer comprised LR, RF, XGBoost, and LightGBM. LR was chosen as a linear baseline to capture straightforward additive relationships and provide interpretability [25]. RF, a bagging ensemble method, was used for its ability to reduce variance and handle high-dimensional feature spaces effectively [26,27]. XGBoost and LightGBM, both gradient boosting frameworks, were selected for their superior performance on tabular data and their capacity to model complex nonlinear dependencies through iterative error correction [28]. To ensure reproducibility and avoid the potential pitfalls of overfitting during exhaustive hyperparameter searching on a limited dataset, we used a set of fixed, predefined hyperparameters based on established practices for imbalanced clinical datasets rather than performing a grid search. For instance, tree-based models were configured with 500 estimators to ensure sufficient learning capacity, while regularization parameters were set to prevent overfitting. A

detailed list of all hyperparameters for each algorithm is provided in [Multimedia Appendix 3](#).

Class imbalance was handled exclusively using foldwise RUS within the training pipeline to prevent information leakage. To avoid mathematically distorting the base learners and causing severe overprediction, no additional algorithm-level class weighting was concurrently applied. Specifically, class weights were explicitly disabled for LR, RF, and the stacking metalearner; `scale_pos_weight` was dynamically set to 1 for XGBoost; and the `is_unbalance` flag was disabled for LightGBM.

The stacking process was implemented using a 5-fold cross-validation scheme within the training set. In this procedure, the training data were divided into 5 folds. For each fold, the base models were trained on the remaining 4 folds and then generated predictions for the holdout fold. These out-of-fold predictions from all 4 base learners were then concatenated to construct a new feature matrix, which served as the input for the metalearner. We selected LR as the metalearner to linearly combine the probabilistic outputs of the base models. This choice was motivated by the need for a calibrated and interpretable final score that effectively weighs the confidence of each base classifier. The metalearner was trained on the full set of out-of-fold predictions, enabling it to learn the optimal combination of base model outputs that minimized the overall prediction error.

The final E-NUTRIC model output was a continuous probability score ranging from 0 to 1, representing the risk of malnutrition. To facilitate clinical decision-making, a

definitive classification was determined using a probability threshold of 0.5. No early stopping mechanisms were used during training to maintain consistency across all experiments. All modeling was implemented using Python (version 3.10; Python Software Foundation) with the scikit-learn, xgboost, and lightgbm libraries.

Statistical Analysis and Model Evaluation

Baseline characteristics of the full analytic cohort were rigorously compared between patients with and with no malnutrition risk. Given the inherent long-tail, nonnormal distributions characteristic of large-scale, real-world EHR data, all continuous variables were uniformly summarized as medians with IQRs. Consequently, continuous physiological and laboratory parameters were compared using the nonparametric Mann-Whitney *U* test. Categorical variables—including binary clinical indicators, specific interventions, and one-hot encoded features—were expressed as absolute frequencies and percentages (n, %). Differences in categorical proportions were evaluated using the Pearson Chi-square test when the expected frequency in any cell of a 2×2 contingency table was less than 5. All statistical tests were 2-sided, and a *P* value of <.05 was considered statistically significant.

Model performance was comprehensively evaluated on the held-out test set using a suite of metrics designed for imbalanced classification tasks. The primary evaluation metric was the AUROC, which assesses the model's discriminative ability across all possible thresholds. To rigorously compare the predictive performance of the E-NUTRIC model against the clinical baseline (mNUTRIC score) and individual ML algorithms, we used the DeLong test for correlated receiver operating characteristic curves [29]. In instances where the DeLong test was inconclusive or yielded unstable estimates due to sample size constraints, we used a nonparametric bootstrap method with 1000 resamples to estimate the CIs and statistical significance of the AUROC differences. Additionally, given the low prevalence of malnutrition in our cohort (4384/51,483, 8.5%), we prioritized the AUPRC as a key indicator of model performance, as it is less sensitive to class imbalance than AUROC [13]. We also reported standard classification metrics including Accuracy, Sensitivity, Specificity, and F_1 -score based on a probability threshold of 0.5.

To enhance the clinical utility and interpretability of the E-NUTRIC model, we integrated SHAP into our evaluation framework [30]. While stacking ensembles provide superior predictive performance, they lack native transparency because of their complex, heterogeneous architecture. To

obtain interpretable and stable local explanations, SHAP analysis was applied specifically to the most predictive component of our ensemble—XGBoost, which was the best-performing individual tree-based base learner. This approach elucidates the critical nonlinear feature interactions driving the risk predictions without compromising computational efficiency. This analysis provided both global feature importance rankings and local explanations for individual patient predictions, offering transparency into the model's decision-making process.

Furthermore, to assess the reliability of the model's risk probabilities, we conducted a comprehensive calibration analysis. For the calibration analysis, mNUTRIC ordinal scores were first converted to predicted probabilities. This was achieved by fitting a univariate LR model on the training set and evaluating its outputs on the held-out test set. For our ML framework, because RUS intrinsically shifts the background prevalence, raw probability outputs from the final Stacking model inherently overestimate risk. Therefore, we performed a post hoc Platt recalibration. The recalibration model was fitted on an independent split within the training dataset and evaluated on the same test set, yielding the Recalibrated E-NUTRIC probabilities. The calibration curve serves as a critical diagnostic for identifying over- or underestimation biases in the predicted risks [31]. All statistical analyses and model development were performed using Python (version 3.10) with the scikit-learn, scipy, shap, and lifelines libraries.

Ethical Considerations

The study was approved by the Institutional Review Board of the Beth Israel Deaconess Medical Center (2001-P-001699/14) and the Massachusetts Institute of Technology (no. 0403000206), and individual consent for this retrospective analysis was waived.

Results

Baseline Characteristics of the Study Population

The final study cohort comprised 51,483 patients, among whom 4384 (8.5%) were identified as being at high risk of malnutrition (Risk Group), while 47,099 (91.5%) were classified as being at no risk of malnutrition (No Risk Group). A summary of key demographic and clinical characteristics is presented in Table 1.

Table 1. Baseline demographic, anthropometric, clinical, vital sign, and laboratory characteristics of the full analytic cohort stratified by malnutrition risk status. A total of 51,483 patients were enrolled, of whom 47,099 were nonmalnourished and 4384 had malnutrition^a.

Variable	Malnutrition risk group (n=4384)	No risk group (n=47,099)	<i>P</i> value
Demographics			
Age (years), median (IQR)	68.00 (57.00-78.00)	67.00 (56.00-78.00)	.002
Sex, n (%)			.02
Male	2436 (55.6)	27,030 (57.4)	

Variable	Malnutrition risk group (n=4384)	No risk group (n=47,099)	<i>P</i> value
Female	1948 (44.4)	20,069 (42.6)	
Ethnicity, n (%)			<.001
Asian	147 (3.4)	1365 (2.9)	
Black	450 (10.3)	4073 (8.6)	
Hispanic	135 (3.1)	1653 (3.5)	
White	2801 (63.9)	30,974 (65.8)	
Other/Unknown	851 (19.4)	9034 (19.2)	
Anthropometrics			
BMI (kg/m ²), median (IQR)	25.19 (20.99-30.06)	28.34 (24.76-32.89)	<.001
Weight (kg), median (IQR)	69.53 (57.30-85.00)	80.30 (68.00-95.35)	<.001
Clinical parameters and scores			
ICU ^b type, n (%)			<.001
CCU ^c	324 (7.4)	5404 (11.5)	
CVICU ^d	208 (4.7)	10,418 (22.1)	
MICU ^e	1260 (28.7)	7990 (17.0)	
SICU ^f	2339 (53.4)	18,893 (40.1)	
Other ^g	253 (5.8)	4394 (9.3)	
APACHE II ^h score, median (IQR)	22.00 (17.00-27.00)	18.00 (13.00-23.00)	<.001
SOFA ⁱ score, median (IQR)	5.00 (3.00-8.00)	4.00 (2.00-6.00)	<.001
Charlson Comorbidity Index, median (IQR)	6.00 (4.00-8.00)	5.00 (3.00-7.00)	<.001
Pre-ICU length of stay (days), median (IQR)	0.09 (0.04-2.14)	0.08 (0.03-0.75)	<.001
Laboratory values (24 hours), mean			
Albumin (g/dL), median (IQR)	2.74 (2.37-3.14)	3.10 (2.70-3.50)	<.001
Creatinine (mg/dL), median (IQR)	0.91 (0.62-1.55)	0.90 (0.70-1.27)	.32
BUN ^j (mg/dL), median (IQR)	23.55 (14.75-39.03)	18.25 (12.93-28.00)	<.001
Glucose (mg/dL) (Chemistry), median (IQR)	124.83 (108.51-152.05)	120.50 (106.75-142.76)	<.001
Potassium (mEq/L) (Chemistry), median (IQR)	4.06 (3.85-4.30)	4.08 (3.85-4.32)	.02
Lymphocytes (Absolute), median (IQR)	1.02 (0.64-1.52)	1.40 (0.89-2.04)	<.001
Phosphate (mg/dL), median (IQR)	3.37 (2.95-3.91)	3.30 (2.87-3.77)	<.001
Magnesium (mg/dL), median (IQR)	2.03 (1.92-2.17)	2.05 (1.93-2.20)	<.001
Cholesterol (mg/dL), median (IQR)	117.00 (86.00-152.75)	150.00 (119.00-185.00)	<.001

^aContinuous variables are expressed as median (IQR) and compared using the Mann-Whitney *U* test. Categorical variables are expressed as counts and percentages, n (%), and compared using the chi-square test. *P* values are 2-tailed, and statistical significance was established at *P*<.05.

^bICU: intensive care unit.

^cCCU: coronary care unit.

^dCVICU: cardiovascular intensive care unit.

^eMICU: medical intensive care unit.

^fSICU: surgical intensive care unit.

^gICU type categories in [Table 1](#) were partially aggregated for readability. The SICU category includes MICU/SICU, Neuro SICU, SICU, and TSICU; the Other category includes Neuro Intermediate, Neuro Stepdown, and Other. Detailed ICU subtype distributions are provided in [Multimedia Appendix 4](#).

^hAPACHE II: Acute Physiology and Chronic Health Evaluation II.

ⁱSOFA: Sequential Organ Failure Assessment.

^jBUN: blood urea nitrogen.

The 2 groups were comparable in terms of age and gender distribution. However, significant disparities were observed in anthropometric measurements and clinical severity indices. Patients in the malnutrition risk group had a significantly lower BMI and lower admission weight. Clinical severity was substantially higher in the risk group, as indicated by elevated Acute Physiology and Chronic Health Evaluation II scores, Charlson Comorbidity Index, and SOFA scores. The mNUTRIC score was also significantly higher in the risk group.

Physiological parameters recorded within the first 24 hours further distinguished the 2 groups. The risk group demonstrated signs of hemodynamic instability, characterized by higher heart rates and respiratory rates. Laboratory analyses revealed significant nutritional and inflammatory differences; specifically, minimum serum albumin levels were lower in the risk group. Similarly, total cholesterol levels and lymphocyte counts were significantly reduced in the risk group (*P*<.001), while renal function markers such as blood urea nitrogen were elevated (*P*<.001). These findings indicate

that the malnutrition risk group had a higher burden of critical illness and metabolic disturbances upon ICU admission. A comprehensive comparison of all analyzed features is provided in [Multimedia Appendix 4](#).

Predictive Performance Comparison

The proposed E-NUTRIC model demonstrated superior predictive performance compared with both the existing clinical scoring system and the individual ML algorithms. As detailed in [Table 2](#), the E-NUTRIC model achieved the highest discrimination with an AUROC of 0.875 (95% CI 0.864-0.885). This represents a substantial and statistically significant improvement over the standard mNUTRIC score, which yielded an AUROC of 0.635 (95% CI 0.617-0.652, $P < .001$). The receiver operating characteristic

curves presented in [Figure 3A](#) illustrate this disparity, showing that the E-NUTRIC model maintains a higher true-positive rate across all decision thresholds compared with the linear clinical baseline.

In the context of the imbalanced dataset, the AUPRC provided a more informative assessment of model utility. The E-NUTRIC model achieved an AUPRC of 0.424, which is more than 3 times that of the mNUTRIC score. [Figure 3](#) highlights this performance gap, showing that the ensemble model maintains precision better as recall increases, whereas the performance of the mNUTRIC score degrades rapidly, approaching the no-skill baseline. Among the individual base learners, XGBoost and LightGBM exhibited strong performance with AUROC of 0.871 and 0.866, respectively.

Table 2. Performance comparison of Ensemble-NUTRIC (E-NUTRIC) with baseline models and clinical scores on the test set.

Model	AUROC ^a (95% CI)	AUPRC ^b	Accuracy	F_1 -score	Sensitivity	Specificity
E-NUTRIC	0.875 (0.864-0.885) ^c	0.424	0.785	0.384	0.790	0.784
LightGBM ^d	0.866 (0.854-0.877)	0.414	0.770	0.372	0.800	0.767
XGBoost ^e	0.871 (0.860-0.882)	0.422	0.774	0.380	0.812	0.770
LR ^f	0.856 (0.844-0.869)	0.374	0.783	0.377	0.772	0.784
Random Forest	0.844 (0.831-0.855)	0.343	0.723	0.334	0.813	0.715
mNUTRIC ^g Score	0.635 (0.617-0.652)	0.126	0.142	0.164	0.989	0.063

^aAUROC: area under the receiver operating characteristic curve.

^bAUPRC: area under the precision-recall curve.

^cThe proposed E-NUTRIC model demonstrated statistically significant improvement in AUROC compared with the clinical baseline mNUTRIC score ($P < .001$) and the best-performing single base learner, XGBoost ($P = .018$), evaluated via the DeLong test.

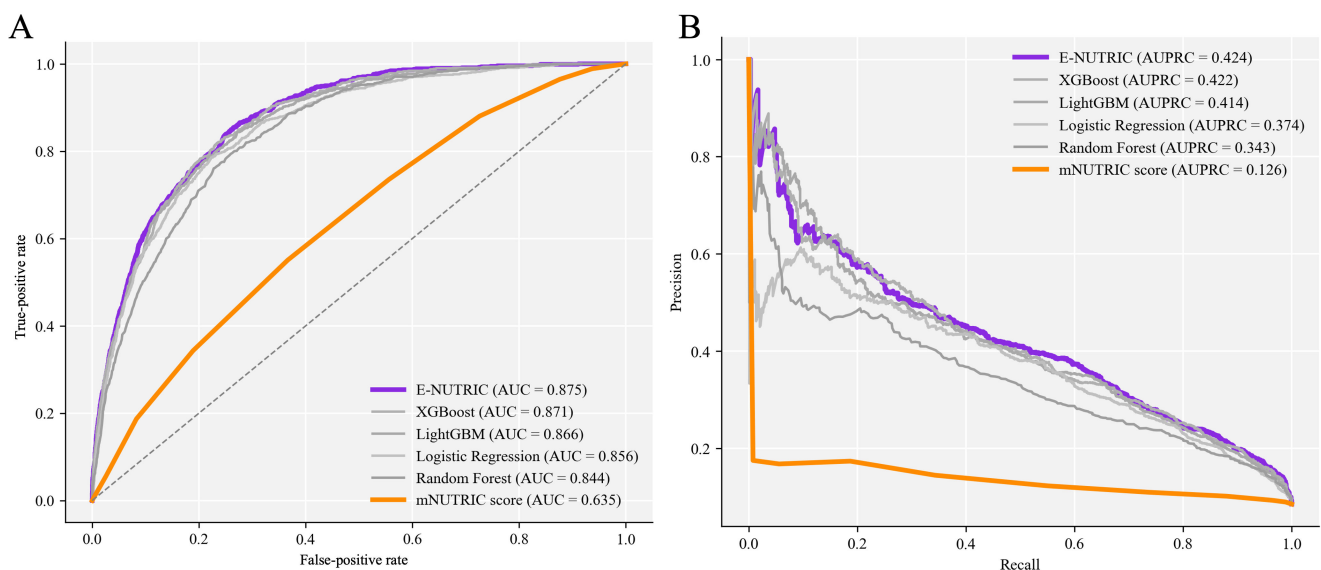
^dLightGBM: Light Gradient Boosting Machine.

^eXGBoost: Extreme Gradient Boosting.

^fLR: Logistic Regression.

^gmNUTRIC: modified NUTRIC.

Figure 3. Discriminative performance of models and score. Receiver operating characteristic curves (A) and precision-recall curves (B). The proposed E-NUTRIC (purple line) significantly outperforms the clinical benchmark mNUTRIC score (orange line). Individual machine learning base learners (XGBoost, LightGBM, Random Forest, and Logistic Regression) are shown in light gray for comparison, demonstrating the superior performance of the ensemble approach. AUC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; E-NUTRIC: Ensemble-NUTRIC; LightGBM: Light Gradient Boosting Machine; mNUTRIC: modified NUTRIC; XGBoost: Extreme Gradient Boosting.



At the standard probability threshold of 0.5, the E-NUTRIC model demonstrated a balanced classification profile with an accuracy of 0.785, a sensitivity of 0.790, and a specificity of 0.784. In contrast, the mNUTRIC score showed extremely high sensitivity but negligible specificity, indicating a tendency to overpredict risk and generate excessive false positives. The superior F_1 -score of the E-NUTRIC model versus mNUTRIC further underscores its robustness in correctly identifying true malnutrition cases while minimizing false positives. Detailed performance curves for all individual base models are provided in [Multimedia Appendix 5](#).

Model Interpretability and Feature Analysis

To elucidate the decision-making logic of the predictive framework, we used SHAP values to quantify the contribution of individual features. Given the heterogeneous architecture of the final stacking ensemble, SHAP was applied to its most predictive tree-based component—the

XGBoost base learner—to ensure accurate mathematical attribution and computational stability.

Figures 4A and 4B present the global feature importance rankings alongside their directional impacts on the model’s output. Minimum serum albumin emerged as the most dominant predictor, followed strongly by admission weight, minimum potassium levels, and specific ICU admission location. The summary plot delineates distinct clinical phenotypes: lower physiological values (blue dots) of albumin, weight, and blood potassium are aggressively associated with positive SHAP values, increasing the predicted risk of malnutrition. Notably, admission to cardiovascular intensive care unit exerted a distinctly protective effect on the risk score, whereas prolonged pre-ICU length of stay and higher Charlson Comorbidity Index scores were strongly positively correlated with malnutrition risk. This adequately reflects the compounding impact of chronic disease burden and extended hospitalization on nutritional depletion.

Figure 4. SHAP global feature importance and summary analysis derived from the XGBoost component. (A) The bar plot ranks the top 20 features by their mean absolute SHAP value, indicating their overall mathematical contribution to the model’s predictions. (B) The summary plot visualizes the distribution of SHAP values across individual patients (represented by dots). Note the color gradient: red indicates a high feature value, and blue indicates a low feature value. A positive SHAP value on the x-axis drives the model toward a higher predicted risk of malnutrition. BUN: blood urea nitrogen; CVICU: cardiovascular intensive care unit; ICU: intensive care unit; MICU: medical intensive care unit; SICU: surgical intensive care unit; SHAP: Shapley Additive Explanations.

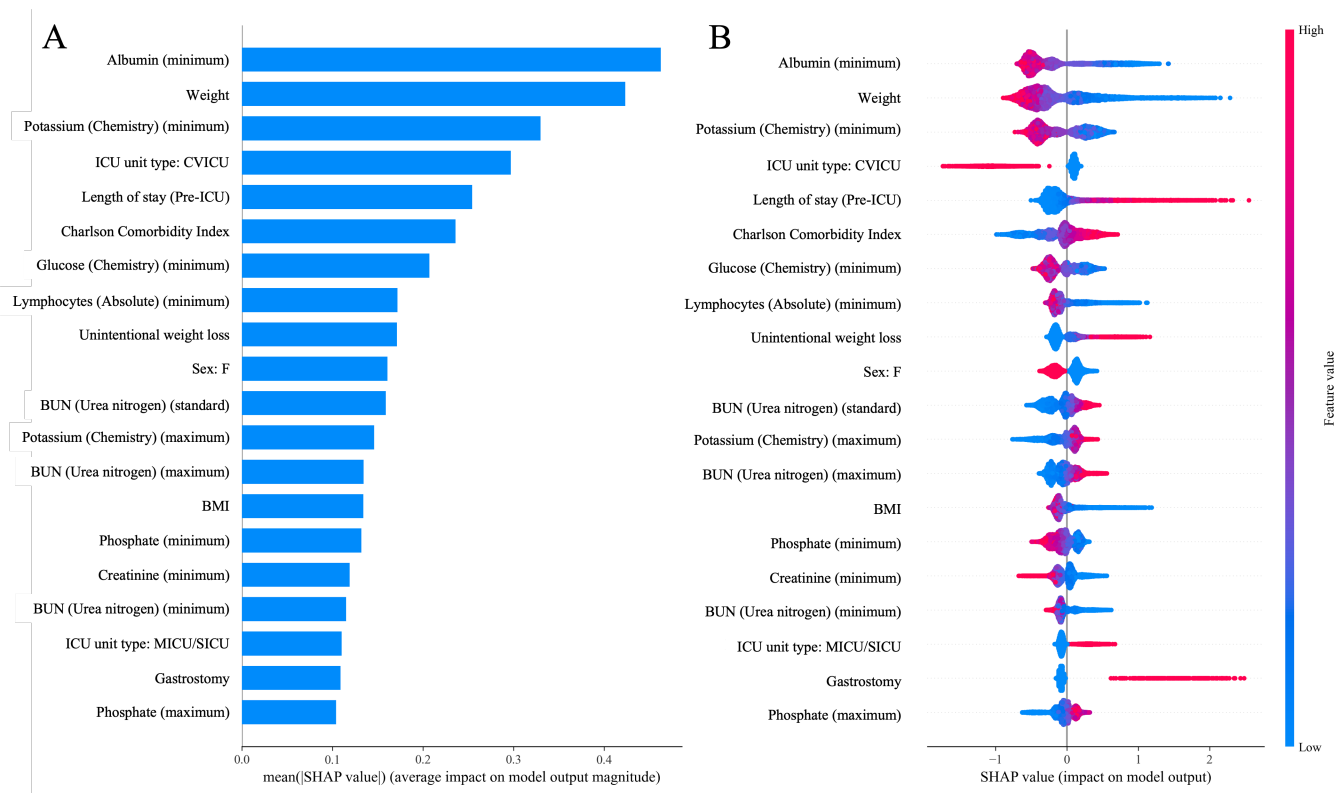


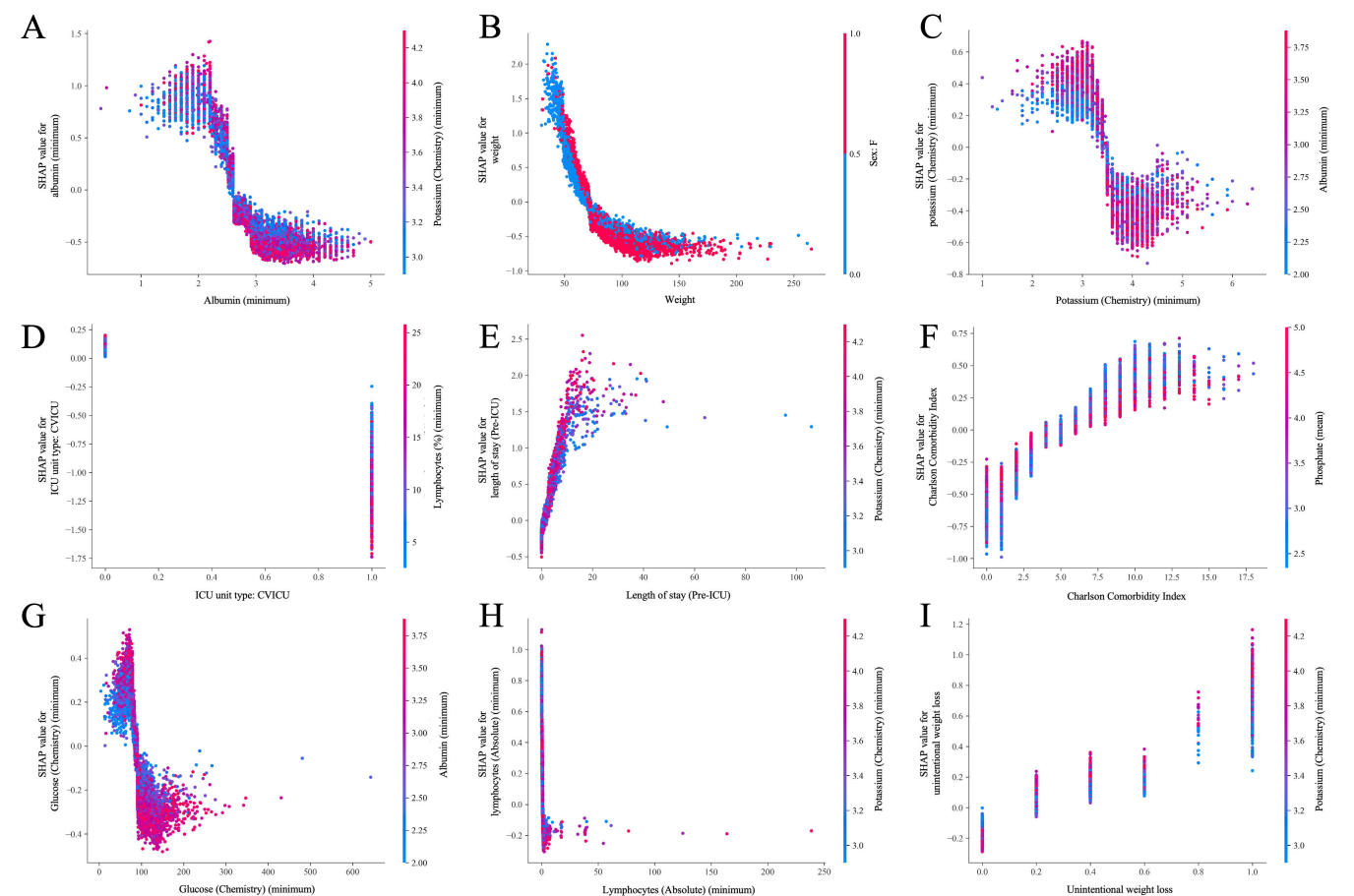
Figure 5A-I further delineate the granular, nonlinear relationships for the top 9 predictive features via SHAP dependence plots. A critical threshold is observed for serum albumin: the risk contribution escalates sharply as levels fall below approximately 2.8 g/dL, corroborating its clinical utility as a potent negative acute-phase reactant. Similarly, the weight dependence plot explicitly demonstrates

a steep surge in risk for patients weighing less than 70 kg, which plateaus at higher baseline weights. Electrolyte and metabolic profiles reveal that acute physiological stress—specifically early hypokalemia (<3.5 mEq/L) and hypoglycemia (<100 mg/dL)—disproportionately compounds the predicted risk. Furthermore, severe early lymphopenia (low absolute lymphocytes) and the documented presence of

unintentional weight loss exhibited definitive risk spikes. Collectively, these interpretability analyses confirm that the model effectively captured clinically and biologically plausible signals, accurately prioritizing established markers of severe catabolism, structural depletion, and diminished physiological reserve without relying on arbitrary data artifacts. Furthermore, a global SHAP interaction analysis

(refer to [Multimedia Appendix 6](#) for details) revealed strong second-order feature dependencies, most notably between absolute admission weight and female gender. This demonstrates the ensemble's capacity to automatically learn sex-specific physiological thresholds for body mass, highlighting a critical advantage of nonlinear ML architectures over traditional additive clinical scores.

Figure 5. SHAP partial dependence plots showing nonlinear clinical thresholds for the top 9 features: (A) minimum albumin, (B) admission weight, (C) minimum potassium, (D) CVICU admission, (E) pre-ICU length of stay, (F) Charlson Comorbidity Index, (G) minimum glucose, (H) minimum absolute lymphocyte count, and (I) unintentional weight loss. These plots illustrate the marginal effect of each variable on predicted malnutrition risk. The x-axis represents the raw feature value from the electronic health record, and the y-axis represents the corresponding SHAP value. The color bar in each panel indicates the interacting feature value, illustrating multidimensional model behavior. Notable clinically relevant nonlinear thresholds are observed natively in albumin, weight, and absolute lymphocytes. CVICU: cardiovascular intensive care unit; ICU: intensive care unit; SHAP: Shapley Additive Explanations.

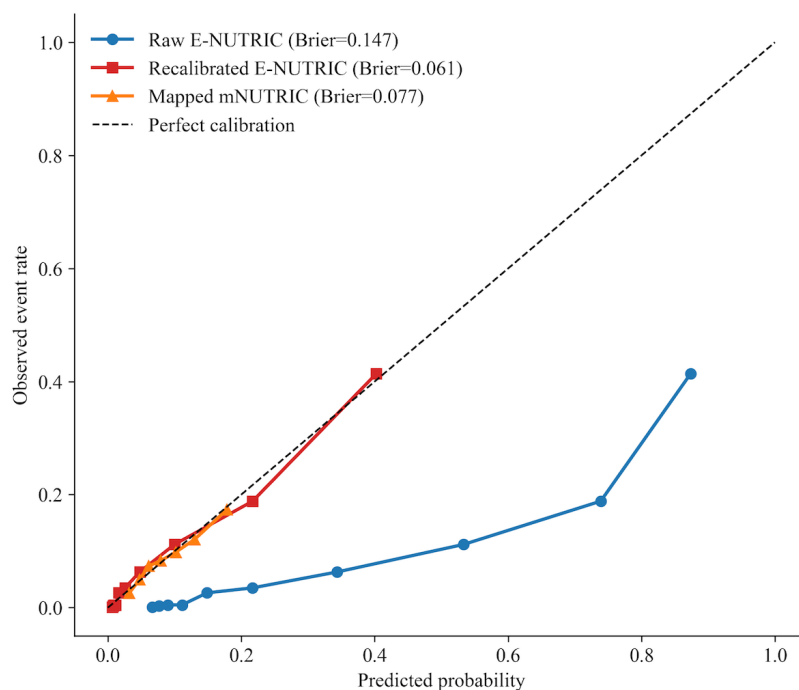


Risk Stratification and Calibration Analysis

To evaluate the reliability of the risk predictions across the probability spectrum, a comprehensive calibration analysis was performed on the test set ([Figure 6](#)). As expected, the raw E-NUTRIC probabilities—generated directly from the ensemble trained on the RUS balanced dataset—exhibited

structural overestimation of absolute risk, deviating systematically below the perfect calibration diagonal (Brier score=0.1471; expected calibration error [ECE]=0.2347). To correct this prevalence-induced probability shift, Platt scaling was applied. The resulting Recalibrated E-NUTRIC model demonstrated excellent alignment with true clinical event rates (Brier score=0.0615; ECE=0.0095), accurately tracking expected outcomes across all risk deciles.

Figure 6. Calibration plots on the test set demonstrating dynamic risk stratification. The diagonal dashed gray line represents perfect calibration, where predicted probabilities perfectly match observed clinical frequencies. The Raw E-NUTRIC probabilities (blue circles) display the expected overestimation inherent to algorithms trained on artificially balanced resampled datasets. Following Platt scaling, the Recalibrated E-NUTRIC (red squares) rigorously corrects this shift, accurately mapping risk across the entire spectrum. The mapped mNUTRIC score (orange triangles)—derived via logistic transformation—is provided for clinical baseline comparison. Notably, while mNUTRIC predictions are severely compressed into the lower probability tier (<0.2), E-NUTRIC effectively spans the full dynamic range to identify critical, high-risk phenotypes. E-NUTRIC: Ensemble-NUTRIC; mNUTRIC: modified NUTRIC.



Following recalibration, we compared the E-NUTRIC ensemble against the baseline mNUTRIC score. To ensure a mathematical equivalent comparison, the ordinal mNUTRIC scores were first mapped to continuous probabilities using a training-set logistic transformation (yielding a Brier score of 0.0765 and an ECE of 0.0045 on the test set). While the mapped mNUTRIC achieved tight calibration metrics, the resulting calibration curve highlights a fundamental limitation of the traditional clinical scoring system: its predicted probabilities are severely compressed, clustering almost entirely below a threshold of 0.20. This indicates that the conventional score lacks the dynamic range necessary to differentiate between moderate-risk and genuinely high-risk patients, effectively categorizing the entire cohort into a low-probability band. In stark contrast, the Recalibrated E-NUTRIC model uses the full probability space. Unlike the mNUTRIC score, which plateaus prematurely, our model effectively isolates a high-risk subgroup of patients at the upper tail of the distribution, providing clinicians with a highly discriminative and granular tool for identifying patients who acutely require intensive nutritional interventions.

Discussion

Principal Findings

In this study, we developed and internally validated the E-NUTRIC model, a stacking ensemble ML framework designed to predict malnutrition risk among critically ill

patients using routine EHR data. Our evaluation indicates that the E-NUTRIC ensemble provides superior discriminative capacity compared with the traditional mNUTRIC score, achieving an AUROC of 0.875 compared with 0.635 for the clinical baseline. Notably, in the context of the naturally imbalanced prevalence of ICU malnutrition, the model yielded a substantial, more than 3-fold increase in the AUPRC. This enhancement addresses a recurring limitation in existing additive clinical scoring systems, which frequently lack the dynamic range necessary to effectively stratify patients at the upper end of the risk spectrum.

While individual tree-based base learners such as XGBoost and LightGBM also exhibited robust performance, the level 1 stacking strategy provided incremental improvement in the stabilization of predictive reliability. This supports the premise that integrating diverse functional algorithms can effectively synthesize complementary clinical signals within complex datasets. Furthermore, to ensure accurate mathematical attribution and avoid the methodological artifacts associated with explaining metalearners, we applied SHAP analysis specifically to the highly predictive XGBoost component. This analytical approach elucidated the nonlinear feature relationships driving the predictions, identifying minimum serum albumin, admission weight, early hypokalemia, and specific ICU unit types as the most influential predictors of malnutrition risk.

Comparison With Prior Work

Previous research in nutritional assessment has predominantly relied on static scoring systems such as the Geriatric Nutrition Risk Index [32], the Controlling Nutrition Status score [33], and the Subjective Global Assessment [9]. While effective in general wards or outpatient oncology settings [34,35], these tools often fail to account for the rapid physiological changes characteristic of the ICU environment. The mNUTRIC score was specifically designed for critically ill patients, yet our findings align with recent studies suggesting its limited discriminatory power without the inclusion of inflammatory markers such as interleukin-6, which are rarely routinely available [36].

In the field of clinical artificial intelligence, recent efforts have applied deep learning and time series analysis to predict patient deterioration [37,38]. However, many of these models prioritize raw predictive accuracy over interpretability and often involve substantial computational demands that limit bedside deployment. Our study differs from such approaches by using a stacking ensemble approach on aggregated clinical features. Unlike single-model approaches or complex deep neural networks, our framework strikes a balance between predictive power and computational efficiency. The superior performance of E-NUTRIC over the mNUTRIC score validates the assertion that nonlinear modeling techniques are necessary to capture the complex interactions between physiological stress, organ failure, and nutritional depletion [39].

Clinical Implications and Biological Plausibility

The interpretability analysis via SHAP values not only confirms the model's reliance on clinically relevant features but also provides biological validation of the algorithm's learning process. Minimum serum albumin emerged as the single most important predictor. Although traditionally viewed as a nutritional marker, hypoalbuminemia in the ICU is a potent indicator of the inflammatory response and capillary leak syndrome, both of which accelerate catabolism and nutritional depletion [40]. The sharp increase in risk when albumin falls below 2.8 g/dL, as observed in our partial dependence plots, aligns with the threshold often cited for severe malnutrition-inflammation complex syndrome.

Similarly, the model identified admission weight and minimum potassium levels as critical features. The nonlinear relationship observed with absolute admission weight—where risk escalates rapidly below 70 kg—was shown in our interaction analysis to be strongly modulated by female gender, dynamically capturing the phenotype of the frail patient with sarcopenia who lacks the metabolic reserve to withstand critical illness [41]. Beyond electrolytes such as potassium—which capture signals related to severe fluid-electrolyte shifts [42]—the model effectively leveraged granular one-hot encoded features, identifying the distinct protective baseline status of cardiovascular intensive care unit admissions and the compounding risk of documented unintentional weight loss and early severe lymphopenia. Unlike the

mNUTRIC score, which is heavily weighted by the SOFA score and thus primarily reflects organ failure, E-NUTRIC incorporates these metabolic and anthropometric granularities, offering a more holistic view of the patient's nutritional status.

Furthermore, the calibration analysis revealed a crucial clinical utility finding: the mNUTRIC score effectively “caps” risk prediction at approximately 20%, treating the vast majority of patients as low risk. In stark contrast, following Platt scaling correction for prevalence-induced shifts, the recalibrated E-NUTRIC model accurately spans the full dynamic probability spectrum. This capability allows for the identification of a high-risk cohort that would otherwise be missed by standard screening, enabling clinicians to direct aggressive nutritional interventions—such as early parenteral nutrition or high-protein delivery—to the patients who are most likely to benefit, thereby optimizing resource allocation.

Limitations

Despite these promising results, our study is subject to several limitations. First, the model was developed and validated using the MIMIC-IV database, which represents a single-center cohort at the US tertiary care center. While the sample size is large and diverse, the care protocols and patient demographics may not fully generalize to community hospitals or different health care systems globally. To mitigate this, we used a robust cross-validation scheme and a holdout test set, but external validation on datasets such as the eICU Collaborative Research Database [43] remains a necessary next step to confirm generalizability.

Second, the definition of malnutrition in our study relied on *ICD-10* diagnostic codes. While this is a standard approach in large-scale EHR research, administrative codes typically have high specificity but low sensitivity, potentially leading to an underestimation of malnutrition prevalence. We attempted to address the resulting class imbalance by using RUS and optimizing for AUPRC, yet some in the control group patients may have had undiagnosed malnutrition. Future studies should ideally use prospectively collected data based on the Global Leadership Initiative on Malnutrition criteria to provide a more rigorous ground truth [44].

Third, the retrospective nature of the study necessitates the use of imputation for missing data. Although our preliminary experiments demonstrated that KNN imputation preserved data integrity better than other methods, any imputation introduces a degree of synthetic variability. We mitigated this by selecting KNN imputation based on preliminary experiments, but the potential for bias cannot be entirely eliminated.

Fourth, while SHAP values provide powerful associations, they do not imply causality. The feature importance rankings indicate which variables drive the prediction, but they do not prove that modifying a feature (eg, correcting hypokalemia) will reduce malnutrition risk. Clinical judgment remains essential in interpreting these outputs [45].

Future Directions

Building on the E-NUTRIC model, future research should focus on prospective validation in a multicenter clinical setting. Integrating this model into an EHR-based clinical decision support system would allow evaluation of its real-world impact on clinical workflow and patient outcomes. Specifically, a randomized controlled trial comparing model-guided nutritional intervention versus standard care would be the gold standard for establishing clinical utility. Additionally, future iterations of the model could incorporate unstructured data, such as nursing notes or dietary intake logs, using natural language processing to further enhance predictive accuracy [46].

Conclusions

This study presents E-NUTRIC, a stacking ensemble ML model designed for the early prediction of malnutrition risk in

ICU patients. By integrating robust predictive algorithms with SHAP-based interpretability, E-NUTRIC offers an interpretable approach to risk stratification. Specifically, compared with the mNUTRIC score, the proposed model demonstrates superior discriminative performance and provides a finer-grained probability spectrum, effectively capturing key metabolic, anthropometric, and demographic features associated with malnutrition. Although prospective external validation remains necessary, E-NUTRIC demonstrates substantial potential as a clinical decision support tool to facilitate timely, targeted, and accurate nutritional interventions for critically ill patients at high risk.

Acknowledgments

The authors declare the use of generative artificial intelligence in the writing process. According to the GAIDeT taxonomy (2025), the "Proofreading and editing" procedure was delegated to the Gemini language model under human supervision, restricted to the scope of grammatical error correction and linguistic phrasing enhancement. The authors subsequently undertook critical appraisal, verification, and editorial revision of the model-generated output as necessitated and hereby affirm their assumption of complete and sole responsibility for the scientific integrity and content of the publication.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this paper: The research was supported by the CAMS Innovation Fund for Medical Sciences (CIFMS) (grant 2021-12M-1-056), the National Natural Science Foundation of China (grant 72074222), and Central High-Level Hospital Clinical Research Special Project (grant 2022-PUMCH-A084).

Data Availability

The datasets analyzed in this study are available in the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (version 3.1). Access to the database is restricted to credentialed users who have completed the required training in human research subjects protection, in accordance with PhysioNet's data use agreement. The source code used for data preprocessing, model development, and statistical analysis is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JH, AF, XZ

Methodology: XZ, JH, AF, KY

Software: XZ, KY, TH

Validation: AF, PL

Formal analysis: XZ, KY, TH, JH

Investigation: XZ

Data curation: JH, XZ

Writing – original draft: XZ

Writing – review & editing: JH, AF

Visualization: XZ, PL, TH

Supervision: JH, AF

Project administration: JH, AF

Funding acquisition: JH

Conflicts of Interest

None declared.

Multimedia Appendix 1

Comprehensive data dictionary of the 296 predictors included in the Ensemble-NUTRIC analytic matrix.

[\[DOCX File \(Microsoft Word File\), 50 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Optimization of data preprocessing strategies.

[\[DOCX File \(Microsoft Word File\), 21 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Hyperparameter settings and model configuration.

[\[DOCX File \(Microsoft Word File\), 26 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Comprehensive baseline characteristics and feature statistics.

[\[DOCX File \(Microsoft Word File\), 66 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Comprehensive model performance analysis.

[\[DOCX File \(Microsoft Word File\), 287 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Shapley Additive Explanations global feature interaction network plot.

[\[DOCX File \(Microsoft Word File\), 380 KB-Multimedia Appendix 6\]](#)

References

1. Popkin BM, Corvalan C, Grummer-Strawn LM. Dynamics of the double burden of malnutrition and the changing nutrition reality. *Lancet*. Jan 4, 2020;395(10217):65-74. [doi: [10.1016/S0140-6736\(19\)32497-3](https://doi.org/10.1016/S0140-6736(19)32497-3)] [Medline: [31852602](https://pubmed.ncbi.nlm.nih.gov/31852602/)]
2. Wells JC, Sawaya AL, Wibaek R, et al. The double burden of malnutrition: aetiological pathways and consequences for health. *Lancet*. Jan 4, 2020;395(10217):75-88. [doi: [10.1016/S0140-6736\(19\)32472-9](https://doi.org/10.1016/S0140-6736(19)32472-9)] [Medline: [31852605](https://pubmed.ncbi.nlm.nih.gov/31852605/)]
3. Pironi L, Sasdelli AS, Ravaioli F, et al. Malnutrition and nutritional therapy in patients with SARS-CoV-2 disease. *Clin Nutr*. Mar 2021;40(3):1330-1337. [doi: [10.1016/j.clnu.2020.08.021](https://doi.org/10.1016/j.clnu.2020.08.021)] [Medline: [32900518](https://pubmed.ncbi.nlm.nih.gov/32900518/)]
4. Pohlenz-Saw JAE, Merriweather JL, Wandrag L. (Mal)nutrition in critical illness and beyond: a narrative review. *Anaesthesia*. Jun 2023;78(6):770-778. [doi: [10.1111/anae.15951](https://doi.org/10.1111/anae.15951)] [Medline: [36644786](https://pubmed.ncbi.nlm.nih.gov/36644786/)]
5. Kaegi-Braun N, Mueller M, Schuetz P, Mueller B, Kutz A. Evaluation of nutritional support and in-hospital mortality in patients with malnutrition. *JAMA Netw Open*. Jan 4, 2021;4(1):e2033433. [doi: [10.1001/jamanetworkopen.2020.33433](https://doi.org/10.1001/jamanetworkopen.2020.33433)] [Medline: [33471118](https://pubmed.ncbi.nlm.nih.gov/33471118/)]
6. Guan X, Chen D, Xu Y. Clinical practice guidelines for nutritional assessment and monitoring of adult ICU patients in China. *J Intensive Med*. Apr 2024;4(2):137-159. [doi: [10.1016/j.jointm.2023.12.002](https://doi.org/10.1016/j.jointm.2023.12.002)] [Medline: [38681796](https://pubmed.ncbi.nlm.nih.gov/38681796/)]
7. Kondrup J, Rasmussen HH, Hamberg O, Stanga Z, Ad Hoc ESPEN Working Group. Nutritional risk screening (NRS 2002): a new method based on an analysis of controlled clinical trials. *Clin Nutr*. Jun 2003;22(3):321-336. [doi: [10.1016/S0261-5614\(02\)00214-5](https://doi.org/10.1016/S0261-5614(02)00214-5)] [Medline: [12765673](https://pubmed.ncbi.nlm.nih.gov/12765673/)]
8. Rahman A, Hasan RM, Agarwala R, Martin C, Day AG, Heyland DK. Identifying critically-ill patients who will benefit most from nutritional therapy: further validation of the “modified NUTRIC” nutritional risk assessment tool. *Clin Nutr*. Feb 2016;35(1):158-162. [doi: [10.1016/j.clnu.2015.01.015](https://doi.org/10.1016/j.clnu.2015.01.015)] [Medline: [25698099](https://pubmed.ncbi.nlm.nih.gov/25698099/)]
9. Zhang X, Zhang J, Du Y, et al. The clinical application value of phase angle of six parts in nutritional evaluation of tumor patients. *Support Care Cancer*. Oct 2022;30(10):7983-7989. [doi: [10.1007/s00520-022-07240-x](https://doi.org/10.1007/s00520-022-07240-x)] [Medline: [35759048](https://pubmed.ncbi.nlm.nih.gov/35759048/)]
10. Song X, He Y, Bai J, Zhang J. A nomogram based on nutritional status and A²DS² score for predicting stroke-associated pneumonia in acute ischemic stroke patients with type 2 diabetes mellitus: a retrospective study. *Front Nutr*. 2022;9:1009041. [doi: [10.3389/fnut.2022.1009041](https://doi.org/10.3389/fnut.2022.1009041)] [Medline: [36313103](https://pubmed.ncbi.nlm.nih.gov/36313103/)]
11. Wang X, Yao K, Huang Z, et al. An artificial intelligence malnutrition screening tool based on electronic medical records. *Clin Nutr ESPEN*. Aug 2025;68:153-159. [doi: [10.1016/j.clnesp.2025.03.178](https://doi.org/10.1016/j.clnesp.2025.03.178)] [Medline: [40311925](https://pubmed.ncbi.nlm.nih.gov/40311925/)]
12. Lee JM, Hauskrecht M. Recent context-aware LSTM for clinical event time-series prediction. Presented at: Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME; Jun 26-29, 2019:13-23; Poznan, Poland. [doi: [10.1007/978-3-030-21642-9_3](https://doi.org/10.1007/978-3-030-21642-9_3)]
13. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. *Artif Intell Rev*. 2024;57(10):273. [doi: [10.1007/s10462-024-10884-2](https://doi.org/10.1007/s10462-024-10884-2)]
14. Gao J, Zhu Y, Wang W, et al. A comprehensive benchmark for COVID-19 predictive modeling using electronic health records in intensive care. *Patterns (N Y)*. Apr 12, 2024;5(4):100951. [doi: [10.1016/j.patter.2024.100951](https://doi.org/10.1016/j.patter.2024.100951)] [Medline: [38645764](https://pubmed.ncbi.nlm.nih.gov/38645764/)]

15. Bohlen L, Rosenberger J, Zschech P, Kraus M. Leveraging interpretable machine learning in intensive care. *Ann Oper Res*. Apr 2025;347(2):1093-1132. [doi: [10.1007/s10479-024-06226-8](https://doi.org/10.1007/s10479-024-06226-8)]
16. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. Jan 3, 2023;10(1):1. [doi: [10.1038/s41597-022-01899-x](https://doi.org/10.1038/s41597-022-01899-x)] [Medline: [36596836](https://pubmed.ncbi.nlm.nih.gov/36596836/)]
17. Liu Y, Xu Y, Guo L, et al. Development and external validation of machine learning models for the early prediction of malnutrition in critically ill patients: a prospective observational study. *BMC Med Inform Decis Mak*. Jul 3, 2025;25(1):248. [doi: [10.1186/s12911-025-03082-9](https://doi.org/10.1186/s12911-025-03082-9)] [Medline: [40611154](https://pubmed.ncbi.nlm.nih.gov/40611154/)]
18. Benito GV, Goldberg X, Brachowicz N, et al. Machine learning for anxiety and depression profiling and risk assessment in the aftermath of an emergency. *Artif Intell Med*. Nov 2024;157:102991. [doi: [10.1016/j.artmed.2024.102991](https://doi.org/10.1016/j.artmed.2024.102991)] [Medline: [39383706](https://pubmed.ncbi.nlm.nih.gov/39383706/)]
19. Baniecki H, Sobieski B, Szatkowski P, Bombinski P, Biecek P. Interpretable machine learning for time-to-event prediction in medicine and healthcare. *Artif Intell Med*. Jan 2025;159:103026. [doi: [10.1016/j.artmed.2024.103026](https://doi.org/10.1016/j.artmed.2024.103026)] [Medline: [39579416](https://pubmed.ncbi.nlm.nih.gov/39579416/)]
20. Boulitsakis Logothetis S, Green D, Holland M, Al Moubayed N. Predicting acute clinical deterioration with interpretable machine learning to support emergency care decision making. *Sci Rep*. Aug 21, 2023;13(1):13563. [doi: [10.1038/s41598-023-40661-0](https://doi.org/10.1038/s41598-023-40661-0)] [Medline: [37604974](https://pubmed.ncbi.nlm.nih.gov/37604974/)]
21. Halder RK, Uddin MN, Uddin M, Aryal S, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *J Big Data*. 2024;11(1). [doi: [10.1186/s40537-024-00973-y](https://doi.org/10.1186/s40537-024-00973-y)]
22. Wang X, Chen Y, Hua L, Wang D, Zhang X, Wang L. Development and validation of a LASSO-based predictive model for inadvertent hypothermia in ICU patients. *Front Med (Lausanne)*. 2025;12:1596030. [doi: [10.3389/fmed.2025.1596030](https://doi.org/10.3389/fmed.2025.1596030)] [Medline: [40901524](https://pubmed.ncbi.nlm.nih.gov/40901524/)]
23. Garouani M, Barhrouj A, Teste O. XStacking: an effective and inherently explainable framework for stacked ensemble learning. *Inf Fusion*. Dec 2025;124:103358. [doi: [10.1016/j.inffus.2025.103358](https://doi.org/10.1016/j.inffus.2025.103358)]
24. Ashika T, Hannah Grace G. Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: an optimized RST-ML approach. *Front Digit Health*. 2025;7:1609308. [doi: [10.3389/fgth.2025.1609308](https://doi.org/10.3389/fgth.2025.1609308)] [Medline: [40613076](https://pubmed.ncbi.nlm.nih.gov/40613076/)]
25. Hua Y, Stead TS, George A, Ganti L. Clinical risk prediction with logistic regression: best practices, validation techniques, and applications in medical research. *Acad Med Surg*. [doi: [10.62186/001c.131964](https://doi.org/10.62186/001c.131964)]
26. Liu B, Mazumder R. Randomization can reduce both bias and variance: a case study in random forests. *J Mach Learn Res*. 2025;26:1-49. [doi: [10.48550/arXiv.2402.12668](https://doi.org/10.48550/arXiv.2402.12668)]
27. Olaniran OR, Olaniran SF, Alzahrani ARR, Alharbi NM, Alzahrani ARR. Random forest adaptation for high-dimensional count regression. *Mathematics (Basel)*. 2025;13(18):3041. [doi: [10.3390/math13183041](https://doi.org/10.3390/math13183041)]
28. Yıldız AY, Kalayci A. Gradient boosting decision trees on medical diagnosis over tabular data. Presented at: 2025 IEEE International Conference on AI and Data Analytics (ICAD); Jun 24-25, 2025; Medford, MA. [doi: [10.1109/ICAD65464.2025.11114069](https://doi.org/10.1109/ICAD65464.2025.11114069)]
29. Zhu H, Liu S, Xu W, Dai J, Benbouzid M. Linearithmic and unbiased implementation of DeLong's algorithm for comparing the areas under correlated ROC curves. *Expert Syst Appl*. Jul 2024;246:123194. [doi: [10.1016/j.eswa.2024.123194](https://doi.org/10.1016/j.eswa.2024.123194)]
30. Kilic ME, Tufekcioglu OA, Yilancioglu YR, Turan OE, Ozcan EE. Explainable AI for specific arrhythmia detection: SHAP-based insights from multi-lead ECG data. *J Med Biol Eng*. Apr 2025;45(2):314-324. [doi: [10.1007/s40846-025-00949-0](https://doi.org/10.1007/s40846-025-00949-0)]
31. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. Jan 8, 2024;384:e074819. [doi: [10.1136/bmj-2023-074819](https://doi.org/10.1136/bmj-2023-074819)] [Medline: [38191193](https://pubmed.ncbi.nlm.nih.gov/38191193/)]
32. Honda Y, Nagai T, Iwakami N, et al. Usefulness of geriatric nutritional risk index for assessing nutritional status and its prognostic impact in patients aged ≥ 65 years with acute heart failure. *Am J Cardiol*. Aug 15, 2016;118(4):550-555. [doi: [10.1016/j.amjcard.2016.05.045](https://doi.org/10.1016/j.amjcard.2016.05.045)] [Medline: [27324158](https://pubmed.ncbi.nlm.nih.gov/27324158/)]
33. Zhu S, Zhao H, Zheng M, Peng J. The impact of malnutrition on atrial fibrillation recurrence post ablation. *Nutr Metab Cardiovasc Dis*. Mar 10, 2021;31(3):834-840. [doi: [10.1016/j.numecd.2020.12.003](https://doi.org/10.1016/j.numecd.2020.12.003)] [Medline: [33549446](https://pubmed.ncbi.nlm.nih.gov/33549446/)]
34. Park JH, Kim E, Seol EM, et al. Prediction model for screening patients at risk of malnutrition after gastric cancer surgery. *Ann Surg Oncol*. Aug 2021;28(8):4471-4481. [doi: [10.1245/s10434-020-09559-3](https://doi.org/10.1245/s10434-020-09559-3)] [Medline: [33481124](https://pubmed.ncbi.nlm.nih.gov/33481124/)]
35. Rietveld SCM, Witvliet-van Nierop JE, Ottens-Oussoren K, van der Peet DL, de van der Schueren MAE. The prediction of deterioration of nutritional status during chemoradiation therapy in patients with esophageal cancer. *Nutr Cancer*. 2018;70(2):229-235. [doi: [10.1080/01635581.2018.1412481](https://doi.org/10.1080/01635581.2018.1412481)] [Medline: [29324061](https://pubmed.ncbi.nlm.nih.gov/29324061/)]

36. Leoni MLG, Moschini E, Beretta M, Zanella M, Nolli M. The modified NUTRIC score (mNUTRIC) is associated with increased 28-day mortality in critically ill COVID-19 patients: internal validation of a prediction model. *Clin Nutr ESPEN*. Apr 2022;48:202-209. [doi: [10.1016/j.clnesp.2022.02.014](https://doi.org/10.1016/j.clnesp.2022.02.014)] [Medline: [35331492](https://pubmed.ncbi.nlm.nih.gov/35331492/)]
37. Li Z, Li J, Kuang K, et al. SimTA++: simple attention neural network for clinical asynchronous time series. *Neural Netw*. Nov 2025;191:107735. [doi: [10.1016/j.neunet.2025.107735](https://doi.org/10.1016/j.neunet.2025.107735)] [Medline: [40570801](https://pubmed.ncbi.nlm.nih.gov/40570801/)]
38. Chauhan VK, Thakur A, O'Donoghue O, Rohanian O, Molaei S, Clifton DA. Continuous patient state attention model for addressing irregularity in electronic health records. *BMC Med Inform Decis Mak*. May 3, 2024;24(1):117. [doi: [10.1186/s12911-024-02514-2](https://doi.org/10.1186/s12911-024-02514-2)] [Medline: [38702692](https://pubmed.ncbi.nlm.nih.gov/38702692/)]
39. Alberda C, Gramlich L, Jones N, et al. The relationship between nutritional intake and clinical outcomes in critically ill patients: results of an international multicenter observational study. *Intensive Care Med*. Oct 2009;35(10):1728-1737. [doi: [10.1007/s00134-009-1567-4](https://doi.org/10.1007/s00134-009-1567-4)] [Medline: [19572118](https://pubmed.ncbi.nlm.nih.gov/19572118/)]
40. Evans DC, Corkins MR, Malone A, et al. The use of visceral proteins as nutrition markers: an ASPEN position paper. *Nutr Clin Pract*. Feb 2021;36(1):22-28. [doi: [10.1002/ncp.10588](https://doi.org/10.1002/ncp.10588)] [Medline: [33125793](https://pubmed.ncbi.nlm.nih.gov/33125793/)]
41. Jensen GL, Cederholm T, Correia M, et al. GLIM criteria for the diagnosis of malnutrition: a consensus report from the global clinical nutrition community. *JPEN J Parenter Enteral Nutr*. Jan 2019;43(1):32-40. [doi: [10.1002/jpen.1440](https://doi.org/10.1002/jpen.1440)] [Medline: [30175461](https://pubmed.ncbi.nlm.nih.gov/30175461/)]
42. Friedli N, Stanga Z, Culkin A, et al. Management and prevention of refeeding syndrome in medical inpatients: an evidence-based and consensus-supported algorithm. *Nutrition*. Mar 2018;47:13-20. [doi: [10.1016/j.nut.2017.09.007](https://doi.org/10.1016/j.nut.2017.09.007)] [Medline: [29429529](https://pubmed.ncbi.nlm.nih.gov/29429529/)]
43. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data*. Sep 11, 2018;5(1):180178. [doi: [10.1038/sdata.2018.178](https://doi.org/10.1038/sdata.2018.178)] [Medline: [30204154](https://pubmed.ncbi.nlm.nih.gov/30204154/)]
44. Iba T, Levy JH, Warkentin TE, et al. Diagnosis and management of sepsis-induced coagulopathy and disseminated intravascular coagulation. *J Thromb Haemost*. Nov 2019;17(11):1989-1994. [doi: [10.1111/jth.14578](https://doi.org/10.1111/jth.14578)] [Medline: [31410983](https://pubmed.ncbi.nlm.nih.gov/31410983/)]
45. Chen CH, Pai KC, Hsieh HM, Chan YJ, Hsu HL, Wang CY. Artificial intelligence assisted nutritional risk evaluation model for critically ill patients: integration of explainable machine learning in intensive care nutrition. *Asia Pac J Clin Nutr*. Jun 2025;34(3):343-352. [doi: [10.6133/apjcn.202506_34\(3\).0009](https://doi.org/10.6133/apjcn.202506_34(3).0009)] [Medline: [40419395](https://pubmed.ncbi.nlm.nih.gov/40419395/)]
46. Di Martino F, Delmastro F, Dolciotti C. Explainable AI for malnutrition risk prediction from m-Health and clinical data. *Smart Health (Amst)*. Dec 2023;30:100429. [doi: [10.1016/j.smhl.2023.100429](https://doi.org/10.1016/j.smhl.2023.100429)]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
E-NUTRIC: Ensemble-NUTRIC
ECE: expected calibration error
EHR: electronic health record
ICD-10: *International Statistical Classification of Diseases, Tenth Revision*
ICD-9: *International Classification of Diseases, Ninth Revision*
ICU: intensive care unit
KNN: k-nearest neighbor
LightGBM: Light Gradient Boosting Machine
LR: Logistic Regression
MIMIC-IV: Medical Information Mart for Intensive Care
ML: machine learning
mNUTRIC: modified NUTRIC
RF: Random Forest
RUS: random undersampling
SHAP: Shapley Additive Explanations
SOFA: Sequential Organ Failure Assessment
XGBoost: Extreme Gradient Boosting

Edited by Andrew Coristine, Arriel Benis; peer-reviewed by Burcu Ozek, Subhash Nerella, Weihua Li; submitted 21.May.2025; final revised version received 21.Apr.2026; accepted 21.Apr.2026; published 03.Jun.2026

Please cite as:

Zhang X, Fang A, Lou P, Yao K, Huang T, Hu J
Development and Interpretability Analysis of a Stacking Ensemble Model for Early Prediction of Nutritional Risk in Intensive Care Unit Patients: Retrospective Cohort Study
JMIR Med Inform 2026;14:e77872
URL: <https://medinform.jmir.org/2026/1/e77872>
doi: [10.2196/77872](https://doi.org/10.2196/77872)

© Xu Zhang, An Fang, Pei Lou, Kuanda Yao, Tianci Huang, Jiahui Hu. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 03.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.