Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
DOI: 10.31466/kfbd.1678766
ISSN (Online): 2564-7377
https://dergipark.org.tr/tr/pub/kfbd
Research Article / Araştırma Makalesi

Improving ICU Mortality Prediction via Meta-Learning and Explainable AI: A MetaCost and LIME Approach

Meta-Öğrenme ve Açıklanabilir Yapay Zekâ ile Yoğun Bakım Ölüm Tahmininin İyileştirilmesi: MetaCost ve LIME Yaklaşımı

Fahrettin KAYA[1]*

Abstract

This study aimed to enhance mortality prediction for Intensive Care Unit (ICU) patients using a Meta-Learning approach and to evaluate the explainability of individual predictions using the LIME (Local Interpretable Model-agnostic Explanations) method. This study analyzed 428 patient records from the MIMIC-III database, including 48 variables including demographics, laboratory results (e.g., Anion gap, Urea nitrogen), and comorbidities. The dataset was imbalanced, with 15% mortality and 85% survival. To address this issue, machine learning models (e.g., Gradient Boosting, Random Forest) were adapted using the MetaCost algorithm, which is a meta-learning method. Performance was evaluated using metrics suited for imbalanced data, such as Average Precision (AP), recall, F2 score, and the Matthews correlation coefficient (MCC). Feature importance was validated statistically, and LIME was applied for per-patient interpretability. Univariate analysis identified 24 statistically significant features ($P<0.01$) differentiating between deceased and surviving patients. The MetaCost-enhanced Gradient Boosting model achieved the best performance, with an AUC of 0.91, AP of 0.75, recall of 0.86, F2 score of 0.85, and MCC of 0.79. The MetaCost algorithm effectively improves ICU mortality prediction accuracy, while LIME enhances interpretability at the individual patient level. This approach can make clinical decision support systems more transparent and reliable. However, further validation on diverse datasets is required to confirm these findings.

Keywords: ICU mortality prediction, Machine learning, Logistic regression, Explainable AI.

Öz

Bu çalışma, yoğun bakım ünitesindeki (YBÜ) hastalar için mortalite tahminini geliştirmek amacıyla bir Meta-Öğrenme yaklaşımı kullanmayı ve bireysel tahminlerin açıklanabilirliğini LIME yöntemiyle değerlendirmeyi amaçlamıştır. Çalışmada, MIMIC-III veri tabanından alınan 428 hasta kaydı analiz edilmiştir. Veriler; demografik bilgiler, laboratuvar sonuçları (örneğin, Anyon açığı, Üre azotu) ve yandaş hastalıklar gibi 48 değişken içermektedir. Veri seti dengesizdir; mortalite oranı %15, sağkalım oranı ise %85'tir. Bu dengesizliği gidermek amacıyla, Meta-Öğrenme yöntemi olan MetaCost algoritması kullanılarak Makine Öğrenmesi modelleri (örneğin, Gradient Boosting, Random Forest) uyarlanmıştır. Model performansı, dengesiz veri setlerine uygun metriklerle değerlendirilmiştir: Ortalama Kesinlik (AP), duyarlılık (recall), F2 skoru ve Matthews korelasyon katsayısı (MCC). Özellik önem dereceleri istatistiksel olarak doğrulanmış, bireysel hasta düzeyinde açıklanabilirlik için LIME yöntemi uygulanmıştır. Tek değişkenli analiz sonucunda, ölen ve hayatta kalan hastaları ayırt eden 24 istatistiksel olarak anlamlı değişken ($P<0.01$) belirlenmiştir. MetaCost ile güçlendirilmiş Gradient Boosting modeli en iyi performansı göstermiştir; AUC: 0.91, AP: 0.75, duyarlılık: 0.86, F2 skoru: 0.85 ve MCC: 0.79. MetaCost algoritması, YBÜ mortalite tahminin de doğruluğu artırmada etkili bulunmuştur. LIME yöntemi ile modelin bireysel hasta düzeyinde yorumlanabilirliğini artırmıştır. Bu yaklaşım, klinik karar destek sistemlerini daha şeffaf ve güvenilir hale getirebilir. Ancak, bulguların doğrulanması için farklı veri setlerinde ek çalışmalara ihtiyaç vardır.

Anahtar Kelimeler: Yoğun bakım mortalite tahmini, Makine öğrenmesi, Lojistik regresyon, Açıklanabilir AI.

[1]Kahramanmaraş Sütçü İmam University Andırın Vocational School, Computer Technologies Department, Kahramanmaraş, Türkiye

***Corresponding Author/Sorumlu Yazar:** fkaya@ksu.edu.tr

Received/Geliş: 17.04.2025 **Accepted/Kabul:** 29.09.2025 **Published/Yayın:** 24.02.2026

Kaya, F. (2026). Improving ICU Mortality Prediction via Meta-Learning and Explainable AI: A MetaCost and LIME Approach. *Karadeniz Fen Bilimleri Dergisi*, 16(1), 140-155. https://doi.org/10.31466/kfbd.1678766

1. Introduction

Intensive Care Units (ICUs) are critical medical settings where patients with severe conditions are treated to improve their survival. Mortality in ICUs is determined by the interaction of numerous factors, and analysis of these factors benefits significantly from advancements in computer technology and the availability of large datasets. Machine Learning (ML), in particular, has shown potential in supporting early diagnosis and personalized treatment approaches in ICUs by leveraging data to provide predictive insights and decision support (Núñez Reiz et al., 2018). ML algorithms are typically designed to perform well on balanced datasets. However, real-world data from intensive care settings often exhibit imbalanced distributions, which can compromise the accuracy of predictions of rare events, such as mortality. Traditional ML approaches offer "balanced" techniques to address this issue; however, these solutions are not always sufficient. This challenge can be mitigated using cost-sensitive learning methods, which consider the costs of misclassification to improve model performance (Ling and Sheng, 2008).

To further enhance model performance and optimize predictions, such as mortality prediction, Meta-Learning approaches provide a flexible and effective framework. In addition to model optimization, explainability is crucial because clinical decision-support systems must be interpretable. Tools like LIME enhance the interpretability of patient-specific predictions, enabling healthcare professionals to adopt model recommendations with greater confidence (Ribeiro et al., 2016; Çanga Boğa et al., 2024; Önder et al., 2025).

This study aims to develop a methodological framework for mortality prediction in ICU datasets by addressing the challenge of class imbalance and providing patient-specific explanations. To achieve this, we employ a cost-sensitive learning strategy (MetaCost) to enhance predictive accuracy and an explainability tool (LIME) to ensure interpretability of model outputs. The proposed framework is designed not to claim clinical novelty, but rather to contribute methodologically by delivering reliable and transparent results that can support decision-making in healthcare analytics.

Fundamental Machine Learning Algorithms:

ML is a critical tool for analyzing healthcare data and developing clinical decision support systems. Logistic regression, particularly in binary classification problems, establishes a linear relationship based on logit transformation and offers the advantage of interpretability (Eratlı ŞY and Şahin, 2020; Yavuz, 2023). Decision trees partition data into subgroups and apply to both numerical and categorical variables, although they are prone to overfitting (Quinlan, 1986). Random Forest improves generalization by combining multiple decision trees (Breiman, 2001). XGBoost focuses on sequential decision trees to correct errors, demonstrating efficacy in predicting critical conditions, such as sepsis (Chen and Guestrin, 2016). Similarly, the Gradient Boosting Classifier enhances

accuracy through sequential weak learners, employing regularization techniques to mitigate overfitting (Friedman, 2000).

Performance Metrics:

Various performance metrics are employed to evaluate the effectiveness of ML algorithms, particularly in healthcare, where the costs of false positives (FP) and false negatives (FN) are critical. A false negative outcome may result in severe cases being undetected and untreated, whereas a false positive outcome could lead to unnecessary interventions and anxiety. In this context, true positives (TP) represent cases correctly classified as deceased, whereas true negatives (TN) correspond to survivors accurately identified by the model. Important performance metrics for imbalanced datasets include recall, F2 score, Matthews Correlation Coefficient (MCC), and Average Precision (AP). The F2 score is particularly relevant in scenarios where the cost of false negatives is high because it places greater emphasis on recall, thereby offering a more accurate assessment of model performance. Average Precision (AP) evaluates the ability to distinguish the positive class by measuring the area under the precision-recall curve, providing reliable results for imbalanced datasets (Saito and Rehmsmeier, 2015; Çelik and Yilmaz, 2021; Çanga and Boğa, 2020; Çanga and Boğa, 2022). These methods are essential tools for evaluating ML algorithms in healthcare applications, particularly for improving predictive accuracy in clinical settings. These metrics can be summarized as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$F2 = (1+2^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(2^2 \cdot \text{Precision}) + \text{Recall}} \tag{4}$$

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n \tag{5}$$

here:

P_n : Precision value at the n-th recall threshold.

R_n : Recall value at the n-th recall threshold.

The Matthews correlation coefficient (MCC) provides an effective evaluation in imbalanced datasets because it considers FP, FN, TP, and TN in a balanced manner. This metric ranges from -1 to 1, where 1 indicates perfect classification (Chicco et al., 2021; Powers, 2020).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

Additionally, in the literature, performance metrics such as AUC, kappa statistic, and F1 score are widely used as general performance indicators.

Interpretability and Transparency in Machine Learning:

LIME (Local Interpretable Model-agnostic Explanations) is a technique designed to explain individual predictions made by complex machine learning models. Rather than attempting to explain the model as a whole, LIME focuses on interpreting the reasoning behind specific predictions. The proposed method perturbs the input data of a particular instance to create a local neighborhood of similar examples. Then, it uses these perturbed data points to train a simpler, interpretable model (such as a linear regression) that approximates the decision boundary of the original model within a local region. This approach allows for a detailed explanation of why a particular prediction was made, thereby helping users understand which features influenced the outcome most strongly. In clinical practice, such local explanations have been shown to align with clinicians' decision-making needs — for instance, Tonekaboni et al. (2019) demonstrated that LIME-style explanations improve clinician trust and usability when model outputs are contextualized for real-world diagnostic workflows. Thus, the proposed method enables greater transparency in machine learning applications, particularly in critical areas like healthcare, where model decisions must be understandable to clinicians. This is further supported by Lundberg et al. (2020), who applied local interpretability methods (including LIME) to clinical risk prediction models in nephrology and critical care, confirming that feature-level explanations significantly enhance model adoption in hospital settings. By providing local explanations, LIME enhances the practical trustworthiness and usability of complex predictive models in practice (Ribeiro et al., 2016).

2. Materials and Methods

In this study, we utilized a previously published subset of the Medical Information Mart for Intensive Care III (MIMIC-III) database (Johnson et al., 2016), which was prepared and shared by Zhou et al. (2023). This dataset includes 1,177 adult patients with a diagnosis of heart failure (HF), identified by manual review of ICD-9 codes, admitted to the intensive care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. Patients without available records for left ventricular ejection fraction (LVEF) or N-terminal pro-brain natriuretic peptide (NT-proBNP) had already been excluded in the original dataset construction.

For the present analysis, additional preprocessing was performed. Specifically, patient records containing missing values in any of the study variables were excluded, resulting in a final cohort of 428 patients with complete data across 48 predictor variables and one outcome variable (in-hospital

mortality). Among these patients, 85% survived and 15% died, reflecting the class imbalance that was addressed during model development.

Access to the MIMIC-III database and the use of data extracted from it were granted upon completion of the NIH Protecting Human Research Participants training and approval in accordance with data use policies (Certificate Record ID: 61633342). As the dataset is publicly available and fully de-identified, no additional institutional ethics approval was required.

In this study, the MetaCost algorithm, which was proposed by (Domingos, 1999), was used to improve the performance of imbalanced data structures. MetaCost is a cost-sensitive learning method that optimizes any base classifier via labeling based on a cost matrix. The simplified steps of the MetaCost algorithm are summarized as follows:

Input:

- A training dataset S
- A base classifier L
- A cost matrix C, where $C(i,j)$represents the cost of predicting class i when the true class is j

Steps:

1. Resampling:
   - Generate m resampled datasets $S_1, S_2, \ldots, S_m$ from the original dataset S.
   - Each S_i is created by randomly sampling $n = |S|$ from S with replacement.
2. Model Training:
   - For each resampled dataset S_i, train a model M_i using the base classifier L.
3. Probability Estimation:
   - For each example $x \in S$ and for each class j, estimate the class probability $P(j|x)$ by aggregating predictions from all M_i. for which x was not in S_i
   - If L supports probabilistic output, use those probabilities directly. Otherwise, use the fraction of models for which x was not in S_i predicting class j.
4. Cost-sensitive Relabeling:
   - For each example x and each possible class label i, compute the expected cost (risk function):
   - $R(i|x) = \sum_j P(j|x)C(i,j)$
   - Reassign x the label i^* that minimizes the expected cost:

$$i^* = \arg\min_i R(i \mid x)$$

5. Final Model Training:
   - Using the newly relabeled dataset, train the final model with the base classifier L.

Output:

- A cost-sensitive classifier trained on relabeled data, minimizing the expected classification cost.

For binary classification, the cost matrix was defined to reflect the relative importance of errors in the minority and majority classes as follows:

$$C = \begin{pmatrix} C_{TN} & C_{FN} \\ C_{FP} & C_{TP} \end{pmatrix} = \begin{pmatrix} C(0,0) & C(0,1) \\ C(1,0) & C(1,1) \end{pmatrix} \quad (7)$$

The correct classifications $C_{TN} = 0$ and $C_{TP} = 0$ were assigned a cost of 0 in the cost matrix. C_{FN} and C_{FP} values were determined approximately based on the ratio of the number of observations in a class to the total number of observations. Thus, C_{FN} was set higher than C_{FP}, prioritizing errors in the minority class. In the case of three-class classification, the cost matrix can be defined as a 3x3 matrix, as shown in Equation (7). Additionally, the Python code of this algorithm is provided in Appendix-1.

2.1. Statistical analysis

A total of 48 features were analyzed between the deceased and surviving groups based on the outcome variable. For categorical variables, frequencies and percentages were calculated, and p-values were determined using the chi-square test. For continuous variables, the range, mean ± standard deviation were calculated; normality and homogeneity of variances were tested. Continuous variables meeting these assumptions were analyzed using the Two independent sample t-test, whereas those not meeting the assumptions were analyzed using the Mann–Whitney U test. Additionally, p-values from the univariate logistic regression analyses were calculated for all variables. All analyses were performed using Python 3.10 and relevant packages.

3. Findings and Discussion

3.1 Statistical results

Table 1. Descriptive and inferential statistics of categorical and continuous features by outcome group

Features	Survived patients (outcome=0) n= 363 (85%)	Dead patients (outcome=1) n=65 (15%)	p-value	p-value*
Renal failure, n (%)	145 (40)	12(18)	0.002 Φ	0.002
Deficiency anemias, n (%)	133(37)	11(17)	0.003 Φ	0.003
depression, n (%)	53(15)	3(5)	0.046 Φ	0.038
atrialfibrillation, n (%)	148(41)	33(51)	0.17 Φ	0.13
COPD, n (%), n (%)	34(9)	3(5)	0.31 Φ	0.22
Hyperlipemia	121(33)	26(40)	0.37 Φ	0.3
diabetes, n (%)	170(47)	26(40)	0.38 Φ	0.31
Male, n (%)	178(49)	29(45)	0.6 Φ	0.5
hypertensive, n (%)	257(71)	44(68)	0.72 Φ	0.61
CHD with no MI, n (%)	24(7)	5(8)	0.96 Φ	0.75
Anion gap	13.77 ± 2.34, (7.78- 25.27)	16.06 ± 3.33, (10.33- 24.88)	<0.001 Ω	<0.001
Urea nitrogen	35.28 ± 19.65, (6.22- 109.13)	50.71 ± 26.8, (18.22- 161.75)	<0.001 Θ	<0.001
Bicarbonate	27.44 ± 5.41, (15.07- 47.67)	23.37 ± 5.58, (12.86- 40.57)	<0.001 Ω	<0.001
Blood calcium	8.51 ± 0.53, (6.97- 10.95)	8.12 ± 0.63, (6.7- 9.72)	<0.001 Ω	<0.001
Lymphocyte	13.28 ± 8.09, (1.45- 60.5)	8.96 ± 7.22, (0.97- 40.57)	<0.001 Θ	<0.001
Leukocyte	10.22 ± 4.36, (2.06- 33.34)	14.22 ± 8.83, (3.67- 64.75)	<0.001 Ω	<0.001
Lactic acid	1.77 ± 0.86, (0.5- 6.45)	2.6 ± 1.39, (0.75- 6.72)	<0.001 Θ	<0.001
Urine output	2054 ± 1292, (0- 8820)	1438 ± 1243, (134- 6700)	<0.001 Ω	<0.001
PT	16.73 ± 6.06, (10.1- 50.02)	21.17 ± 12.1, (11.18- 71.27)	<0.001 Θ	<0.001
Platelets	248.26 ± 113.48, (30.32- 1028)	191 ± 101.7, (30.36- 584.64)	<0.001 Ω	<0.001
INR	1.54 ± 0.7, (0.9- 5.42)	2.05 ± 1.39, (0.94- 8.34)	<0.001 Θ	<0.001
PH	7.38 ± 0.06, (7.17- 7.56)	7.35 ± 0.07, (7.16- 7.48)	<0.001 Ω	<0.001
heart rate	83.98 ± 16.13, (36- 135.71)	90.96 ± 15.82, (59.08- 126.38)	0.002 Ω	0.002
NT-proBNP	10890 ± 12400, (50- 63198)	17425 ± 19558, (694- 118928)	0.002 Θ	0.002
Creatinine	1.63 ± 1.28, (0.27- 15.53)	1.83 ± 0.81, (0.7- 4.41)	0.002 Θ	0.25
Basophils	0.41 ± 0.33, (0.1- 3.0)	0.33 ± 0.32, (0.1- 2.0)	0.002 Ω	0.068
age	71.68 ± 13.43, (35- 98)	77.17 ± 12.52, (40- 99)	0.003 Ω	0.003
temperature	36.75 ± 0.65, (34.32- 39.13)	36.49 ± 0.68, (34.61- 38.25)	0.003 Ω	0.004
Blood potassium	4.15 ± 0.37, (3.24- 5.82)	4.31 ± 0.45, (3.4- 5.52)	0.003 Ω	0.004
Respiratory rate	20.84 ± 4.29, (12- 40.9)	22.41 ± 4.47, (12.36- 33.85)	0.006 Ω	0.008
Neutrophils	79.71 ± 9.63, (36.4- 96.6)	83.39 ± 12.25, (28.33- 97.1)	0.007 Ω	0.007
Systolic blood pressure	118.18 ± 16.78, (85.06- 180.7)	112.12 ± 16.32, (88.52- 174.24)	0.008 Ω	0.008
Chloride	102.37 ± 5.16, (85.64- 116.73)	103.97 ± 5.26, (90.5- 117.32)	0.023 Θ	0.023
Diastolic blood pressure	60.65 ± 9.99, (32.81- 95.91)	57.67 ± 8.95, (24.74- 78.32)	0.025 Θ	0.026
RDW	15.93 ± 2.02, (12.09- 23.82)	16.53 ± 2.56, (12.85- 29.05)	0.038 Ω	0.041
glucose	153.35 ± 50.71, (69.1- 369)	170.49 ± 63.02, (77.12- 342.5)	0.038 Θ	0.018
BMI	31.28 ± 9.96, (13.67- 83.26)	28.51 ± 7.12, (15.82- 65.12)	0.055 Θ	0.034
PCO2	45.72 ± 13.02, (19.5- 98.6)	43.51 ± 12.98, (26.33- 92)	0.11 Θ	0.21
EF	48.4 ± 12.76, (15- 75)	46.77 ± 14.54, (20- 75)	0.18 Θ	0.35
Magnesium ion	2.13 ± 0.24, (1.41- 4.07)	2.18 ± 0.27, (1.8- 3.25)	0.23 Θ	0.10
Creatine kinase	238 ± 963, (10.25- 16829)	941 ± 5322, (12- 42987)	0.36 Θ	0.15

MCV	89.57 ± 6.68, (64.62- 116.71)	90.38 ± 6.88, (74- 105)	0.37 Ω	0.36
SP O2	96.17 ± 2.37, (87.47- 100)	95.93 ± 3.47, (75.92- 99.82)	0.47 Ω	0.47
Blood sodium	139.38 ± 3.76, (121.05- 149.13)	139.04 ± 4.8, (124.46- 149)	0.52 Ω	0.52
MCH	29.39 ± 2.64, (18.12- 40.31)	29.58 ± 2.57, (23.18- 35.2)	0.59 Θ	0.58
MCHC	32.82 ± 1.39, (28.02- 37.01)	32.75 ± 1.34, (29.35- 35.75)	0.70 Ω	0.70
RBC	3.58 ± 0.65, (2.15- 6)	3.56 ± 0.68, (2.22- 6.15)	0.86 Θ	0.83
hematocrit	31.83 ± 5.42, (21.15- 55.42)	31.9 ± 5.4, (22.2- 45.73)	0.92 Θ	0.92

Renal failure: Kidney failure, deficiency anemias: Deficiency anemias, depression: Depression, atrial fibrillation: Atrial fibrillation, COPD: Chronic Obstructive Pulmonary Disease, Hyperlipemia: High levels of lipids in the blood, diabetes: Diabetes, Gender: Gender, hypertensive: Hypertension, CHD with no MI: Coronary Heart Disease without Myocardial Infarction , Anion gap: Anion gap (mEq/L), Urea nitrogen: Blood urea nitrogen level (mg/dL), Bicarbonate: Blood bicarbonate level (mmol/L), Blood calcium: Blood calcium level (mg/dL), Lymphocyte: Lymphocyte count (%), Leukocyte: White blood cell count (10^9/L), Lactic acid: Blood lactic acid level (mmol/L), Urine output: Volume of urine output (mL), PT: Prothrombin Time (clotting) (seconds), Platelets: Platelet count (10^9/L), INR: International Normalized Ratio (for blood clotting) (Ratio), PH: Blood pH level (pH units), heart rate: Heart rate (beats/min), NT-proBNP: N-terminal pro b-type Natriuretic Peptide (pg/mL), Creatinine: Blood creatinine level (mg/dL), Basophils: Basophil count (%), age: Age (years), temperature: Body temperature (°C), Blood potassium: Blood potassium level (mmol/L), Respiratory rate: Respiratory rate (breaths/min), Neutrophils: Neutrophil count (%), Systolic blood pressure: Systolic blood pressure (mmHg), Chloride: Blood chloride level (mmol/L), Diastolic blood pressure: Diastolic blood pressure (mmHg), RDW: Red Cell Distribution Width (%), glucose: Blood glucose level (mg/dL), BMI: Body Mass Index (kg/m²), PCO2: Partial Pressure of Carbon Dioxide (mmHg), EF: Ejection Fraction (%), Magnesium ion: Blood magnesium ion level (mg/dL), Creatine kinase: Creatine kinase enzyme level (U/L), MCV: Mean Corpuscular Volume (fL), SpO_2: Blood oxygen saturation (%), Blood sodium: Blood sodium level (mmol/L), MCH: Mean Corpuscular Hemoglobin (pg), MCHC: Mean Corpuscular Hemoglobin Concentration (g/dL), RBC: Red blood cell count (10^{12}/L), hematocrit: Hematocrit (%), Ω: Two independent sample t-test, Θ: Mann-Whitney U test, Φ: Chi Square test, p-value*: Logistic Regression

According to the findings in Table 1, 40% of the surviving patients (n=363) exhibited renal failure, compared with 18% of the deceased individuals (n=65). Similarly, deficiency anemia was observed in 37% of surviving and 17% of deceased patients. Statistical analyses indicated that these two features were significant predictors of mortality ($P<0.01$). For other features (e.g., depression, atrial fibrillation), no significant differences were observed between the deceased and surviving patients ($P>0.05$).

Univariate analyses (Two independent sample t-tests, Mann–Whitney U tests, and logistic regression) revealed significant differences in several continuous variables between deceased and surviving patients. Significant biochemical, hematological, and physiological parameters included anion gap, urine output, bicarbonate, creatinine, lymphocyte percentage, white blood cell count (WBC), and prothrombin time (PT), Lactic acid, Platelets, INR, PH, and heart rate (all $P < 0.001$). For example, the anion gap and urine output were markedly higher and lower, respectively, in deceased patients than in survivors. Additionally, variables such as heart rate and NT-proBNP levels were significantly different between the groups ($P< 0.01$). The full statistical details and descriptive statistics are presented in Table 1.

3.2 Performance evaluation of the machine learning algorithm models

During model development, the class imbalance in the training data (85% survivors vs. 15% non-survivors) was considered. To account for the higher prevalence of survivors, misclassification costs were set to [[0, 7], [1, 0]], **assigning a penalty of 7 for misclassifying non-survivors (i.e., predicting them as survivors) and 1 for misclassifying survivors.** This choice was made manually as a

heuristic starting point, reflecting the imbalance in the dataset due to the high computational cost and time required for systematic optimization. Future studies could incorporate systematic optimization approaches, such as grid search, gradient-based search, simulated annealing, or genetic algorithms, to refine the cost matrix and potentially improve predictive performance. Table 2 presents the performance metrics for the standard, balanced, and MetaCost-adapted models of the selected machine learning algorithms evaluated on the test dataset.

Table 2. Performance Metrics of Machine Learning Models for ICU Mortality Prediction

Models	Accuracy	Precision	Recall	F2 Score	AUC Score	AP	MCC	Confusion Matrix
Gradient Boosting	0.86	1.00	0.14	0.17	0.93	0.79	0.35	[72, 0] [12, 2]
Balanced Gradient Boosting	0.86	0.56	0.71	0.68	0.88	0.57	0.55	[64, 8] [4, 10]
MetaCost Gradient Boosting	**0.94**	**0.80**	**0.86**	**0.85**	**0.91**	**0.75**	**0.79**	**[69, 3] [2, 12]**
MetaCost Gradient Boosting*	0.94	0.94	0.94	0.94	0.91	0.75	0.79	[69, 3] [2, 12]
Logistic Regression	0.90	0.78	0.50	0.54	0.86	0.69	0.57	[70, 2] [7, 7]]
Balanced Logistic Regression	0.81	0.45	0.71	0.64	0.85	0.71	0.46	[60, 12] [4, 10]
MetaCost Logistic Regression	0.83	0.48	0.71	0.65	0.87	0.68	0.48	[61, 11] [4, 10]
Random Forest	0.86	0.67	0.29	0.32	0.88	0.53	0.37	[70, 2] [10, 4]
Balanced Random Forest	0.86	0.75	0.21	0.25	0.83	0.54	0.35	[71, 1] [11, 3]
Balanced Random Forest*	0.86	0.85	0.86	0.84	0.83	0.54	0.35	[71, 1] [11, 3]
MetaCost Random Forest	0.91	0.75	0.64	0.66	0.92	0.7	**0**.64	[69, 3] [5, 9]
Decision Trees	0.80	0.40	0.43	0.42	0.65	0.26	0.30	[63, 9] [8, 6]
Balanced Decision Trees	0.76	0.27	0.29	0.28	0.57	0.19	0.13	[61, 11] [10, 4]
MetaCost Decision Trees	0.83	0.47	0.64	0.60	0.75	0.36	0.45	[62, 10] [5, 9]
XGBoost	0.86	0.67	0.29	0.32	0.87	0.64	0.37	[70, 2] [10, 4]
XGBoost*	0.86	0.84	0.86	0.85	0.87	0.64	0.37	[70, 2] [10, 4]
Balanced XGBoost	0.88	0.67	0.57	0.59	0.90	0.66	0.55	[68, 4] [6, 8]
MetaCost XGBoost	0.88	0.67	0.57	0.59	0.91	0.71	0.55	[68, 4] [6, 8]

*Weighted Precision, Recall, F2 Score performance values

According to the performance metrics presented in Table 2, MetaCost algorithm-based models significantly improved mortality prediction for ICU patients compared with standard machine

learning models, particularly in terms of recall, F2, and MCC scores, which are critical for imbalanced datasets. The MetaCost gradient boosting model demonstrated superior mortality prediction performance. This model correctly classified 69 surviving patients and misclassified 3, while correctly predicting 12 deceased patients and misclassifying 2. Its performance metrics include precision (80%), recall (86%), F2 score (85%), AUC (91%), AP (75%), and MCC (0.79).

3.3. Feature importance analysis

The importance of variables identified by models trained using the MetaCost algorithm was analyzed. The logistic regression model was used to evaluate the variables based on their coefficients, and the results of the other models were analyzed using feature importance scores. The results are presented in Figure 1.

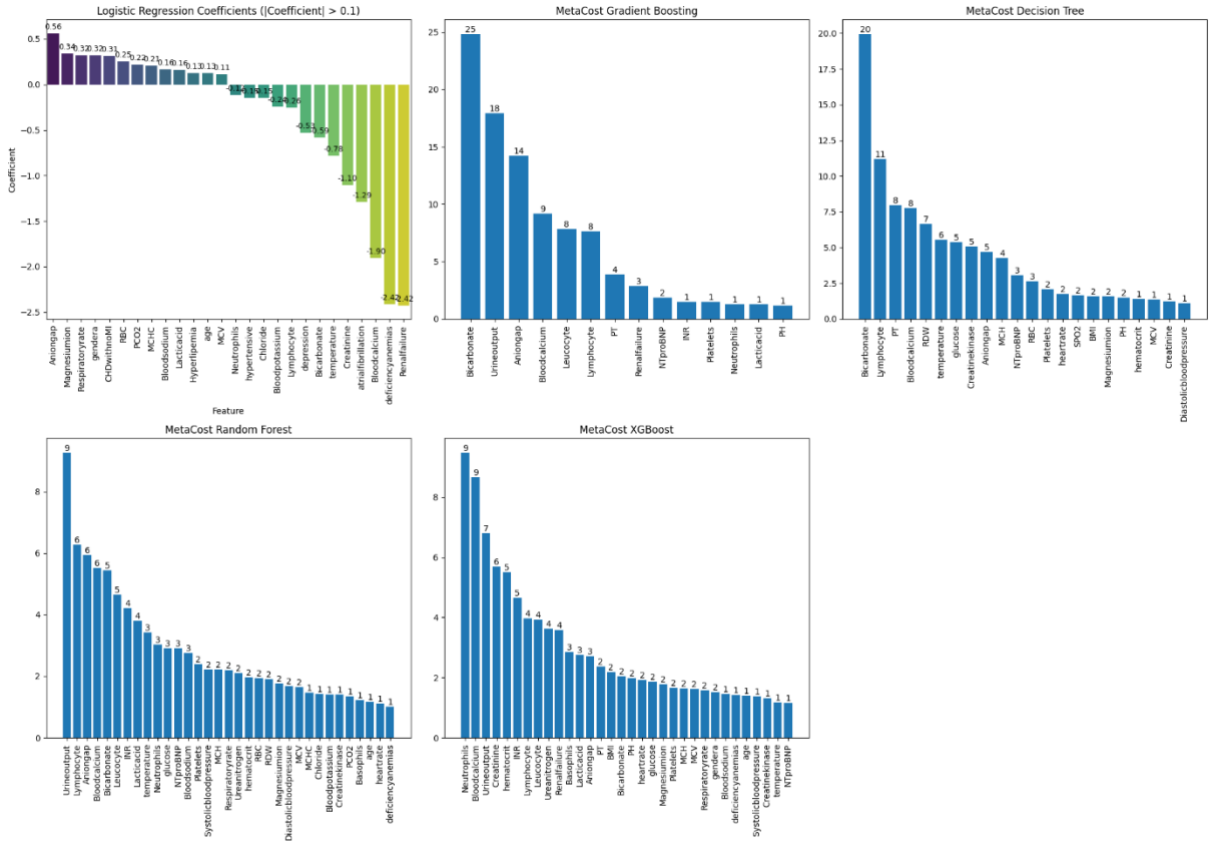

Figure 1. Feature Importance and Coefficients Assigned to Patient Characteristics by Different MetaCost Models

In Figure 1, variables with absolute coefficient values greater than 0.5 in the Logistic Regression model are compared with variables assigned an importance score of 1% or higher by other models.

The MetaCost gradient boosting model was assigned high importance to fewer patient characteristics in classification tasks. For instance, Bicarbonate was assigned an importance of 25%, urine output of 18%, and anion gap of 14%. Similarly, the MetaCost Decision Trees model identified bicarbonate (25%), Leukocyte (11%), and PT (8%) as significant variables.

In contrast, the MetaCost XGBoost and MetaCost Random Forest models assigned scores to a broader range of patient characteristics, and the importance scores of these two models were notably similar. The Logistic Regression model emphasized variables such as Anion Gap, Renal Failure, Age, and respiratory rate and assigned them high coefficients.

The LIME tool was used to enhance the interpretability of the best-performing MetaCost gradient boosting model. The decision mechanisms of this model, which was developed for ICU mortality prediction, were examined in detail using two examples presented in Figure 2.

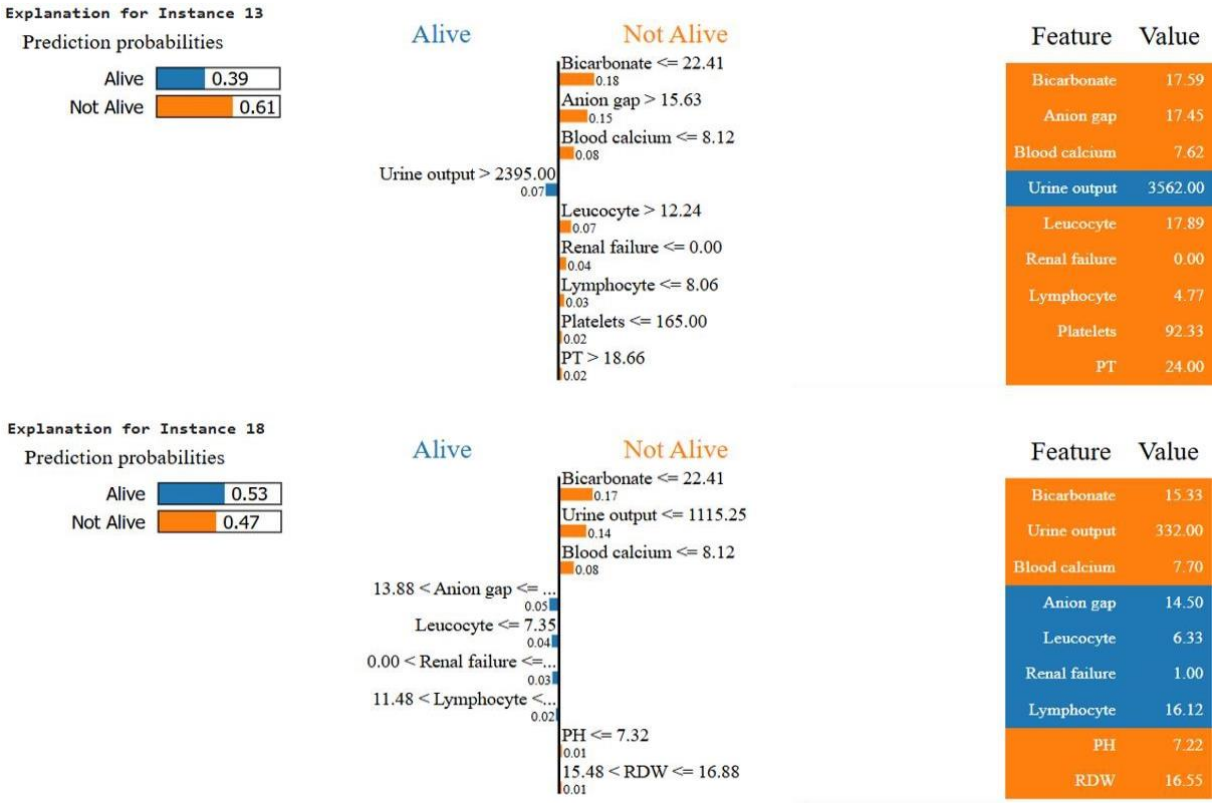

Figure 2. Patient-Level Interpretation of the MetaCost Model Using LIME

Figure 2 illustrates how two patient characteristics in the test data contributed to individual predictions and how these contributions influenced the classification as 'Not Alive' (Negative) or 'Alive' (Positive). The graph visualizes the variables the model focused on and the actual values of the patient characteristics in the context of the prediction probabilities: 61% for “Not Alive” in Instance 13 and 53% for “Alive” in Instance 18. For instance, in Instance 13, Bicarbonate (0.18) and

Anion Gap (0.15) had a positive contribution, whereas urine output (0.07) had a negative impact on mortality prediction.

This study provides a detailed analysis of the advantages of the MetaCost algorithm and LIME for ICU mortality prediction. The results demonstrate that the basic machine learning models adapted with MetaCost improved prediction accuracy on imbalanced datasets.

Evaluating imbalanced datasets using only AUC and AP metrics can lead to misleading results. For instance, in Table 2, where the gradient boosting model achieved the highest AUC (93%) and AP (79%) scores, but identified only two deceased patients, resulting in an MCC value of 0.35. These results indicate that the classification performance of the proposed model on imbalanced data was low. Consequently, the MCC was found to be a more effective metric for evaluating model performance in such cases. In contrast, the MetaCost Gradient Boosting model trained on the same dataset achieved the best performance, with an MCC value of 0.79.

In addition, metrics such as precision, recall, F1, and F2 scores, can produce misleading results when calculated as "weighted" on imbalanced datasets. For example, in Table 2, due to weighting, the precision of the XGBoost* model increased from 67% to 84%, recall from 29% to 86%, and the F2 score from 32% to 85%. This indicates that weighted metrics cannot provide sufficient insight into imbalanced datasets when used alone. Consistent with these findings, Li et al. (2021) and Serim (2023) used similar datasets and obtained comparable results. However, the models adapted using MetaCost demonstrated superior performance.

The findings presented in Tables 1 and 2 revealed that the key patient features influencing mortality prediction were consistent with both statistical tests ($p < 0.001$) and model scores. These observations align with prior evidence. For instance, elevated Anion Gap (AG) levels, which are significantly higher in deceased patients, are consistent with Cheng et al. (2020) , who identified high AG levels as a strong predictor of ICU mortality. Similarly, reduced urine output and low bicarbonate levels were associated with acute kidney injury and prolonged intensive care unit stays, as reported by Mandelbaum et al. (2011) and Libório et al. (2015). Creatinine levels align with those of Bagshaw et al. (2008), emphasizing their role in acute kidney injury and mortality.

The association between lymphocyte percentage and leukocyte counts and mortality is also consistent with the findings of Zhang et al. (2024) , who reported that low lymphocyte counts were correlated with higher mortality rates. Additionally, significant differences in prothrombin time (PT) and NT-proBNP levels among deceased patients support findings from Jorgensen et al. (1992) and Meyer et al. (2007), which linked disseminated intravascular coagulation (DIC) and elevated mortality risk. The relationship between hypocalcemia (low blood calcium) and mortality is consistent with the findings of Melchers and van Zanten (2023).

As shown in Figure 1, the global ranking of patient features in the MetaCost Gradient Boosting model indicates that bicarbonate, Urine Output, and the anion gap are the most influential features. These contributions can be positive or negative. The LIME tool allows the direction of these contributions to be determined on a patient-specific (local) level. For instance, in Figure 2, the features contributing to the prediction for Instance 18 include bicarbonate, urine output, and Blood Bicarbonate Level. Decision-makers can increase the 'Alive' prediction probability (0.53) for critical cases such as Instance 18. The feature boundaries in the graph provide guidance for making adjustments. By reducing 'positive' effects or amplifying controllable 'negative' effects, interventions can be made to improve the patient's survival probability.

Although excluding incomplete records ensured consistency across all study variables and strengthened internal validity, it also reduced the final sample size and may limit the generalizability of our findings. Patients with missing data might have had systematically different characteristics compared with those included in the analysis. Therefore, the results should be interpreted with caution, as they may not fully represent the broader population of ICU-admitted heart failure patients in the original MIMIC-III cohort.

The cost matrix [[0, 7], [1, 0]] was manually specified to mitigate the effects of class imbalance due to the high computational cost and time constraints of systematic optimization methods. While this approach encourages the model to be more cautious in classifying the majority (survivor) class, the chosen values directly influence performance and may not represent the optimal configuration. As highlighted in previous research (Lawson, 2009), cost matrices can be optimized using search-based methods (e.g., grid search, simulated annealing, genetic algorithms), and future work may explore such strategies to further refine model performance. In addition, due to interpretable programming languages like Python and R, the outputs of source codes can be integrated into a web interface (Kaya et al., 2019). For instance, all graphs and explanations generated by LIME for a specific case (e.g., instance 18) can be transferred to a web page, allowing clinicians or healthcare professionals to easily view and interpret them. Such integration can enhance the usability and accessibility of clinical decision support systems.

4. Conclusions and Recommendations

In conclusion, the MetaCost algorithm offers an effective approach for enhancing classification performance in healthcare datasets characterized by imbalanced distributions involving mortality by adapting fundamental machine learning algorithms to be cost-sensitive. The integration of the LIME method, which provides explanations at the individual patient level, enables clinical decision support systems to become more reliable and interpretable. These methods not only improve predictive

performance but also provide a methodological basis for integrating interpretability into healthcare analytics. However, further validation studies on diverse patient groups and larger datasets are necessary to strengthen the generalizability and reliability of this approach.

Acknowledgements

We thank the PhysioNet team for providing access to the MIMIC-III database. We also acknowledge Zhou et al. (2023) for sharing their pre-processed data and feature selection, which supported this study. We also sincerely thank Zhou, J., Li, F., Song, Y., et al. (2023) for sharing their pre-processed data and feature selections, which significantly facilitated the implementation of this study.

Authors' Contributions

This manuscript was solely authored by FK, who was responsible for the conception, design, analysis, interpretation of the data, and writing of the manuscript.

Statement of Research and Publication Ethics

This study was conducted using the MIMIC-III database, which consists of publicly available and de-identified health records. Access to the dataset was obtained after completion of the required training on human subjects research, provided by the National Institutes of Health (Certificate Record ID: 61633342). In accordance with institutional and international ethical guidelines, no additional ethical approval was required for this analysis.

References

Bagshaw, S. M., George, C., Bellomo, R., and ANZICS Database Management Committee. (2008). A comparison of the RIFLE and AKIN criteria for acute kidney injury in critically ill patients. *Nephrology Dialysis Transplantation, 23*(5), 1569–1574. https://doi.org/10.1093/ndt/gfn009

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Çanga, D., and Boğa, M. (2020). Determination of the effect of some properties on egg yield with regression analysis method bagging Mars and R application. *Turkish Journal of Agriculture - Food Science and Technology, 8*(8), 1705–1712. https://doi.org/10.24925/turjaf.v8i8.1705-1712.3468

Çanga, D., and Boğa, M. (2022). Detection of correct pregnancy status in lactating dairy cattle using MARS data mining algorithm. *Turkish Journal of Veterinary & Animal Sciences, 46*(6), 809–819. https://doi.org/10.55730/1300-0128.4257

Çanga Boğa, D., Boğa, M., and Tırınk, C. (2024). Comparison of nonlinear functions to define the growth in intensive feedlot system with XGBoost algorithm. *Turkish Journal of Agriculture - Food Science and Technology, 12*(8), 1408–1416. https://doi.org/10.24925/turjaf.v12i8.1408-1416.6562

Çelik, S., and Yilmaz, O. (2021). The relationship between the coat colors of Kars shepherd dog and its morphological characteristics using some data mining methods. *International Journal of Livestock Research, 11*(1), 53–61. https://doi.org/10.5455/ijlr.20200604

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cheng, B., Li, D., Gong, Y., Ying, B., and Wang, B. (2020). Serum anion gap predicts all-cause mortality in critically ill patients with acute kidney injury: Analysis of the MIMIC-III database. *Disease Markers, 2020*, Article 6501272. https://doi.org/10.1155/2020/6501272

Chicco, D., Tötsch, N., and Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining, 14*, 13. https://doi.org/10.1186/s13040-021-00244-z

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining* (pp. 155–164). Association for Computing Machinery.

Eratlı, Ş. Y., and Şahin, M. (2020). Investigation of factors affecting the achievement of university students with logistic regression analysis: School of Physical Education and Sport example. *SAGE Open, 10*(1). https://doi.org/10.1177/2158244020902082

Friedman, J. (2000). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035. https://doi.org/10.1038/sdata.2016.35

Jorgensen, M., Gustafsen, K., Ernst, S., and Thostrup, J. S. (1992). Disseminated intravascular coagulation in critically ill patients: Laboratory diagnosis. *Intensive Care World, 9*(3), 108–113.

Kaya, F., Korkmaz, F., and Efe, E. (2019). Advanced machine learning techniques for predictive modeling. In *International Symposium on Advanced Engineering Technologies (ISADET)*, Kahramanmaraş, Turkey.

Lawson, M. J. (2009). *The impact of cost matrix selection on cost-sensitive learning: An empirical study* (Doctoral dissertation, Virginia Polytechnic Institute and State University). VTechWorks. https://vtechworks.lib.vt.edu/handle/10919/29623

Li, F., Xin, H., Zhang, J., Fu, M., Zhou, J., and Lian, Z. (2021). Prediction model of in-hospital mortality in intensive care unit patients with heart failure: Machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open, 11*(7), e044779. https://doi.org/10.1136/bmjopen-2020-044779

Libório, A. B., Noritomi, D. T., Leite, T. T., de Melo Bezerra, C. T., de Faria, E. R., and Kellum, J. A. (2015). Increased serum bicarbonate in critically ill patients: A retrospective analysis. *Intensive Care Medicine, 41*(3), 479–486. https://doi.org/10.1007/s00134-015-3649-9

Ling, C. X., and Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. In C. Sammut (Ed.), *Encyclopedia of Machine Learning*. Springer. https://www.csd.uwo.ca/~xling/papers/cost_sensitive.pdf

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mandelbaum, T., Scott, D. J., Lee, J., Mark, R. G., Malhotra, A., Waikar, S. S., Howell, M. D., and Talmor, D. (2011). Outcome of critically ill patients with acute kidney injury using the Acute Kidney Injury Network criteria. *Critical Care Medicine, 39*(12), 2659–2664. https://doi.org/10.1097/CCM.0b013e31822823e7

Melchers, M., and van Zanten, A. R. H. (2023). Management of hypocalcaemia in the critically ill. *Current Opinion in Critical Care, 29*(4), 330–338. https://doi.org/10.1097/MCC.0000000000001059

Meyer, B., Huelsmann, M., Wexberg, P., Delle Karth, G., Berger, R., Moertl, D., and Pacher, R. (2007). N-terminal pro-B-type natriuretic peptide is an independent predictor of outcome in an unselected cohort of critically ill patients. *Critical Care Medicine, 35*(10), 2268–2273. https://doi.org/10.1097/01.CCM.0000284507.59405.53

Núñez Reiz, A., Armengol de la Hoz, M. A., and Sánchez García, M. (2018). Big data analysis and machine learning in intensive care units. *Medicina Intensiva (English Edition), 43*(7), 416–426. https://doi.org/10.1016/j.medine.2019.06.012

Önder, H., Tirink, C., Yakubets, T., Getya, A., Matvieiev, M., Kononenko, R., ... and Kaya, F. (2025). Predicting live weight for female rabbits of meat crosses from body measurements using LightGBM, XGBoost and support vector machine algorithms. *Veterinary Medicine and Science, 11*(1), e70149. https://doi.org/10.1002/vms3.70149

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*. https://arxiv.org/abs/2010.16061

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Serim, A. B. Ö. (2023). Characterization of mortality prediction: An ensemble learning analysis using the MIMIC-III dataset. *Journal of Scientific Reports-A, (054)*, 364–384.

Tonekaboni, S., Joshi, S., Goldenberg, A., and Duvenaud, D. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of Machine Learning for Healthcare* (Vol. 106, pp. 359–374). http://proceedings.mlr.press/v106/tonekaboni19a.html

Yavuz, E. (2023). Determining the factors affecting the achievement status of high school students by logistic regression. *ResearchGate, 400*, 155–171.

Zhang, J., Zhao, Q., Liu, S., Yuan, N., and Hu, Z. (2024). Clinical predictive value of the CRP-albumin-lymphocyte index for prognosis of critically ill patients with sepsis in intensive care unit: A retrospective single-center observational study. *Frontiers in Public Health, 12*, 1395134. https://doi.org/10.3389/fpubh.2024.1395134

Zhou, J., Li, F., Song, Y., et al. (2023). *Prediction model of in-hospital mortality in intensive care unit patients with heart failure: Machine learning-based, retrospective analysis of the MIMIC-III database* [Dataset]. Dryad. https://doi.org/10.5061/dryad.0p2ngf1zd