TURKISH JOURNAL OF EMERGENCY MEDICINE

Original Article

Access this article online

Quick Response Code:

Website: https://turkjemergmed.com/

DOI: 10.4103/tjem.tjem_161_24

Improved outcome prediction in acute pancreatitis with generated data and advanced machine learning algorithms

Murat Özdede[1]*, Ali Batur[2], Alp Eren Aksoy[2]

Departments of [1]Internal Medicine and [2]Emergency Medicine, Faculty of Medicine, Hacettepe University, Ankara, Türkiye
**Corresponding author*

Abstract:

OBJECTIVES: Traditional scoring systems have been widely used to predict acute pancreatitis (AP) severity but have limitations in predictive accuracy. This study investigates the use of machine learning (ML) algorithms to improve predictive accuracy in AP.

METHODS: A retrospective study was conducted using data from 101 AP patients in a tertiary hospital in Türkiye. Data were preprocessed, and synthetic data were generated with Gaussian noise addition and balanced with the ADASYN algorithm, resulting in 250 cases. Supervised ML models, including random forest (RF) and XGBoost (XGB), were trained, tested, and validated against traditional clinical scores (Ranson's, modified Glasgow, and BISAP) using area under the curve (AUC), F1 score, and recall.

RESULTS: RF outperformed XGB with an AUC of 0.89, F1 score of 0.82, and recall of 0.82. BISAP showed balanced performance (AUC = 0.70, F1 = 0.44, and recall = 0.85), whereas the Glasgow criteria had the highest recall but lower precision (AUC = 0.70, F1 = 0.38, and recall = 0.95). Ranson's admission criteria were the least effective (AUC = 0.53, F1 = 0.42, and recall = 0.39), probable because it lacked the 48th h features.

CONCLUSION: ML models, especially RF, significantly outperform traditional clinical scores in predicting adverse outcomes in AP, suggesting that integrating ML into clinical practice could improve prognostic assessments.

Keywords:

Acute pancreatitis, machine learning, mortality, outcomes, prognosis, scoring methods

Introduction

Acute pancreatitis (AP) is a serious and sudden inflammatory condition of the pancreas, often leading to significant abdominal emergencies. Although most cases of AP are self-limiting, approximately one-fifth develop severe AP, which can lead to a mortality rate of up to 30%.[1] Therefore, accurate severity prediction is essential for anticipating outcomes.

Over the years, various scoring systems have been developed to predict the severity of AP.[1,2] Ranson's criteria, one of the oldest and most widely used systems, has been extensively validated, followed by the Glasgow-Imrie criteria and its modified versions.[3-5] The Acute Physiology and Chronic Health Examination (APACHE) II, a complex score which was originally developed to estimate intensive care unit (ICU) mortality, has been extensively used for AP severity prediction due to its ability to be calculated at any time during a

Submitted: 12-08-2024
Revised: 30-10-2024
Accepted: 31-10-2024
Published: 02-01-2025

ORCID:
MÖ: 0000-0002-6981-1210
AB: 0000-0002-2057-3215
AEA: 0009-0001-9808-5326

Address for correspondence:
Dr. Murat Özdede, Hacettepe Mh. 06230 Ankara, Türkiye.
E-mail: mozdede@gmail.com

How to cite this article: Özdede M, Batur A, Aksoy AE. Improved outcome prediction in acute pancreatitis with generated data and advanced machine learning algorithms. Turk J Emerg Med 2025;25:32-40.

Box-ED section

What is already known about the study topic?

- Traditional scoring systems are commonly used to predict severity in acute pancreatitis but have limitations in predictive accuracy.

How is this study structured?

- This retrospective study analyzed data from 101 acute pancreatitis patients. Tree-based ensemble methods were trained and tested on generated data and validated on the original dataset to assess their predictive accuracy against traditional scoring systems.

What does this study tell us?

- The study found that the random forest and XGBoost classifier models significantly outperformed traditional clinical scores in predicting adverse outcomes in acute pancreatitis. Key predictive features identified included serum glucose, lactate, albumin, blood urea nitrogen (BUN), and age.

What is the conflict on the issue? Is it important for readers?

- While traditional scoring systems are widely used due to their simplicity, they may oversimplify risk assessments, potentially leading to less accurate predictions. This study highlights the need for incorporating advanced ML models into clinical practice to improve the accuracy of severity predictions, ultimately leading to better patient outcomes.

patient's hospital stay, often outperforming older ones.[6-8] The Bedside Index of Severity in AP (BISAP), developed more recently, offers similar advantages to APACHE II by eliminating the need for a 48-h interval. The BISAP score has less variables which are cost-effective and can be done in emergency setting.

Recently, there has been increasing interest in using machine learning (ML) models to improve predictive accuracy in emergency medicine, including for the prognosis of AP, where early studies have shown superior performance compared to traditional scoring systems.[9,10] This study aims to demonstrate how ML, combined with innovative data augmentation techniques, can significantly improve prognostic models in AP, comparing traditional scoring systems.

The study aims to identify more effective prognostic markers in AP and prove their efficacy by comparing them with existing classical risk scoring systems. Thus, it is aimed to be able to determine outcomes such as intensive care admission or death at an early stage.

Methods

Study design and definitions

This study is a single-center, retrospective cohort study. Data from patients diagnosed with AP in a tertiary emergency department in Turkey between January 2018 and September 2022 were used. Inclusion criteria required data matching at least two of the following: (1) characteristic abdominal pain, (2) serum pancreatic enzyme levels at least three times the normal upper limit (pancreatic amylase >159 U/L and lipase >201 U/L), and (3) characteristic findings of AP on contrast-enhanced computed tomography (CT) scans. Patients with missing data required for traditional scoring systems and patients with unknown outcomes who could not be followed up in the hospital were excluded from the study. A sample size was not calculated for the study, and all patients who met the criteria during the specified period were included in the study.

Detailed medical histories were collected, including comorbidities, previous episodes of pancreatitis, symptomatology, admission vitals, arterial blood gas analysis, hematocrit, biochemistry, complete blood count studies, baseline and longitudinal measurements of amylase, lipase, and pancreatic amylase levels as well as coagulation parameters and cardiac enzymes. The trajectories of the subsequent measurements of pancreatic enzymes were also analyzed through percent change and average real variability calculations as formulated below:

$$ARV = \frac{1}{n-1}\sum_{i=2}^{n} | X_i - X_{i-1} |$$

The BISAP score, modified Glasgow criteria, and Ranson's admission criterion were calculated from the data. AP severity was graded according to the revised Atlanta classification. Local complications were identified from radiology reports.

Dates of discharge, ward and ICU hospitalizations, or death were recorded. The composite outcome of ICU admission and/or death was chosen based on its practical relevance in emergency medicine, where physicians must rapidly identify high-risk patients and make timely decisions regarding their care. This binary classification of adverse outcomes reflects the critical needs of real-world clinical settings, where time-sensitive decisions regarding patient disposition are essential.

Preprocessing data

The preprocessing involved removing noisy, duplicate, or incomplete records, resulting in 101 cases. Missing data were imputed using random

forest (RF) techniques, and skewed data were normalized to a range of −1 to 1. To address the dataset's imbalance and limited case numbers, data generation was employed, resulting in a synthetic dataset of 250 cases. New data points were created by selectively adding Gaussian Noise to continuous variables, carefully calibrated to resemble the original data while introducing slight variations for diversity. To further balance the dataset, the ADASYN (Adaptive Synthetic Sampling) algorithm was applied, focusing on difficult-to-learn examples to reduce bias. The final dataset achieved a 65:35 balance between majority and minority classes, chosen to preserve data authenticity while avoiding excessive artificiality. Each step of synthesis and balancing was carefully monitored through heat maps, principal component analysis (PCA), summary statistics, and manual adjustments to ensure the generated data were realistic and representative [Figure 1].

Model development and validation

Supervised ML models were developed using tree-based ensemble methods, namely random forest (RF) and XGBoost (XGB), known for their robustness in classification tasks. RF was chosen for its ability to handle high-dimensional data, while XGB was selected for its gradient boosting feature, which sequentially builds models to correct errors, leading to high accuracy. XGB also includes regularization parameters to prevent overfitting.

Feature selection was performed before training to avoid computational exhaustion and identify the most informative features. Information gain, RF, and logistic regression analyses were used for feature selection.

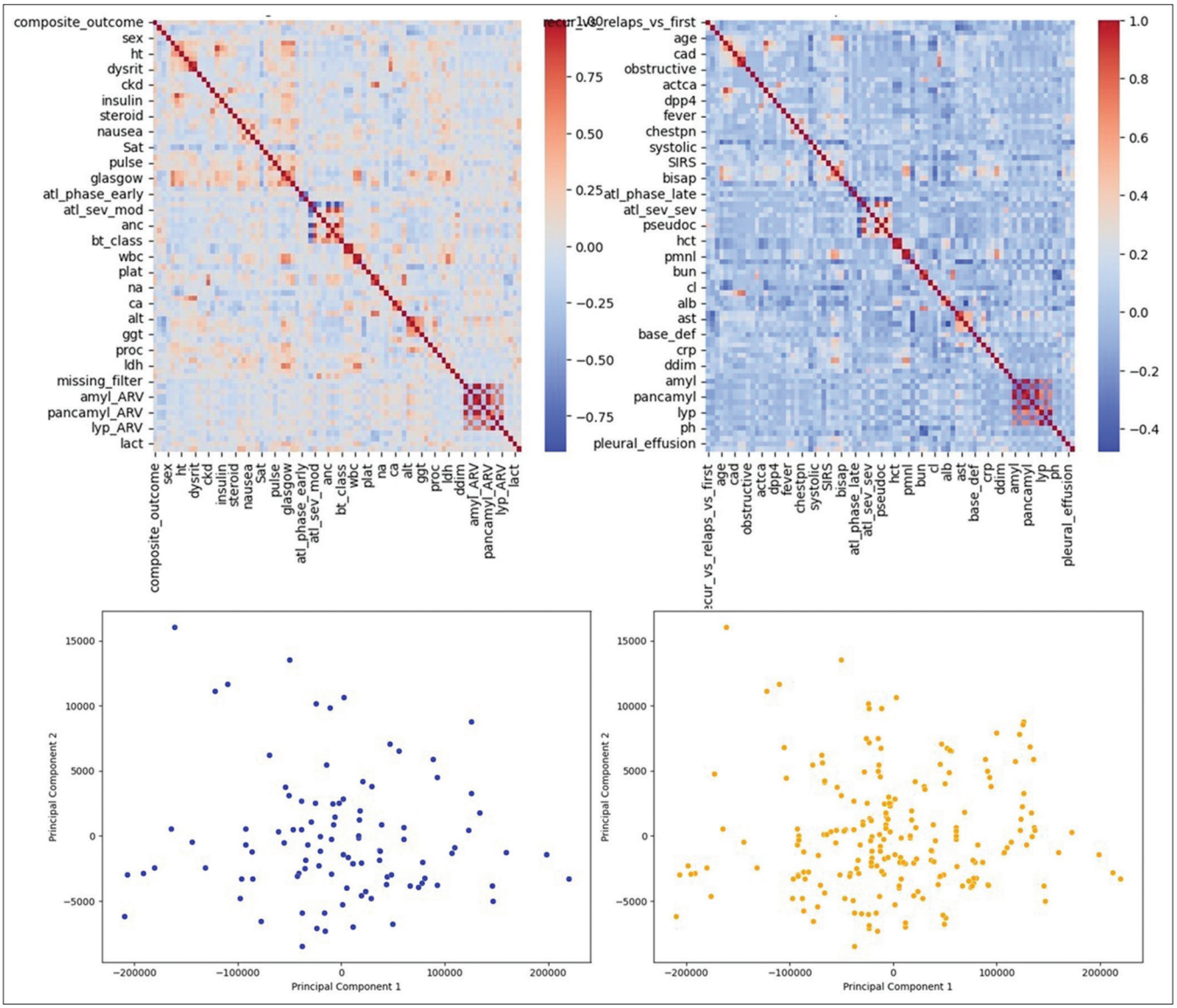

Figure 1: Heat maps and principal component analysis plots. The left side represents the original data set and the right side represents the generated data set

After feature selection, hyperparameter tuning was conducted using RandomizedSearchCV, identifying optimal parameters for RF (n_estimators = 100, min_samples_split = 5, min_samples_leaf = 1, max_depth = 10) and XGB (subsample = 0.8, n_estimators = 200, max_depth = 3, learning_rate = 0.1), which were trained on 70% of the generated data, with 30% reserved for testing. Validation on the original dataset compared model performance metrics (area under the curve [AUC], recall, F1) against traditional clinical scores (Ranson's admission, modified Glasgow, and BISAP). Metrics such as positive predictive value, negative predictive value, and accuracy were not used, as they may provide a skewed picture in imbalanced datasets, where a high accuracy can still result in poor performance in identifying minority class events. "SHAP (SHapley Additive exPlanations) values" were plotted to highlight feature importance, enhancing the transparency of model decisions.

Descriptive analyses were performed before the normalization and generation of the data, and due to the relatively small sample size and skewed data, non-parametric tests (Mann-Whitney U for continuous variables and Chi-square or Pearson's test for categorical variables) were employed for statistical comparisons. All statistical analyses, data-mining, and ML coding were performed using Python (v3.12.5), utilizing packages such as pandas, scikit-learn, XGB, and SHAP.

Ethical approval for this study was obtained from Hacettepe University Ethics Committee in Turkey on the date of December 27, 2022, with approval number of GO 22/1317.

Results

Descriptive statistics

This cohort primarily consisted of middle-aged adults, with a slightly higher representation of male patients [Table 1]. A significant sex disparity was noted, with a higher proportion of females in the adverse outcome group (29.3%) compared to the total cohort (19.8%) ($P = 0.048$), while age differences were not statistically significant. The recurrence of pancreatitis and biliary origin showed no significant variation between groups. Most patients presented with acute edematous pancreatitis, and CT findings were comparable across groups.

Notably, certain comorbidities, including congestive heart failure, dysrhythmia, chronic kidney disease, and active cancer, were more prevalent in the adverse outcome group. These patients also exhibited more severe clinical presentations, as reflected by higher Glasgow and BISAP scores. In addition, this group had significantly lower oxygen saturation (SpO2, $P = 0.015$), higher respiratory rates ($P = 0.001$), and a greater incidence of fever ($P = 0.005$) and dyspnea ($P = 0.005$), indicating a more critical initial state. Although the sample size is limited, there was a trend toward more severe pancreatitis and local complications, such as acute necrotic collections and walled-off necrosis, in the adverse outcome group.

Significant laboratory differences included higher blood urea nitrogen (BUN) levels ($P = 0.06$), lower albumin ($P = 0.08$), and higher initial glucose levels ($P = 0.01$) in the adverse outcome group [Table 2]. Elevated troponin ($P = 0.03$) and D-dimer ($P = 0.046$) levels suggested myocardial stress and a hypercoagulable state, respectively. In addition, this group had lower pO2 ($P = 0.04$), pCO2 ($P = 0.04$), and bicarbonate levels ($P = 0.038$), indicating respiratory compromise and metabolic acidosis, while elevated lactate ($P = 0.005$) pointed to tissue hypoxia. Interestingly, neither pancreatic enzyme levels nor their trajectories showed significant differences across outcome groups.

Supervised machine learning models

Initial feature selection included "fever," "troponin," "lactate," "glucose," "albumin," "BUN," "age," and "CKD." However, "fever," "troponin," and "CKD" were later excluded due to minimal predictive contribution. As a result of ML, the top ranking features of prognostic were "lactate," "glucose," followed by "BUN," and "age."

The RF model demonstrated superior performance across key metrics, with an AUC of 0.89, F1 score of 0.82, and recall of 0.82, outperforming the XGB model, which still exhibited competitive performance (AUC = 0.85, F1 = 0.77, and recall = 0.77). To benchmark these models, traditional clinical scoring systems – BISAP, modified Glasgow, and Ranson's admission criteria – were evaluated on the original dataset, with thresholds adjusted to optimize recall and specificity.

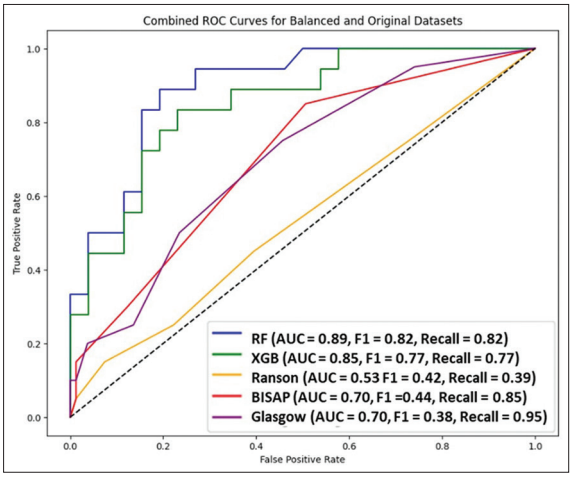

Figure 2: Receiver operating characteristic curves for machine learning models and clinical scores

Table 1: Baseline clinical characteristics of the cohort stratified by outcome group

	Total, *n* (%)	No event, *n* (%)	Adverse event, *n* (%)	*P*
Total cohort	101 (100)	81 (80.2)	20 (19.8)	
Demographics				
Female sex	41 (40.6)	29 (70.7)	12 (29.3)	0.048
Age (median/IQR)	53 (27)	52 (27)	58.5 (28)	0.13
Features and classification				
Recurring pancreatitis	23 (22.8)	21 (91.3)	2 (8.7)	0.13
Origin				
Biliary pancreatitis	48 (47.5)	39 (81.3)	9 (18.8)	0.8
Nonbiliary pancreatitis	53 (52.5)	42 (79.2)	11 (20.8)	
Pleural effusion	3 (100)	1 (33.3)	2 (66.7)	0.1
CT classification				
Acute edematous pancreatitis	69 (68.3)	54 (78.3)	15 (21.7)	0.42
Necrotizing pancreatitis	5 (5)	3 (60)	2 (40)	
Infected pancreatic necrosis	1 (1)	1 (100)	0	
Nonsignificant findings	26 (25.7)	23 (88.5)	3 (11.5)	
Comorbidities and medication history				
Diabetes mellitus	23 (22.8)	17 (73.9)	6 (26.1)	0.4
Arterial hypertension	34 (33.7)	24 (70.6)	10 (29.4)	0.08
Coronary artery disease	11 (10.9)	9 (81.8)	2 (18.2)	0.8
Congestive heart failure	5 (5)	2 (40)	3 (60)	0.05
Dysrhythmia	4 (4)	1 (25)	3 (75)	0.024
Chronic kidney disease	5 (5)	2 (40)	3 (60)	0.05
History of malignancy	9 (8.9)	5 (55.6)	4 (44.4)	0.07
Active cancer	6 (5.9)	2 (33.3)	4 (66.7)	0.003
Metformin	20 (19.8)	14 (70)	6 (30)	0.2
Insulin	7 (6.9)	4 (57.1)	3 (42.9)	0.113
Incretin-based antidiabetics	4 (4)	2 (50)	2 (50)	0.12
Chemotherapeutics	6 (6)	5 (66.7)	2 (33.3)	0.4
Steroid therapy	5 (5)	3 (60)	2 (40)	0.25
Vital findings (median/IQR)				
Body temperature (celsius)	36.6 (1.13)	36.6 (1.2)	36.5 (0.92)	0.33
Pulse oximetry (%)	97 (3)	97 (2)	95 (4)	0.015
Pulse rate (/min)	83.5 (25)	83 (25)	87 (44)	0.63
Respiratory rate (/min)	20 (4)	18 (2)	20.5 (4)	0.001
Systolic blood pressure (mmHg)	130 (24)	130 (24)	131.5 (43)	0.75
Symptoms on admission				
Fever	12 (11.9)	6 (50)	6 (50)	0.005
Dyspnea	4 (4)	1 (25)	3 (75)	0.005
Fatigue	15 (14.9)	10 (66.7)	5 (33.3)	0.15
Abdominal pain	100 (99)	80 (80)	20 (20)	0.62
Back pain	43 (42.6)	33 (76.7)	10 (23.3)	0.45
Nausea	60 (59.4)	47 (46.5)	13 (21.7)	0.57
Anorexia	24 (23.8)	18 (75)	6 (25)	0.47
Severity scores (median/IQR)				
Ranson's criteria (on admission)	1 (2)	1 (2)	1 (2.5)	0.625
Modified Glasgow criteria	2 (2)	1 (2)	2.5 (2.5)	0.005
BISAP score	1 (1)	1 (1)	2 (2)	0.002
Atlanta severity				
Mild	95 (94.1)	77 (81.1)	18 (18.9)	0.53
Moderate	4 (4)	3 (75)	1 (25)	
Severe	2 (2)	1 (50)	1 (50)	
Local complications				
Acute pancreatic fluid collection	61 (60.4)	50 (82)	11 (18)	0.6
Acute necrotic collection	5 (5)	3 (60)	2 (40)	0.26

Contd...

Table 1: Contd...

	Total, *n* (%)	No event, *n* (%)	Adverse event, *n* (%)	*P*
Pseudocysts	9 (8.9)	7 (77.8)	2 (22.2)	0.85
Walled of necrosis	5 (5)	3 (60)	2 (40)	0.26
Prognosis				
Exitus	5 (5)	NA	NA	NA
Need for intensive care unite	19 (18.9)	NA	NA	NA

IQR: Interquartile range, CT: Computed tomography, BISAP: Bedside Index of Severity in Acute Pancreatitis, NA: Not available

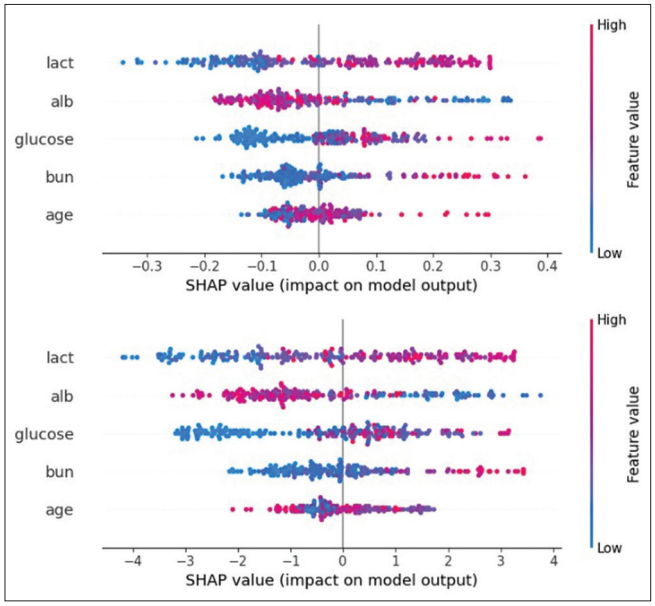

Figure 3: SHapley Additive exPlanations plots of the top five ranking features in random forest and XGBoost models. SHAP: SHapley Additive exPlanations, lac: Lactate, alb: Albumin, bun: Blood urea nitrogen

Among these, BISAP provided balanced performance, with higher precision and an F1 score indicating reliable prediction (AUC = 0.70, F1 = 0.44, and recall = 0.85). The modified Glasgow criteria had the highest recall (AUC = 0.70, F1 = 0.38, and recall = 0.95) but reduced precision. Ranson's criteria were the least effective predictor (AUC = 0.53, F1 = 0.42, and recall = 0.39). Receiver operating characteristic (ROC) curve analysis showed that both ML models significantly outperformed traditional scoring systems [Figure 2]. SHAP plots highlighted lactate, albumin, and glucose levels as the top contributors to the models' predictive power [Figure 3].

Discussion

This study demonstrates that ML algorithms significantly outperform traditional clinical scoring systems in predicting outcomes in AP, particularly in the fast-paced environment of emergency medicine, where rapid decision-making is critical. As a result of the current study, predictors that provide more successful results than traditional scoring systems in predicting outcomes such as intensive care hospitalization and death associated with AP were identified. These predictors can easily determine the intensive care requirements and mortality risks of AP in the emergency department in the early period.

Traditional clinical scores, inherently, exhibit more rigid, angular ROC curves, reflecting their limited, less-nuanced nature. This characteristic indicates the granularity and lower resolution inherent in traditional scores, which rely on a small number of ordinal points, thus relying on those may lead to oversimplified judgments.

While traditional scoring systems like Ranson's criteria are still valued for their simplicity, they are less practical in emergency settings due to requirements such as the 48-h data collection period. This delay hinders their predictive accuracy when timely decisions are necessary. With the reliance on data that may not be immediately available, such as fluid sequestration over 48 h, their predictive performance was still shown to be inferior to APACHE-II and BISAP.[11] In our study, we were unable to grasp its full performance as we were limited by the lack of certain 48-h data points, such as fluid sequestration and base deficit. Similarly, the modified Glasgow criteria, which rely on capturing the worst data points within the first 48 h, present challenges in emergency settings where comprehensive data flow may not be available.[1,12]

BISAP, developed and validated in large cohorts, demonstrated a more balanced performance in our study, and it effectively predicts severity, particularly in early-stage predictions of organ failure and in-hospital mortality, despite its simplicity.[7,13] However, recent studies suggest that BISAP may underperform compared to newer scoring systems like the WL and the Chinese Simple Scoring System score.[14]

Our analysis highlights the importance of BUN as a predictor of AP severity. BUN, a key component of both BISAP and Glasgow scores, consistently correlates with poor outcomes, reflecting critical factors such as kidney perfusion, plasma volume, and catabolic processes.[15] Serum albumin levels, another significant feature identified by our ML models, are associated with nutritional status, inflammation, and capillary permeability, all influencing

Table 2: Laboratory values of the cohort stratified by outcome group

	Total	No event	Event	*P*
Complete blood count (median/IQR)				
Hemoglobin (g/dL)	13.8 (2.9)	13.9 (2.85)	13.5 (3.8)	0.67
Hematocrit (%)	40.9 (7.95)	41 (7.25)	40.55 (13)	0.56
Hematocrit change over 48 h (%)	−6.25 (9.74)	−6.24 (8.84)	−6.64 (15.84)	0.74
Leukocyte (/mm^3)	10,100 (4250)	10,100 (4100)	10,100 (5100)	0.7
Neutrophil (/mm^3)	7800 (4500)	7790 (4555)	7825 (5090)	0.68
Lymphocyte (/mm^3)	1300 (1005)	1400 (1065)	1200 (975)	0.14
Thrombocytes (×10^9/L)	235,782 (90,000)	240,000 (8600)	223,500 (120,500)	0.8
Biochemistry (median/IQR)				
BUN (mg/dL)	14 (6.77)	13.51 (6.43)	15.61 (13.8)	0.06
Creatinine (mg/dL)	0.72 (0.3)	0.7 (0.28)	0.76 (0.93)	0.57
Sodium (mEq/L)	136 (5)	136 (5)	134.5 (5)	0.18
Potassium (mEq/L)	3.92 (0.48)	3.92 (0.46)	4.04 (1.56)	0.27
Chloride (mEq/L)	101 (6)	102 (7)	100 (6.75)	0.11
Calcium (mEq/L)	9.38 (0.7)	9.37 (0.64)	9.38 (0.79)	0.34
Calcium change over 48 h (%)	−6.8 (6.81)	−6.8 (6.66)	−7.4 (9.92)	0.36
Phosphate	3.24 (0.86)	3.24 (0.88)	3.23 (1.2)	0.8
Albumin (mg/dL)	4.04 (0.59)	4.06 (0.51)	3.72 (0.99)	0.08
Glucose (mg/dL)	126 (78)	119 (68)	156 (98.75)	0.01
Uric acid (mg/dL)	5.2 (2.6)	5.19 (2.23)	5.22 (3.28)	0.9
ALT (U/L)	77 (248)	77 (253.5)	62 (247.25)	0.84
AST (U/L)	127 (295)	130 (290)	95 (484.25)	0.94
AST change over 48 h (%)	−26.19 (44.16)	−27.58 (41.71)	−22.7 (60.4)	0.8
ALP (U/L)	126 (141)	126 (143)	127.5 (173)	0.76
ALP change over 48 h (%)	−11.42 (17.65)	−11.68 (17.04)	−9.45 (20.45)	0.73
GGT (U/L)	181.5 (475)	212.5 (489.5)	108.5 (417.25)	0.65
GGT change over 48 h (%)	−13.15 (24.26)	−11.52 (23.21)	−16.48 (28.99)	0.65
LDH (U/L)	259 (287)	259 (287)	268.75 (788)	0.53
Troponin (ng/mL)	3.4 (3.9)	4.2 (7.9)	7.3 (21.2)	0.03
C-reactive protein (mg/dL)	8.16 (9.95)	8.38 (10.3)	7.27 (12)	0.54
Procalcitonin (µ/L)	0.22 (4.1)	0.19 (1.6)	0.29 (6.62)	0.33
D-dimer (ng/mL)	1.98 (3.32)	1.55 (3.36)	3.94 (4.36)	0.046
Blood gases (median/IQR)				
pH	7.4 (0.04)	7.4 (0.04)	7.41 (0.1)	0.96
pO_2	65 (16.6)	65 (16.45)	57.9 (20.35)	0.04
pCO_2	39.6 (6.9)	40.52 (5.8)	37.3 (9.5)	0.04
$cHCO_3$	23.5 (2.95)	23.65 (2.75)	22.3 (3.5)	0.038
Lactate	1.72 (1.16)	1.67 (0.94)	2.15 (1.07)	0.005
Base deficit	−0.67 (6.04)	−0.1 (5.61)	−1.4 (7.76)	0.18
Pancreatic enzymes and their trajectory indicators (median/IQR)				
Amylase	699 (1368)	611 (1221.5)	892.5 (1273.5)	0.2
Amylase change (%)	−68.96 (46)	−68.96 (51)	−69.64 (34)	0.98
Amylase ARV	468 (1224.5)	399 (971.5)	609 (1247)	0.27
Pancreatic amylase	506 (962)	474 (940)	643 (847)	0.36
Pancreatic amylase change (%)	−74.32 (49)	−75.87 (50)	−71.13 (48)	0.74
Pancreatic amylase ARV	341 (843.5)	335 (835)	487 (845)	0.77
Lipase	1526 (2577.5)	1404 (3038.5)	2060.5 (1682)	0.66
Lipase change (%)	−84.26 (45)	−83.22 (59)	−86.43 (28)	0.78
Lipase ARV	1021 (2584)	789 (2762.5)	1629.5 (1831.5)	0.76

ALT: Alanine transaminase, AST: Aspartate transaminase, ALP: Alkaline phosphatase, ARV: Average real variability, GGT: Gamma-glutamyl transferase, LDH: Lactate dehydrogenase, IQR: Interquartile range, BUN: Blood urea nitrogen

AP outcomes. It was demonstrated that albumin with BUN correlates best with oxidative stress significantly in AP patients.[16] Low albumin levels can also indicate poor nutritional status, increased catabolism, heightened inflammation, and reduced hepatic albumin synthesis.[17] Serum lactate, a marker of tissue hypoxia, also emerged as a top predictor in our study. Elevated lactate levels often signal severe disease or complications like pancreatic necrosis.[18] However, the most striking feature might be serum glucose levels, a component of Ranson's

criteria which illustrates a strong linear relationship with adverse events. Elevated glucose levels may not only reflect poor metabolic control, often due to pre-existing diabetes, and higher inflammatory stress but also contribute to the severity by promoting oxidative stress and tissue damage, creating a cyclical "chicken-and-egg" scenario.[19]

This study adds to the growing body of research implementing ML algorithms for AP severity prediction. A XGB-based model, trained on patients from emergency and ward data, was demonstrated to predict severe AP with a significantly higher precision and accuracy, compared with BISAP and HAPS scores.[20] In one study, ML was implemented to predict acute kidney injury in AP patients, XGB performed best, though RF also showed strong area under the ROC values.[16] A large study using publicly available databases augmented with synthetic data found RF to be the most effective in many scenarios.[21] Another observational study highlighted XGB's superior performance, identifying glucose and albumin as key features, though it incorporated CTSI into the model, which may limit its applicability.[22] A systematic review further confirmed that ML models often surpass traditional scores in classification tasks for AP.[10]

Limitations

Our study is inherently limited by its retrospective design, single-center dataset, and relatively small sample size, which may affect the generalizability of the findings. In addition, the absence of complete 48-h data hindered our ability to fully calculate Ranson's criteria and reliably assess the Glasgow criteria, potentially underestimating their predictive performance.

While our study demonstrates the potential of ML models to predict adverse outcomes in AP, the generalizability of the findings may be limited due to the single-center dataset and the relatively small sample size. Future studies with larger, multicenter datasets and prospective designs are needed to validate these findings and explore the potential of ML algorithms in broader clinical settings.

Conclusion

The ML models, particularly RF, significantly outperformed traditional clinical scores in predicting adverse outcomes in AP. While BISAP and modified Glasgow showed some utility, their overall effectiveness was lower, particularly compared to the ML approaches. These findings suggest that integrating advanced ML models into clinical practice could enhance the accuracy and reliability of predicting adverse outcomes in AP. At last, "lactate," "glucose," "BUN," and "age" predictors that evaluate the prognosis of AP much more successfully than traditional scoring methods can be used in emergency departments.

Author contribution statement

The manuscript has been read and approved by all authors. Conceptualization: MO and AB, Data curation: MO, AB, and AEA, Formal analysis: MO, AB, and AEA, Methodology: MO and AB, Software: MO, Supervision: MO and AB, Validation: MO, Visualization: MO, AB, and AEA, Writing – Original Draft: MO, Writing – Review and Editing: AB and AEA.

Conflicts of interest

None Declared.

Ethical approval

This study was conducted in accordance with international and national regulations, aligning with the Declaration of Helsinki, the Human Tissue Act 2004, and the Turkish Data Protection Law. Ethical approval for this study was obtained from Hacettepe University Ethics Committee in Türkiye on the date of 27th December 2022, with approval number of GO 22/1317.

Funding

None.

References

1. Blamey SL, Imrie CW, O'Neill J, Gilmour WH, Carter DC. Prognostic factors in acute pancreatitis. Gut 1984;25:1340-6.
2. Simoes M, Alves P, Esperto H, Canha C, Meira E, Ferreira E, *et al*. Predicting acute pancreatitis severity: Comparison of prognostic scores. Gastroenterology Res 2011;4:216-22.
3. Ong Y, Shelat VG. Ranson score to stratify severity in acute pancreatitis remains valid – Old is gold. Expert Rev Gastroenterol Hepatol 2021;15:865-77.
4. Taylor SL, Morgan DL, Denson KD, Lane MM, Pennington LR. A comparison of the Ranson, Glasgow, and APACHE II scoring systems to a multiple organ system score in predicting patient outcome in pancreatitis. Am J Surg 2005;189:219-22.
5. Ranson JH, Rifkind KM, Roses DF, Fink SD, Eng K, Spencer FC. Prognostic signs and the role of operative management in acute pancreatitis. Surg Gynecol Obstet 1974;139:69-81.
6. Chauhan R, Saxena N, Kapur N, Kardam D. Comparison of modified Glasgow-Imrie, Ranson, and apache II scoring systems in predicting the severity of acute pancreatitis. Pol Przegl Chir 2022;95:6-12.
7. Papachristou GI, Muddana V, Yadav D, O'Connell M, Sanders MK, Slivka A, *et al*. Comparison of BISAP, Ranson's, APACHE-II, and CTSI scores in predicting organ failure, complications, and mortality in acute pancreatitis. Off J Am Coll Gastroenterol ACG 2010;105:435-41.
8. Larvin M, McMahon MJ. APACHE-II score for assessment and monitoring of acute pancreatitis. Lancet 1989;2:201-5.
9. Jin X, Ding Z, Li T, Xiong J, Tian G, Liu J. Comparison of MPL-ANN and PLS-DA models for predicting the severity of patients with acute pancreatitis: An exploratory study. Am J Emerg Med 2021;44:85-91.
10. Zhou Y, Ge YT, Shi XL, Wu KY, Chen WW, Ding YB, *et al.* Machine learning predictive models for acute pancreatitis: A systematic review. Int J Med Inform 2022;157:104641.
11. Mikó A, Vigh É, Mátrai P, Szakó L, Csupor D, Bajor J, *et al*. Computed tomography severity index versus other indices in

the prediction of severity and mortality in acute pancreatitis: A predictive accuracy meta-analysis. Systematic review. Front Physiol 2019;10:1002. [doi: 10.3389/fphys.2019.01002].
12. Steinberg WM. Predictors of severity of acute pancreatitis. Gastroenterol Clin North Am 1990;19:849-61.
13. Hagjer S, Kumar N. Evaluation of the BISAP scoring system in prognostication of acute pancreatitis – A prospective observational study. Int J Surg 2018;54:76-81.
14. Güzel YE, Çolak N, Okuv AC, Teymuroğlu S, Teke Mİ. Comparing prognostic scoring systems in acute pancreatitis: Bedside index of severity in acute pancreatitis, WL, and Chinese simple scoring system scores. Turk J Emerg Med 2024;24:165-71.
15. Wu BU, Johannes RS, Sun X, Conwell DL, Banks PA. Early changes in blood urea nitrogen predict mortality in acute pancreatitis. Gastroenterology 2009;137:129-35.
16. Yin M, Zhang R, Zhou Z, Liu L, Gao J, Xu W, *et al.* Automated machine learning for the early prediction of the severity of acute pancreatitis in hospitals. Front Cell Infect Microbiol 2022;12:886935.
17. Soeters PB, Wolfe RR, Shenkin A. Hypoalbuminemia: Pathogenesis and clinical significance. J Parenter Enteral Nutr 2019;43:181-93.
18. Zeng J, Wan J, He W, Zhu Y, Zeng H, Liu P, *et al*. Prognostic value of arterial lactate metabolic clearance rate in moderate and severe acute pancreatitis. Dis Markers 2022;2022:9233199. doi: 10.1155/2022/9233199.
19. Wang D, Lu J, Zhang P, Hu Z, Shi Y. Relationship between blood glucose levels and length of hospital stay in patients with acute pancreatitis: An analysis of MIMIC-III database. Clin Transl Sci 2023;16:246-57.
20. Thapa R, Iqbal Z, Garikipati A, Siefkas A, Hoffman J, Mao Q, *et al.* Early prediction of severe acute pancreatitis using machine learning. Pancreatology 2022;22:43-50.
21. Hameed MA, Alamgir Z. Improving mortality prediction in acute pancreatitis by machine learning and data augmentation. Comput Biol Med 2022;150:106077.
22. Zhou Y, Han F, Shi XL, Zhang JX, Li GY, Yuan CC, *et al.* Prediction of the severity of acute pancreatitis using machine learning models. Postgrad Med 2022;134:703-10.