Chirurgia (2026) 121: 268-277
No. 3, May - June
Copyright© Celsius
http://dx.doi.org/10.21614/chirurgia.3345

COLOSSUS-AI: A Pilot Study on Artificial Intelligence–Driven Prediction of Postoperative Outcomes in Colorectal Surgery

Marian Botoncea[1,2], Cătălin Dumitru Cosma[1,2]*, Cosmin-Lucian Nicolescu[1,2], Vlad-Olimpiu Butiurca[1,2], Dragoș Molnar[1,2], Călin Molnar[1,2]

[1]Faculty of Medicine, George Emil Palade University of Medicine, Pharmacy, Sciences and Technology of Târgu Mureș, Romania
[2]General Surgery Clinic No.1, County Emergency Clinical Hospital of Târgu-Mureș, Romania

***Corresponding author:**
Cătălin Dumitru Cosma, MD
38 Gheorghe Marinescu St
540139, Târgu Mureș, Romania
E-mail: catalin.cosma@umfst.ro

Abbreviations:
AI: Artificial intelligence;
ASA: American Society of Anesthesiologists;
AUC: Area under the curve;
BMI: Body mass index;
COLOSSUS-AI: Artificial intelligence-based colorectal surgery outcome prediction model
eGFR: Estimated glomerular filtration rate;
ICU: Intensive care unit;
IQR: Interquartile range;
ML: Machine learning;
NLR: Neutrophil-to-lymphocyte ratio;
NPV: Negative predictive value;
PLR: Platelet-to-lymphocyte ratio;
PPV: Positive predictive value;
ROC: Receiver operating characteristic;
ROC-AUC: Area under the receiver operating characteristic curve;
SD: Standard deviation;
SHAP: SHapley Additive exPlanations
TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence.

Received: 09.05.2026
Accepted: 18.06.2026

Rezumat

COLOSSUS-AI: Studiu pilot privind predicția rezultatelor postoperatorii după chirurgia colorectală cu ajutorul inteligenței artificiale

Introducere: Complicațiile postoperatorii reprezintă o cauză importantă de morbiditate după chirurgia colorectală, fiind asociate cu spitalizare prelungită, reintervenții, internare în terapie intensivă şi mortalitate crescută. Identificarea precoce a pacienților cu risc crescut rămâne dificilă, în special din cauza heterogenității populației chirurgicale colorectale. Inteligența artificială şi metodele de învățare automată pot contribui la integrarea variabilelor clinice, biologice şi operatorii într-un model predictiv individualizat. *Scop:* Scopul acestui studiu a fost dezvoltarea şi evaluarea internă a unui model preliminar de învățare automată, denumit COLOSSUS-AI, pentru predicția complicațiilor postoperatorii precoce după chirurgia colorectală.
Material şi Metodă: A fost realizat un studiu retrospectiv, observațional, unicentric, care a inclus 310 pacienți supuşi intervențiilor chirurgicale colorectale. Datele demografice, clinice, biologice, specifice bolii, operatorii şi postoperatorii au fost colectate retrospectiv şi organizate într-o bază de date structurată cu 84 de variabile. Criteriul principal de evaluare a fost apariția unui eveniment postoperator advers. Au fost evaluate trei modele predictive: regresie logistică, algoritm tip random forest şi amplificare gradientală. Performanța modelelor a fost apreciată prin analiza curbelor ROC, aria de sub curbă, acuratețe, sensibilitate şi specificitate.
Rezultate: Evenimentele adverse postoperatorii au fost înregistrate la 120 de pacienți (38,7%). Acestea au fost asociate cu vârsta mai înaintată, internarea în regim de urgență, valori crescute ale markerilor inflamatori, raport neutrofile-limfocite crescut, albumină preoperatorie scăzută, afectarea parametrilor funcției renale, fistula anastomotică şi mortalitatea intraspitalicească. Dintre modelele evaluate, amplificarea gradientală a avut cea mai bună performanță predictivă, cu o arie de sub curba ROC de 0,886, acuratețe de 81,7%, sensibilitate de 72,2% şi specificitate de 87,7%. Algortitmul tip random forest a obținut o arie de sub curbă de 0,841, iar regresia logistică de 0,768.
Concluzii: Studiul pilot COLOSSUS-AI sugerează că datele perioperatorii colectate în mod curent pot fi utilizate pentru dezvoltarea unor modele exploratorii de învățare automată destinate predicției evenimentelor adverse postoperatorii după chirurgia

colorectală. Deşi amplificarea gradientală a demonstrat cea mai bună performanță preliminară, rezultatele trebuie interpretate cu prudență, fiind necesare validări externe pe loturi multicentrice mai mari înainte de implementarea clinică.

Cuvinte cheie: chirurgie colorectală, inteligență artificială, învățare automată, complicații postoperatorii, model predictiv; fistulă anastomotică

Abstract

Background: Postoperative complications remain a major source of morbidity after colorectal surgery and are associated with prolonged hospitalization, reintervention, intensive care unit admission, and increased mortality. Early identification of high-risk patients remains challenging due to the heterogeneity of colorectal surgical populations. Artificial intelligence and machine learning methods may support individualized risk prediction by integrating clinical, biological, and operative variables. *Aim:* This study aimed to develop and internally evaluate a preliminary machine learning model, named COLOSSUS-AI, for predicting early postoperative adverse outcomes after colorectal surgery.

Material and Methods: This retrospective, single-center, observational pilot study included 310 patients undergoing colorectal surgery. Demographic, clinical, laboratory, disease-specific, operative, and postoperative data were retrospectively collected and organized into a structured database comprising 84 variables. The primary endpoint was postoperative adverse outcome. Three predictive models were evaluated: logistic regression, random forest, and gradient boosting. Model performance was assessed using receiver operating characteristic curve analysis, area under the curve, accuracy, sensitivity, and specificity.

Results: Postoperative adverse outcomes occurred in 120 patients (38.7%). Adverse outcomes were associated with older age, urgent or emergency admission, increased inflammatory markers, higher neutrophil-to-lymphocyte ratio, lower preoperative albumin, impaired renal function parameters, anastomotic leak, and in-hospital mortality. Among the evaluated models, gradient boosting achieved the best predictive performance, with an area under the receiver operating characteristic curve of 0.886, accuracy of 81.7%, sensitivity of 72.2%, and specificity of 87.7%. Random forest achieved an area under the curve of 0.841, while logistic regression achieved an area under the curve of 0.768.

Conclusions: The COLOSSUS-AI pilot study suggests that routinely collected perioperative data can be used to develop exploratory machine learning models for predicting postoperative adverse outcomes after colorectal surgery. Although gradient boosting showed the best preliminary performance, these findings should be interpreted cautiously. External validation in larger multicenter cohorts is required before clinical implementation.

Keywords: colorectal surgery, artificial intelligence, machine learning, postoperative complications, predictive model, anastomotic leak, gradient boosting

Introduction

Postoperative complications remain a significant challenge in colorectal surgery and are associated with increased morbidity, prolonged hospitalization, readmission, reoperation, mortality, and healthcare costs (1-4). Despite advances in perioperative care, surgical technique, minimally invasive approaches, and enhanced recovery protocols, the early identification of patients at increased postoperative risk remains difficult (5).

Colorectal surgery involves a heterogeneous patient population, including patients operated for benign disease, colorectal cancer, inflammatory bowel disease, diverticular disease, obstruction, perforation, or other emergency conditions. This heterogeneity makes postoperative risk prediction complex (6). Traditional risk scores and conventional statistical models can provide useful estimates, but they are often based on a limited number of a priori defined variables and may not fully capture nonlinear interactions between clinical, biological and operative parameters (7).

Artificial intelligence and machine learning methods have gained increasing attention in surgical research because of their capacity to process complex datasets and identify hidden patterns associated with adverse outcomes (8,9). In colorectal surgery, these approaches may support individualized risk stratification, early warning systems, and clinical decision-support tools (10,11). However, many available models remain limited by retrospective design, small datasets, insufficient validation, and lack of clinical interpretability (12).

The COLOSSUS-AI project was designed as a pilot study to assess the feasibility of developing an

artificial intelligence–based predictive model using routinely collected perioperative data from patients undergoing colorectal surgery. The aim of this study was to develop and internally evaluate a preliminary machine learning model for predicting early postoperative complications after colorectal resection.

Material and methods

Study Design

This was a retrospective, single-center, observational pilot study conducted in a tertiary surgical unit at the General Surgery Clinic No. 1, County Emergency Clinical Hospital of Târgu Mureş, affiliated with the George Emil Palade University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Romania.

Consecutive patients who underwent colorectal surgical procedures during the study period were identified from the institutional electronic medical record system. Clinical, laboratory, operative, and postoperative data were retrospectively extracted and entered into a structured database. The study was designed as a feasibility and proof-of-concept analysis intended to evaluate the potential of machine learning techniques for postoperative risk prediction.

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and received approvals from the Ethics Committee (no. 29331/21.10.2025).

Patient Selection

Adult patients (≥18 years) who underwent elective, urgent, or emergency colorectal surgery were considered eligible for inclusion. Both benign and malignant colorectal diseases were included.

Patients were excluded if essential perioperative data were unavailable, if postoperative outcome documentation was incomplete, or if duplicate records were identified. Cases with missing information regarding the primary endpoint were excluded from predictive modeling.

Data Collection

Clinical information was extracted from the institutional electronic medical record system and organized into a structured database comprising 84 variables for 310 patients.

Variables were grouped into five predefined domains:

1. Demographic variables (age, sex, body mass index, ASA score);
2. Clinical variables (comorbidities, smoking status, previous abdominal surgery);
3. Laboratory variables (hemoglobin, leukocyte count, neutrophil count, lymphocyte count, platelet count, albumin, creatinine, urea, C-reactive protein);
4. Disease-specific variables (diagnosis, tumor location, disease stage, emergency versus elective presentation);
5. Operative variables (procedure type, operative approach, operative time, blood loss, anastomosis creation, protective stoma formation).

Derived variables were calculated when appropriate, including the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), estimated glomerular filtration rate (eGFR), and length of hospital stay.

Outcome Definitions

The primary endpoint was postoperative adverse outcome, defined as the occurrence of at least one clinically relevant postoperative event during the index hospitalization. For the purpose of this study, postoperative adverse outcome included documented anastomotic leak, suspected digestive postoperative complication, need for reintervention during the same admission, intensive care unit admission or postoperative intensive care requirement, and in-hospital mortality.

Anastomotic leak was defined as a clinically, radiologically, endoscopically, or intraoperatively documented defect of the intestinal anastomosis requiring conservative, interventional, or surgical management. Reintervention was defined as any unplanned surgical or interventional procedure performed during the same hospitalization as a consequence of postoperative deterioration or complication. In-hospital mortality was defined as death occurring during the index admission.

Postoperative complications were classified according to the Clavien–Dindo classification whenever sufficient documentation was available. Major complications were considered Clavien–Dindo grade III or higher. Because of the limited number of isolated events, particularly anastomotic leak and reintervention, the primary endpoint was analyzed as a composite outcome to provide sufficient event frequency for exploratory predictive modeling.

Secondary endpoints included anastomotic leak, intensive care unit admission, reintervention during the same admission, and in-hospital mortality.

Data Preprocessing

Before model development, the dataset underwent a standardized preprocessing workflow. Missing values, inconsistencies, and biologically implausible outliers were identified and reviewed. Variables with excessive missingness were excluded from predictive modeling when appropriate.

Categorical variables were transformed using numerical encoding techniques. Continuous variables were standardized when required by the selected algorithms. Feature selection was performed using a combination of clinical relevance and exploratory univariate analysis.

To prevent information leakage, outcome variables were excluded from the predictor matrix before model training.

Machine Learning Model Development

Predictive modeling was performed using Python and the Scikit-learn machine learning library. The primary endpoint used for model development was postoperative adverse outcome.

Before model training, all outcome-related variables were excluded from the predictor matrix to reduce the risk of information leakage. The dataset was then divided into training and testing subsets using a stratified random split, with 70% of patients allocated to the training set and 30% to the testing set. Stratification was performed according to the primary endpoint in order to preserve the proportion of postoperative adverse outcomes in both subsets.

Model development, including preprocessing, imputation, encoding, scaling when required, feature selection, and parameter optimization, was performed using the training dataset. The testing dataset was kept separate and was used only for final internal performance assessment.

Three predictive approaches were evaluated: logistic regression, random forest, and gradient boosting. Logistic regression was used as an interpretable baseline model and was implemented with balanced class weighting to account for class imbalance. Random forest and gradient boosting were evaluated as exploratory tree-based machine learning models because of their ability to model nonlinear relationships and interactions between perioperative variables.

Model performance was assessed on the independent testing subset using receiver operating characteristic curve analysis. The area under the receiver operating characteristic curve was used as the main discrimination metric. Additional performance metrics included accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

No external validation cohort was available; therefore, the reported model performance represents internal validation only. The results were considered exploratory and hypothesis-generating rather than definitive evidence of clinical applicability.

The simplified workflow of the COLOSSUS-AI exploratory prediction model, including data preprocessing, model development, and internal validation, is presented in *Fig. 1*.

Model Evaluation

Model performance was assessed on the testing

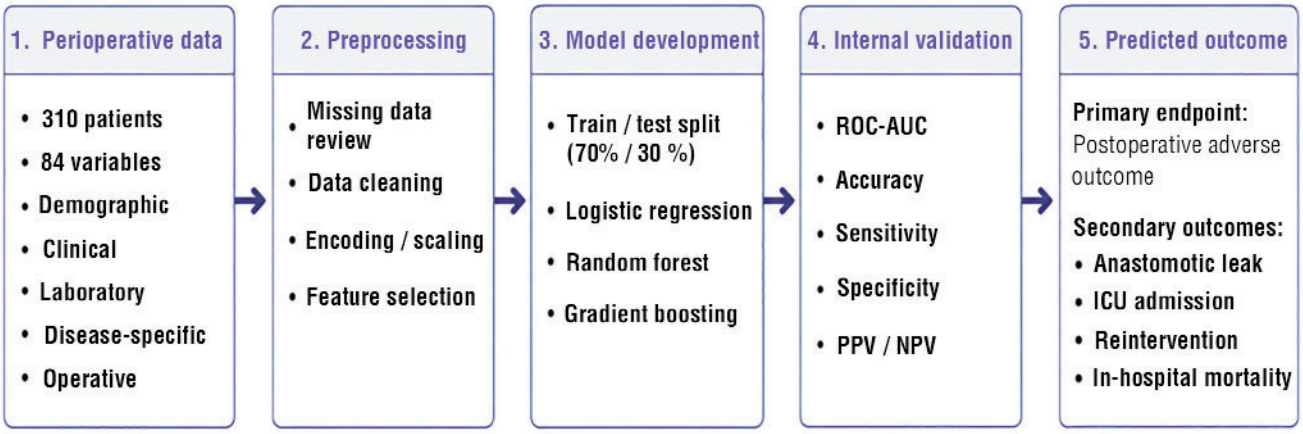

Figure 1. Simplified COLOSSUS-AI workflow for predicting postoperative adverse outcomes after colorectal surgery, including data preprocessing, model development, and internal validation.

dataset using receiver operating characteristic curve analysis. The area under the receiver operating characteristic curve was used as the main measure of model discrimination. Additional performance metrics included accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.

Given the exploratory nature of this pilot study and the limited sample size, model calibration and clinical utility were not considered definitive endpoints. Calibration was assessed descriptively where applicable, but no formal recalibration was performed. Decision-curve analysis was not performed in this pilot phase. Therefore, clinical applicability was interpreted cautiously and limited to preliminary risk stratification rather than direct clinical decision-making.

Future validation studies should include calibration plots, calibration intercept and slope, Brier score, and decision-curve analysis to determine whether the model provides clinically meaningful benefit across relevant risk thresholds.

Statistical Analysis

Descriptive statistics were used to summarize patient characteristics, operative variables, and postoperative outcomes.

Continuous variables were expressed as mean ± standard deviation or median with interquartile range (IQR), according to distribution. Categorical variables were reported as frequencies and percentages.

Comparisons between patients with and without postoperative adverse outcomes were performed using the Student's t-test or Mann–Whitney U test for continuous variables and the χ^2 test or Fisher's exact test for categorical variables, as appropriate.

Statistical analyses were performed using Python with the Scikit-learn library for machine learning procedures. Descriptive and comparative analyses were verified using EasyMedStat. A two-sided p value <0.05 was considered statistically significant.

Results

Patient Characteristics

A total of 310 patients undergoing colorectal surgery were included in the final analysis. The cohort comprised 186 men (60.0%) and 124 women (40.0%). The mean age was 67.3 ± 10.9 years, while the median age was 68 years (interquartile range (IQR): 61–75).

Malignant disease was recorded in 293 patients (94.5%). Regarding admission type, 183 patients (59.0%) were admitted electively, 73 patients (23.5%) urgently, and 54 patients (17.4%) as emergency cases. Overall, urgent or emergency admissions accounted for 127 patients (41.0%). The median length of hospital stay was 8 days (IQR: 7–10), with a mean duration of 9.4 ± 7.0 days.

Table 1. Baseline and operative characteristics of the study cohort

Variable	Value
Total patients	310
Male sex	186 (60.0%)
Female sex	124 (40.0%)
Age, mean ± SD	67.3 ± 10.9 years
Age, median (IQR)	68 (61–75) years
Malignant disease	293 (94.5%)
Elective admission	183 (59.0%)
Urgent or emergency admission	127 (41.0%)
Length of hospital stay, mean ± SD	9.4 ± 7.0 days
Length of hospital stay, median (IQR)	8 (7–10) days
Open surgical approach	211 (68.1%)
Laparoscopic surgical approach	59 (19.0%)
Converted procedure	23 (7.4%)
Anastomosis performed	171 (55.2%)
Protective stoma	47 (15.2%)
Adhesiolysis performed	199 (64.2%)
Peritoneal adhesions documented	280 (90.3%)

Abbreviations: SD, standard deviation; IQR, interquartile range.

Baseline and operative characteristics of the study cohort are summarized in *Table 1.*

Operative Characteristics

The most frequently performed procedures were right hemicolectomy in 85 patients (27.4%), rectosigmoid resection in 69 patients (22.3%), left hemicolectomy in 27 patients (8.7%), sigmoid resection in 24 patients (7.7%), Hartmann procedure in 11 patients (3.5%), and abdominoperineal rectal amputation in 11 patients (3.5%).

The surgical approach was open in 211 patients (68.1%), laparoscopic in 59 patients (19.0%), and converted in 23 patients (7.4%). An anastomosis was performed in 171 patients (55.2%), while a protective stoma was created in 47 patients (15.2%). Adhesiolysis was performed in 199 patients (64.2%), and peritoneal adhesions were documented in 280 patients (90.3%).

Postoperative Outcomes

The composite primary endpoint included clinically relevant postoperative digestive complications, anastomotic leak, reintervention, intensive care unit involvement, or in-hospital mortality.

Postoperative adverse outcomes were recorded in 120 of 310 patients (38.7%). Documented anastomotic leak occurred in 7 patients (2.3%), while suspected digestive postoperative complications were recorded in

123 patients (39.7%). Reintervention during the same admission was required in 5 patients (1.6%). Intensive care unit involvement was mentioned in 149 cases (48.1%). In-hospital mortality occurred in 27 patients (8.7%). Postoperative outcomes are summarized in *Table 2*.

Table 2. Postoperative outcomes

Outcome	Result
Postoperative adverse outcome	120 (38.7%)
Documented anastomotic leak	7 (2.3%)
Suspected digestive postoperative complication	123 (39.7%)
Reintervention during the same admission	5 (1.6%)
Intensive care unit involvement	149 (48.1%)
In-hospital mortality	27 (8.7%)

Note: Postoperative adverse outcome was the primary endpoint used for predictive modeling.

Comparison According to Postoperative Adverse Outcome

Patients with postoperative adverse outcomes were significantly older than those without adverse outcomes, with a median age of 70 years (IQR: 63–77) compared with 67 years (IQR: 60–73), respectively (p=0.008).

Urgent or emergency admission was significantly more frequent among patients with adverse outcomes. In the adverse outcome group, 63 of 120 patients (52.5%) were admitted urgently or as emergency cases, compared with 64 of 190 patients (33.7%) in the group without adverse outcomes (p=0.002).

Preoperative inflammatory markers were also associated with adverse outcomes. Patients with adverse outcomes had a higher median white blood cell count, 8.38 $\times 10^9$/L (IQR: 6.52–10.25), compared with 7.81$\times 10^9$/L (IQR: 6.21–9.31) in patients without adverse outcomes (p=0.022). Median neutrophil count was also higher in the adverse outcome group, 5.75 $\times 10^9$/L (IQR: 4.17–8.79), compared with 5.28 $\times 10^9$/L (IQR: 3.96–6.39) (p=0.008).

The neutrophil-to-lymphocyte ratio was significantly higher in patients with adverse outcomes, with a median value of 4.05 (IQR: 2.82–6.14), compared with 3.21 (IQR: 2.33–4.69) in patients without adverse outcomes (p<0.001). Preoperative albumin was significantly lower in the adverse outcome group, with a median value of 3.15 g/dL (IQR: 2.69–3.52), compared with 3.66 g/dL (IQR: 3.20–3.99) in patients without adverse outcomes (p<0.001).

Renal function parameters also differed significantly between groups. Patients with adverse outcomes had a lower median estimated glomerular filtration rate, 75.7 mL/min/1.73 m^2 (IQR: 61.5–93.0), compared with 82.1 mL/min/1.73 m^2 (IQR: 72.6–93.9) in patients without adverse outcomes (p=0.006). Median preoperative urea was higher in the adverse outcome group, 34.6 mg/dL (IQR: 27.8–42.8), compared with 30.8 mg/dL (IQR: 25.5–37.2) in patients without adverse outcomes (p<0.001).

Length of hospital stay was longer in patients with adverse outcomes, with a median of 9 days (IQR: 7–12), compared with 8 days (IQR: 7–9.75) in patients without adverse outcomes; however, this difference did not reach statistical significance (p=0.056).

Documented anastomotic leak was significantly more frequent in patients with adverse outcomes, occurring in 6 of 120 patients (5.0%) compared with 1 of 190 patients (0.5%) without adverse outcomes (p=0.015). In-hospital mortality occurred exclusively in the adverse outcome group, being recorded in 27 of 120 patients (22.5%) (p<0.001).

Comparative analysis according to postoperative adverse outcome is presented in *Table 3*.

Exploratory Predictive Modeling

Exploratory predictive modeling was performed using

Table 3. Comparison according to postoperative adverse outcome

Variable	No adverse outcome n=190	Adverse outcome n=120	p value
Age, median (IQR), years	67 (60–73)	70 (63–77)	0.008
Urgent/emergency admission	64 (33.7%)	63 (52.5%)	0.002
White blood cell count, median (IQR), X10 /L	7.81 (6.21–9.31)	8.38 (6.52–10.25)	0.022
Neutrophil count, median (IQR), X10 /L	5.28 (3.96–6.39)	5.75 (4.17–8.79)	0.008
Neutrophil-to-lymphocyte ratio, median (IQR)	3.21 (2.33–4.69)	4.05 (2.82–6.14)	<0.001
Albumin, median (IQR), g/dL	3.66 (3.20–3.99)	3.15 (2.69–3.52)	<0.001
eGFR, median (IQR), mL/min/1.73 m^2	82.1 (72.6–93.9)	75.7 (61.5–93.0)	0.006
Urea, median (IQR), mg/dL	30.8 (25.5–37.2)	34.6 (27.8–42.8)	<0.001
Length of hospital stay, median (IQR), days	8 (7–9.75)	9 (7–12)	0.056
Anastomotic leak	1 (0.5%)	6 (5.0%)	0.015
In-hospital mortality	0 (0%)	27 (22.5%)	<0.001

Abbreviations: IQR, interquartile range; eGFR, estimated glomerular filtration rate.

routinely available demographic, clinical, laboratory, disease-specific, and operative variables. Logistic regression was used as an interpretable baseline model, while random forest and gradient boosting were evaluated as machine learning approaches.

Receiver operating characteristic curve analysis demonstrated good discriminative performance of the evaluated models (*Fig. 2*). Gradient boosting achieved the highest area under the curve (AUC = 0.886), followed by random forest (AUC = 0.841) and logistic regression (AUC = 0.768).

Gradient boosting also showed the best overall classification performance, with an accuracy of 81.7%, sensitivity of 72.2%, and specificity of 87.7%. Random forest achieved an accuracy of 74.2%, sensitivity of 58.3%, and specificity of 84.2%, whereas logistic regression achieved an accuracy of 68.8%, sensitivity of 61.1%, and specificity of 73.7%. Predictive model performance is summarized in *Table 4*.

Overall, the predictive models demonstrated moderate-to-good preliminary discrimination, with gradient boosting showing the best overall performance for prediction of postoperative adverse outcomes. However, given the retrospective single-center design and pilot nature of the dataset, these findings should be interpreted as exploratory and require external validation before clinical implementation.

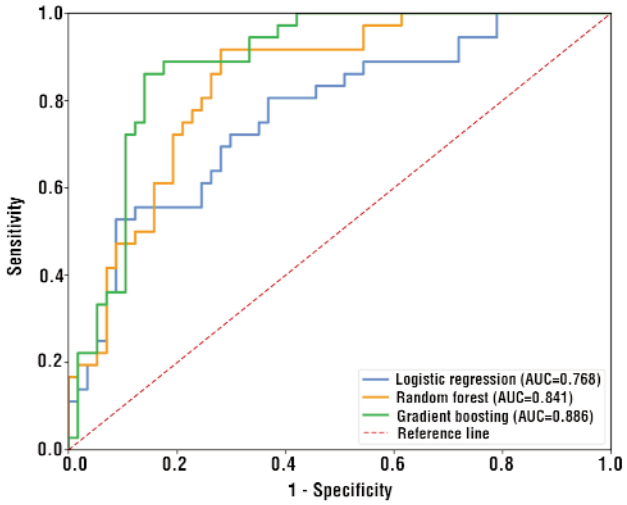

Figure 2. ROC curve analysis of the COLOSSUS-AI models. Receiver operating characteristic curves of logistic regression, random forest, and gradient boosting for prediction of postoperative adverse outcomes after colorectal surgery

Summary of Main Findings

In this retrospective pilot cohort, postoperative adverse outcomes occurred in 38.7% of patients undergoing colorectal surgery. Adverse outcomes were associated with older age, urgent or emergency admission, higher inflammatory markers, increased neutrophil-to-lymphocyte ratio, lower preoperative albumin, impaired renal function parameters, documented anastomotic leak, and in-hospital mortality. The COLOSSUS-AI exploratory modeling framework demonstrated feasible predictive performance, with the best-performing model, gradient boosting, achieving an AUC of 0.886.

Discussion

This retrospective pilot study evaluated the feasibility of developing an artificial intelligence–based predictive model for early postoperative adverse outcomes after colorectal surgery using routinely available perioperative data. In a cohort of 310 patients, postoperative adverse outcomes occurred in 38.7% of cases. Adverse outcomes were associated with older age, urgent or emergency admission, increased inflammatory markers, higher neutrophil-to-lymphocyte ratio, lower preoperative albumin, impaired renal function parameters, documented anastomotic leak, and in-hospital mortality. Exploratory predictive modeling demonstrated moderate-to-good discrimination, with gradient boosting achieving the highest area under the receiver operating characteristic curve among the evaluated models.

The present findings are consistent with recent literature showing that postoperative risk prediction after colorectal surgery is clinically relevant but methodologically challenging. Colorectal surgery includes a heterogeneous patient population, with differences in indication, urgency, tumor location, nutritional reserve, inflammatory status, operative complexity, and postoperative management. This heterogeneity limits the performance of simple risk

Table 4. Predictive model performance

Model	AUC	Accuracy	Sensitivity	Specificity
Logistic regression	0.768	68.8%	61.1%	73.7%
Random forest	0.841	74.2%	58.3%	84.2%
Gradient boosting	0.886	81.7%	72.2%	87.7%

Abbreviations: AUC, area under the receiver operating characteristic curve.

stratification tools and supports the exploration of multidimensional predictive approaches. Recent medical literature suggests that machine learning models may improve postoperative outcome prediction in colorectal surgery, but their clinical utility remains limited by heterogeneity of datasets, variable outcome definitions, and insufficient external validation (13-15).

In this study, urgent or emergency admission was significantly associated with postoperative adverse outcomes. This finding is clinically expected, as non-elective colorectal surgery is frequently performed in patients with obstruction, perforation, peritonitis, bleeding, advanced disease, or impaired physiological reserve. Similar observations have been reported in studies evaluating postoperative complications and anastomotic leak, where emergency presentation and disease severity are repeatedly associated with worse outcomes (11,15). The inclusion of admission status in predictive modeling is therefore justified, particularly in mixed colorectal cohorts including both elective and emergency surgery.

Inflammatory and nutritional markers were also strongly associated with adverse outcomes in the present cohort. Patients with postoperative adverse outcomes had higher leukocyte and neutrophil counts, higher neutrophil-to-lymphocyte ratio, and lower albumin levels. These findings support the concept that preoperative systemic inflammation and impaired nutritional reserve are important components of postoperative vulnerability. Recent machine learning studies in colorectal surgery have similarly emphasized the importance of routinely available clinical and laboratory variables for predicting adverse postoperative outcomes, including anastomotic leak and major complications (11,14).

Anastomotic leak was uncommon in the present dataset, occurring in 2.3% of patients, but it was significantly more frequent among patients with postoperative adverse outcomes. Although the low event rate prevented robust modeling of anastomotic leak as an independent primary endpoint, its clinical importance remains substantial. Anastomotic leak is one of the most severe complications after colorectal surgery and is associated with reintervention, sepsis, prolonged hospitalization, impaired quality of life, and mortality (15,16). Recent studies continue to focus specifically on AI-based anastomotic leak prediction, including models based on preoperative clinical variables, large multicenter datasets, and integrative machine learning approaches (11,15).

The COLOSSUS-AI models demonstrated encouraging preliminary performance. Logistic regression achieved an AUC of 0.768, while random forest and gradient boosting achieved AUC values of 0.841 and 0.886, respectively. These findings are broadly consistent with recent colorectal AI literature, where ensemble-based models frequently show promising discrimination for postoperative complications and anastomotic leak prediction (11,14,17). However, the degree of benefit over regression-based models varies between studies, and model performance may depend on endpoint definition, data quality, case-mix, event rate, and validation strategy (14,18).

Among the evaluated models, gradient boosting showed the best performance, with the highest AUC, accuracy, sensitivity, and specificity. This may be explained by the ability of boosting algorithms to model nonlinear relationships and interactions between multiple perioperative variables. However, this result should be interpreted cautiously. In relatively small retrospective datasets, high internal performance may reflect overfitting, especially when many candidate predictors are available. Therefore, although the AUC of 0.886 is promising, it should be considered exploratory until tested in independent external cohorts.

The need for cautious interpretation is reinforced by current methodological standards for prediction model studies. The TRIPOD+AI statement emphasizes transparent reporting of prediction model development and validation, regardless of whether regression or machine learning methods are used (19). It also highlights the importance of clear outcome definition, predictor selection, missing-data handling, validation strategy, and reporting of performance metrics (19). In this context, the present study should be viewed as an initial feasibility model rather than a definitive clinical decision-support tool.

Compared with studies focused exclusively on anastomotic leak, COLOSSUS-AI used a broader composite endpoint of postoperative adverse outcome. This approach has both advantages and disadvantages. The main advantage is that it provides a higher event rate, allowing more stable exploratory modeling in a pilot cohort. The main limitation is that a composite endpoint may group clinically different events, such as infection, leak, intensive care unit admission, reintervention, or death. Future versions of the model should therefore evaluate separate endpoints when sample size allows, particularly anastomotic leak, major Clavien–Dindo complications (20), intensive care unit admission, and mortality.

The clinical value of COLOSSUS-AI lies in its potential to support postoperative risk stratification. A validated model could help identify patients requiring closer monitoring, early biological reassessment, intensified surveillance, nutritional correction, or

earlier imaging when clinical deterioration occurs. Recent literature suggests that AI-based prediction tools may become clinically useful only when they are externally validated, interpretable, and integrated into surgical workflows (15,18,19). Therefore, further development of COLOSSUS-AI should focus not only on discrimination, but also on calibration, explainability, and clinical utility.

This study has several limitations. First, the retrospective design introduces potential selection bias, information bias, and missing-data bias. Second, this was a single-center study, which limits generalizability to other institutions with different patient populations, surgical practices, perioperative protocols, and documentation systems. Third, although the dataset included 310 patients and 84 variables, the sample size remains limited for machine learning, particularly for low-frequency endpoints such as anastomotic leak and reintervention. Fourth, the models were internally evaluated only, without external validation. No independent external cohort was available, and therefore the reported AUC values may overestimate real-world performance. Future studies should validate the COLOSSUS-AI model in larger external and multicenter cohorts. Another limitation is that calibration and clinical utility were not comprehensively assessed. Therefore, although the models demonstrated promising discrimination, their ability to provide reliable individual risk estimates and clinically useful decision support remains unproven. Future research must involve calibration analysis, decision-curve analysis, feature importance, and explainability methods like SHAP values.

Despite these limitations, the study has several strengths. It uses real-world clinical data from a tertiary surgical unit, includes both elective and emergency colorectal cases, evaluates multiple perioperative domains, and compares an interpretable regression model with machine learning approaches. The study also follows the current direction of colorectal AI research by emphasizing feasibility, internal validation, transparent performance reporting, and the need for external validation before clinical implementation (14,19).

Future development of the COLOSSUS-AI project should focus on multicenter dataset expansion, prospective validation, improved endpoint standardization, and integration of explainable AI methods. Additional predictors such as frailty indices, detailed comorbidity scores, intraoperative hemodynamic parameters, radiological findings, and enhanced recovery compliance may further improve model performance. Ultimately, the model should be evaluated not only for discrimination but also for calibration, clinical utility, and impact on postoperative decision-making.

In conclusion, this pilot study suggests that routinely collected perioperative data can be used to develop exploratory machine learning models for predicting postoperative adverse outcomes after colorectal surgery. Gradient boosting achieved the best preliminary performance, but the findings require cautious interpretation. Larger, multicenter, externally validated studies are necessary before COLOSSUS-AI can be considered for clinical implementation.

Conclusions

This retrospective pilot study suggests that routinely collected perioperative data can be used to develop exploratory machine learning models for predicting early postoperative adverse outcomes after colorectal surgery. In the present cohort, postoperative adverse outcomes were associated with older age, urgent or emergency admission, increased inflammatory markers, higher neutrophil-to-lymphocyte ratio, lower preoperative albumin, impaired renal function parameters, anastomotic leak, and in-hospital mortality.

Among the evaluated predictive models, gradient boosting showed the best preliminary performance, achieving the highest area under the receiver operating characteristic curve. These findings support the feasibility of the COLOSSUS-AI framework as a pilot approach for postoperative risk stratification in colorectal surgery.

However, the results should be interpreted cautiously due to the retrospective single-center design, limited sample size, and lack of external validation. Further multicenter studies with larger datasets, standardized endpoint definitions, calibration assessment, and explainable artificial intelligence methods are required before clinical implementation.

Author's Contributions

Conceptualization, M.B. and C.D.C.; methodology, M.B., C.D.C., and C.-L.N.; software and data processing, M.B. and V.-O.B.; validation, C.D.C., D.M., and C.M.; formal analysis, M.B. and V.-O.B.; investigation, M.B., C.-L.N., and V.-O.B.; data curation, M.B., C.-L.N., and V.-O.B.; writing-original draft preparation, M.B.; writing-review and editing, C.D.C., C.-L.N., V.-O.B., D.M., and C.M.; visualization, M.B. and V.-O.B.; supervision, D.M. and C.M.; project administration, M.B. and C.D.C. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors declare that ChatGPT-5 (OpenAI, San Francisco, CA, USA) was utilized for language editing and to help with the initial conceptual design of Figure 1. The authors evaluated, changed, approved the final figure content, scientific meaning, workflow organization and interpretation. No data, analysis, conclusions or interpretation of any kind was produced by AI and included in this study. The authors independently conceived, verified, and approved all scientific material, study design, methods, data interpretation and results. Statistical computations and visualizations related to descriptive and comparative analyses were verified using EasyMedStat. Machine learning procedures and ROC curve analysis were performed using Python and the Scikit-learn library.

Conflicts of Interest

The authors declare no conflict of interests

Funding

This research received no external funding.

Ethical Considerations

This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and received approvals from the Ethics Committee (no. 29331/21.10.2025). Due to the retrospective and anonymized nature of the study, the requirement for individual informed consent was waived by the Ethics Committee.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request. Due to ethical and privacy restrictions, the dataset is not publicly available.

References

1. Marginean SS, Zurzu M, Garofil D, Tigora A, Paic V, Bratucu M, et al. Evaluation of Key Risk Factors Associated with Postoperative Complications in Colorectal Cancer Surgery. J. Mind Med. Sci. 2025;12:22.
2. Mascarenhas M, Mendes F, Fonseca F, Carvalho E, Santos A, Cavadas D, et al. A Novel Deep Learning Model for Predicting Colorectal Anastomotic Leakage: A Pioneer Multicenter Transatlantic Study. Clin Med. 2025;14(15):5462.
3. Louis M, Johnston SA, Churilov L, Ma R, Christophi C, Weinberg L. Financial burden of postoperative complications following colonic resection: A systematic review. Medicine (Baltimore). 2021;100(27):e26546.
4. Andras D, Lazar AM, Crețoiu D, Berghea F, Georgescu DE, Grigorean V, et al. Analyzing postoperative complications in colorectal cancer surgery: a systematic review enhanced by artificial intelligence. Front Surg 2024;11:1452223.
5. Talmor D, Kelly B. How to better identify patients at high risk of postoperative complications? Curr Opin Crit Care. 2017;23(5):417-423.
6. Pata F, Tutino R, Picciariello A, Cantarella F. Editorial: Colorectal surgery and proctology: past, present, and future. Front Surg. 2024;11:1373867.
7. Elengickal D, Nizich M, Toma M. Stratified Procedural Risk Assessment in Colorectal Surgery: A Comparative Analysis of Statistical and Machine Learning Approaches Using Combined Surgical Approach and Operative Duration Categories. Surgeries 2026;7:42.
8. Singh V, Vasisht S, Hashimoto DA. Artificial intelligence in surgery: what is needed for ongoing innovation. Surgery (Oxford) 2025;43:129–134.
9. Ahmed MM, Othman ZK, Adebayo UO, Kasimieh O, Okesanya OJ, Musa SS, et al. Artificial intelligence and machine learning approaches for patient safety in complex surgery: a review. Patient Saf Surg. 2025; 19(1):33.
10. Fu Y-Y, Jiao Y, Liu Y-H, Dong S-S. Application and challenges of artificial intelligence in predicting perioperative complications of colorectal cancer. World J Gastrointest Surg. 2025;17(7):106340.
11. Taha-Mehlitz S, Däster S, Bach L, Ochs V, von Flüe M, Steinemann D, et al. Modern Machine Learning Practices in Colorectal Surgery: A Scoping Review. J Clin Med. 2022;11(9):2431.
12. Fındık M. Research trends and ethical perspectives on explainable artificial intelligence in emergency medicine: a bibliometric analysis. Scand J Trauma Resusc Emerg Med. 2026;34(1):26.
13. Celotto F, Capelli G, Ferrari S, Scarpa M, Pucciarelli S, Spolverato G. Application and use of artificial intelligence in colorectal cancer surgery: where are we? Art Int Surg 2024;4:348-63.
14. Mohamedahmed AY, Zaman S, Agrof M, Adam MA, Husain N, Yassin NA. Systematic review and meta-analysis of the role of machine learning in predicting postoperative complications following colorectal surgery: how far has machine learning come? Int J Surg. 2025;111(11):8550-8562.
15. Ochs V, Taha-Mehlitz S, Lavanchy JL, Steinemann D, Müller BP, Rosenberg R, et al. Development of a Clinical Prediction Model for Anastomotic Leakage in Colorectal Surgery. JAMA Netw Open. 2025;8(10):e2538267.
16. Gielen AHC, Heuvelings DJI, Sylla P, van Loon Y-T, Melenhorst J, Bouvy ND, et al. Impact of Anastomotic Leakage After Colorectal Cancer Surgery on Quality of Life: A Systematic Review. Dis Colon Rectum. 2025;68(2):154-170.
17. Celotto F, Bao QR, Capelli G, Spolverato G, Gumbs AA. Machine learning and deep learning to improve prevention of anastomotic leak after rectal cancer surgery. World J Gastrointest Surg. 2025;17(1):101772.
18. Cira K, Neumann P-A. Machine Learning for Anastomotic Leak Prediction in Colorectal Surgery—Between Validation and Clinical Implementation. JAMA Netw Open. 2025;8(10):e2538274.
19. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.
20. Clavien PA, Barkun J, de Oliveira ML, Vauthey JN, Dindo D, Schulick RD, et al. The Clavien-Dindo Classification of Surgical Complications: Five-Year Experience. Ann Surg. 2009;250(2):187-96.