Chirurgia (2026) 121: 371-382
No. 4, July - August

http://dx.doi.org/10.21614/chirurgia.3376

PRIMARCH-APP AI: Development and Internal Validation of an Explainable Machine Learning Model for Preoperative Prediction of Complicated Acute Appendicitis

Dragoș Călin Molnar[1,2], Cătălin Dumitru Cosma[1,2]*, Marian Botoncea[1,2], Adrian Tudor[1,2], Vlad Olimpiu Butiurcă[1,2], Călin Molnar[1,2], Bogdan Suciu Andrei[1,2]

[1]George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Târgu Mureș, Romania
[2]General Surgery I, Emergency County Hospital of Târgu Mureș, Romania

***Corresponding author:**
Catalin Cosma, MD
General Surgery I,
Emergency County Hospital,
Târgu Mureș, Romania
E-mail: catalin.cosma@umfst.ro

Abbreviations:
AI: Artificial Intelligence;
AIR: Appendicitis Inflammatory Response;
AUC: Area Under the Curve;
BMI: Body Mass Index;
CI: Confidence Interval;
CRP: C-Reactive Protein;
CT: Computed Tomography;
CV: Cross-Validation;
FN: False Negative;
FP: False Positive;
ICU: Intensive Care Unit;
IQR: Interquartile Range;
LR: Logistic Regression;
ML: Machine Learning;
NLR: Neutrophil-to-Lymphocyte Ratio;
NPV: Negative Predictive Value;
PPV: Positive Predictive Value;
PRIMARCH-APP AI: PRIMARCH Appendicitis Artificial Intelligence;
RF: Random Forest;
RLQ: Right Lower Quadrant,
ROC: Receiver Operating Characteristic;

Received: 09.06.2026
Accepted: 07.08.2026

Rezumat

PRIMARCH-APP AI: Dezvoltarea și validarea internă a unui model explicabil de învățare automată pentru predicția preoperatorie a apendicitei acute complicate

Introducere: Identificarea precoce a apendicitei acute complicate poate îmbunătăți stratificarea preoperatorie a riscului şi poate sprijini procesul decizional chirurgical. Scopul studiului a fost dezvoltarea şi validarea internă a unui model explicabil de învățare automată pentru predicția apendicitei acute complicate utilizând variabile preoperatorii disponibile în practica clinică curentă.
Materiale şi Metodă: În acest studiu prospectiv observațional au fost incluşi 199 de pacienți adulți supuşi apendicectomiei în regim de urgență. Variabilele demografice, clinice, biologice, scorurile clinice şi parametrii imagistici colectați preoperator au fost utilizați pentru dezvoltarea modelelor de regresie logistică, random forest şi XGBoost. Performanța a fost evaluată prin validare încrucişată în cinci subseturi şi pe un set de testare intern stratificat. Explicabilitatea modelului a fost evaluată utilizând SHapley Additive exPlanations (SHAP)
Rezultate: Apendicita acută complicată a fost identificată la 57/199 pacienți (28,6%). Random forest şi XGBoost au obținut cea mai mare valoare ROC-AUC pe setul de testare (0,934). Modelul random forest a prezentat cea mai favorabilă performanță globală, cu o acuratețe de 87,5%, sensibilitate de 72,7%, specificitate de 93,1% şi un scor Brier de 0,091. Analiza SHAP a identificat frecvența cardiacă, scorul Alvarado, sensibilitatea la decompresiune, durata simptomatologiei şi diametrul apendicelui printre cei mai importanți predictori.
Concluzii: PRIMARCH-APP AI a demonstrat o capacitate discriminatorie ridicată şi o predicție clinic interpretabilă a apendicitei acute complicate utilizând date preoperatorii disponibile în practica curentă. Validarea externă şi prospectivă multicentrică este necesară înaintea implementării clinice.

Cuvinte cheie: apendicită acută, apendicită complicată, învățare automată, inteligență artificială, inteligență artificială explicabilă, SHAP, predicția riscului

ROC-AUC: Area Under the Receiver Operating Characteristic Curve;
SD: Standard Deviation;
SHAP: SHapley Additive exPlanations;
TN: True Negative;
TP: True Positive;
US: Ultrasonography;
WBC: White Blood Cell Count;
XAI: Explainable Artificial Intelligence;
XGBoost: Extreme Gradient Boosting.

Abstract

Background: Early identification of complicated acute appendicitis may improve preoperative risk stratification and support surgical decision-making. This study aimed to develop and internally validate an explainable machine-learning model for predicting complicated acute appendicitis using routinely available preoperative variables.

Methods: In this prospective observational cohort study, 199 adult patients undergoing emergency appendectomy were included. Demographic, clinical, laboratory, scoring, and imaging variables collected before surgery were used to develop logistic regression, random forest, and XGBoost models. Performance was evaluated using five-fold cross-validation and a stratified held-out test set. Model explainability was assessed using SHapley Additive exPlanations (SHAP).

Results: Complicated acute appendicitis was identified in 57/199 patients (28.6%). Random forest and XGBoost achieved the highest test ROC-AUC (0.934). Random forest demonstrated the best overall probability performance, with 87.5% accuracy, 72.7% sensitivity, 93.1% specificity, and a Brier score of 0.091. SHAP analysis identified heart rate, Alvarado score, rebound tenderness, symptom duration, and appendiceal diameter among the most influential predictors.

Conclusions: PRIMARCH-APP AI demonstrated high discrimination and clinically interpretable prediction of complicated acute appendicitis using routinely available preoperative data. External and prospective multicenter validation is required before clinical implementation.

Keywords: acute appendicitis, complicated appendicitis, machine learning, artificial intelligence, explainable artificial intelligence, SHAP, risk prediction

Introduction

Acute appendicitis remains one of the most frequent causes of acute abdominal pain requiring emergency surgical assessment and continues to represent a substantial component of the global emergency surgery burden (1,2). Although appendectomy is among the most standardized procedures in general surgery, the clinical spectrum of acute appendicitis is heterogeneous, ranging from uncomplicated inflammation confined to the appendix to gangrene, perforation, periappendiceal abscess, or diffuse peritonitis. Contemporary guidelines increasingly recognize uncomplicated and complicated acute appendicitis as clinically distinct entities, since this distinction may influence therapeutic strategy, urgency of intervention, antimicrobial management, and expected postoperative course (3,4).

Accurate preoperative identification of complicated acute appendicitis therefore remains clinically relevant. Clinical examination and conventional laboratory parameters provide important diagnostic information but may be insufficient when interpreted independently. Several scoring systems, including the Alvarado score, Appendicitis Inflammatory Response (AIR) score, and Adult Appendicitis Score (AAS), have been developed to improve diagnostic stratification (5-9). Nevertheless, these scores were primarily designed to estimate the probability of acute appendicitis rather than specifically predict complicated disease. Dedicated models combining clinical and radiological variables have subsequently demonstrated improved discrimination between uncomplicated and complicated appendicitis, although their performance and generalizability remain variable (10-12).

Cross-sectional imaging, particularly contrast-enhanced computed tomography (CT), has further improved the characterization of appendiceal inflammation. Findings such as appendiceal wall defects, periappendiceal fluid, extraluminal gas, abscess formation, marked periappendiceal inflammatory changes, and appendicoliths have been associated with complicated disease (13-19). Similarly, systemic inflammatory markers may provide additional information regarding disease severity. Beyond conventional leukocyte count and C-reactive protein (CRP), derived indices such as the neutrophil-to-lymphocyte ratio and other composite inflammatory biomarkers have been investigated as potential predictors of complicated appendicitis (20-26). However, the multidimensional relationships between demographic characteristics, clinical presentation, laboratory inflammation, estab-

lished clinical scores, and imaging findings may be difficult to capture using conventional threshold-based approaches.

Machine learning (ML) provides a potential framework for integrating these heterogeneous variables and identifying nonlinear interactions that may not be readily apparent using traditional statistical methods. Recent studies have applied ML algorithms to the diagnosis or severity stratification of acute appendicitis, including models specifically developed to predict perforation or complicated disease (27-31). Nevertheless, predictive performance alone is insufficient for meaningful clinical translation. In surgical decision-making, clinicians must also understand which variables drive individual predictions, whether predicted probabilities are appropriately calibrated, and whether the model provides clinically useful information beyond existing assessment strategies. Explainable artificial intelligence (XAI), particularly SHapley Additive exPlanations (SHAP), offers a method for quantifying the contribution of individual predictors at both population and patient levels, potentially improving transparency and clinical interpretability of complex ML models (32,37).

Against this background, the PRIMARCH-APP AI project was developed as an explainable, clinically oriented prediction framework using information available during the preoperative assessment of patients with acute appendicitis. The present study aimed to develop and internally validate an explainable machine learning model for the preoperative prediction of complicated acute appendicitis, integrating demographic, clinical, laboratory, inflammatory, clinical-score, and imaging variables. In addition to discrimination, model evaluation was designed to assess calibration and interpretability, with the objective of identifying the variables most strongly influencing predictions and establishing a reproducible framework for subsequent external and prospective validation. The study and model-development process were structured according to contemporary recommendations for transparent reporting of multivariable prediction models incorporating machine-learning methods (32-36).

Material and Methods

Study Design and Setting

This prospective observational cohort study was conducted in the First Department of General Surgery, Emergency County Clinical Hospital Târgu Mures, Romania. Consecutive adult patients presenting with acute appendicitis and undergoing emergency appendectomy during the predefined study period were prospectively screened for eligibility and enrolled according to the study protocol.

Clinical, laboratory, imaging, operative, and post-operative variables were prospectively collected using a standardized data collection protocol and entered into the dedicated anonymized PRIMARCH-APP AI database. The prediction component of the study was specifically designed to reproduce the information available to the surgeon during preoperative assessment. Consequently, only variables available before operative intervention were eligible for inclusion as candidate predictors, whereas intraoperative and postoperative variables were retained exclusively for outcome determination and secondary descriptive analyses.

The study was designed as a prospective cohort study for the development and internal validation of a clinical prediction model based on machine-learning methods. Model development and reporting followed the principles of the TRIPOD+AI statement and contemporary methodological recommendations for clinical prediction models, including appropriate assessment of model discrimination, calibration, internal validation, and risk of overfitting (32-36).

Study Population

Consecutive adult patients evaluated and treated for acute appendicitis during the study period were considered for inclusion. Patients were eligible if they were aged ≥18 years, had a clinical and/or imaging diagnosis of acute appendicitis, underwent emergency appendectomy during the index hospitalization, and had the predefined preoperative clinical and laboratory assessment available for analysis.

Patients were excluded when the final diagnosis did not confirm acute appendicitis, when appendectomy was performed as an interval or elective procedure following previous conservative treatment, when appendiceal neoplasia represented the principal pathological process, or when insufficient information was available to establish the primary study outcome. Patients with insufficient preoperative information for prediction-model development were also excluded from the final analytical cohort (*Fig. 1*).

Definition of Complicated Acute Appendicitis

The primary outcome was the presence of complicated acute appendicitis, analyzed as a binary endpoint: uncomplicated versus complicated acute appendicitis.

The final classification was established using operative findings documented during emergency appendectomy. Complicated acute appendicitis was

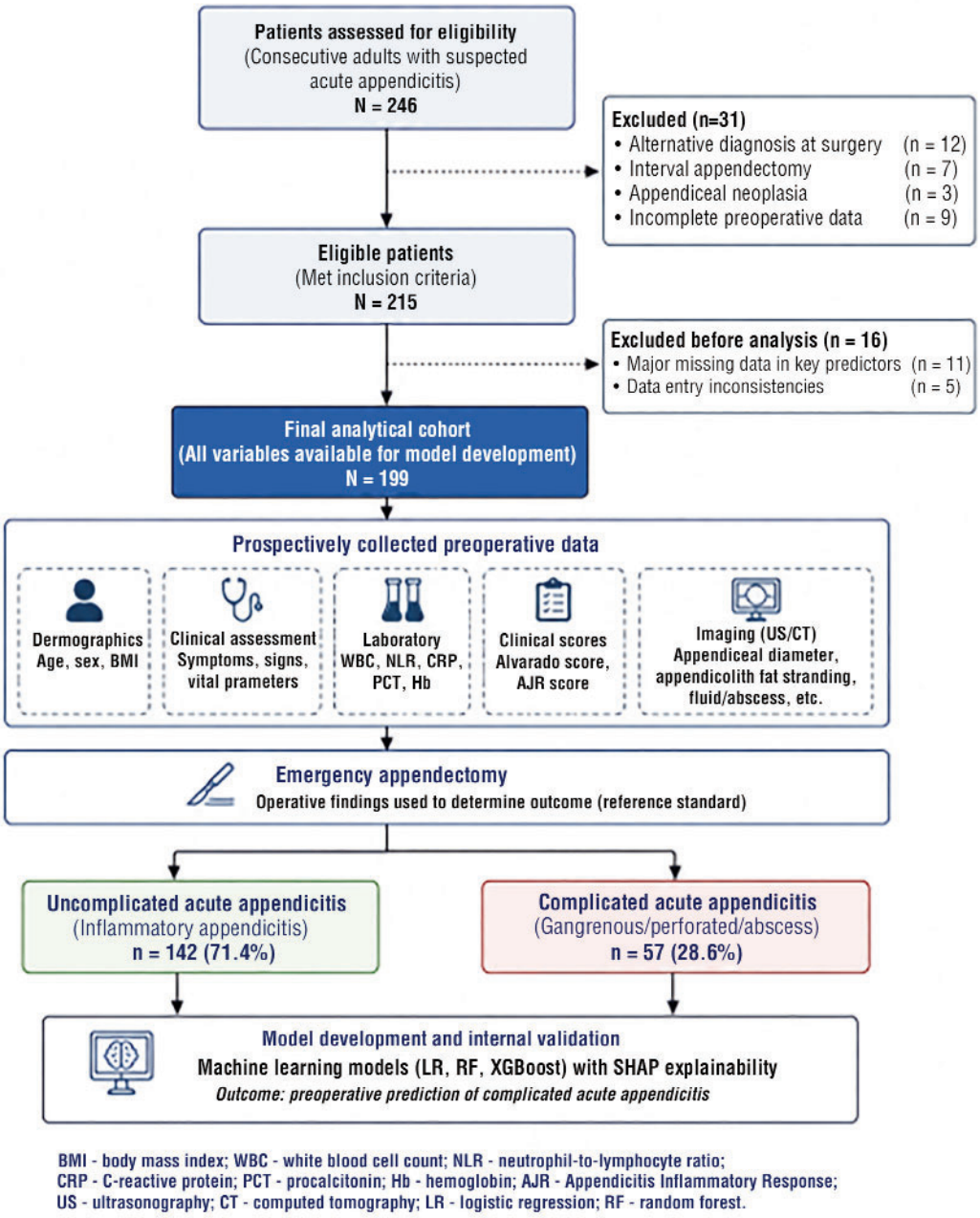


Figure 1. PRIMARCH-APP AI study flow diagram: patient selection, prospective data collection, and cohort formation.

Flow diagram illustrating prospective patient enrollment, eligibility assessment, cohort formation, and outcome classification. Preoperative demographic, clinical, laboratory, scoring, and imaging variables were collected before emergency appendectomy. Operative findings served as the reference standard for classification into uncomplicated or complicated acute appendicitis. The final analytical cohort was subsequently used for machine-learning model development, internal validation, and SHAP-based explainability analysis. SHAP, SHapley Additive exPlanations.

defined by the presence of advanced inflammatory disease, including gangrenous appendicitis, appendiceal perforation, or periappendiceal abscess, whereas acute appendicitis without these findings was categorized as uncomplicated. The classification was established according to predefined criteria consistent with contemporary concepts and WSES recommendations regarding uncomplicated and complicated acute appendicitis (3,4).

The outcome was determined only after surgical exploration and was therefore unavailable to the prediction model at the intended time of application.

Intraoperative findings were used to establish the reference outcome but were not included among candidate predictors.

Data Collection and Candidate Predictors

Data were prospectively collected according to a predefined standardized protocol. Candidate predictors were selected before model development on the basis of their clinical relevance, availability during routine emergency assessment, and previously reported association with the diagnosis or severity of acute appendicitis.

The dataset incorporated demographic characteristics, clinical presentation, physical examination findings, laboratory inflammatory parameters, established appendicitis scores, and preoperative imaging findings. All variables intended for prediction were required to be available before operative intervention.

Demographic variables included age, sex, and body mass index. Clinical variables included duration of symptoms before admission, migration of pain to the right lower quadrant, right lower quadrant tenderness, rebound tenderness, generalized peritonitis, fever, heart rate, and nausea or vomiting.

Established clinical scoring systems included the Alvarado score and the Appendicitis Inflammatory Response (AIR) score. These scores were recorded prospectively and were considered potential predictors because they integrate multiple components of the clinical and inflammatory presentation of acute appendicitis (5-9).

Laboratory parameters included white blood cell count, neutrophil percentage, lymphocyte percentage, C-reactive protein, procalcitonin, and hemoglobin concentration. The neutrophil-to-lymphocyte ratio (NLR) was derived from differential leukocyte counts and evaluated as an additional inflammatory predictor because of its previously demonstrated association with complicated appendicitis (20,21).

Preoperative Imaging Assessment

Preoperative imaging was performed according to routine clinical indications and included abdominal ultrasonography and/or computed tomography (CT). Imaging characteristics were prospectively recorded in the study database when available.

The predefined imaging variables included maximum appendiceal diameter, appendiceal wall defect or discontinuity, presence of an appendicolith, periappendiceal fluid, abscess formation, phlegmon, and the severity of periappendiceal fat stranding. Fat stranding was categorized according to its recorded severity. Radiologist diagnostic confidence was additionally recorded where available.

Imaging characteristics were included because CT findings provide important information for differentiating uncomplicated from complicated appendicitis, particularly when secondary signs of transmural inflammation, perforation, or peri-appendiceal inflammatory extension are present (12-19).

Only imaging information available before surgery was eligible for inclusion in the predictive models.

Prevention of Data Leakage

Strict temporal separation between predictor variables and outcome information was maintained throughout model development. This represented a predefined methodological principle of the PRIMARCH-APP AI framework.

Variables derived from operative or postoperative information were excluded from the predictor matrix. Specifically, operative approach, intraoperative appendiceal appearance, peritoneal contamination, intraoperatively identified appendicolith, operative duration, postoperative complications, Clavien-Dindo classification, intensive care unit admission, reoperation, length of hospital stay, and other postoperative outcomes were not provided to the machine-learning algorithms.

These variables were retained in the prospective database for outcome classification and descriptive analyses but were not used for preoperative prediction. This approach ensured that the performance of the model reflected realistically available information at the intended point of clinical application.

Data Preprocessing and Quality Control

Before model development, the prospectively collected database underwent a predefined quality-control procedure. Patient identifiers were removed before computational analysis. Variables were examined for missing values, inconsistent coding, implausible observations, duplicate entries, and discrepancies in measurement units.

Continuous variables were converted to standardized numerical formats. Binary and categorical variables were encoded according to the predefined PRIMARCH-APP AI data dictionary. Clinical variables recorded with accompanying units or textual descriptors were standardized before analysis.

The extent and pattern of missing data were assessed for each candidate predictor. Variables with substantial missingness or insufficient reliability were considered for exclusion before model training.

Missing values among retained predictors were handled using preprocessing procedures appropriate to the respective variable type. Importantly, preprocessing parameters were derived exclusively from the model-development data and subsequently applied to validation observations, thereby preventing information leakage. Feature engineering was limited to clinically interpretable transformations and pre-defined derived variables. No postoperative information was used to generate engineered features.

Machine-Learning Model Development

The primary modeling objective was the preoperative prediction of complicated acute appendicitis. Three supervised classification algorithms were evaluated: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). Logistic regression was included as an interpretable statistical benchmark, whereas RF and XGBoost were selected to capture potentially nonlinear relationships and interactions among predictors.

Only variables available before operative intervention were included in the predictor matrix. The analytical dataset was divided using a stratified random 80:20 training/test split, preserving the distribution of complicated and uncomplicated appendicitis. A fixed random state of 42 was used to ensure reproducibility. The held-out test subset remained isolated from model fitting and was used exclusively for final internal performance assessment.

Missing continuous variables were imputed using the median value calculated from the development data, whereas missing categorical variables were imputed using the most frequent category. Categorical variables were transformed using one-hot encoding, with previously unseen categories ignored during transformation. Continuous and binary predictors were standardized for logistic regression using standard scaling after imputation. Numerical standardization was not applied to RF or XGBoost.

Model hyperparameters were predefined within the development framework. Logistic regression was fitted using balanced class weights and a maximum of 2,000 iterations. RF was constructed using 500 trees, a minimum leaf size of two observations, square-root feature sampling, and balanced class weights. XGBoost was configured with 500 estimators, a maximum tree depth of four, a learning rate of 0.05, subsampling of 0.80, column subsampling of 0.80, and log-loss as the evaluation metric.

Internal Validation and Model Performance

Model stability within the development cohort was assessed using five-fold stratified cross-validation, with shuffling and a fixed random state of 42. ROC-AUC was used as the primary cross-validation performance measure. Following model development, final predictive performance was evaluated on the independent held-out 20% test subset.

Test-set performance was quantified using the ROC-AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and Brier score. Classification metrics were calculated using a probability threshold of 0.50. The Brier score was used to assess overall probability accuracy, with lower values indicating better agreement between predicted probabilities and observed outcomes.

Explainable Artificial Intelligence Analysis

Model interpretability was evaluated using SHapley Additive exPlanations (SHAP). Following comparison of the candidate algorithms, RF was selected as the principal PRIMARCH-APP AI model based on its overall combination of discrimination and probability accuracy. SHAP values were calculated using a tree-based explainer applied to the final RF classifier. Global predictor importance was quantified as the mean absolute SHAP value across observations in the held-out test subset.

Statistical Analysis

Descriptive statistical analysis was performed before machine-learning model development. Continuous variables were assessed for distribution and summarized as mean ± standard deviation when approximately normally distributed or median and interquartile range when non-normally distributed. Categorical variables were reported as absolute frequencies and percentages.

Baseline characteristics were compared between patients with uncomplicated and complicated acute appendicitis. Continuous variables were compared using the independent-samples t test or Mann-Whitney U test, according to distribution. Categorical variables were analyzed using the χ^2 test or Fisher's exact test, as appropriate. All statistical tests were two-sided, and a p value <0.05 was considered statistically significant. Univariate statistical significance was not used as the sole criterion for predictor inclusion, because potentially informative interactions and nonlinear relationships may be identified by machine-learning algorithms even when conventional univariate associations are limited. The computational analysis

was performed using Python, employing established libraries for data processing, statistical analysis, machine learning, model evaluation, and SHAP-based explainability. The complete computational workflow, including preprocessing procedures, model specifications, hyperparameter settings, internal validation procedures, and analysis code, will be provided as supplementary material/appendices to facilitate transparency and reproducibility.

Ethical Considerations

Patient data were anonymized before computational analysis, and no directly identifiable information was incorporated into the machine-learning dataset. The study protocol and prospective data collection procedure were submitted to the appropriate institutional ethics committee: 4243/16.03.2026.

Results

Study Population and Outcome Distribution

A total of 199 consecutive adult patients prospectively enrolled between 9 January 2024 and 1 May 2026 were included in the current analytical cohort. According to the predefined operative reference standard, 142 patients (71.4%) were classified as having uncomplicated acute appendicitis, whereas 57 patients (28.6%) had complicated acute appendicitis.

The complicated group included patients with gangrenous, perforated, or abscess-associated appendicitis. Within the overall cohort, operative assessment classified 110 patients as inflammatory appendicitis, 46 as gangrenous appendicitis, 26 as perforated appendicitis, and 17 as appendicitis associated with abscess formation.

The mean age of the entire cohort was 43.0 ± 17.6 years. Patients with complicated appendicitis were significantly older than those with uncomplicated disease (53.9 ± 13.4 vs. 38.6 ± 17.2 years; p<0.001). Sex distribution was comparable between groups, with no significant association between sex and complicated disease (p=0.875).

Clinical and Laboratory Characteristics

Several preoperative clinical characteristics differed substantially according to appendicitis severity. Symptom duration was significantly longer among patients with complicated appendicitis, with a median duration of 23 hours (IQR 18-33) compared with 11 hours (IQR 8-18) in the uncomplicated group (p<0.001). Heart rate was also higher in complicated disease (104.8 ± 9.6 vs. 89.4 ± 10.3 beats/min; p<0.001).

Signs indicating more advanced inflammatory involvement were markedly enriched among patients with complicated appendicitis. Rebound tenderness was present in 77.2% (44/57) of complicated cases compared with 14.8% (21/142) of uncomplicated cases (p<0.001). Generalized peritonitis occurred exclusively within the complicated group, affecting 22/57 patients (38.6%) (p<0.001). Fever ≥38°C was observed in 94.7% of complicated compared with 69.7% of uncomplicated cases (p<0.001). Nausea or vomiting was also more frequent in complicated disease (89.5% vs. 71.1%; p=0.005).

Both conventional appendicitis scores demonstrated higher values in patients with complicated disease. The median Alvarado score was 8 (IQR 8-9) in the complicated group compared with 7 (IQR 7-8) in uncomplicated appendicitis (p<0.001). Similarly, the median AIR score was 8 (IQR 7-8) versus 7 (IQR 6-7), respectively (p<0.001).

Inflammatory laboratory parameters also differed between groups, although their separation was less pronounced than that observed for several clinical variables. Median WBC count was 16.3 × 10^3/μL in complicated versus 14.1 × 10^3/μL in uncomplicated appendicitis (p=0.013). Median NLR was 10.71 and 7.59, respectively (p=0.017) (*Table 1*).

Among imaging findings, complicated appendicitis was particularly associated with increased appendiceal diameter, appendicolith, periappendiceal fluid, abscess formation, and severe periappendiceal fat stranding. These differences demonstrated clinically meaningful separation between uncomplicated and complicated disease across multiple preoperative domains.

Preoperative Imaging Characteristics

Computed tomography represented the predominant imaging modality, being performed in 172/199 patients (86.4%). Imaging findings demonstrated important differences between uncomplicated and complicated disease. The median appendiceal diameter was 13 mm (IQR 10-15) among patients with complicated appendicitis compared with 9 mm (IQR 7-12) in uncomplicated cases (p<0.001).

An appendicolith was identified preoperatively in 34/57 patients (59.6%) with complicated appendicitis compared with 41/142 (28.9%) with uncomplicated disease (p<0.001). Periappendiceal fluid was present in 82.5% (47/57) of complicated cases compared with 45.1% (64/142) of uncomplicated cases (p<0.001). Abscess formation represented one of the imaging characteristics most strongly associated with complicated disease, occurring in 22/57 patients (38.6%)

Table 1. Baseline clinical, laboratory, and imaging characteristics according to appendicitis severity

Variable	Overall (n=199)	Uncomplicated (n=142)	Complicated (n=57)	p-value
Demographic characteristics				
Age, years	43.0 ± 17.6	38.6 ± 17.2	53.9 ± 13.4	<0.001
Male sex, n (%)	109 (54.8%)	77 (54.2%)	32 (56.1%)	0.806
Clinical presentation				
Symptom duration, h	14 (9–22)	11 (8–18)	23 (18–33)	<0.001
Heart rate, beats/min	93.8 ± 12.2	89.4 ± 10.3	104.8 ± 9.6	<0.001
Rebound tenderness, n (%)	65 (32.7%)	21 (14.8%)	44 (77.2%)	<0.001
Generalized peritonitis, n (%)	22 (11.1%)	0 (0.0%)	22 (38.6%)	<0.001
Fever ≥ 38˚C, n (%)	153 (76.9%)	99 (69.7%)	54 (94.7%)	<0.001
Nausea/vomiting, n (%)	152 (76.4%)	101 (71.1%)	51 (89.5%)	0.006
Clinical scores				
Alvarado score	8 (7–8)	7 (7–8)	8 (8–9)	<0.001
AIR score*	7 (6–8)	7 (6–7)	8 (7–8)	<0.001
Laboratory parameters				
WBC, X10^3/μL	15.0 (12.1–17.2)	14.1 (11.9–16.8)	16.3 (13.0–18.4)	0.013
NLR	8.1 (4.6–13.3)	7.6 (4.3–11.9)	10.7 (6.0–15.1)	0.017
CRP, mg/L†	25.6 (5.0–145.5)	10.2 (2.7–43.2)	216.4 (72.6–248.8)	<0.001
Preoperative imaging characteristics				
Appendix diameter, mm	10 (8–13)	9 (7–12)	13 (10–15)	<0.001
Appendicolith, n (%)	75 (37.7%)	41 (28.9%)	34 (59.6%)	<0.001
Periappendiceal fluid, n (%)	111 (55.8%)	64 (45.1%)	47 (82.5%)	<0.001
Abscess, n (%)	25 (12.6%)	3 (2.1%)	22 (38.6%)	<0.001
Severe fat stranding, n (%)	22 (11.1%)	3 (2.1%)	19 (33.3%)	<0.001

compared with only 3/142 patients (2.1%) with uncomplicated appendicitis (p<0.001). Severe periappendiceal fat stranding was likewise substantially more frequent in complicated disease (33.3% vs. 2.1%).

These findings confirmed that the prospective cohort contained clinically meaningful separation between uncomplicated and complicated appendicitis across demographic, clinical, inflammatory, scoring, and imaging domains.

Machine-Learning Model Performance

The prospectively collected preoperative variables were subsequently used to develop and internally evaluate the PRIMARCH-APP AI prediction framework. Logistic regression, random forest, and XGBoost classifiers were compared using an identical preprocessing and validation strategy.

During five-fold cross-validation within the development cohort, all three approaches demonstrated high discrimination. Mean ROC-AUC was 0.935 for logistic regression, 0.945 for random forest, and 0.936 for XGBoost. Performance remained high when the models were evaluated on the stratified held-out internal test set. Logistic regression achieved a ROC-AUC of 0.897, whereas both random forest and XGBoost achieved a ROC-AUC of 0.934. Among the evaluated models, the random forest classifier demonstrated the most favorable overall balance between discrimination, classification performance, and probability accuracy. On the held-out test cohort, random forest achieved an accuracy of 87.5%, sensitivity of 72.7%, specificity of 93.1%, positive predictive value of 80.0%, negative predictive value of 90.0%, and F1-score of 0.762. Its Brier score was 0.091, lower than that observed for logistic regression (0.114) and XGBoost (0.113). Logistic regression demonstrated a higher sensitivity of 81.8%, with specificity of 89.7%, accuracy of 87.5%, and F1-score of 0.783. XGBoost demonstrated specificity of 93.1%, but sensitivity was lower at 63.6%, with an overall accuracy of 85.0% (*Table 2*).

Explainability Analysis

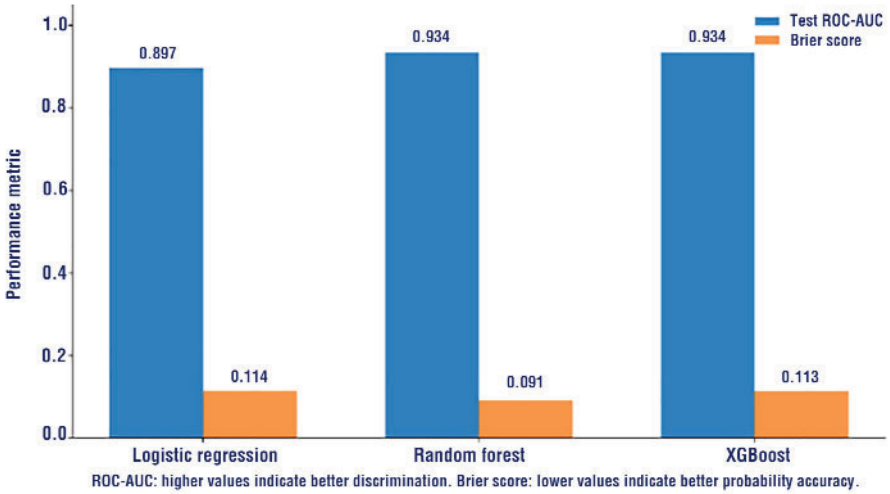
Explainability analysis was performed for the selected random forest model using SHapley Additive exPlanations (SHAP). The objective was to identify the preoperative variables contributing most strongly to predictions of complicated appendicitis and to determine the direction and magnitude of their influence at both cohort and individual-patient levels. The descriptive analyses already indicated that older age, longer symptom duration, higher heart rate, higher Alvarado and AIR scores, rebound tenderness, generalized peritonitis, increased appendiceal diameter, appendicolith, periappendiceal fluid, abscess formation, and severe fat stranding represented prominent candidate determinants of complicated disease (*Fig. 3, Table 3*).

Table 2. Comparative performance of machine-learning models for preoperative prediction of complicated acute appendicitis

Model	CV ROC-AUC	Test ROC-AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	F1-score	Brier score
Logistic regression	0.935	0.897	81.8	89.7	87.5	75.0	92.9	0.783	0.114
Random forest	0.945	0.934	72.7	93.1	87.5	80.0	90.0	0.762	0.091
XGBoost	0.936	0.934	63.6	93.1	85.0	77.8	87.1	0.700	0.113

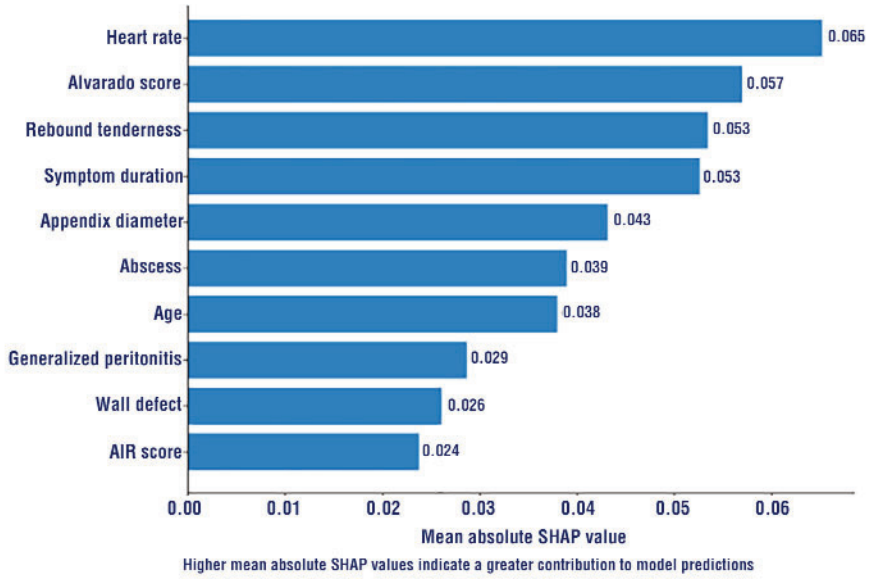
Model performance was assessed using five-fold cross-validation (CV) within the development cohort and subsequently evaluated in the held-out internal test set. Values in bold indicate the best performance for each metric. CV, cross-validation; NPV, negative predictive value; PPV, positive predictive value; ROC-AUC, area under the receiver operating characteristic curve; XGBoost, extreme gradient boosting.

Figure 2. Discrimination and probability accuracy of the PRIMARCH-APP AI machine-learning models



Comparative performance of the PRIMARCH-APP AI prediction models. Discrimination and probability accuracy of logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost) evaluated in the held-out internal test cohort. RF and XGBoost achieved the highest ROC-AUC (0.934), whereas RF demonstrated the lowest Brier score (0.091), indicating the lowest overall probabilistic prediction error. Higher ROC-AUC values indicate better discrimination, whereas lower Brier scores indicate better probability accuracy. LR, logistic regression; RF, random forest; ROC-AUC, area under the receiver operating characteristic curve; XGBoost, extreme gradient boosting.

Figure 3. Global explainability of the PRIMARCH-APP AI model using SHapley additive exPlanations (SHAP)



Global feature importance for the selected random forest model is expressed as the mean absolute SHAP value across observations in the held-out internal test subset. Higher mean absolute SHAP values indicate a greater overall contribution of the respective preoperative variable to model predictions. Heart rate, Alvarado score, rebound tenderness, symptom duration, and appendiceal diameter were the most influential predictors, followed by abscess formation, age, generalized peritonitis, appendiceal wall defect, and AIR score. AIR, Appendicitis Inflammatory Response; SHAP, SHapley Additive exPlanations.

Table 3. Principal preoperative predictors contributing to the prediction of complicated acute appendicitis in the PRIMARCH-APP AI framework

Predictor	Predictor domain	Association with complicated appendicitis	Clinical interpretation
Age	Demographic	Higher age ↑	Increasing age was associated with a greater probability of complicated disease
Symptom duration	Clinical	Longer duration ↑	Prolonged evolution before presentation was associated with progression toward complicated appendicitis
Rebound tenderness	Clinical	Presence ↑	Signs of localized peritoneal irritation were associated with increased probability of complicated disease
Generalized peritonitis	Clinical	Presence ↑	Diffuse peritoneal irritation indicated advanced local inflammatory disease
CRP	Laboratory	Higher values ↑	Greater systemic inflammatory response was associated with complicated appendicitis
NLR	Laboratory	Higher values ↑	Increased neutrophil-predominant systemic inflammation was associated with greater disease severity
Appendix diameter	Imaging	Larger diameter ↑	Greater appendiceal dilatation was associated with advanced appendiceal inflammation
Appendicolith	Imaging	Presence ↑	Appendicolith was associated with an increased probability of complicated appendicitis
Periappendiceal fluid	Imaging	Presence ↑	Local inflammatory fluid was associated with extension of the inflammatory process
Severe fat stranding	Imaging	Presence/severity ↑	Extensive periappendiceal inflammatory changes were associated with complicated disease

Principal preoperative predictors contributing to the prediction of complicated acute appendicitis in the PRIMARCH-APP AI framework. Candidate predictors comprise demographic, clinical, laboratory, and imaging variables available before surgical intervention. The direction of association indicates whether increasing values or the presence of the respective feature was associated with a higher probability of complicated acute appendicitis. Final model-specific feature importance and individual predictor contributions were assessed using SHapley Additive exPlanations (SHAP). CRP, C-reactive protein; NLR, neutrophil-to-lymphocyte ratio; SHAP, SHapley Additive exPlanations.

Discussion

The present prospective study describes the development and internal validation of PRIMARCH-APP AI, an explainable machine-learning framework designed to predict complicated acute appendicitis using information available before surgery. Logistic regression, random forest, and XGBoost all demonstrated high discriminatory performance. Random forest and XGBoost achieved a ROC-AUC of 0.934 in the held-out internal test cohort, while random forest provided the most favorable overall balance between discrimination and probability accuracy, with a specificity of 93.1%, accuracy of 87.5%, and Brier score of 0.091. These results suggest that routinely available preoperative variables can provide substantial information regarding appendicitis severity when integrated within a multivariable ML framework.

The distinction between uncomplicated and complicated acute appendicitis has become increasingly relevant because disease severity may influence operative timing, antimicrobial treatment, resource allocation, and expected postoperative evolution (1-4). Conventional clinical scores, including the Alvarado and AIR scores, provide useful diagnostic stratification but were not primarily developed to estimate the individual probability of complicated disease (5-9). Previous models combining clinical and radiological parameters have improved this differentiation, supporting the concept that appendicitis severity is better characterized through integration of multiple information domains rather than isolated predictors (10-12).

Imaging findings were particularly relevant in the present cohort. Complicated appendicitis was associated with increased appendiceal diameter and higher frequencies of appendicolith, periappendiceal fluid, abscess formation, and severe fat stranding. These observations are consistent with previous studies demonstrating that structural appendiceal abnormalities and secondary inflammatory CT signs contribute substantially to differentiation between uncomplicated and complicated disease (12-19). Laboratory markers, including NLR and CRP, provide complementary information regarding systemic inflammatory burden and have also been associated with appendicitis severity (20-26). An advantage of ML is that these variables can be interpreted simultaneously with demographic characteristics, clinical findings, established scores, and imaging features rather than through isolated diagnostic thresholds.

Recent studies have demonstrated the feasibility of applying ML to the prediction of perforated or complicated appendicitis (27-31). Our findings support this emerging evidence while emphasizing that algorithmic complexity does not necessarily guarantee superior clinical performance. Logistic regression achieved a ROC-AUC of 0.897 and the highest sensitivity among the evaluated models. Random forest was selected as the principal PRIMARCH-APP AI model because it combined high discrimination with greater specificity and the lowest Brier score. This comparison is important because clinically oriented AI models should demonstrate meaningful advantages beyond simply applying more complex computational techniques.

Explainability represents an important component

of the proposed framework. SHAP analysis identified heart rate, Alvarado score, rebound tenderness, symptom duration, and appendiceal diameter as the most influential predictors in the current model, followed by abscess formation, age, generalized peritonitis, appendiceal wall defect, and AIR score. These findings are clinically coherent and reflect three major dimensions of complicated disease: systemic inflammatory response, progression over time, and local anatomical severity. SHAP-based interpretation therefore provides more than a predicted probability by demonstrating which characteristics contribute to an individual patient's estimated risk (37).

The intended role of PRIMARCH-APP AI is not to replace surgical judgment or radiological assessment but to provide an additional tool for preoperative risk stratification. Identification of patients with a high probability of complicated appendicitis could potentially support emergency department prioritization, timing of operative intervention, antimicrobial planning, operating-room allocation, and perioperative counseling. Nevertheless, the sensitivity of the selected random forest model indicates that it should not be considered an autonomous rule-out system, and classification thresholds may require adjustment according to the intended clinical application.

Several methodological strengths should be emphasized. The cohort was prospectively assembled using a standardized data-collection framework, and predictors were restricted to information available before surgery, minimizing the risk of target leakage. The study compared ML algorithms against an interpretable logistic-regression benchmark and assessed not only discrimination but also probability accuracy and model explainability, consistent with contemporary recommendations for clinical prediction-model development (32-36).

The present study has several important limitations. First, its single-center design, moderate sample size (n=199), and relatively limited number of complicated appendicitis events (n=57) increase the risk of model overfitting and restrict the generalizability of the observed performance estimates. Although stratified cross-validation and a held-out internal test set were used to reduce optimistic performance estimation, these procedures cannot substitute for validation in an independent population. Consequently, the reported discriminatory performance should be interpreted as preliminary and specific to the current development environment.

Second, some candidate variables were not uniformly available across the entire cohort and required predefined missing-data handling procedures, which may have influenced model estimates. Third, preoperative imaging was performed according to routine clinical practice rather than a centralized study-specific imaging protocol. Consequently, imaging-derived predictors may be affected by interobserver variability related to radiological interpretation, particularly for findings such as periappendiceal fat stranding, appendiceal wall abnormalities, and radiologist confidence. Interobserver agreement was not formally quantified in the present study and should be specifically assessed in future external and multicenter validation studies.

Conclusion

PRIMARCH-APP AI demonstrated promising internally validated performance for the preoperative prediction of complicated acute appendicitis using routinely available clinical, laboratory, and imaging variables. The explainable machine-learning framework provided clinically interpretable risk estimates; however, the present findings should be considered preliminary and should not be interpreted as evidence of clinical utility or generalizability. External and temporal validation in independent populations, followed by prospective multicenter evaluation, is required before clinical implementation.

Conflict of Interest

The authors declare no conflicts of interest.

Artificial Intelligence Use Statement

Artificial intelligence tools were used during manuscript preparation solely for language refinement, editorial assistance, and improvement of textual clarity. All scientific content, study design, data collection, statistical and machine-learning analyses, interpretation of results, and final conclusions were reviewed and validated by the authors. The authors take full responsibility for the content of the manuscript.

References

1. Bhangu A, Søreide K, Di Saverio S, Assarsson JH, Drake FT. Acute appendicitis: modern understanding of pathogenesis, diagnosis, and management. Lancet. 2015;386(10000):1278-87.
2. Ferris M, Quan S, Kaplan BS, Molodecky N, Ball CG, Chernoff GW, et al. The global incidence of appendicitis: a systematic review of population-based studies. Ann Surg. 2017;266(2):237-41.
3. Di Saverio S, Podda M, De Simone B, Ceresoli M, Augustin G, Gori A, et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. World J Emerg Surg. 2020;15(1):27.
4. Podda M, Ceresoli M, De Simone B, Fugazzola P, Pata F, Balla A, et al. Diagnosis and

treatment of acute appendicitis: 2025 edition of the WSES Jerusalem guidelines. JAMA Surg. 2026;161(3):283-295.
5. Andersson M, Andersson RE. The Appendicitis Inflammatory Response Score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score. World J Surg. 2008;32(8):1843-9.
6. de Castro SMM, Ünlü Ç, Steller EP, van Wagensveld BA, Vrouenraets BC. Evaluation of the Appendicitis Inflammatory Response Score for patients with acute appendicitis. World J Surg. 2012;36(7):1540-5.
7. Sammalkorpi HE, Mentula P, Leppäniemi A. A new adult appendicitis score improves diagnostic accuracy of acute appendicitis - a prospective study. BMC Gastroenterol. 2014;14:114.
8. Kollár D, McCartan DP, Bourke M, Cross KS, Dowdall J. Predicting acute appendicitis? A comparison of the Alvarado score, the Appendicitis Inflammatory Response score and clinical assessment. World J Surg. 2015;39(1):104-9.
9. Andersson M, Kolodziej B, Andersson RE, STRAPPSCORE Study Group. Randomized clinical trial of Appendicitis Inflammatory Response score-based management of patients with suspected appendicitis. Br J Surg. 2017;104(11):1451-61.
10. Atema JJ, van Rossem CC, Leeuwenburgh MM, Stoker J, Boermeester MA. Scoring system to distinguish uncomplicated from complicated acute appendicitis. Br J Surg. 2015;102(8):979-90.
11. Imaoka Y, Itamoto T, Takakura Y, Suzuki T, Ikeda S, Urushihara T. Validity of predictive factors of acute complicated appendicitis. World J Emerg Surg. 2016;11:48.
12. Avanesov M, Wiese NJ, Karul M, Guerreiro H, Keller S, Busch P, et al. Diagnostic prediction of complicated appendicitis by combined clinical and radiological appendicitis severity index (APSI). Eur Radiol. 2018;28(9):3601-10.
13. Kim HY, Park JH, Lee YJ, Lee SS, Jeon JJ, Lee KH. Systematic review and meta-analysis of CT features for differentiating complicated and uncomplicated appendicitis. Radiology. 2018;287(1):104-15.
14. Bom WJ, Scheijmans JCG, Salminen P, Boermeester MA. Discriminating complicated from uncomplicated appendicitis by ultrasound imaging, computed tomography or magnetic resonance imaging: systematic review and meta-analysis of diagnostic accuracy. BJS Open. 2021;5(2):zraa030.
15. Iamwat J, Teerasamit W, Apisarnthanarak P, Noppakunsomboon N, Kaewlai R. Predictive ability of CT findings in the differentiation of complicated and uncomplicated appendicitis: a retrospective investigation of 201 patients undergoing appendectomy at initial admission. Insights Imaging. 2021;12(1):143.
16. Lin HA, Tsai HW, Chao CC, Lin SF. Periappendiceal fat-stranding models for discriminating between complicated and uncomplicated acute appendicitis: a diagnostic and validation study. World J Emerg Surg. 2021;16:52.
17. Kaewlai R, et al. Validation of scoring systems for the prediction of complicated appendicitis in adults using clinical and computed tomographic findings. Insights Imaging. 2023;14:191.
18. Mällinen J, Vaarala S, Mäkinen M, Lietzén E, Grönroos J, Ohtonen P, et al. Appendicolith appendicitis is clinically complicated acute appendicitis - is it histopathologically different from uncomplicated acute appendicitis? Int J Colorectal Dis. 2019;34:1393-400.
19. Nikkolo C, Muuli M, Kirsimägi Ü, Lepner U. Appendicolith as a sign of complicated appendicitis: a myth or reality? A retrospective study. Eur Surg Res. 2025;66(1):1-8.
20. Hajibandeh S, Hajibandeh S, Hobbs N, Mansour M. Neutrophil-to-lymphocyte ratio predicts acute appendicitis and distinguishes between complicated and uncomplicated appendicitis: a systematic review and meta-analysis. Am J Surg. 2020;219(1):154-63.
21. Khan A, Riaz M, Kelly ME, Khan W, Waldron R, Barry K, et al. Prospective validation of neutrophil-to-lymphocyte ratio as a diagnostic and management adjunct in acute appendicitis. Ir J Med Sci. 2018;187(2):379-84.
22. Hou J, Feng W, Liu W, Hou J, Die X, Sun J, et al. The use of the ratio of C-reactive protein to albumin for the diagnosis of complicated appendicitis in children. Am J Emerg Med. 2022;52:148-54.
23. Saridas A, Vural N, Duyan M, Guven HC, Ertas E, Cander B. Comparison of the ability of newly inflammatory markers to predict complicated appendicitis. Open Med (Wars). 2024;19(1):20241002.
24. Feier CVI, Motoc A, Muntean C, Vonica RC, Gaborean V, Olariu S, et al. Systemic inflammatory indices and age-dependent severity in acute appendicitis: a retrospective cohort study. Front Immunol. 2025;16:1620459.
25. Xu T, Zhang Q, Zhao H, Meng Y, Wang F, Li Y, et al. A risk score system for predicting complicated appendicitis and aid decision-making for antibiotic therapy in acute appendicitis. Ann Palliat Med. 2021;10(6):6133-6144.
26. Zarog M, O'Leary P, Kiernan M, Bolger J, Tibbitts P, Coffey S, et al. Circulating fibrocyte percentage and neutrophil-lymphocyte ratio are accurate biomarkers of uncomplicated and complicated appendicitis: a prospective cohort study. Int J Surg. 2023;109:343-51.
27. Akbulut S, Yagin FH, Cicek IB, Koc C, Colak C, Yilmaz S. Prediction of perforated and nonperforated acute appendicitis using machine learning-based explainable artificial intelligence. Diagnostics (Basel). 2023;13(6):1173.
28. Byun J, Park S, Hwang SM. Diagnostic algorithm based on machine learning to predict complicated appendicitis in children using CT, laboratory, and clinical features. Diagnostics (Basel). 2023;13(5):923.
29. Males I, Boban Z, Kumric M, Vrdoljak J, Berkovic K, Pogorelic Z, et al. Applying an explainable machine learning model might reduce the number of negative appendectomies in pediatric patients with a high probability of acute appendicitis. Sci Rep. 2024;14:12772.
30. Wei W, Tongping S, Jiaming W. Construction of a clinical prediction model for complicated appendicitis based on machine learning techniques. Sci Rep. 2024; 14:16473.
31. Chen S, Xia J, Xu B, Huang Y, Teng M, Pan J. Risk prediction and effect evaluation of complicated appendicitis based on XGBoost modeling. BMC Gastroenterol. 2025; 25(1):295.
32. 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.
33. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8.
34. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
35. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part II-binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96.
36. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74.
37. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.
38. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
39. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2016. p. 785-94.
40. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019; 380(14):1347-58.