

A hybrid machine learning approach to improve the diagnostic accuracy of acute appendicitis

✉ Betül Keskinliç Yağız,¹ ✉ Yasemin Keskin,² ✉ Metin Yalaza,³ ✉ Şiyar Ersöz⁴

¹Department of General Surgery, Samsun Gazi State Hospital, Samsun-Türkiye

²Department of General Surgery, University of Health Sciences, Dr. Abdurrahman Yurtaslan Oncology T&R Hospital, Ankara-Türkiye

³Department of Surgical Oncology, University of Health Sciences, Dr. Abdurrahman Yurtaslan Oncology T&R Hospital, Ankara-Türkiye

⁴Department of General Surgery, Ankara University Faculty of Medicine, Ankara-Türkiye

ABSTRACT

BACKGROUND: Acute appendicitis is the most common diagnosis considered in patients presenting to the emergency department with right lower quadrant pain. However, atypical presentations often lead to unnecessary surgeries and increased healthcare costs. This study aimed to improve diagnostic accuracy in acute appendicitis using a hybrid machine learning (ML) model.

METHODS: A retrospective analysis was performed on 395 patients who underwent appendectomy for suspected acute appendicitis between 2020 and 2024 at Ankara University Faculty of Medicine, Department of General Surgery. Demographic, clinical, laboratory, and radiological variables were collected. ML algorithms, including NaiveBayes, MultilayerPerceptron, IBk, AdaBoost, RandomForest, and a hybrid model combining NaiveBayes, AdaBoost, and RandomForest, were applied. The dataset was evaluated using 10-fold cross-validation, repeated 1,000 times. Accuracy, F-measure, Matthews Correlation Coefficient (MCC), receiver operating characteristic (ROC) area, and precision-recall curve (PRC) area were used as performance criteria.

RESULTS: Among the 395 patients, 52.9% were male, with a mean age of 37.3 ± 15.6 years. Histopathological examination confirmed acute appendicitis in 341 (86.3%) patients and negative appendectomy in 54 (13.7%) patients. The diagnostic accuracy of the Alvarado score at a cut-off value of ≥ 6 was 79.0%. Among the ML algorithms, the hybrid model achieved the best performance, with 92.9% accuracy, 93% F-measure, 70.4% MCC, 90.8% ROC area, and 93.4% PRC area. This model correctly predicted 95.6% of acute appendicitis cases and 75.9% of negative appendectomy cases.

CONCLUSION: The hybrid ML model demonstrated superior diagnostic accuracy compared to the Alvarado score for acute appendicitis. Integration of such models into clinical practice could reduce negative appendectomy rates and enhance patient management by enabling faster and more reliable diagnosis.

Keywords: Acute appendicitis; hybrid model; diagnostic accuracy; negative appendectomy; machine learning.

INTRODUCTION

Acute appendicitis is the most common diagnosis considered in patients presenting to the emergency department with right iliac fossa pain. Studies show that approximately 7% of people will experience this condition at some point in their lives.^[1] Although proper diagnosis is usually straightforward

in patients presenting with classical symptoms such as migratory abdominal pain and nausea, atypical presentations often lead to unnecessary tests and delays in treatment. The challenge of making a proper and timely diagnosis remains a significant risk for patients, leading to unnecessary surgical operations, prolonged hospital stays, and increased medical costs.^[2] Globally, the rate of negative acute appendicitis was

Cite this article as: Keskinliç Yağız B, Keskin Y, Yalaza M, Ersöz Ş. A hybrid machine learning approach to improve the diagnostic accuracy of acute appendicitis. *Ulus Travma Acil Cerrahi Derg* 2026;32:128-136.

Address for correspondence: Şiyar Ersöz

Department of General Surgery, Ankara University Faculty of Medicine, Ankara, Türkiye

E-mail: drsiyar@hotmail.com

Ulus Travma Acil Cerrahi Derg 2026;32(2):128-136 DOI: 10.14744/tjtes.2025.56275

Submitted: 24.09.2025 Revised: 16.11.2025 Accepted: 02.12.2025 Published: 09.02.2026

OPEN ACCESS This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



previously reported to be between 15% and 30%, but it has fallen below 10% in recent years due to the increased availability of laparoscopy and advanced radiological tools.^[3-5]

Early recognition and interpretation of symptoms and signs in the emergency department are essential for optimizing the benefits of diagnostic imaging, reducing hospital stay, and preventing unnecessary surgical interventions. In this regard, clinical scoring systems, imaging techniques such as ultrasonography and computed tomography, diagnostic laparoscopy, and computer-aided diagnostic systems have been developed. Some researchers use only laboratory parameters in predictive scoring models, while others combine laboratory results with clinical signs and symptoms.^[6-8] The most commonly used clinical scoring system, the Alvarado score, is sensitive enough to rule out appendicitis but lacks specificity, leading to a significant rate of false-positive cases.^[8,9] The limitations of the Alvarado score and similar models include limited sensitivity in complex cases and a lack of universal applicability across different patient populations. In addition, there may be a lack of correlation between different diagnostic methods and clinical observations. Therefore, the need for a reliable, quick, and user-friendly scoring system to aid in the preoperative diagnosis of appendicitis still remains.

Artificial neural networks (ANN) and machine learning (ML) models have often outperformed the Alvarado scoring system in predicting acute appendicitis.^[10,11] Unlike appendicitis scoring systems based on summing scores for important clinical parameters, ML algorithms can process the complex, nonlinear correlations and interactions between factors. Apart from traditional statistical techniques, ML is a subfield of artificial intelligence aimed at making predictions about new observations by learning from existing data. However, a significant limitation of many ML models is the lack of transparency, interpretability, and explainability. To overcome these shortcomings, hybrid models have recently gained increased attention in clinical research. Our hypothesis is that an ML model can detect acute appendicitis in individuals with right iliac fossa pain with higher accuracy than the Alvarado scoring system.

MATERIALS AND METHODS

Patient Selection

In this study, patients who underwent appendectomy for suspected acute appendicitis at the Department of General Surgery, Ankara University Faculty of Medicine, between 2020 and 2024 were retrospectively evaluated. Ethical approval for the study was obtained from the Ankara University Human Research Ethical Committee (Approval No: İ03-225-25). All methods were conducted in accordance with relevant guidelines and regulations, including institutional ethical standards and the Declaration of Helsinki. As this was a retrospective study based on the analysis of patient data from institutional databases, the requirement for informed consent was waived.

Patients aged > 18 years with suspected acute appendicitis who

underwent emergency surgery were included in the study. Patients under 18 years of age, pregnant patients, those who underwent incidental appendectomy, and those whose histopathological results revealed neuroendocrine tumors or appendiceal mucinous neoplasms were excluded from the study.

Data Collection

Demographic characteristics (age and sex), symptom duration, presence of migratory pain to the right iliac fossa, anorexia, nausea/vomiting, tenderness in the right iliac fossa, rebound pain, elevated temperature (fever), leukocytosis, left shift (>75% neutrophils), Alvarado scores, laboratory parameters (total bilirubin, C-reactive protein [CRP], white blood cell [WBC] count, hemoglobin [Hb], lymphocyte count, platelet [PLT] count, red blood cell distribution width [RDW], neutrophil count, neutrophil percentage, platelet-to-neutrophil ratio [PNR], neutrophil-to-lymphocyte ratio [NLR], white cell nucleated region [WNR]), radiological findings including appendiceal diameter and periappendiceal fat stranding, and histopathological results were retrospectively evaluated. The Alvarado score was calculated based on eight parameters, with a total score ranging from 1 to 10 [7]. Radiological and pathological evaluations in this study were conducted independently by multiple specialists rather than a single observer, in accordance with the methodological design of the study.

Study Design

A total of 395 patients were included, of whom 150 (38%) underwent open appendectomy and 245 (62%) underwent laparoscopic appendectomy (three-port technique). Patients were divided into two groups (negative appendectomy and acute appendicitis) based on histological results. Negative appendectomy was defined as the absence of inflammatory cell infiltration. The aim of this study was to develop ML models to predict negative appendectomy and acute appendicitis (based on histological results) and to evaluate the performance of these models.

Statistical Analysis

SPSS Statistics for Windows, version 11.5 (IBM Corp., Armonk, NY, USA) was used for data analysis. Mean \pm standard deviation (SD) for continuous variables and frequency (percentage) for categorical variables were used for descriptive statistical analysis. The Mann-Whitney U test was used to evaluate differences between continuous variables with two categories when normal distribution assumptions were not met. The chi-square test was used to examine relationships between two categorical variables. The receiver operating characteristic (ROC) curve was used to determine the cut-off value for the Alvarado score. Statistical significance was set at 0.05.

All ML analyses were conducted using the R programming language with the RWeka and e1071 packages. Variable importance was assessed using the InfoGain and GainRatioAttributeEval tests. ML classification methods, including NaiveBayes, MultilayerPerceptron, k-nearest neighbors (IBk), AdaBoost, RandomForest (RF), and a hybrid model (Na-

Table 1. Demographic characteristics of the study population (n=395)

Sex, n (%)	
Male	209 (52.9)
Female	186 (47.1)
Age, years	
Mean±SD	37.27±15.65
Symptom duration, n (%)	
<24 hours	183 (46.3)
24-48 hours	128 (32.4)
≥48 hours	84 (21.3)
Migration of pain to the right lower quadrant, n (%)	
Present	312 (79.0)
Absent	83 (21.0)
Anorexia, n (%)	
Present	234 (59.2)
Absent	161 (40.8)
Nausea/vomiting, n (%)	
Present	214 (54.2)
Absent	181 (45.8)
Right lower quadrant tenderness, n (%)	
Present	382 (96.7)
Absent	13 (3.3)
Rebound, n (%)	
Present	261 (66.1)
Absent	134 (33.9)
Fever (≥37.3°C), n (%)	
Present	22 (5.6)
Absent	373 (94.4)
Left shift (>75% neutrophils), n (%)	
Present	333 (84.3)
Absent	62 (15.7)
Alvarado score	
Mean±SD	6.87±1.75
Total bilirubin, mg/dL	
Mean±SD	0.83±1.06
CRP, mg/dL	
Mean±SD	57.06±74.43
WBC, 103/μl	
Mean±SD	13.70±4.79
Hb, g/dL	
Mean±SD	13.89±2.00
Lymphocyte count, 103/μl	
Mean±SD	1.91±0.88
PLT count, 103/μl	
Mean±SD	260.05±71.47

RDW	
Mean±SD	13.29±1.67
Neutrophil count, 103/μl	
Mean±SD	10.67±4.55
Neutrophil, %	
Mean±SD	76.96±18.41
PNR	
Mean±SD	30.06±19.06
NLR	
Mean±SD	7.32±5.88
WNR	
Mean±SD	1.37±0.40
Appendiceal diameter, mm	
Mean±SD	10.02±2.93
Periappendiceal fat stranding, n (%)	
Present	362 (91.6)
Absent	33 (8.4)
Histopathological diagnosis, n (%)	
Negative appendectomy	54 (13.7)
Acute appendicitis	341 (86.3)

CRP: C-reactive protein; WBC: White blood cell count; PLT: Platelet count; RDW: Red blood cell distribution width; PNR: Platelet-to-neutrophil ratio; NLR: Neutrophil-to-lymphocyte ratio; WNR: White cell nucleated ratio.

iveBayes + AdaBoost + RandomForest), were applied. The dataset was evaluated using 10-fold cross-validation, and all analyses were repeated 1,000 times. Performance metrics included accuracy, F-measure, Matthews correlation coefficient (MCC), ROC area, and precision-recall curve (PRC) area.

RESULTS

Of the 395 patients included in the study, 52.9% were male, with an average age of 37.27±15.65 years (range: 18–79). The duration of symptoms was less than 24 hours in 183 (46.3%) patients, between 24 and 48 hours in 128 (32.4%), and more than 48 hours in 84 (21.3%) patients. Migration of pain to the right lower quadrant was present in 79.0% of patients, anorexia in 59.2%, nausea/vomiting in 54.2%, tenderness in the right lower quadrant in 96.7%, rebound pain in 66.1%, fever in 5.6%, and left shift in 84.3%.

The mean Alvarado score was 6.87±1.75. The mean values for total bilirubin, CRP, WBC, Hb, lymphocyte count, PLT count, RDW, and neutrophil count were 0.83±1.06, 57.06±74.43, 13.70±4.79, 13.89±2.00, 1.91±0.88, 260.05±71.47, 13.29±1.67, and 10.67±4.55, respectively. The mean appendiceal diameter detected radiologically was 10.02±2.93 mm, and periappendiceal fat stranding was observed in 362 (91.6%) patients. Histopathological examination revealed inflammatory changes consistent with acute appendicitis in 341 (86.3%) patients, while 54 (13.7%) had normal appendi-

Table 2. Demographic, clinical, laboratory, and radiological characteristics of the study groups

	Negative Appendectomy (n=54,13.7%)	Acute Appendicitis (n=341,86.3%)	p value
Sex, n (%)			
Male	23 (11.0)	186 (89.0)	0.102 ^a
Female	31 (16.7)	155 (83.3)	
Age, years			
Mean±SD	35.00±14.59	37.63±15.80	0.247 ^c
Symptoms duration, n (%)			
<24 hours	20 (10.9)	163 (89.1)	0.278 ^a
24-48 hours	19 (14.8)	109 (85.2)	
>48 hours	15 (17.9)	69 (82.1)	
Migration of pain to the right lower quadrant, n (%)			
Present	38 (12.2)	274 (87.8)	0.094 ^a
Absent	16 (19.3)	67 (80.7)	
Anorexia, n (%)			
Present	27 (11.5)	207 (88.5)	0.137 ^a
Absent	27 (16.8)	134 (83.2)	
Nausea/vomiting, n (%)			
Present	24 (11.2)	190 (88.8)	0.122 ^a
Absent	30 (16.6)	151 (83.4)	
Right lower quadrant tenderness, n (%)			
Present	50 (13.1)	332 (86.9)	0.087 ^b
Absent	4 (30.8)	9 (69.2)	
Rebound, n (%)			
Present	28 (10.7)	233 (89.3)	0.017 ^a
Absent	26 (29.4)	108 (80.6)	
Fever (≥37.3°C), n (%)			
Present	2 (9.1)	20 (90.9)	0.520 ^a
Absent	52 (13.9)	321 (86.1)	
Left shift (>75% neutrophils), n (%)			
Present	32 (9.6)	301 (90.4)	<0.001 ^a
Absent	22 (35.5)	40 (64.5)	
Alvarado score			
Mean±SD	5.46±1.80	7.09±1.64	<0.001 ^c
Total bilirubin, mg/dL			
Mean±SD	0.57±0.51	0.87±1.12	<0.001 ^c
CRP, mg/dL			
Mean±SD	25.34±36.85	62.09±77.60	<0.001 ^c
WBC, 10 ³ /μl			
Mean±SD	10.33±4.33	14.23±4.65	<0.001 ^c
Hb, g/dL			
Mean±SD	13.46±2.16	13.96±1.97	0.124 ^c
Lymphocyte count, 10 ³ /μl			
Mean±SD	2.02±0.79	1.89±0.89	0.149 ^c

PLT, 10 ³ /μl			
Mean±SD	267.39±84.90	258.88±69.18	0.617 ^c
RDW			
Mean±SD	13.15±1.04	13.31±1.75	0.656 ^c
Neutrophil count, 10 ³ /μl			
Mean±SD	7.34±4.27	11.18±4.37	<0.001 ^c
Neutrophil, %			
Mean±SD	69.07±17.25	78.18±18.30	<0.001 ^c
PNR			
Mean±SD	45.12±21.93	27.72±17.48	<0.001 ^c
NLR			
Mean±SD	5.02±6.37	7.67±5.73	<0.001 ^c
WNR			
Mean±SD	1.54±0.40	1.35±0.40	<0.001 ^c
Appendiceal diameter, mm			
Mean±SD	8.43±2.53	10.27±2.91	<0.001 ^c
Periappendiceal fat stranding, n (%)			
Present	40 (11.0)	322 (89.0)	<0.001 ^b
Absent	14 (42.4)	19 (57.6)	

CRP: C-reactive protein; WBC: White blood cell count; PLT: Platelet count; RDW: Red blood cell distribution width; PNR: Platelet-to-neutrophil ratio; NLR: Neutrophil-to-lymphocyte ratio; WNR: White cell nucleated ratio; SD: Standard derivation. ^cChi-square test; ^bFishers's exact test; ^aMann-Whitney U test.

ceal tissue and were classified as the negative appendectomy group. The demographic characteristics of the study population are presented in Table 1.

Table 2 presents a comparison of demographic, clinical, laboratory, and radiological characteristics between the groups. No significant differences were observed between the groups in terms of age, sex, symptom duration, right-sided localized pain, anorexia, nausea/vomiting, tenderness in the right lower quadrant, and fever. Rebound was present in 10.7% of patients in the negative appendectomy group and 89.3% of those in the acute appendicitis group ($p=0.017$). A left shift was present in 9.6% of patients in the negative appendectomy group and 90.4% of those in the acute appendicitis group ($p<0.001$). The mean Alvarado score was significantly higher in the acute appendicitis group than in the negative appendectomy group (7.09 ± 1.64 vs. 5.46 ± 1.80 , respectively; $p<0.001$). Mean levels of total bilirubin, CRP, WBC, neutrophil count, neutrophil %, and NLR were significantly higher in the acute appendicitis group ($p<0.001$ for all). No significant differences were observed between the groups in laboratory parameters such as Hb, lymphocyte count, PLT count, and RDW. The mean appendiceal diameter was significantly larger in the acute appendicitis group compared to the negative appendectomy group (10.27 ± 2.91 vs. 8.43 ± 2.53 , respectively; $p<0.001$). The rate of periappendiceal fat stranding was 11% in the negative appendectomy group and 89% in the acute appendicitis group ($p<0.001$). The mean PNR and WNR values were significantly higher in the negative appendectomy group ($p<0.001$ for both).

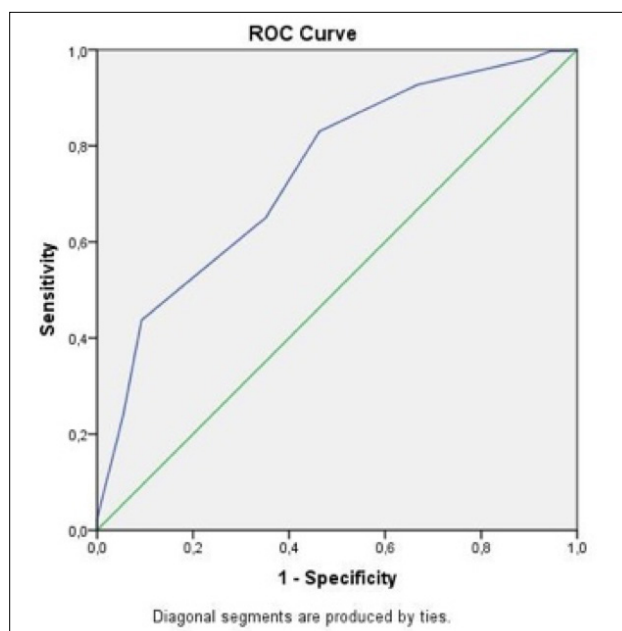


Figure 1. Receiver operating characteristic (ROC) curve of the Alvarado score for histopathological results.

In the ROC analysis of the Alvarado score for histopathological results, the area under the curve was 0.744, and the cut-off value for the Alvarado score was 6. The sensitivity and specificity values for this threshold were 0.830 and 0.537, respectively. The ROC curve is shown in Figure 1. For an Alvarado score ≥ 6 , the accuracy was 78.99%, with true-positive

and true-negative results of 283 (91.9%) and 29 (33.3%) respectively.

In the selection of variables for the ML model, the importance of variables and their effect on the outcome variable were evaluated using the InfoGain and GainRatioAttributeEval tests (Fig. 2). When variable importance and clinical relevance were evaluated in combination, the model consisted of the variables neutrophil count, WBC, NLR, PNR, appendiceal diameter, neutrophil percentage, WNR, total bilirubin, and CRP. As a result, 10 variables (nine independent variables and one dependent [pathology]) were included in the overall model, and ML analyses were performed using these variables.

NaiveBayes, MultilayerPerceptron, IBk, AdaBoost, and RandomForest methods were used to evaluate prediction performance. When the results of these methods were assessed, the accuracy for predicting acute appendicitis was high, whereas the accuracy for predicting negative appendectomy was low. To improve overall accuracy, a hybrid model approach (which has recently been introduced in the literature and involves

the combination of multiple methods) was applied, consisting of NaiveBayes, AdaBoost, and RandomForest (Table 3).

Among the ML models, the best performance was achieved using the hybrid model. When the performance criteria of the hybrid model were examined, the F-measure was 93%, the MCC was 70.4%, the ROC area was 90.8%, and the PRC area was 93.4%. Based on the hybrid model created, the overall accuracy was 92.9%. Additionally, this model correctly predicted 95.6% of patients diagnosed with acute appendicitis and 75.9% of patients with negative appendectomy (Table 3).

DISCUSSION

Our study aimed to evaluate the applicability of ML models for the diagnosis of acute appendicitis. In this study, the hybrid model consisting of NaiveBayes, AdaBoost, and RandomForest methods was found to predict acute appendicitis with high accuracy. This finding is consistent with previous studies in the literature, supporting the superior performance of hybrid models.^[12,13]

Table 3. Machine learning performance metrics for pathology prediction

	Accuracy	F-Measure	MCC	ROC Area	PRC Area
NaiveBayes					
Negative appendectomy	0.630	0.430	0.331	0.749	0.376
Acute appendicitis	0.795	0.858			0.926
Overall	0.772	0.799			0.851
MultilayerPerceptron					
Negative appendectomy	0.296	0.372	0.314	0.723	0.402
Acute appendicitis	0.953	0.923			0.929
Overall	0.863	0.848			0.857
IBk					
Negative appendectomy	0.352	0.342	0.235	0.650	0.227
Acute appendicitis	0.889	0.892			0.902
Overall	0.815	0.817			0.810
AdaBoost					
Negative appendectomy	0.278	0.375	0.340	0.779	0.450
Acute appendicitis	0.968	0.930			0.947
Overall	0.873	0.854			0.879
RandomForest					
Negative appendectomy	0.278	0.375	0.340	0.779	0.450
Acute appendicitis	0.968	0.930			0.947
Overall	0.873	0.854			0.879
Hybrid Model					
Negative appendectomy	0.759	0.745	0.704	0.908	0.763
Acute appendicitis	0.956	0.959			0.961
Overall	0.929	0.930			0.934

MCC: Matthews correlation coefficient; PRC: Precision-recall curve.

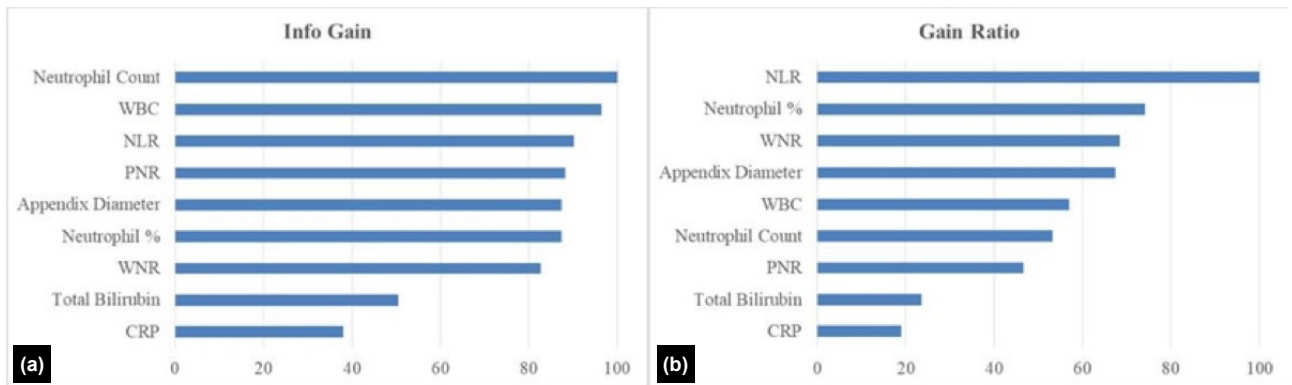


Figure 2. Feature importance results for the pathology groups. (a) InfoGainAttributeEval; (b) GainRatioAttributeEval.

Due to the lack of consistent international guidelines, the diagnosis and treatment of appendicitis are primarily based on clinical findings, standard laboratory tests, and radiological imaging modalities. The persistence of negative appendectomy rates indicates that the clinical diagnosis of acute appendicitis remains a significant challenge. The use of laparoscopy has been shown to reduce negative appendectomy rates.^[14,15] This reduction may be achieved by leaving the appendix in situ in cases where pathologies mimicking appendicitis are identified or when a macroscopically normal-appearing appendix is observed. With the increasing use of laparoscopy, the laparoscopic appendicitis score has been employed to help surgeons standardize macroscopic evaluation and has been reported to be effective in reducing the rate of negative appendectomy without missing acute appendicitis.^[4] To address these diagnostic challenges, various scoring systems have been developed to support accurate diagnosis of acute appendicitis. Nevertheless, considerable variability has been reported across studies in terms of diagnostic sensitivity and accuracy.^[16-18]

The Alvarado score is considered an effective tool in the diagnosis of acute appendicitis.^[7,8] However, it is influenced by physical examination findings, which depend on the experience of the attending physician, and its sensitivity is lower in women. False-positive and false-negative results may lead to unnecessary medical procedures and wasted resources. In this context, integrating laboratory parameters, as well as clinical and radiological findings, into ML models has improved diagnostic accuracy. A systematic review by Issaiy et al.^[10] reported that ANN and ML models demonstrate high performance in diagnosing acute appendicitis. Males et al.^[19] showed that their ML model spared 17% of pediatric patients from unnecessary surgery while missing required surgery in only 0.3% of cases. Similarly, Akbulut et al.^[20] used CatBoost with SHAP to accurately distinguish between perforated and non-perforated appendicitis in a large adult population. In another study, Navaei et al.^[21] achieved 94.5% accuracy using a Random Forest model that integrated clinical, laboratory, and imaging variables in pediatric cohorts. These findings are consistent with our results and highlight the benefit of combining

diverse clinical features with interpretable machine learning frameworks. Incorporating tools such as SHAP or LIME could further enhance model transparency and facilitate physician acceptance, which remains a critical barrier to the clinical adoption of AI-supported decision-making tools.

Studies using ML models provide an essential foundation for future research. However, the performance and generalizability of these models depend on the quality and diversity of the datasets used. Therefore, future studies should aim to include larger and more heterogeneous patient populations, explore different ML algorithms, and integrate additional variables, such as time-series changes or serial laboratory trends, to further enhance model performance. Variable selection in ML models significantly affects model success. For example, in a study by Schipper et al.,^[22] an ML model based on vital signs, physical examination, and patient history achieved an area under the ROC curve (AUROC) of 0.919, which increased to 0.923 when laboratory data were included. In this study, we found that parameters such as neutrophil count, WBC, NLR, PNR, appendiceal diameter, neutrophil percentage, WNR, total bilirubin, and CRP played a crucial role in the model's performance (Fig. 2).

The performance of various ML models, including NaiveBayes, MultilayerPerceptron, IBk, AdaBoost, and RandomForest, was evaluated. Most ML models demonstrated high accuracy, particularly in the diagnosis of acute appendicitis, except for the NaiveBayes model. The fact that the prediction results for patients diagnosed with negative appendectomy were lower than expected indicates that current models need improvement. The hybrid model leverages the strengths of different ML algorithms, resulting in more balanced and accurate results. Based on accuracy, F-measure, and MCC values, the best-performing ML model was the hybrid model consisting of NaiveBayes, AdaBoost, and RandomForest. In our study, the hybrid model achieved a pathology prediction accuracy of 92.9%. These findings highlight the potential of ML algorithms in medical diagnostics. Such models can facilitate accurate and rapid diagnoses, thereby reducing negative appendectomy rates and improving patient outcomes.

ML models have been shown to provide higher accuracy and reliability than traditional methods in the diagnosis of acute appendicitis. Hsieh et al. demonstrated that ML and artificial intelligence (ANN) achieved higher accuracy than the Alvarado score (accuracy of RF, ANN, and Alvarado score was 96%, 91%, and 80%, respectively).^[23] In a study using a decision tree model, Kang et al.^[24] showed that the ML model had a higher diagnostic value than the Alvarado score (area under the curve [AUC]: 0.850 vs. 0.695). This finding demonstrates that the Alvarado score is a clinically useful tool in the diagnosis of acute appendicitis. ROC analysis of the Alvarado score revealed an AUC of 0.744 and a cutoff value of 6. At this cut-off, sensitivity, specificity, and accuracy were 83%, 53.7%, and 78.99%, respectively. As the specificity was not satisfactory, it is evident that the Alvarado scoring system has shortcomings and requires additional diagnostic methods to improve its accuracy. ML models offer a novel approach that can compensate for the shortcomings of existing scoring systems. In our study, ML methods such as NaiveBayes, MultilayerPerceptron, IBk, AdaBoost, and RandomForest were used. Accuracy above 80% was achieved with the MultilayerPerceptron, IBk, AdaBoost, and RandomForest models. However, it was observed that these models could not predict patients diagnosed with negative appendectomy with high accuracy. The NaiveBayes model achieved more accurate predictions for patients with negative appendectomy compared to the other models. The highest diagnostic accuracy for both acute appendicitis and negative appendectomy was achieved with the hybrid model. Notably, 75.9% of patients with negative appendectomy were correctly predicted using the hybrid model consisting of NaiveBayes, AdaBoost, and RandomForest. In this study, the hybrid model achieved a high accuracy of 92.9%. Our results show that a hybrid ML model can be used to reduce negative appendectomy rates.

Nevertheless, this study has several limitations. The relatively small sample size of 395 patients may limit the representativeness of the findings, and the retrospective design may introduce bias due to incomplete or inconsistent clinical documentation. In addition, the single-center nature of the study restricts generalizability. It should also be noted that radiological and pathological assessments were not performed by a single reviewer. Instead, multiple specialists independently conducted the evaluations, reflecting routine multidisciplinary clinical practice. Although this approach may introduce some interobserver variability, it enhances the generalizability of the findings by more closely mirroring real-world conditions. Future studies employing standardized and blinded assessments by designated reviewers may further reduce variability and improve diagnostic consistency. In our study, we focused on comparing the hybrid learning model specifically with the Alvarado score as a widely used reference. In future research, comparing machine learning models with existing diagnostic scoring systems for acute appendicitis may further improve diagnostic accuracy and support clinical decision-making processes.

CONCLUSION

This study evaluated the effectiveness of ML models for diagnosing acute appendicitis. Our findings demonstrate that the ML model developed using the hybrid approach provides higher diagnostic accuracy for acute appendicitis. The accuracy (92.9%) of the hybrid model was higher than that of the traditional Alvarado scoring system (78.99%). These results indicate that the hybrid model significantly contributes to reducing negative appendectomy rates. Integrating ML models into clinical practice has the potential to improve patient management by enabling faster and more accurate diagnoses, particularly in time-sensitive conditions such as acute appendicitis.

Ethics Committee Approval: This study was approved by the Ankara University Human Research Ethics Committee (Date: 18.03.2025, Decision No: İ03-225-25).

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept: B.Y., Y.K., M.Y.; Design: B.Y., Y.K., M.Y.; Supervision: B.K., Ş.E.; Materials: Y.K., Ş.E.; Data collection and/or processing: B.K., Y.K.; Analysis and/or interpretation: B.K., M.Y.; Literature review: B.K., Ş.E.; Writing: B.K., Y.K.; Critical review: B.K., Ş.E.

Conflict of Interest: None declared.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

- Dixon F, Singh A. Acute appendicitis. *Surgery* 2023;41:418–36. [CrossRef]
- Eldar S, Nash E, Sabo E, Matter I, Kunin J, Mogilner JG, et al. Delay of surgery in acute appendicitis. *Am J Surg* 1997;173:194–8. [CrossRef]
- Mock K, Lu Y, Friedlander S, Kim DY, Lee SL. Misdiagnosing adult appendicitis: clinical, cost, and socioeconomic implications of negative appendectomy. *Am J Surg* 2016;212:1076–2. [CrossRef]
- Gelpke K, Hamminga JTH, van Bastelaar JJ, de Vos B, Bodegom ME, Heineman E, et al. Reducing the negative appendectomy rate with the laparoscopic appendicitis score; a multicenter prospective cohort and validation study. *Int J Surg* 2020;79:257–64. [CrossRef]
- Unlü C, de Castro SM, Tuynman JB, Wüst AF, Steller EP, van Wagenveld BA. Evaluating routine diagnostic imaging in acute appendicitis. *Int J Surg* 2009;7:451–5. [CrossRef]
- Samuel M. Pediatric appendicitis score. *J Pediatr Surg* 2002;37:877–81. [CrossRef]
- Alvarado A. A practical score for the early diagnosis of acute appendicitis. *Ann Emerg Med* 1986;15:557–64. [CrossRef]
- 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;15:27. [CrossRef]
- Ohle R, O'Reilly F, O'Brien KK, Fahey T, Dimitrov BD. The Alvarado score for predicting acute appendicitis: a systematic review. *BMC Med* 2011;9:139. [CrossRef]
- Issaiy M, Zarei D, Saghazadeh A. Artificial intelligence and acute appendicitis: A systematic review of diagnostic and prognostic models. *World J Emerg Surg* 2023;18:59. [CrossRef]
- Prabhudesai SG, Gould S, Rekhraj S, Tekkis PP, Glazer G, Ziprin P. Artificial neural networks: useful aid in diagnosing acute appendicitis. *World J Surg* 2008;32:305–9; discussion 310–1. [CrossRef]

12. Kavitha M, Gnanaswar G, Dinesh R, Sai Rohith Y, Sai Suraj R. Heart Disease Prediction using Hybrid machine Learning Model. 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, p. 1329–33. [CrossRef]
13. Dogan K, Selcuk T. A novel deep learning approach for the automatic diagnosis of acute appendicitis. *J Clin Med* 2024;13:4949. [CrossRef]
14. Kotaluoto S, Ukkonen M, Pauniah SL, Helminen M, Sand J, Rantanen T. Mortality related to appendectomy; a population based analysis over two decades in Finland. *World J Surg* 2017;41:64–9. [CrossRef]
15. Henriksen SR, Christophersen C, Rosenberg J, Fonnes S. Varying negative appendectomy rates after laparoscopic appendectomy: a systematic review and meta-analysis. *Langenbecks Arch Surg* 2023;408:205. [CrossRef]
16. Korkut M, Bedel C, Karancı Y, Avcı A, Duyan M. Accuracy of Alvarado, Eskelinen, Ohmann, RIPASA and Tzanakis scores in diagnosis of acute appendicitis; A cross-sectional study. *Arch Acad Emerg Med* 2020;8:e20.
17. Sharma P, Jain A, Shankar G, Jinkala S, Kumbhar US, Shamanna SG. Diagnostic accuracy of Alvarado, RIPASA and Tzanakis scoring system in acute appendicitis: A prospective observational study. *Trop Doct* 2021;51:475–81. [CrossRef]
18. Gonullu E, Bayhan Z, Capoglu R, Mantoglu B, Kamburoglu B, Harman-tepe T, et al. Diagnostic accuracy rates of appendicitis scoring systems for the stratified age groups. *Emerg Med Int* 2022;2022:2505977. [CrossRef]
19. 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. [CrossRef]
20. 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;13:1173. [CrossRef]
21. Navaei M, Doogchi Z, Gholami F, Tavakoli MK. Leveraging machine learning for pediatric appendicitis diagnosis: A retrospective study integrating clinical, laboratory, and imaging data. *Health Sci Rep* 2025;8:e70756. [CrossRef]
22. Schipper A, Belgers P, O'Connor R, Jie KE, Dooijes R, Bosma JS, et al. Machine-learning based prediction of appendicitis for patients presenting with acute abdominal pain at the emergency department. *World J Emerg Surg* 2024;19:40. [CrossRef]
23. Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery* 2011;149:87–93. [CrossRef]
24. Kang HJ, Kang H, Kim B, Chae MS, Ha YR, Oh SB, et al. Evaluation of the diagnostic performance of a decision tree model in suspected acute appendicitis with equivocal preoperative computed tomography findings compared with Alvarado, Eskelinen, and adult appendicitis scores: A STARD compliant article. *Medicine (Baltimore)* 2019;98:e17368. Erratum in: *Medicine (Baltimore)* 2019;98:e18733. [CrossRef]

ORİJİNAL ÇALIŞMA - ÖZ

Akut apandisit tanısında tanısal doğruluğu artırmaya yönelik hibrit makine öğrenmesi yaklaşımı

AMAÇ: Akut apandisit, acil serviste sağ alt kadrans ağrısı ile başvuran hastalarda en sık değerlendirilen tanıdır. Ancak atipik bulgular tanısal güçlükler yaratmakta, gereksiz operasyonlara ve maliyet artışına yol açabilmektedir. Bu çalışmanın amacı, hibrit makine öğrenmesi (ML) modeli kullanılarak akut apandisit tanısında doğruluğun artırılmasıdır.

GEREÇ VE YÖNTEM: 2020-2024 yılları arasında Ankara Üniversitesi Tıp Fakültesi Genel Cerrahi Anabilim Dalı'nda akut apandisit ön tanısı ile opere edilen 395 hasta retrospektif olarak incelendi. Demografik, klinik, laboratuvar ve radyolojik veriler değerlendirildi. NaiveBayes, MultilayerPerceptron, lBk, AdaBoost, RandomForest ve bu modellerin kombinasyonundan oluşturulan Hibrit Model uygulandı. Veriler 10-kat çapraz doğrulama ile analiz edildi. Doğruluk, F-measure, Matthews Korelasyon Katsayısı (MCC), ROC ve PRC alanı performans kriteri olarak kullanıldı.

BULGULAR: Çalışmaya dahil edilen hastaların %52.9'u erkek olup ortalama yaş 37.3 ± 15.6 yıl idi. Histopatolojik incelemede 341 (%86.3) olgu akut apandisit, 54 (%1.7) olgu negatif apendektomi olarak raporlandı. Alvarado skoru ≥ 6 için tanısal doğruluk %79.0 bulundu. ML algoritmaları içinde en iyi sonuç Hibrit Model ile elde edildi. Bu model %92.9 doğruluk oranı, %93 F-measure, %70.4 MCC, %90.8 ROC ve %93.4 PRC alanı sağladı. Ayrıca, akut apandisitli olguların %95.6'sı ve negatif apendektomi grubunun %75.9'u doğru tahmin edildi.

SONUÇ: Hibrit ML modeli, akut apandisit tanısında Alvarado skoruna kıyasla daha yüksek doğruluk sunmaktadır. Bu yaklaşımın klinik kullanıma entegrasyonu, negatif apendektomi oranlarını azaltarak hasta yönetimini iyileştirebilir.

Anahtar sözcükler: Akut apandisit; hibrit model; makine öğrenimi; negatif apendektomi; tanısal doğruluk.

Ulus Travma Acil Cerrahi Derg 2026;32(2):128-136 DOI: 10.14744/tjtes.2025.56275