Turkish Journal of Pediatric Surgery 2026;40(2):53-64
https://doi.org/10.62114/JTAPS.2026.248

Turkish Journal of Pediatric Surgery

Original Article

OPEN ACCESS

Differentiating acute appendicitis in children in the emergency department: A decision tree-based model

Şenay Kurtuluş[1], Fatih Battal[2], Sait Can Yücebaş[3], Merve Köseoğlu[4]

Abdominal pain is one of the most common reasons for pediatric emergency department visits. It may result from self-limiting conditions or from acute appendicitis (AA), which requires prompt surgical intervention. However, diagnosing AA in children remains challenging due to the overlap of clinical findings with nonspecific abdominal pain (NAP). Inadequate assessment may lead to unnecessary surgical exploration, whereas delayed diagnosis increases the risk of complications such as perforation, periappendicitis, wound infection, and intra-abdominal adhesions.[1,2]

Several clinical scoring systems have been developed to support the diagnosis of AA. These tools are generally sensitive in excluding AA in low-risk patients, thereby reducing unnecessary imaging and negative appendectomy

Received: March 18, 2025
Accepted: June 02, 2026
Published online: June 29, 2026

Correspondence: Şenay Kurtuluş, MD.

E-mail: senaykurtulus@yahoo.com

[1]Department of Pediatric Surgery, Çanakkale Onsekiz Mart University Faculty of Medicine, Çanakkale, Türkiye

[2]Department of Pediatrics, Çanakkale Onsekiz Mart University Faculty of Medicine, Çanakkale, Türkiye

[3]Computer Engineering, Çanakkale Onsekiz Mart University Faculty of Engineering, Çanakkale, Türkiye

[4]Department of Pediatrics, Hakkari Statement Hospital, Hakkari, Türkiye

Citation:

Kurtuluş Ş, Battal F, Yücebaş SC, Köseoğlu M. Differentiating acute appendicitis in children in the emergency department: A decision tree-based model. Turkish J Ped Surg 2026;40(2):53-64. doi: 10.62114/JTAPS.2026.248.

Abstract

Objectives: This study aims to develop a clinically applicable and interpretable decision tree-based model using routine clinical and laboratory parameters to differentiate acute appendicitis (AA) from nonspecific abdominal pain (NAP) in pediatric patients.

Patients and methods: This retrospective study included 708 pediatric patients presenting with abdominal pain between January 2019 and December 2023. Demographic characteristics, physical examination findings, laboratory parameters, and Alvarado scores (AS) were analyzed. The dataset was divided into training and held-out internal validation sets using stratified sampling to preserve class distribution. Model performance was evaluated using stratified five-fold cross-validation on the training set, while the held-out internal validation set was kept completely unseen during model development. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique applied exclusively within training folds. The dataset was partitioned using stratified sampling, and a decision tree model based on the classification and regression tree algorithm with an entropy splitting criterion was constructed. Model performance was evaluated on the held-out internal validation set using area under the curve (AUC), precision-recall (PR)-AUC, and F1-score.

Results: A total of 708 patients categorized into two groups: NAP group (296 males, 307 females; mean age: 10.20 ± 3.81 years, range, 4 to 17 years) and AA group (68 males, 37 females; mean age: 11.21 ± 3.51 years, range, 5 to 17 years) presenting with abdominal pain were recruited. Patients with AA had significantly higher white blood cell counts, neutrophil percentages, neutrophil-to-lymphocyte ratios, C-reactive protein levels, and AS than those in the NAP group (all $p < 0.001$). Right lower quadrant (RLQ) tenderness emerged as the most important discriminative variable, followed by the AS, left shift, age, and anorexia. The decision tree model achieved an F1-score of 0.884, an AUC of 0.946, and a PR-AUC of 0.8545. Importantly, AA was associated with a very low likelihood in patients without RLQ tenderness and with low AS, defining a clinically meaningful low-risk subgroup.

Conclusion: Decision tree-based models may serve as valuable adjuncts to clinical judgment in evaluating suspected pediatric AA, particularly by identifying low-risk patients for whom further diagnostic testing or imaging is unlikely to alter clinical management.

Keywords: Abdominal pain, acute appendicitis, Alvarado score, decision tree, pediatric emergency.

rates.[3,4] However, the routine use of abdominal and pelvic computed tomography (CT) has not consistently lowered negative appendectomy rates in children, and concerns about ionizing radiation have led to reconsideration of CT as a first-line diagnostic modality.[5] Ultrasonography (US), although widely used, has variable sensitivity but high specificity. While positive findings can support the diagnosis of AA, negative results do not reliably exclude the disease (sensitivity 71-94%, specificity 81-98%), and its accuracy remains highly operator-dependent. In cases with non-diagnostic US findings, clinical scoring systems have limited positive predictive value, whereas low scores demonstrate high negative predictive value for excluding AA.[6,7]

The Alvarado scoring system (ASS), one of the most widely used tools, incorporates symptoms, physical examination findings, and laboratory parameters.[8,9] It consists of eight components: migration of pain to the right lower quadrant (RLQ), anorexia, nausea and/or vomiting, fever, RLQ tenderness, rebound tenderness, leukocytosis, and left shift. Despite its simplicity, the diagnostic accuracy of the Alvarado score (AS) in pediatric populations remains controversial. A meta-analysis by Bai et al.[10] reported a pooled sensitivity of 76% and specificity of 71% in children, indicating moderate diagnostic performance and supporting its role as an adjunct rather than a standalone diagnostic tool.

In recent years, machine learning approaches have increasingly been applied to clinical decision-making, enabling the integration of multiple variables and the identification of complex, nonlinear relationships.[11] Among these methods, decision tree models are particularly attractive in clinical settings due to their interpretability, ability to model nonlinear interactions, and relatively low complexity.[12,13]

Therefore, this study aimed to develop a decision tree-based diagnostic model using routinely available clinical and laboratory parameters. The primary objective was to construct a clinically applicable, interpretable, and accurate tool for differentiating AA in pediatric patients presenting to the emergency department with abdominal pain.

PATIENTS AND METHODS

This retrospective study was conducted at Çanakkale Onsekiz Mart University Faculty of Medicine, Department of Pediatric Emergency between January 2019 and December 2023. Eligible patients were identified from hospital records. Inclusion criteria comprised all pediatric patients presenting with abdominal pain who underwent clinical and laboratory evaluation for suspected AA. Patients with incomplete clinical history, missing physical examination findings, or unavailable laboratory results were excluded from the analysis. The requirement for informed consent was waived due to the retrospective nature of the study. The study protocol was approved by the Çanakkale Onsekiz Mart University Clinical Research Ethics Committee (Date: 21.06.2023, No.: 2023/09-14). The study was conducted in accordance with the principles of the Declaration of Helsinki.

The diagnosis of AA was confirmed by histopathological examination in patients who underwent surgery. For non-operated patients, the diagnosis of NAP was established based on clinical follow-up and resolution of symptoms without surgical intervention.

Collected variables included demographic characteristics (age, sex), clinical findings (presence of migratory pain, anorexia, nausea/vomiting, RLQ tenderness, and rebound tenderness), and laboratory parameters. Laboratory data comprised body temperature, white blood cell (WBC) count, neutrophil and lymphocyte percentages, neutrophil-to-lymphocyte ratio (NLR), and C-reactive protein (CRP) levels. In addition, the AS was calculated for each patient, as summarized in Table 1.

Key variables were operationally defined. In particular, “left shift” was defined as an increase in immature neutrophil forms or an elevated neutrophil percentage beyond the normal reference range, consistent with institutional laboratory standards.

No imputation methods were applied for missing data, as patients with incomplete records were excluded. A decision tree-based machine learning model was developed to differentiate patients with AA using the collected clinical and laboratory parameters.

TABLE 1
The Alvarado score (AS)

Component	Score
Clinical symptoms	
Migration of pain to the RLQ	1
Anorexia	1
Nausea/vomiting	1
Physical examination	
RLQ tenderness	2
Rebound tenderness	1
Fever (> 37.5 °C)	1
Laboratory findings	
Leukocytosis, WBC > 10×10^9 /L	2
Elevated neutrophil count > 75%	1
Total	10

RLQ, right lower quadrant; WBC, white blood count.

Data partitioning strategy

To ensure an unbiased evaluation of model performance, the dataset was divided into a training set (80%, n = 566) and a held-out internal validation set (20%, n = 142). Stratified random sampling with a fixed random seed (42) was applied to preserve the original class distribution in both subsets.

The held-out internal validation set remained completely independent and was not used at any stage of model training or optimization. Within the training set, model performance was assessed using stratified five-fold cross-validation to ensure robust internal validation while maintaining class balance across folds.

A summary of the data partitioning scheme is presented in Table 2.

Class imbalance

The dataset exhibited a significant class imbalance, with a majority-to-minority ratio of 5.74:1. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.[14] The SMOTE was implemented exclusively within the training folds during stratified five-fold cross-validation to prevent data leakage. For each training fold, synthetic minority-class samples were generated by interpolating between existing minority samples and their k-nearest neighbors (k = 5) until a balanced class distribution (1:1) was achieved. The SMOTE was not applied to the validation folds or the held-out internal validation set at any stage of the analysis.

To evaluate the distributional similarity between the original and synthetic data, the Kolmogorov-Smirnov test was performed. The results of this analysis are presented in Table 3.

The majority of features (12/17) did not show statistically significant differences in the Kolmogorov-Smirnov test ($p \geq 0.05$), indicating good distributional similarity between the original and synthetic datasets. Most features exhibiting significant differences were binary variables. This discrepancy is likely attributable to the nature of SMOTE, which may generate fractional values for binary variables, thereby altering their original distribution. This limitation has been previously reported in the literature.[15]

To further assess the quality of the synthetic minority samples, dimensionality reduction and correlation structure analyses were performed, as shown in Figures 1 and 2. Visual inspection demonstrated substantial overlap between synthetic and real minority samples in the feature space. Additionally, correlation analysis revealed a high degree of structural similarity between the two

TABLE 2
Data partitioning strategy

Partition	Total (n)	NAP (n)	AA (n)	Usage
Training + cross-validation	566	482	84	Model training and internal 5-fold cross-validation
External validation	142	121	21	Held-out evaluation only
Total	708	603	105	-

NAP, nonspecific abdominal pain.

TABLE 3
Kolmogorov-Smirnov test results comparing original and synthetic data

	Real	Synthetic		
	Mean ± SD	Mean ± SD	KS statistic	*p*
Age (year)	11.19 ± 3.55	11.15 ± 2.94	0.144	0.182
Sex	1.63 ± 0.48	1.67 ± 0.40	0.222	0.007
Migration to RLQ	0.36 ± 0.48	0.33 ± 0.41	0.188	0.033
Anorexia	0.87 ± 0.34	0.85 ± 0.29	0.164	0.089
Nausea/vomiting	0.81 ± 0.40	0.76 ± 0.35	0.218	0.009
RLQ tenderness	2.00 ± 0.00	2.00 ± 0.00	0.000	1.000
Fever	0.08 ± 0.26	0.04 ± 0.15	0.075	0.890
Rebound tenderness	0.69 ± 0.46	0.69 ± 0.38	0.219	0.008
Leukocytosis	1.82 ± 0.57	1.87 ± 0.42	0.074	0.897
Left shift	0.82 ± 0.38	0.86 ± 0.34	0.063	0.970
Alvarado score	7.55 ± 1.03	7.54 ± 0.81	0.165	0.085
Temperature (°C)	36.65 ± 0.58	36.60 ± 0.34	0.071	0.917
WBC (×10³/μL)	15.27 ± 4.34	15.26 ± 3.40	0.108	0.496
Neutrophil (%)	80.70 ± 9.18	81.71 ± 7.80	0.077	0.864
Lymphocyte (%)	11.99 ± 7.36	11.17 ± 6.16	0.081	0.825
NLR	10.63 ± 9.24	10.67 ± 7.88	0.090	0.724
CRP (mg/L)	7.18 ± 10.34	7.43 ± 9.69	0.120	0.368

SD, standard deviation; RLQ, right lower quadrant; WBC, white blood count; NLR, neutrophil-to-lymphocyte ratio; CRP, C-reactive protein. Most variables showed no statistically significant differences between the original and synthetic datasets ($p \geq 0.05$), indicating good distributional similarity. Significant differences observed in some binary variables may be attributed to the generation of fractional values by SMOTE, a known limitation of the method.

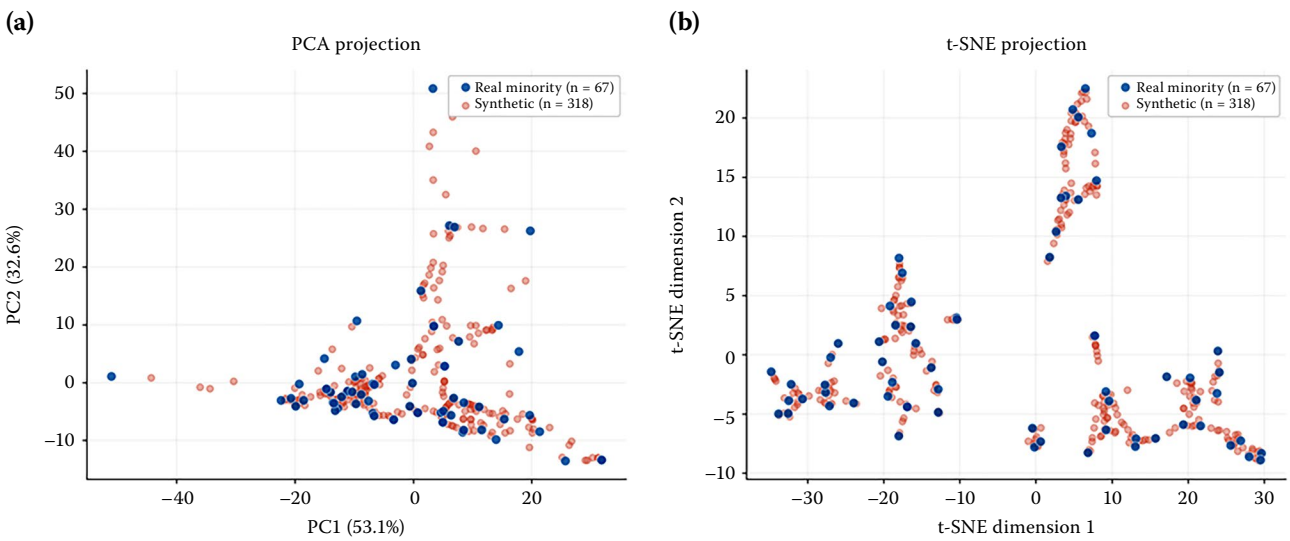

Figure 1. PCA and t-SNE projections of real and synthetic minority samples. Two-dimensional projections of real and SMOTE-generated synthetic minority samples using **(a)** PCA and **(b)** t-SNE. In both projections, synthetic samples (red) closely overlap with real minority samples (blue), indicating that SMOTE preserves the underlying data distribution and local structure of the minority class.

PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding; SMOTE, Synthetic Minority Over-sampling Technique.

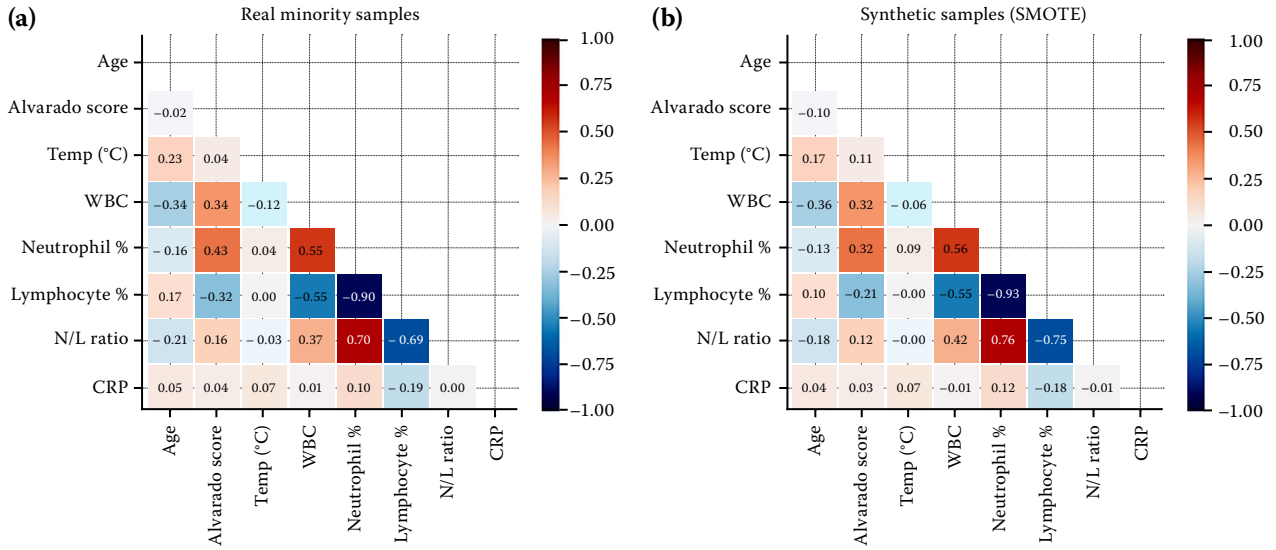

Figure 2. Correlation Structure of Real and Synthetic Minority Samples. Correlation matrices of real **(a)** and SMOTE-generated synthetic **(b)** minority samples. Each cell represents the Pearson correlation coefficient between variable pairs. The high similarity between the matrices (Pearson r = 0.9903 for the lower triangular elements) indicates that SMOTE effectively preserved the underlying inter-variable dependency structure of the minority class.

SMOTE, Synthetic Minority Over-sampling Technique; WBC, white blood cell count; CRP, C-reactive protein.

datasets. These findings suggest that SMOTE effectively preserved the overall feature space topology without introducing artificial clustering. Furthermore, the correlation matrix comparison between the real and synthetic minority samples (Figure 2) revealed substantial preservation of the inter-feature correlation structure, as indicated by a Pearson correlation coefficient of 0.9903 between the lower triangular elements of the matrices. These findings suggest that SMOTE effectively preserves both the geometric distribution and the inter-variable dependency structure of the minority class. To further assess its impact, model performance was evaluated under two scenarios: Scenario A (no sampling) and Scenario B (with resampling).

Decision tree method

In decision trees, variables are assigned to nodes, and branching occurs according to their values, linking each node to subsequent lower-level nodes. The selection of splitting variables and the generation of branches are governed by mathematical measures known as splitting criteria.[12] In this study, entropy and the information gain ratio were employed as splitting criteria to determine the most informative variable at each node. Entropy measures the level of uncertainty within the dataset, whereas the information gain ratio quantifies the reduction in entropy achieved by a split, enabling the selection of the variable with the greatest discriminatory power.

Entropy (dataset):

$$Info(S) = \sum_{i=1}^{m} pi \log_2 pi$$

Entropy after splitting by an attribute A:

$$Entropy(A) = \sum_{j=1}^{l} \frac{|Sj|}{|S|} \cdot Info(Sj)$$

Information gain:

$$Gain(A) = Info(S) - Entropy(A)$$

Entropy measures the level of uncertainty within the dataset. When the data are partitioned based on a given attribute, the resulting entropy

TABLE 4
Demographic and clinical characteristics of the study population (n = 708)

Variables	NAP (n = 603)					AA (n = 105)					
	n	%	Mean ± SD	Median	Min-Max	n	%	Mean ± SD	Median	Min-Max	*p*
Age (year)			10.20 ± 3.81	10	4-17			11.21 ± 3.51	11	5-17	0.147
Sex											0.003
Girl	307	50.9				37	35.2				
Boy	296	46.1				68	64.8				
Alvarado score			3.04 ± 1.44	3	1-7			7.55 ± 0.99	5	5-9	< 0 .001
Body temperature (°C)			11.00 ± 4.36	36.4	36-39.2			36.65 ± 0.55	36.6	35.9-39	0.002
WBC (x10^3/µl)			11.00 ± 4.36	10.1	3.3-28.5			14.89 ± 4.19	15.0	4.4-28.6	< 0.001
Neutrophils differential count (%)			67.06 ± 15.47	68.5	2.6-94.5			80.28 ± 9.32	81.6	40.1-94.7	< 0.001
Lymphocyte differential count (%)			23.83 ± 13.54	22.4	1-83			12.58 ± 7.53	10.6	2-46.3	< 0.001
NLR			5.05 ± 5.48	3.1	0.2-43.7			10.21 ± 8.64	7.97	0.87-47.2	< 0.001
CRP (mg/L)			1.68 ± 3.64	0.35	0.01-31.7			9.25 ± 16.65	3.70	0.06-91	< 0.001

NAP, nonspecific abdominal pain; AA, acute appendicitis; SD, standard deviation; WBC, white blood cell count; NLR, neutrophil-to-lymphocyte ratio; CRP, C-reactive protein.

is calculated as the weighted sum of the entropies of the resulting subsets. The reduction in entropy achieved by this partitioning defines the information gain, with higher values indicating better class separation and greater discriminative power for node splitting.

As the uncertainty associated with a variable increases, its entropy rises, and its information gain decreases, indicating a lower contribution to the classification decision. Therefore, the variable with the highest information gain is selected as the root node of the tree. After the initial split, entropy and information gain are recalculated for each subset, and the same procedure is recursively applied to subsequent nodes. The process continues until no further informative splits can be made, and the samples reaching each terminal node (leaf) are assigned to a class.

Decision tree modeling was performed using Scikit-learn (v1.8.0). To reduce overfitting, post-pruning was applied by imposing a minimum samples-per-leaf constraint (min_samples_leaf = 10). The model was trained under two scenarios (Scenario A and Scenario B) using the training dataset, and its performance was subsequently evaluated on an independent held-out validation set. The study was conducted and reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines.

Statistical analysis

Statistical analysis was performed using the PASW version 18.0 software (SPSS Inc., Chicago, IL, USA). Continuous variables are presented as mean ± standard deviation (SD) or median (min-max), as appropriate. The normality of data distribution was assessed using the Kolmogorov-Smirnov and Shapiro-Wilk tests. Non-normally distributed variables were compared using the Mann-Whitney U test. A p-value < 0.05 was considered statistically significant.

RESULTS

A total of 708 patients were included and categorized into two groups: the NAP group (296 males, 307 females; mean age: 10.20 ± 3.81 years, range, 4 to 17 years) and the AA group (68 males, 37 females; mean age: 11.21 ± 3.51 years, range, 5 to 17 years) presenting with abdominal

pain were recruited. The mean age did not differ significantly between the groups (p = 0.147). Male sex was significantly associated with AA, with 68 of 105 patients (64.8%) being male in the AA group compared with 296 of 603 patients (49.1%) in the NAP group (p = 0.003). Laboratory findings demonstrated significant differences between the groups. Patients with AA had higher WBC counts (14.9 ± 0.4 *vs.* 11.0 ± 0.2 × 10^3/μL), neutrophil percentages (80.3 ± 0.9% *vs.* 67.1 ± 0.6%), NLR (10.2 ± 0.8 *vs.* 5.04 ± 0.2), and CRP levels (9.2 ± 1.6 *vs.* 1.7 ± 0.15 mg/L) compared with the NAP group (p < 0.001 for all). Conversely, lymphocyte percentage was significantly lower in the AA group (12.6 ± 0.7% *vs.* 23.8 ± 0.6%, p < 0.001).

Body temperature was slightly but significantly higher in the AA group compared with the NAP group (36.6 ± 0.05 °C *vs.* 36.5 ± 0.18 °C, p = 0.002). Demographic and laboratory characteristics of the study population are summarized in Table 4.

The AS was significantly higher in the AA group than in the NAP group (7.6 ± 0.9 *vs.* 3.0 ± 0.05, p < 0.001). Right lower quadrant tenderness showed the strongest association with AA, being present in 96.3% of patients with AA compared with 3.7% in the NAP group (p < 0.001). Rebound tenderness was also significantly more frequent in the AA group (62.0% *vs.* 6.2%, p < 0.001).

Additionally, anorexia, nausea/vomiting, leukocytosis, and left shift were significantly more

TABLE 5
Distribution of AS components in patients with AA and NAP (n = 708)

	NAP (n = 603)					AA (n = 105)					
	n	%	Mean ± SD	Median	Min-Max	n	%	Mean ± SD	Median	Min-Max	*p*
AS			3.0 ± 0.05	3	1-7			7.6 ± 0.9	5	5-9	< 0.001
Migration to RLQ											
Yes	280	46.4				45	42.9				0.550
No	323	84.6				60	57.1				
Anorexia											
Yes	334	55.4				93	88.6				< 0.001
No	269	44.6				12	11.4				
Nausea-vomiting											
Yes	332	55.1				87	82.9				< 0.001
No	271	44.9				18	6.2				
RLQ tenderness											
Yes	4	0.7				103	98.1				< 0.001
No	599	99.3				2	1.9				
Rebound tenderness											
Yes	41	6.8				67	63.8				< 0.001
No	562	93.2				37	36.2				
Fever											
Yes	53	8.8				9	8.6				0.938
No	550	91.2				96	91.4				
Leukocytosis (> 10 × 10^3/μL)											
Yes	281	46.6				96	91.4				< 0.001
No	332	53.4				9	8.6				
Neutrophil shift, (> 75%)											
Yes	222	36.8				82	78.1				< 0.001
No	381	63.2				23	21.9				

AS, Alvarado score; NAP, nonspecific abdominal pain; SD, standard deviation; RLQ, right lower quadrant.

common in patients with AA than in those with NAP (all $p < 0.001$). In contrast, pain migration to the RLQ and fever did not differ significantly between the groups ($p = 0.550$ and $p = 0.938$, respectively). The distribution of AS components between the groups is presented in Table 5.

Decision tree modeling was performed using a Classification and Regression Trees (CART)-based algorithm with an entropy splitting criterion. Stratified 5-fold cross-validation was applied, ensuring that class proportions were preserved in each fold, with approximately 113 samples per fold (including ~17 AA cases per fold). To address class imbalance, the SMOTE was applied exclusively within the training folds, thereby preventing data leakage into the validation folds. Synthetic samples were generated by interpolation between existing minority-class observations and their five nearest neighbors until a balanced 1:1 class distribution was achieved. No oversampling was applied to the validation folds or the held-out internal validation dataset. The final decision tree structure is presented in Figure 3.

Model performance was evaluated under two scenarios (Scenario A: no resampling; Scenario B: with SMOTE) using an independent held-out internal validation dataset without resampling. The performance results are summarized in Table 6. These results indicate that performance metrics were highly similar across both scenarios. This

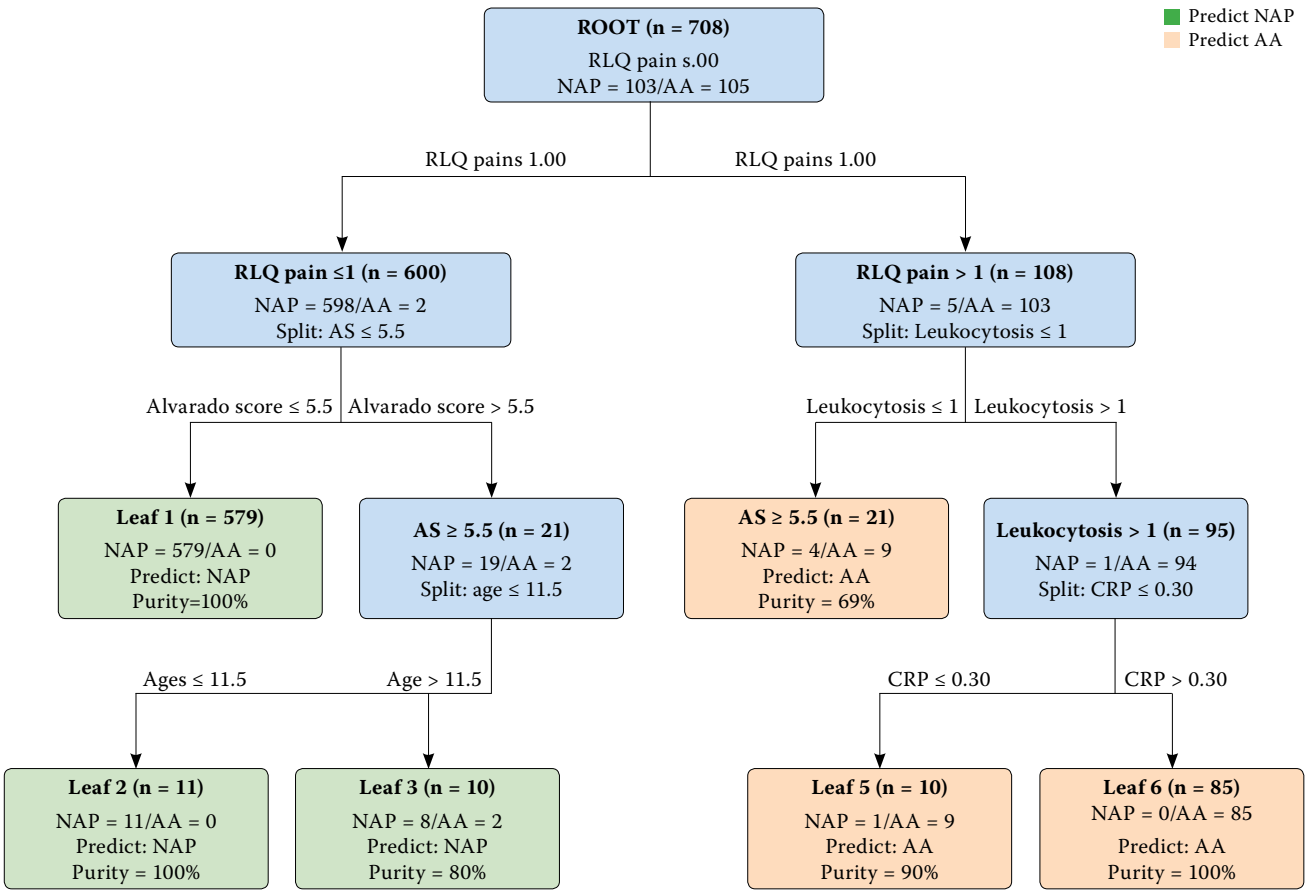

Figure 3. Decision tree structure for the classification of acute appendicitis and non-specific abdominal pain. Decision tree model for differentiating AA from NAP. The root node is defined by RLQ pain, indicating its strongest discriminative role. Subsequent splits are based on Alvarado score, age, leukocytosis, and CRP, forming a hierarchical decision structure. Terminal nodes (leaves) represent the final classification outcomes, with node purity indicating classification accuracy within each subgroup. The model demonstrates high discriminative performance, particularly in branches defined by RLQ pain and inflammatory markers.

AA, acute appendicitis; NAP, non-specific abdominal pain; RLQ, right lower quadrant; CRP, C-reactive protein.

TABLE 6
Performance metrics of the decision tree model on the held-out internal validation dataset under Scenario A (no SMOTE) and Scenario B (with SMOTE)

	Scenario A-No SMOTE	Scenario B-SMOTE
AUC-ROC	0.946	0.939
Sensitivity	0.905	0.905
Specificity	0.975	0.975
Accuracy	0.965	0.965
Balanced Accuracy	0.940	0.940
PPV	0.864	0.864
NPV	0.983	0.983
F1-score	0.884	0.884

SMOTE, Synthetic Minority Over-sampling Technique; AUC-ROC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; F1-score, harmonic mean of precision and recall.

may be attributed to the strong discriminative power of RLQ pain, which emerged as the most influential feature in the model. The confusion matrix corresponding to the held-out internal validation results is presented in Table 7.

In the context of class-imbalanced datasets, the precision-recall area under the curve (PR-AUC) is an important performance metric, as it focuses on the minority class. Performance values exceeding the baseline prevalence indicate meaningful predictive ability.[16] In this study, the prevalence of the minority class (AA) was 0.1479 and was used as the baseline for a random classifier. The PR-AUC was 0.8545 in Scenario A and 0.7899 in Scenario B, corresponding to more than a fourfold relative improvement over the baseline. These findings indicate strong discriminative performance beyond chance.

For interpretability, a final decision tree was trained using the entire training dataset and subsequently visualized. Model performance, however, was evaluated exclusively on the held-out internal validation dataset without retraining or modification. The structure of the final decision tree is presented in Figure 3.

TABLE 7
Confusion matrix of the decision tree model on the held-out internal validation dataset

	Predicted AA	Predicted NAP
Actual AA	19	2
Actual NAP	3	118

AA, acute appendicitis; NAP, nonspecific abdominal pain.

The decision tree partitions patients based on RLQ pain, followed by the AS, age, leukocytosis, and CRP. The RLQ pain emerges as the most discriminative feature at the root node, while subsequent splits incorporate both clinical and laboratory variables to refine classification.

At the root level, RLQ pain provides a strong initial separation between NAP and AA. On the left branch, the absence of RLQ pain is strongly associated with non-appendicitis cases. Within this branch, low AS defines a highly homogeneous subgroup of NAP cases, indicating that the combination of low clinical scores and absence of RLQ pain offers strong discriminative value.

On the right branch, the majority of cases are classified as AA, highlighting the strong association between RLQ pain and appendicitis. Within this subgroup, leukocytosis further refines classification, with elevated leukocyte counts reinforcing the predominance of AA cases. C-reactive protein contributes additional granularity by improving risk stratification within these subgroups.

As shown in Figure 3, some terminal splits lead to leaf nodes with the same predicted class but different purity levels. Although these splits do not alter the predicted class label, they improve

subgroup homogeneity and enhance the model's ability to refine class probability estimates.

DISCUSSION

In the present study, a decision tree-based diagnostic model was developed to differentiate AA in pediatric patients presenting with abdominal pain. The most important finding of this study is that RLQ tenderness emerged as the strongest predictor for distinguishing AA, followed by the AS, left shift, age, and anorexia. This hierarchical structure highlights the continued central role of physical examination in pediatric appendicitis, even in the era of advanced imaging and machine-learning-based decision support systems.

In our cohort, the mean AS was significantly higher in the AA group compared with the NAP group (7.6 ± 0.9 *vs.* 3.0 ± 0.05; $p < 0.001$), supporting its diagnostic relevance in pediatric AA. Consistent with the decision tree findings, AA was predominantly identified in patients with RLQ tenderness and higher AS values (e.g., > 6.5). Previous studies have similarly reported that AS ≥ 7 is strongly associated with AA.[14,17]

On the other hand, there is also a meta-analysis showing that ASS has moderate accuracy in the diagnosis of AA in children and can be an auxiliary diagnostic tool but should not be relied upon.[10] Other studies have reported sensitivities of approximately 72% and specificities of 79%, noting that an AS ≥ 6 is insufficient for definitive AA diagnosis, whereas scores below 5 may reliably exclude AA with high sensitivity.[4,18] Based on our decision tree model, AA was associated with a very low likelihood in patients without RLQ tenderness and with an AS ≤ 5.5. This suggests that the model may be particularly effective in identifying children at low risk of AA, thereby reducing unnecessary diagnostic testing or imaging that is unlikely to alter clinical management. Importantly, data-driven thresholds identified by the model may not directly correspond to clinically established cut-offs and should be interpreted as supportive rather than definitive decision boundaries.

Avoiding unnecessary interventions is important in children with suspected AA; however, minimizing false-negative classifications remains equally essential, as delayed diagnosis may increase the risk of perforation and related complications. In the present study, the number of false-negative cases was limited, reflecting the high sensitivity and negative predictive value of the model. Nevertheless, the proposed decision tree should be considered as an adjunct to clinical judgment rather than a replacement for it, particularly in patients with persistent symptoms or evolving clinical findings. In selected low-risk patients with regressing symptoms, the model may help support clinical observation and reassessment instead of immediate imaging or intervention.

Inadequate physical examination and an over-reliance on imaging modalities may lead to diagnostic errors and unnecessary interventions. A cross-sectional study reported that insufficient physical examination resulted in avoidable radiation exposure in 17% of patients.[19] Therefore, physical examination should not be regarded as an increasingly diminishing clinical art. In our study, RLQ tenderness emerged as the most important variable distinguishing AA from NAP, underscoring the enduring diagnostic value of careful clinical assessment.

Laboratory parameters represent simple and readily accessible adjuncts in the diagnosis of AA, particularly in resource-limited settings, and the NLR has been reported to demonstrate higher diagnostic performance. Previous studies have shown a significant association between elevated NLR values and the diagnosis of AA, with a reported threshold value of approximately 2.5 and moderate-to-high diagnostic accuracy. In these studies, the area under the ROC curve has been reported to be around 0.90, with sensitivity and specificity values of approximately 71% and 87%, respectively.[20] Consistent with these findings, the present study demonstrated significantly higher NLR levels in children with AA (10.2 ± 0.8 *vs.* 5.04 ± 0.2). Nevertheless, NLR should not be considered a stand-alone diagnostic marker; rather, its integration into clinical assessment and decision-making algorithms, such as the decision tree model used in this study, may enhance diagnostic confidence and improve risk stratification.

The US is the first-choice imaging option in facilities with experienced radiologists, but this may not always be possible. Due to its varied sensitivities and high specificities, clinicians can diagnose AA

based on positive findings, but negative US findings cannot rule out AA (sensitivity, 71-94%; specificity, 81-98%).[18,21] Nondiagnostic US has been noted to have suboptimal accuracy, with an incidence of approximately 49% and even a 58.1% probability of nondiagnostic US or misreading in overweight children.[22-24] It is not always practical to use imaging modalities for every patient, and distinguishing appendicitis can be difficult even with imaging.

Previous studies have demonstrated that combining thorough clinical evaluation with US yields diagnostic accuracy comparable to that of CT, while significantly reducing radiation exposure.[5,25] Stephen et al.[26] reported that CT did not improve diagnostic accuracy beyond that achieved with history, physical examination, and laboratory findings alone, and emphasized that a negative CT scan does not reliably exclude appendicitis in the presence of strong clinical suspicion. They suggested that focused CT imaging, with a high positive predictive value of 95.6%, may be reserved for selected patients with atypical clinical presentations.

An additional limitation should be acknowledged regarding the combined use of the AS and several of its individual components within the decision tree model. Since the AS already incorporates variables such as RLQ tenderness, anorexia, and left shift, some degree of feature overlap and interdependence may have influenced the hierarchical structure of the model. Nevertheless, the primary aim of the present study was to develop an interpretable, clinically applicable model rather than to establish independent causal relationships between predictors.

Importantly, the increased use of CT in pediatric patients has not been associated with a reduction in negative appendectomy rates.[5,27] Given the potential long-term risks of ionizing radiation, these findings highlight the need for pediatric surgeons to critically reassess the role of CT and to re-emphasize the primacy of clinical examination in the diagnostic algorithm for suspected AA.

Several limitations of this study should be acknowledged. First, the retrospective single-center design, based on data obtained from one pediatric emergency department, may limit the generalizability of the findings to other institutions and patient populations. In addition, the study relied on retrospective chart review, potential selection bias cannot be fully excluded, and the consecutive inclusion of all eligible patients could not be definitively confirmed.

Second, the original dataset was imbalanced. Although this was addressed using SMOTE resampling, synthetic data may not fully capture the characteristics of real minority-class observations.

Third, model performance was evaluated using a held-out validation dataset derived from the same source population. The use of a truly independent external cohort would provide a more robust assessment of model generalizability.

Finally, although post-pruning and minimum leaf-size constraints were applied to mitigate overfitting, this possibility cannot be entirely excluded given the sample size. Therefore, future multicenter prospective studies incorporating independent external validation are warranted to confirm the robustness, reproducibility, and clinical applicability of these findings.

In conclusion, this study demonstrates that a decision tree-based approach can effectively differentiate AA in children presenting to the emergency department with abdominal pain, achieving an F1-score of 0.884, an AUC of 0.946, and a PR-AUC of 0.8545. The proposed model provides a structured and reproducible framework that integrates physical examination findings with laboratory data, potentially improving diagnostic confidence, facilitating more efficient patient triage, and reducing unnecessary investigations. These findings support the potential utility of interpretable machine learning models as adjunctive tools in the clinical evaluation of suspected AA in pediatric patients.

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: F.B., Ş.K., C.S.Y.: Conception; C.S.Y.: Design; Ş.K., M.K.: Data collection or processing; Ş.K., C.S.Y.: Analysis or interpretation, literature search, writing. All authors reviewed the results and approved the final version of the manuscript.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding: The authors received no financial support for the research and/or authorship of this article.

AI Disclosure: The authors declare that artificial intelligence (AI) tools were not used, or were used solely for language editing, and had no role in data analysis, interpretation, or the formulation of conclusions. All scientific content, data interpretation, and conclusions are the sole responsibility of the authors. The authors further confirm that AI tools were not used to generate, fabricate, or 'hallucinate' references, and that all references have been carefully verified for accuracy.

REFERENCES

1. Nasiri S, Mohebbi F, Sodagari N, Hedayat A. Diagnostic values of ultrasound and the Modified Alvarado Scoring System in acute appendicitis. Int J Emerg Med 2012;5:26. doi: 10.1186/1865-1380-5-26.
2. Kırkıl C, Karabulut K, Aygen E, Ilhan YS, Yur M, Binnetoğlu K, et al. Appendicitis scores may be useful in reducing the costs of treatment for right lower quadrant pain. Ulus Travma Acil Cerrahi Derg 2013;19:13-9. doi: 10.5505/tjtes.2013.88714.
3. Neilson IR, Laberge JM, Nguyen LT, Moir C, Doody D, Sonnino RE, Youssef et al. Appendicitis in children: Current therapeutic recommendations. J Pediatr Surg 1990;25:1113-6. doi: 10.1016/0022-3468(90)90742-r.
4. Podda M, Pisanu A, Sartelli M, Coccolini F, Damaskos D, Augustin G, et al. Diagnosis of acute appendicitis based on clinical scores: Is it a myth or reality? Acta Biomed 2021;92:e2021231. doi: 10.23750/abm.v92i4.11666.
5. Martin AE, Vollman D, Adler B, Caniano DA. CT scans may not reduce the negative appendectomy rate in children. J Pediatr Surg 2004;39:886-90. doi: 10.1016/j.jpedsurg.2004.02.034.
6. Harel S, Mallon M, Langston J, Blutstein R, Kassutto Z, Gaughan J. Factors contributing to nonvisualization of the appendix on ultrasound in children with suspected appendicitis. Pediatr Emerg Care 2022;38:e678-82. doi: 10.1097/PEC.0000000000002394.
7. Bal A, Anil M, Nartürk M, Özdemir T, Arikan A, Köylüoğlu G, et al. Importance of clinical decision making by experienced pediatric surgeons when children are suspected of having acute appendicitis: The reality in a high-volume pediatric emergency department. Pediatr Emerg Care 2017;33:e38-42. doi: 10.1097/PEC.0000000000000763.
8. Memon ZA, Irfan S, Fatima K, Iqbal MS, Sami W. Acute appendicitis: Diagnostic accuracy of Alvarado scoring system. Asian J Surg 2013;36:144-9. doi: 10.1016/j.asjsur.2013.04.004.
9. Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;15:557-64. doi: 10.1016/s0196-0644(86)80993-3.
10. Bai S, Hu S, Zhang Y, Guo S, Zhu R, Zeng J. The value of the Alvarado Score for the diagnosis of acute appendicitis in children: A systematic review and meta-analysis. J Pediatr Surg 2023;58:1886-92. doi: 10.1016/j.jpedsurg.2023.02.060.
11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8-17. doi: 10.1016/j.csbj.2014.11.005.
12. Kokol P, Mernik M, Zavrsnik J, Kancler K, Malcić I. Decision trees based on automatic learning and their use in cardiology. J Med Syst 1994;18:201-6. doi: 10.1007/BF00996704.
13. Sezikli İ, Tutan MB, Turhan VB, Özkan MB, Topcu R. Medical management or surgery for acute cholecystitis: Enhancing treatment selection with decision trees. Ulus Travma Acil Cerrahi Derg 2024;30:883-91. doi: 10.14744/tjtes.2024.64796.
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
15. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263-84.
16. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432. doi: 10.1371/journal.pone.0118432.
17. Noori IF, Jabbar AS, Noori AF. Clinical scores (Alvarado and AIR scores) versus imaging (ultrasound and CT scan) in the diagnosis of equivocal cases of acute appendicitis: A randomized controlled study. Ann Med Surg (Lond) 2023;85:676-83. doi: 10.1097/MS9.0000000000000270.
18. Gorter RR, Eker HH, Gorter-Stam MA, Abis GS, Acharya A, Ankersmit M, et al. Diagnosis and management of acute appendicitis. EAES consensus development conference 2015. Surg Endosc 2016;30:4668-90. doi: 10.1007/s00464-016-5245-7.
19. Verghese A, Charlton B, Kassirer JP, Ramsey M, Ioannidis JP. Inadequacies of physical examination as a cause of medical errors and adverse events: A collection of vignettes. Am J Med 2015;128:1322-4.e3. doi: 10.1016/j.amjmed.2015.06.004.
20. Khan SA, Ashraf R, Hassaan N, Naseer M, Azad MH, Javed H. The role of neutrophil-to-lymphocyte ratio in the diagnosis of acute appendicitis. Cureus 2023;15:e51164. doi: 10.7759/cureus.51164.
21. Téoule P, Laffolie J, Rolle U, Reissfelder C. Acute appendicitis in childhood and adulthood. Dtsch Arztebl Int 2020;117:764-74. doi: 10.3238/arztebl.2020.0764.
22. Keller C, Wang NE, Imler DL, Vasanawala SS, Bruzoni M, Quinn JV. Predictors of nondiagnostic ultrasound for appendicitis. J Emerg Med 2017;52:318-23. doi: 10.1016/j.jemermed.2016.07.101.
23. Schuh S, Man C, Cheng A, Murphy A, Mohanta A, Moineddin R, et al. Predictors of non-diagnostic ultrasound scanning in children with suspected appendicitis. J Pediatr 2011;158:112-8. doi: 10.1016/j.jpeds.2010.07.035.
24. Trout AT, Sanchez R, Ladino-Torres MF, Pai DR, Strouse PJ. A critical evaluation of US for the diagnosis of pediatric acute appendicitis in a real-life setting: How can we improve the diagnostic value of sonography? Pediatr Radiol 2012;42:813-23. doi: 10.1007/s00247-012-2358-6.
25. El-Gohary Y, Molina M, Chang J, Dodd A, Miller E, Harrell C, et al. The use of computed tomography versus clinical acumen in diagnosing appendicitis in children: A two-institution international study. J Pediatr Surg 2021;56:1356-61. doi: 10.1016/j.jpedsurg.2020.09.061.
26. Stephen AE, Segev DL, Ryan DP, Mullins ME, Kim SH, Schnitzer JJ, et al. The diagnosis of acute appendicitis in a pediatric population: To CT or not to CT. J Pediatr Surg 2003;38:367-71; discsussion 367-71. doi: 10.1053/jpsu.2003.50110.
27. Russell WS, Schuh AM, Hill JG, Hebra A, Cina RA, Smith CD, et al. Clinical practice guidelines for pediatric appendicitis evaluation can decrease computed tomography utilization while maintaining diagnostic accuracy. Pediatr Emerg Care 2013;29:568-73. doi: 10.1097/PEC.0b013e31828e5718.