Dharma: A novel machine learning framework for pediatric appendicitis—diagnosis, severity assessment and evidence-based clinical decision support.

Anup Thapa[1*], Subash Pahari[2], Shashank Timilsina[1], Binay Chapagain[3]

[1]Department of Emergency Medicine and General Practice, Matri-Sishu Miteri Provincial Hospital, Pokhara, Gandaki, Nepal.

[2]Department of Computer Engineering, Faculty of Science and Technology, Pokhara University, Pokhara, Gandaki, Nepal.

[3]Department of Critical Care Medicine, Tribhuvan University Teaching Hospital, Kathmandu, Bagmati, Nepal.

Corresponding author: ajungthapa11@gmail.com (AT)

Abstract

Acute appendicitis is a common but diagnostically challenging surgical emergency in children. Existing linear scoring systems lack sufficient accuracy for standalone use, while advanced imaging is constrained by risks of sedation, contrast, and radiation. Furthermore, no available tools provide prognostic guidance.

We introduce *Dharma*, a machine learning framework consisting of a clinically grounded imputer and two random forest classifiers for diagnosis and severity assessment. Designed for real-world bedside use, Dharma is openly accessible through a web application.

Dharma achieved excellent diagnostic performance, with an AUC-ROC of 0.98 [0.97–0.99] and accuracy of 93% [91–95]. For prognostic classification, it identified complicated appendicitis with high sensitivity (96% [93–99]) and negative predictive value (97% [94–99]). In cases without appendix visualization—a frequent limitation in resource-constrained settings—Dharma maintained strong performance (AUC-ROC 0.96 [0.93–0.99]), with specificity of 97% [93–100] and PPV of 93% [84–100] at a 44% threshold, and sensitivity of 92% [84–98] with NPV of 95% [91–99] at a 25% threshold. These threshold-dependent trade-offs enable Dharma to support both ruling in and ruling out appendicitis within diverse clinical workflows.

Beyond pediatric appendicitis, Dharma's open-source framework and clinically grounded design provide a generalizable foundation for developing equitable and practical decision-support tools in healthcare.

Keywords: Dharma, pediatric appendicitis, machine learning, clinical decision support system, evidence-based medicine, artificial intelligence, Novel

Author Summary

Accurate diagnosis of pediatric appendicitis remains a long-standing challenge due to the absence of definitive biomarkers, overlapping clinical features, and limitations of current imaging modalities. Ultrasonography is widely used but often fails to visualize the appendix or requires detection of subtle signs that are difficult even for experienced radiologists. Advanced imaging such as CT or MRI, though effective, is not readily accessible in many low- and middle-income countries and carries risks related to radiation or sedation. Traditional clinical scoring systems, while simple to apply, are statistical black-boxes that rely on linear arithmetic models and cannot fully capture the complex, non-linear interactions among symptoms, laboratory values, and imaging findings. Moreover, these scores are limited to diagnosis alone and provide little or no insight into disease severity or prognosis.

To address these gaps, we developed **Dharma**, an interpretable, adaptable, and scalable machine learning–based scoring system that integrates routinely available clinical, laboratory, and radiological data to diagnose pediatric appendicitis and predict its severity. Delivered through an easy-to-use web application, Dharma provides real-time, evidence-based decision support to clinicians, particularly in resource-constrained settings. Beyond appendicitis, its open-access framework and concepts have the potential to inform solutions for diverse healthcare challenges, reflecting the direction of 21st century medicine.

Background

Acute abdominal pain (AAP) is a common complaint of Pediatric Patients which accounts for up to 10 % of total visits to the Emergency Department while acute appendicitis is the major cause of AAP in children over the age of 1 [1,2]. Despite being a routinely encountered clinical dilemma, diagnosis of acute appendicitis still poses a major challenge for the attending healthcare professionals who use patterns of signs, symptoms, laboratory values and radiological findings to make an informed decision, probably, due to the variability in clinical presentation and the frequent absence of classic symptoms in pediatric population. Different scoring systems,

with Alvarado being the oldest and the most common, have been incorporated in the workflow to help the physicians with quantifying the risk of child having acute appendicitis. However, none of the scoring systems provide sufficient sensitivity, specificity or predictive values to be used as an exclusive standard in setting the diagnosis. [3-5].

The increasing reliance on CT and USG for evaluating acute abdominal pain has not been accompanied by a corresponding improvement in appendicitis detection rates [19]. Misdiagnosis rates remain alarmingly high, ranging from 70% to 100% in children aged three years or younger, with a gradual decrease as age increases [21]. This diagnostic challenge is often attributed to the atypical presentation of the disease in younger children and the anatomical variability of the diameter of the appendix. Consequently, the lack of a reliable diagnostic tool may contribute to the persistently high incidence of Negative Appendectomy (NA) worldwide, reported between 5% and 20% in various centers [7-10]. NA is a dreaded outcome for any operating surgeon due to the inherent risks associated with general anesthesia and surgery, particularly in the pediatric population.

Moreover, no traditional inflammatory biomarkers, clinical signs and symptoms, or imaging modalities can reliably predict the progression of uncomplicated to complicated appendicitis [12,42]. This is particularly concerning in pediatric patients, where the incidence of perforated appendix is alarmingly high and increases as the age decreases, reaching over 80% in children younger than three years old [22]. This highlights the urgent need to refine diagnostic and prognostic strategies for pediatric appendicitis.

Machine Learning (ML), a subfield of Artificial Intelligence (AI), involves analyzing data to identify patterns and applying that information to make predictions in similar future scenarios [38], thus, offers a promising avenue for improving diagnostic and prognostic precision, particularly in pediatric patients, by integrating clinical, laboratory, and imaging data into a comprehensive predictive framework.

ML has been applied to assist in the diagnosis of acute appendicitis since the 1990s [27], with models specifically targeting pediatric patients emerging in the early 2010s [28]. These models aim to predict pediatric appendicitis, its complications, and postoperative recovery outcomes [11,24-31]. Despite the development of numerous ML-based tools, none have undergone independent validation, highlighting a critical gap in their clinical applicability.

ML-based tools have the potential to enhance the accuracy of appendicitis diagnosis and severity prediction [39], thereby reducing unnecessary radiation exposure from CT scans, minimizing negative appendectomy rates, and optimizing referral patterns from primary care settings. An effective ML tool should provide reliable predictions even in cases with missing data, ensuring applicability across varied clinical environments, from well-resourced urban hospitals to resource-limited rural settings. Consequently, there is a pressing need for a transparent, interpretable, adaptable and easy-to-use diagnostic tool that can reliably predict or rule out appendicitis and its complications in pediatric patients.

Thus, we defined two primary objectives for this study:

- To develop and evaluate an optimized, generalizable machine learning model for the diagnosis and complication prediction of pediatric appendicitis, and to compare its

performance with conventional diagnostic tools, including the Alvarado Score (AS), Pediatric Appendicitis Score (PAS), Appendicitis Inflammatory Response (AIR) Score, Tzanaki Score and Ultrasonography (USG).

- To deploy the model as an interpretable, web-based diagnostic and prognostic tool to support evidence-based clinical decision-making in the management of suspected appendicitis cases.

Materials And Methods

Ethics Statement

The original dataset was collected in accordance with the ethical guidelines and approval procedures of the University of Regensburg(approval numbers: 18-1063-101, 18-1063_1-101, and 18-1063_2-101). For the present study, only anonymized secondary data were analyzed. In line with these institutional regulations, no identifiable patient information was included, and therefore additional ethical approval and individual informed consent were not required.

Data Source

This study analyzed anonymized secondary data from pediatric patients with suspected appendicitis. In the primary cohort, diagnosis data were available for 780 of 782 patients

admitted with abdominal pain to Children's Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021 [44]. For testing the diagnostic model, 239 unique records from 430 suspected appendicitis cases collected at the same hospital between January 1, 2016, and December 31, 2018, were used [11]. To develop the severity assessment model, the training dataset was augmented with 49 complicated cases from 301 records collected between 2015 and 2022 at the Department of Pediatric Surgery and Pediatric Traumatology, Florence-Nightingale Hospital, Düsseldorf, Germany [47]. All datasets were already anonymized, with patient identifiers removed, and are publicly accessible from the respective sources.

Dataset Description And Preprocessing

The primary dataset consisted of 780 suspected pediatric appendicitis cases with 55 independent variables and three dependent variables: Diagnosis, Treatment Modality and Severity.

- Diagnosis: appendicitis (n = 463) and no appendicitis (n = 317)
- Severity: complications (n = 119) and no complications or no appendicitis (n = 661)
- Treatment Modality: Conservative (n = 483) and Surgical (n = 297)

Advanced radiological features such as fecal impaction, bowel wall thickening, or pathological lymph nodes were excluded from the study, as these findings can be challenging to assess even for experienced radiologists. Our aim was to develop a model that could augment pediatric surgical decision-making, particularly in resource-constrained settings where access to advanced radiology is almost always limited.

Data preprocessing involved standardizing categorical encodings: binary variables were mapped to 0/1, urinary ketones were coded as 0 (absence/trace), 1 (1+), 2 (2+), and 3 (3+), peritonitis was coded as 0 (absent), 1 (local), and 2 (generalized), and stool changes were coded as 0 (normal) and 1 (any alteration, including diarrhea or constipation).

Missing values were present in almost all variables (see Supplementary Table 1). The appendix diameter variable had 36% missingness, yet it was retained because of both the inherent challenges of appendix visualization on ultrasonography (USG) and its well-established diagnostic and pathophysiological relevance in appendicitis.

S1 Table: Missingness in Each of the Feature Columns.

For initial exploratory analyses, statistical tests were conducted on the available data without imputation. For model development, missing values in selected predictors were handled using a custom imputation pipeline, the Dharma_Imputer. This approach incorporated domain knowledge by grouping inflammatory markers (WBC count, neutrophil percentage, CRP, and temperature) and imputing them through iterative imputation with scikit-learn's decision tree regressor. Categorical variables were then imputed separately, with each feature modeled by a dedicated decision tree classifier. The appendix diameter feature was treated distinctly: missing

values were flagged by an indicator column and imputed with a clinically impossible placeholder value of -1, reflecting that non-visualization of the appendix on ultrasound is common and diagnostically meaningful. This strategy preserved the clinical nuance that decision-making differs depending on whether the appendix is visualized or not.

To preserve data integrity, imputations for training, validation, and test sets were performed independently. The trained Dharma_Imputer was also integrated into the web-based clinical decision support tool, enabling clinically informed, real-time handling of missing values.

For the primary classification task (diagnosis), we used the full dataset of 780 cases. Additionally, we extracted 239 unique cases from the 430 suspected appendicitis cases published by Marcinkevics et al. [11] and employed them exclusively for external evaluation of our models and existing linear scoring systems.

For the severity assessment task, the dataset was augmented with the full set of 430 suspected cases from Marcinkevics et al. [11]. After duplicate removal, the combined dataset comprised 1,069 unique patient records. Among these, only patients with a confirmed diagnosis of appendicitis (n = 650) were included for training and testing the severity assessment model.

Treatment modality was not analyzed in this study, as we consider it a clinical decision to be determined by the operating surgeon based on the diagnosis and anticipated risk of complications.

Feature Selection

The Chi-Square test was performed to assess the association between categorical variables and disease state, while Cramér's V measured the strength of these associations. Variables with a significant association ($p < 0.001$) and a moderate to strong effect size (Cramér's $V > 0.10$) were selected for the development of the model. Contralateral rebound tenderness, despite meeting the initial inclusion criteria, was excluded from the model's feature pool as it is not a classically defined or standardized clinical sign in the diagnostic framework of acute appendicitis.

For continuous variables, the Mann–Whitney U test was applied to compare distributions between positive and negative classes, and rank-biserial correlation was calculated to quantify effect size. Features with significant differences ($p < 0.001$) and correlation coefficients > 0.40 were retained for the feature pool.

All statistical analyses were conducted in Python using Pandas, NumPy, and SciPy on the original dataset of 780 suspected appendicitis cases.

To refine feature selection, Recursive Feature Elimination with Cross-Validation (RFECV) was employed to identify the most influential features. Different subsets of selected features were then evaluated independently to determine the optimal combination and arrangement for model development.

Data mining was performed separately for diagnostic and prognostic targets. The final feature sets used in each model are summarized in Table 1.

Table 1. List of Dharma's Features	
Diagnosis	**Complications**
Nausea	Nausea
Appetite Loss	Appetite Loss
Peritonitis	Peritonitis
WBC Count	Urinary Ketones
Neutrophil Percentage	Free Fluids
CRP	CRP
Urinary Ketones	WBC Count
Appendix Diameter	Body Temperature
Free Fluids	Appendix Diameter
	Neutrophil Percentage

Model Development

The Random Forest model was selected for its suitability in handling non-parametric medical data. For diagnosis prediction, the dataset was split into training, validation, and test sets in a 3:1:1 ratio. The model was optimized through hyperparameter tuning using grid search, followed by randomized search and stratified cross-validation, ensuring robust performance. To mitigate overfitting and ensure consistent performance across diverse clinical presentations, 10-fold stratified cross-validation was performed on the 555 bootstrap samples of the training set.

The optimal classification threshold was determined using Youden's Index and the Euclidean Distance method on the Receiver Operating Characteristic (ROC) curve applied to the validation set. We aimed to strike a balance between sensitivity and specificity to minimize both false positives and false negatives in appendicitis diagnosis.

The model's performance was then evaluated on independent test sets using AUC-ROC, sensitivity, specificity, and predictive values to ensure an unbiased assessment. AUC-ROC was chosen as the primary evaluation metric for the diagnostic task, as our goal was to develop a well-balanced model that performs reliably across varying decision thresholds.

For severity assessment, we combined our dataset with a cohort of 430 suspected appendicitis cases from Marcinkevics et al. [11]. After removing duplicates, the combined dataset consisted of 1,069 cases, of which only the confirmed appendicitis cases (n = 650) were included, as complications are only relevant among diagnosed cases. The cohort was split into training, validation, and testing sets in a 3:1:1 ratio.

The severity classes were highly imbalanced, with negative-to-positive ratios of 286:104 in the training set, 98:32 in the validation set, and 106:24 in the test set. The higher prevalence of the negative class in the training set biased the model towards predicting uncomplicated cases. To address this, we augmented the training set with 49 complicated cases from the dataset described by Marcinkevics et al. [47] and applied NearMiss undersampling to reduce the abundance of easily identifiable uncomplicated cases. This resulted in a more balanced dataset with a ratio of 235:153 (1.54:1).

We further adjusted the class weights, assigning the positive class a weight of 5:1, so that Dharma penalized false negatives more heavily than false positives. This deliberate bias toward the positive class was deemed clinically appropriate, as missing complications carries far greater risk like perforation and gangrene formation than falsely flagging non-complicated cases. Since all patients in this cohort were already diagnosed with appendicitis, a false positive would only lead to an appendectomy—an established management even for non-complicated appendicitis—whereas a false negative could delay necessary surgical intervention in complicated cases.

The primary evaluation metric for the complication model was recall (sensitivity) for the positive class, reflecting the clinical priority of minimizing missed complications.

We named the final machine-learning framework **Dharma**, which comprises the clinically grounded *Dharma_Imputer* and two Random Forest classifier models: one for diagnosis and the other for prognosis among positively diagnosed cases.

The hyperparameters used for Dharma's Diagnostic and Severity Assessment models are in Table 2.

Table 2. Hyperparameters for Dharma		
Hyperparameters	**Diagnostic Model**	**Prognostic Model**
n_estimators	555	888
min_samples_split	12	5
min_samples_leaf	1	5
max_depth	35	3
class_weight	"balanced"	{0: 1, 1: 5}
max_features	sqrt	sqrt
random_state	17	17

Model Evaluation

We applied a multi-layer evaluation strategy for Dharma across both diagnostic and prognostic tasks to assess robustness and generalization. First, model performance was estimated using 10-fold stratified cross-validation across 555 bootstrap samples of the training set. Within each

bootstrap sample, Dharma was benchmarked against state-of-the-art models appropriate for our dataset size, namely XGBoost and LightGBM.

The diagnostic model was further evaluated on two independent test sets: (i) 5,555 bootstrap samples of the unaltered test split from the primary dataset, and (ii) 5,555 bootstrap samples of 239 unique cases extracted from Marcinkevics et al. [11]. For the prognostic (complications) model, the test set contained only 18% complicated cases (106:24). Dharma's severity feature model was benchmarked against XGBoost under two settings: one tuned with simple hyperparameters and another with complex hyperparameters highly regularized for high recall.

Model performance was reported in terms of AUC-ROC, sensitivity, specificity, predictive values, and corresponding 95th-percentile confidence intervals. Hyperparameter tuning for all benchmark models was performed using grid search with stratified 10-fold cross-validation to ensure a fair comparison.

The hyperparameter settings of all benchmark models for diagnostic and prognostic tasks are reported in Tables 3 and 4 respectively.

Table 3. Hyperparameters of SOTA models for diagnosis

XGBoost		**LightGBM**	
n_estimators	555	**n_estimators**	333
max_depth	3	**max_depth**	3
min_child_weight	6	**min_child_samples**	3

learning_rate	0.01	**learning_rate**	0.01
gamma	0	**min_gain_to_split**	0
subsample	0.6	**bagging_fraction**	0.6
colsample_bytree	0.6	**feature_fraction**	0.6
reg_alpha	0.1	**lambda_l1**	0
reg_lamda	1	**lambda_l2**	1
random_state	17	**random_state**	17

Table 4. Hyperparameters of SOTA models for complications

Hyperparameters	**XGB(Simple)**	**XGB(Complex)**
n_estimators	111	111
max_depth	3	3
min_child_weight	6	6
scale_pos_weight	7	7
learning_rate	0.01	0.01
gamma	-	1
subsample	-	0.8
colsample_bytree	-	0.6
reg_alpha	-	0.1

reg_lamda	-	10

We further evaluated the robustness of Dharma against different imputation strategies for missing values. For the markers of inflammation (WBC, Neutrophil Percentage, Body temperature, CRP), we applied Iterative Imputer from the Scikit-learn library with decision tree regressor and linear regressor estimators, as well as a K-nearest neighbors (KNN) imputer. For categorical variables, we paired these regression imputation strategies with a decision tree classifier custom-trained for each feature. In addition, we tested simple strategies using mean imputation for continuous variables and mode imputation for categorical variables. Across all methods, no statistically significant difference in model performance was observed, indicating that Dharma's predictive ability was stable irrespective of the imputation approach (Supplementary Table 2).

For the final Dharma_Imputer, we selected a combination of decision tree regressor imputation for continuous variables and decision tree classifier imputation for categorical variables. This approach enables real-time, clinically informed imputation, which is particularly important in real-world settings where missing data are frequent, especially in resource-constrained environments.

S1 Table. Dharma's performance under different imputation strategies.

Explainability Analysis

o examine Dharma's decision-making process, we applied SHAP's Tree Explainer (SHapley Additive exPlanations), a game theory–based approach that quantifies the contribution of each feature relative to a baseline prediction (base SHAP value). A feature's SHAP value indicates its influence in shifting the predicted probability toward or away from this baseline. Global feature contributions were visualized using SHAP summary plots for both diagnostic and prognostic tasks. Additionally, we employed the Random Forest's native *feature_importance* method to assess the relative importance assigned to each variable by the modelsl.

At the individual level, SHAP values were integrated into our web application alongside 95th-percentile confidence intervals derived from the 555 decision trees in Dharma's random forest, providing real-time uncertainty estimates. This integration enables attending physicians to explore the reasoning behind model predictions, thereby enhancing transparency, interpretability, and clinical trust.

Results

Data Mining

A total of 780 pediatric patients (ages 0–18 years; mean age: 11.35 years, 95% CI: 11.10–11.59) with suspected appendicitis were included in the diagnostic prediction cohort, with a male-to-

female ratio of 1.07 (403:377). For complication prediction, 650 confirmed appendicitis cases from the combined dataset, as described earlier, were included.

A Chi-square test was performed to assess the significance of associations, and Cramér's V was used to evaluate the strength of association of the categorical features. Seven features showed significant association with appendicitis diagnosis and six features were significantly associated with complication prediction ($p < 0.001$). Among the diagnostic features, peritonitis demonstrated the strongest association (Cramér's V = 0.37), followed by neutrophilia (Cramér's V = 0.23). For complication prediction, peritonitis again showed the strongest association (Cramér's V = 0.30), followed by urinary ketones (Cramér's V = 0.29), Table 5 and 6.

Table 5. Diagnosis and Categorical Variables (n=780)

Features	Chi-Square	P-value	Cramer's V	Samples
Peritonitis	108.24	<0.001	0.37	773
Neutrophilia	67.17	<0.001	0.30	732
Free Fluids	32.48	<0.001	0.21	719
Ketones in Urine	24.34	<0.001	0.20	582
Contralateral Rebound Tenderness	25.08	<0.001	0.18	767
Nausea/Vomiting	23.19	<0.001	0.17	774
Loss of Appetite	19.24	<0.001	0.16	772
Ipsilateral Rebound Tenderness	8.64	<0.01	0.12	619

Sex	10.78	<0.01	0.12	779
RIF Tenderness	8.18	<0.01	0.10	774
Migratory Pain	6.70	0.01	0.09	773
Coughing Pain	6.30	0.01	0.09	766
Psoas Sign	4.00	0.05	0.07	745
Stool	3.04	0.22	0.06	765
RBC in Urine	2.02	0.57	0.06	576
WBC in Urine	1.25	0.74	0.05	583
Dysuria	1.61	0.20	0.05	753

Table 6. Complications and Categorical variables(n=463)

Features	**Chi-Square**	**P-value**	**Cramer's V**	**Samples**
Peritonitis	41.29	<0.001	0.3	457
Ketones in Urine	26.45	<0.001	0.29	318
Neutrophilia	23.75	<0.001	0.23	434
Nausea/Vomiting	23.17	<0.001	0.22	458
Loss of Appetite	20.56	<0.001	0.21	456
Free Fluids	16.72	<0.001	0.2	430
RBC in Urine	10.65	0.01	0.18	317
Ipsilateral Rebound Tenderness	5.62	<0.05	0.14	305
WBC In Urine	4.9	0.18	0.12	319
Stool	5.12	0.08	0.11	451

RIF Tenderness	2.66	0.1	0.08	458
Psoas Sign	1.58	0.21	0.06	433
Contralateral Rebound Tenderness	1.5	0.22	0.06	451
Migratory Pain	0.59	0.44	0.04	457
Dysuria	0.54	0.46	0.04	437
Sex	0.54	0.46	0.03	462
Coughing Pain	0.01	0.94	0.004	450

A Mann-Whitney U test was conducted to assess the statistical significance of differences in the distributions of continuous variables between positive and negative cases, and rank-biserial correlation was used to evaluate the effect size of these differences, Table 7 and 8. Six variables were significantly associated with appendicitis diagnosis, and five variables were significantly associated with complication prediction ($p < 0.001$). Among the diagnostic variables, appendix diameter demonstrated the strongest effect (rank-biserial correlation = 0.9), followed by C-reactive protein (CRP) level (rank-biserial correlation = 0.46). For complication prediction, CRP level exhibited the strongest effect (rank-biserial correlation = 0.64), followed by white blood cell (WBC) count (rank-biserial correlation = 0.45).

Table 7. Diagnosis and Continuous variables (n=780)

Feature	U_statistic	P_value	Rank-Biserial	Mean (Positive	Mean(Negative	Samples

			correlation	Class)	Class)	
Appendix Diameter	2341	<0.001	0.9	8.7	5.04	498
CRP	38686	<0.001	0.46	44.9	11.72	771
WBC Count	39910	<0.001	0.45	14.28	10.33	776
Neutrophil Percentage	32518.5	<0.001	0.42	76.03	65.6	679
Body Temperature	57410	<0.001	0.21	37.52	37.24	775
BMI	79280	<0.001	-0.15	18.45	19.56	753
Weight	82617	<0.01	-0.13	41.72	45.25	778
Age	80879	0.02	-0.1	11.08	11.72	780
Height	74276.5	0.08	-0.07	146.93	149.51	755
RDW	65536.5	0.21	0.05	13.4	12.87	756
Hemoglobin	72577.5	0.52	-0.03	13.38	13.38	764
RBC Count	70347	0.92	0.004	4.79	4.82	764
Thrombocyte Count	70459.5	0.95	0.003	285.79	284.48	764

Table 8. Complications and Continuous variables(n=463)

Feature	U_statistic	P_value	Rank-Biserial Correlation	Mean (Positive Class)	Mean(Negative Class)	Samples
CRP	6952.5	<0.001	0.64	107.68	24.52	457

WBC Count	10960	<0.001	0.45	17.43	13.23	460
Body Temperature	12010.5	<0.001	0.4	37.93	37.38	459
Neutrophil Percentage	8803.5	<0.001	0.38	82.44	74.14	403
Appendix Diameter	7296	<0.001	0.28	9.53	8.51	371
Weight	24201.5	<0.01	-0.2	38.73	42.75	461
Age	24327	<0.01	-0.2	10.08	11.42	463
Height	21988.5	<0.01	-0.17	141.56	148.79	443
BMI	21284.5	0.01	-0.16	17.92	18.64	441
Hemoglobin	21018.5	0.06	-0.12	13.36	13.39	450
Thrombocyte Count	17298	0.2	0.08	296.65	282.24	450
RBC Count	20250.5	0.23	-0.08	4.75	4.8	450
RDW	17530	0.44	0.05	13.56	13.35	445

Diagnostic Performance

Conventional Tools

All conventional scoring systems—Alvarado Score (AS), Pediatric Appendicitis Score (PAS), Appendicitis Inflammatory Response (AIR) Score, and Tzanaki Score—were significantly associated with appendicitis diagnosis ($p < 0.001$). The discriminatory performance, expressed as AUC-ROC with 95% confidence intervals, was 0.76 [0.71–0.82] for AS, 0.70 [0.64–0.76] for PAS, 0.74 [0.68–0.80] for AIR, and 0.91 [0.87–0.94] for the Tzanaki Score. Median values were consistently higher among confirmed cases compared to negatives (AS: 7 vs. 5; PAS: 5 vs. 4; AIR: 6 vs. 4; Tzanaki: 12 vs. 6), indicating higher scoring trends among positive cases, Table 9.

Table 9. Association of Diagnostic Tools				
Scoring Systems	**P-value**	**Rank-Biserial Correlation**	**Median(Positive Class)**	**Median(Negative Class)**
Alvarado Score	<0.001	0.52	7	5
Pediatric Appendicitis Score	<0.001	0.4	5	4
AIR Score	<0.001	0.48	6	4
Tzanaki Score	<0.001	0.82	12	6

The Pediatric Appendicitis Score (PAS) with a low cut-off of 4 demonstrated the highest sensitivity, 89 [84–93], making it most effective for ruling out the disease. Conversely, the Appendicitis Inflammatory Response (AIR) Score with a cut-off of 9 achieved very high specificity, 99 [97–100], supporting its role in ruling in the disease. A stepwise approach—using PAS (≥ 4) for initial screening and AIR (≥ 9) for final diagnostic confirmation may provide a promising framework for future studies. However, none of these linear scoring systems, when used in isolation, demonstrated sufficient performance to serve as a reliable standalone diagnostic method, Table 10.

Table 10. Performance of Diagnostic tools (n=289)						
Tools	**AUC-ROC**	**Sensitivity %**	**Specificity %**	**PPV %**	**NPV %**	**Accuracy %**

Alvarado Score(≥5)	0.76 [0.71-0.82]	86 [81-91]	45 [35-55]	68 [74-80]	64 [53-75]	72 [66-78]
Alvarado Score(≥7)	0.76 [0.71-0.82]	55 [48-62]	83[76-90]	86 [79-92]	50 [43-58]	65 [60-70]
PAS Score(≥4)	0.70 [0.64-0.76]	89 [84-93]	38 [29-48]	72 [67-78]	65 [53-76]	71 [66-76]
PAS Score(≥6)	0.70 [0.64-0.76]	48 [40-55]	76 [68-85]	79 [71-86]	44 [3]	58 [52-63]
AIR Score(≥5)	0.74 [68-80]	67 [60-73]	67 [57-76]	79 [72-85]	52 [44-61]	67 [61-72]
AIR Score(≥9)	0.74 [68-80]	16 [11-21]	99 [97-100]	97 [89-100]	39 [33-45]	45 [40-51]
Tzanaki Score(≥8)	0.91 [87-94]	85 [80-90]	88 [81-94]	93 [89-97]	76 [68-84]	86 [82-90]
USG(≥6 mm)	0.94 [0.89-0.99]	94 [91-97]	95 [87-100]	99 [97-100]	80 [67-91]	94 [91-97]

Ultrasonography (USG) demonstrated the highest diagnostic performance among the available tools, with an AUC-ROC of 0.94 [0.89–0.99]. However, appendix diameter measurements were unavailable in 34% of suspected cases, reflecting the inherent challenges in appendix visualization. A similar limitation was observed with the Tzanaki Score, which relies heavily on USG findings for its calculation despite its strong discriminatory ability.

Dharma

Dharma demonstrated consistently high diagnostic performance across cross-validation and independent test sets.

In 10-fold cross-validation, the model achieved an AUC-ROC of 0.98 [0.97–0.99] with an overall accuracy of 93% [91–95]. The performance was balanced, with sensitivity 93% [91–95], specificity 93% [90–96], PPV 95% [93–97], and NPV 90% [87–94].

In the first independent test set, the model maintained robust discriminative ability with an AUC-ROC of 0.98 [0.96–0.99] and accuracy of 91% [87–95]. Sensitivity remained high (93% [88–98]) with slightly lower specificity (87% [78–95]), while predictive values were balanced (PPV 93% [88–97], NPV 88% [78–96]).

In the second curated test set, performance was further reinforced, achieving an AUC-ROC of 0.98 [0.96–0.99] and accuracy of 94% [91–97]. The model demonstrated high specificity (98% [95–100]) and PPV (99% [97–100]) while maintaining sensitivity (92% [88–96]) and NPV (87% [80–93]).

Even in the cohort with an unvisualized appendix, Dharma maintained strong discriminatory ability with an AUC-ROC of 0.96 [0.93–0.99]. At a threshold of 44%, it achieved a very high specificity of 97% [93–100] with a PPV of 93% [84–100]. At a threshold of 25%, it achieved a high NPV of 95% [91–99] and sensitivity of 92% [84–98].

Overall, Dharma showed excellent and stable diagnostic accuracy with consistently high AUC (>0.96) across all evaluations, demonstrating both generalizability and clinical robustness. Detailed diagnostic results are summarized in Table 11 and Table 12.

Table 11. Dharma's Diagnostic Performance						
Dharma	**AUC-ROC**	**Specificity**	**PPV**	**Sensitivity**	**NPV**	**Accuracy**
10-fold cross- validation	0.98 [0.97-0.99]	93% [90-96]	95% [93-97]	93% [90-95]	90% [87-94]	93% [91-95]
Split Test Set(n=156)	0.98 [0.96-0.99]	87% [78-95]	93% [88-97]	93% [88-98]	88% [78-96]	91% [87-95]
Curated Test Set(n=289)	0.98 [0.96-0.99]	98% [95-100]	99% [97-100]	92% [88-96]	87% [80-93]	94% [91-97]

Table 12. Unvisualized appendix cohort (n = 144)						
Threshold	**AUC-ROC**	**Specificity**	**PPV**	**Sensitivity**	**NPV**	**Accuracy**
44%	0.96 [0.93-0.99]	97% [93-100]	93% [84-100]	75% [62-86]	87% [80-93]	89% [83-94]

25%	0.96 [0.93-0.99]	89% [83-95]	82% [72-92]	92% [84-98]	95% [91-99]	90%[85-95]

Prognostic Performance

For complication prediction, Dharma achieved a sensitivity of 96% [93–99] and NPV of 97% [94–99] in 10-fold cross-validation, with corresponding specificity 65% [53–75], PPV 65% [59–72], and AUC-ROC 0.92 [0.90–0.95]. This performance profile, characterized by high sensitivity and NPV, reflects the intended design of Dharma's severity assessment module as a screening tool to safely rule out complications and to guide decisions regarding surgical versus conservative management, as well as the urgency of surgery.

In the unaltered independent test set, the model again demonstrated strong recall for the positive class, with sensitivity 88% [68–97] and NPV 95% [85–99]. Specificity was lower (49% [39–59]) with a corresponding PPV of 28% [18–40]. Confidence intervals for this evaluation were calculated using Clopper–Pearson exact binomial estimation, which is appropriate for small sample sizes and rare events. This was necessary given the pronounced class imbalance (106 negative vs. 24 positive cases). The wide confidence interval observed for sensitivity likely reflects the limited number of positive cases (24, 18% of the test set).

Overall, Dharma demonstrated strong prognostic performance for complication prediction, with high sensitivity and NPV supporting its role as a reliable screening tool. Detailed prognostic results are summarized in Table 13.

Table 13. Dharma's Performance for Severity Assessment				
Dharma	**Sensitivity**	**NPV**	**Specificity**	**PPV**
10-fold cross-validation	96% [93-99]	97% [94-99]	65% [53-75]	65% [59-72]
Unaltered Test Set (n=130)	88% [68-97]	95% [85-99]	49% [39-59]	28% [18-40]

Benchmarking Analyses

Diagnostic Value

We compared Dharma, the Imputer–Diagnostic Classifier, against existing clinical scoring systems, ultrasonography (USG), and two state-of-the-art models suited for medium-sized tabular data (XGBoost and LightGBM). Evaluation was performed using 5,555 bootstrap samples of the curated test set (n = 289) and the split test set.

Dharma consistently outperformed all existing scoring systems across every key metric (Table 14). Against SOTA models, Dharma was slightly inferior in the split test set but showed a slight edge in the curated test set (Table 15). Taken together, these findings suggest Dharma performs on par with contemporary machine-learning benchmarks while clearly surpassing conventional diagnostic tools.

Table 14. Dharma minus Conventional tools on 5555 bootstrap samples (n=289)

Tool	AUC-ROC	Accuracy	Sensitivity	Specificity	PPV	NPV
AS (≥5)	0.22 [0.16–0.27]	0.23 [0.17–0.28]	0.06 [0.01–0.11]	0.53 [0.42–0.63]	0.25 [0.19–0.31]	0.23 [0.12–0.34]
AS (≥7)	0.22 [0.16–0.27]	0.29 [0.24–0.35]	0.37 [0.30–0.44]	0.15 [0.07–0.23]	0.13 [0.07–0.20]	0.37 [0.30–0.44]
PAS (≥4)	0.28 [0.22–0.34]	0.23 [0.17–0.29]	0.03 [–0.03–0.09]	0.60 [0.49–0.70]	0.26 [0.21–0.32]	0.22 [0.10–0.34]
PAS (≥6)	0.28 [0.22–0.34]	0.36 [0.30–0.42]	0.44 [0.36–0.52]	0.22 [0.13–0.31]	0.20 [0.13–0.28]	0.43 [0.35–0.50]
AIR (≥5)	0.24 [0.18–0.30]	0.27 [0.21–0.33]	0.25 [0.18–0.32]	0.31 [0.22–0.41]	0.20 [0.14–0.27]	0.35 [0.27–0.43]
AIR (≥9)	0.24 [0.18–	0.49 [0.43–	0.76 [0.69–	–0.01 [–0.04–	0.02 [–0.02–	0.48 [0.41–

	0.30]	0.55]	0.82]	0.02]	0.10]	0.54]
Tzanaki	0.07 [0.04–0.10]	0.08 [0.04–0.12]	0.07 [0.03–0.11]	0.10 [0.04–0.16]	0.06 [0.03–0.10]	0.11 [0.05–0.17]
USG	0.04 [0.00–0.09]	0.01 [0.00–0.03]	0.01 [0.00–0.03]	0.00 [0.00–0.00]	0.00 [0.00–0.00]	0.04 [0.00–0.10]

Table 15. Dharma minus SOTA models in two test sets

Metrics	**Split test set (n=156)**		**Curated test set (n=289)**	
	XGBoost	**LightGBM**	**XGBoost**	**LightGBM**
AUC-ROC	0.00[-0.01-0.00]	0.00[-0.01-0.00]	0.01[0.00-0.02]	0.01[0.00-0.01]
Accuracy	-0.02[-0.04-0]	-0.01[-0.03-0.01]	0.00[-0.01-0.01]	0.00[-0.01-0.01]
Sensitivity	-0.03[-0.07-0]	-0.01[-0.04-0.02]	-0.01[-0.02-0.01]	-0.01[-0.02-0.-01]
Specificity	0.00[0.00-0.00]	0.00[0.00-0.00]	0.01[0.00-0.03]	0.01[0-0.03]
PPV	0.00[-0.01-0]	0.00[0.00-0.00]	0.01[0.00-0.02]	0.01[0-0.02]
NPV	-0.05[-0.11-0]	-0.02[-0.07-0.04]	-0.01[-0.03-0.02]	-0.01[-0.03-0.02]

Prognostic Performance

Dharma was marginally inferior to the XGBoost models in key prognostic metrics, namely sensitivity and NPV, but outperformed them markedly in secondary metrics, including accuracy, specificity, and PPV. A detailed comparison is provided in Table 16.

Table 16. Dharma minus SOTAs for prognosis

Metric	XGBoost (Simple Hyperparameters)	XGBoost (Complex Hyperparameters)
Sensitivity	–0.01 [–0.05–0.01]	–0.03 [–0.07–0.01]
NPV	–0.01 [–0.04–0.02]	–0.01 [–0.05–0.10]
Specificity	0.18 [0.08–0.27]	0.40 [0.29–0.51]
AUC-ROC	0.01 [–0.01–0.02]	0.00 [–0.01–0.01]
Accuracy	0.10 [0.04–0.15]	0.23 [0.17–0.29]

PPV	0.10 [0.05–0.14]	0.18 [0.14–0.23]

SHAP Analysis

SHAP (Shapley Additive Explanations) analysis was performed to interpret the Dharma model's predictions. For the diagnostic model, the base value—the average model output before any feature input—was 0.50, while for the severity assessment model, the base value was 0.73, reflecting a clear bias towards the complicated cases. In the diagnostic task, appendix diameter exerted the greatest influence on predictions, followed by white blood cell (WBC) count, C-reactive protein (CRP), and neutrophil percentage. For the complications model, CRP had the strongest impact, followed by appendix diameter and peritonitis. Interestingly, larger appendix diameters were associated with lower predicted complication risk, suggesting an inverse relationship between diameter and severity that warrants further exploration.

The absence of appendix diameter data reduced the model's diagnostic probability below the base value; subsequent contributions from other features then shifted the prediction upward to reach the final output. Under these circumstances, peritonitis had the strongest influence, followed by CRP, WBC count, and neutrophil percentage. The global feature importance patterns for both diagnostic and prognostic tasks are illustrated in the SHAP summary dot plots (S1–S3 Figs).

S1 Fig. SHAP dot plot for diagnosis prediction

S2 Fig. SHAP dot plot for diagnosis prediction(cases with no appendix diameter data)

S3 Fig. SHAP dot plot for severity assessment

Discussion

In this study, we present Dharma, a novel machine learning framework purpose-built for pediatric appendicitis. Dharma integrates its own native imputer (*Dharma_Imputer*) with a stacked architecture of two Random Forest models. The first model, optimized for balanced diagnostic performance, achieved an AUC-ROC of 0.98 [0.97–0.99] and an accuracy of 93% [91–95] in identifying acute appendicitis.

For patients classified as acute appendicitis by this upstream model, a second Random Forest was deployed to assess prognosis. This prognostic model prioritized recall for the positive class, reaching a sensitivity of 88% [68–97] and a negative predictive value (NPV) of 95% [85–99]. In doing so, Dharma not only provides a highly reliable diagnosis of this common yet challenging pediatric emergency but also serves as a screening tool for potential complications, enabling safe exclusion and early, well-informed management decisions.

In our study, we compared Dharma against widely used linear scoring systems, both older (Alvarado Score, Pediatric Appendicitis Score [PAS]) and newer (Appendicitis Inflammatory Response [AIR] Score and Tzanaki Score). Each scoring system showed strengths and weaknesses: for instance, AIR with a high cut-off of 9 demonstrated excellent specificity for appendicitis diagnosis (99 [97–100]), whereas PAS with a low cut-off of 4 was highly sensitive (89 [84–93]). Using PAS (≤4) for ruling out and AIR (≥9) for ruling in may provide a simple yet pragmatic framework for clinical decision-making.

However, none of these scoring systems alone is sufficiently robust for diagnosing acute appendicitis. Their commonly used intermediate cut-offs have relatively poor specificity (38%–67%), leading to a high rate of false positives and subsequent negative appendectomies. Similar patterns were noted by Sefa Sağ et al. [6], who reported specificities as low as (14.3%–57.1%). These limitations underscore the need for more accurate, balanced and data-driven non-linear approaches such as Dharma.

As expected, ultrasonography (USG) demonstrated excellent diagnostic performance in our cohort, with an AUC-ROC of 0.94 [0.89–0.99]. Using a 6 mm appendix diameter cutoff, sensitivity and specificity reached 95% [87–100] and 95% [87–100], respectively. However, in 34% of cases the appendix could not be visualized. This limitation aligns with prior reports by Duncan et al. [40] and Jenny et al. [41], who described non-visualization rates as high as 71–76% in suspected appendicitis. Failure to identify the appendix may be attributed to anatomical variation, patient habitus, or the operator-dependent nature of USG [13]. Moreover, advanced features beyond diameter—such as peri-appendiceal fluid collection (PALC), fat stranding,

target sign, hyperemia on Color Doppler, or free fluid—are typically recognized only by experienced radiologists, who may be unavailable in primary care or resource-limited contexts. Importantly, USG findings must be integrated with clinical and laboratory parameters for accurate diagnosis [43], and they provide little prognostic information about disease severity. In LMICs and overcrowded emergency departments, even timely USG access may be limited, further constraining its practical utility.

Advanced imaging modalities such as CT and MRI provide high diagnostic accuracy for appendicitis [14–16,18]. However, their use in children is limited by the risks of sedation, contrast exposure, and ionizing radiation [17]. In addition, these modalities remain largely inaccessible in many primary and secondary healthcare settings, particularly in resource-limited regions like ours. Crucially, while they improve diagnostic certainty, they do not reliably differentiate between uncomplicated and complicated appendicitis—an essential distinction for guiding timely and appropriate management [20].

Significant progress has been made in applying machine learning to the diagnosis, severity stratification, and management of pediatric appendicitis [11,24–27]. Yet, many of these studies are highly technical and AI-centric, which makes their predictions difficult for clinicians to interpret and incorporate into day-to-day practice. Unlike traditional bedside scoring systems, machine learning models also require dedicated, user-friendly interfaces to be clinically useful. For example, the image-based approach by Byun et al. [26], although innovative, is computationally intensive and impractical for real-time decision-making in busy emergency departments. Similarly, the online tool by Marcinkevics et al. [11] improved accessibility but

remains dependent on complex ultrasonographic variables requiring expert radiological input. Furthermore, several features included in their dataset were not directly used in the model's decision process, which diminishes transparency and clinical applicability. As a result, such tools are better suited to academic exploration and dataset development rather than to frontline clinical deployment.

When benchmarked against contemporary state-of-the-art machine learning architectures such as XGBoost and LightGBM, Dharma demonstrated broadly comparable performance across both diagnostic and prognostic tasks. On one external test set, Dharma achieved marginally superior discrimination, while on another, it performed slightly below the strongest baseline. Although these differences reached statistical significance, they were clinically negligible, underscoring that Dharma delivers equivalent discriminative power while offering distinct advantages in interpretability, accessibility, and clinical alignment. Thus, Dharma is not merely another high-performing algorithm but a purpose-built, clinically grounded tool designed for real-world use. More broadly, this highlights that future progress in clinical decision support may hinge less on incremental algorithmic optimization and more on designing systems that are clinically aligned, resource-sensitive and interpretable.

Dharma assesses suspected cases of acute appendicitis by analyzing a comprehensive non-linear pattern of clinical features—mirroring the approach of an attending pediatric surgeon. These include symptoms such as nausea or vomiting and loss of appetite; signs like peritonitis; physiological parameters such as body temperature; laboratory findings including white blood cell (WBC) count, percentage of neutrophils, C-reactive protein (CRP), and urinary ketones; as

well as radiological findings such as appendiceal diameter and the presence of free fluid on ultrasonography. These features are already well-established in existing literature for the diagnosis and risk stratification of acute appendicitis [32-37].

While the assessment of signs and symptoms can be particularly challenging in younger children, certain features remain comparatively more accessible. History-based symptoms such as nausea/vomiting and poor feeding are generally easier to elicit, even in preverbal children through caregiver observations. Similarly, signs of peritonitis can often be assessed using surrogate physical findings, including tenderness, rebound tenderness, guarding, and rigidity—either localized to the right iliac fossa (RIF) or generalized across the abdomen. Radiological parameters such as appendiceal diameter and the presence of peritoneal free fluid are also reliably assessable in most pediatric patients, even with screening ultrasonography or Point-Of-Care Ultrasound (POCUS), making them particularly valuable in settings where advanced imaging modalities or experienced radiologists may be unavailable or limited.

A key contribution of this work is the translation of Dharma into a practical, transparent, and interpretable web-based tool. The platform provides Dharma Score, case-level probability estimate with 95% confidence intervals derived from the 555 estimators in the diagnostic random forest, as well as the predicted probability of chances of complications, thereby supporting surgeons in selecting appropriate treatment modalities. To enhance interpretability, Dharma also generates feature-level SHAP explanations for both diagnostic and prognostic tasks, allowing clinicians to understand and trust the model’s reasoning. For real-time deployment, the Dharma_Imputer has been integrated into the web application pipeline to intelligently impute

variables such as CRP, appendix diameter, and free fluid, which are often difficult to assess in resource-limited settings. Other variables have been made mandatory, as they are relatively easier to obtain even under such constraints.

We believe that knowledge should be freely available and openly shared. In alignment with this principle, we have made our source code publicly accessible on GitHub [46], and the web application is freely accessible at dharma-ai.org , promoting transparency, reproducibility, and continuous improvement in medical science. Screenshots of the interface are provided in Supplementary Figures S4–S6.

S4 Fig. Homepage

S5 Fig. Diagnostic and Prognostic Prediction

S6 Fig. Feature level SHAP explanation

Nevertheless, this study has certain limitations that warrant acknowledgment. Foremost, the lack of access to large, high-quality medical datasets remains a significant barrier, limiting the full potential of statistical and machine learning approaches that we believe are essential for advancing 21st-century medicine [23]. We strongly advocate for the open sharing of anonymized clinical datasets to accelerate broader and more impactful research. Another limitation is that our model was developed on single-center data, raising questions about its adaptability across diverse populations. Furthermore, the absence of exclusively histologically confirmed appendicitis cases may have constrained the model's ability to fully optimize pattern recognition

in both diagnostic and prognostic tasks. Future work should prioritize multicenter validation—both prospective and retrospective—including studies in low- and middle-income countries (LMICs) to rigorously assess Dharma's real-world clinical utility and generalizability. We also encourage independent researchers and med-tech innovators to validate Dharma further, and where necessary, calibrate its thresholds to suit the epidemiological and clinical nuances of their own populations.

Despite these limitations, the Dharma represents more than an incremental step in clinical decision support. By integrating readily available clinical, laboratory, and radiological findings into a unified predictive framework, it provides real-time, evidence-based diagnostic and prognostic support for one of the most common pediatric surgical emergencies. With its high specificity and positive predictive value, Dharma has the potential to meaningfully reduce unnecessary negative appendectomies while maintaining patient safety through very low false-negative rates. In doing so, the model could lessen reliance on advanced imaging such as CT or MRI, thereby reducing the risks associated with sedation, contrast exposure, and ionizing radiation in children—offering both clinical and economic benefits.

Importantly, Dharma is designed to function across diverse healthcare settings, from resource-limited facilities to advanced tertiary centers. In primary and secondary care, where even basic imaging and surgical expertise may be unavailable, Dharma leverages accessible clinical and laboratory features to support early recognition, guide timely referrals, and reduce delays in management. In high-volume tertiary emergency departments, it may help prioritize cases when imaging is inconclusive or radiology services are saturated. At advanced surgical centers,

calibrated thresholds optimized for high specificity and PPV allow Dharma to be used as a reliable rule-in tool. Combined with its severity assessment feature, it can also assist in triaging patients for appendectomy and informing decisions between conservative versus surgical management.

Taken together, Dharma illustrates how data-driven tools can complement, rather than replace, clinical judgment. By bridging bedside intuition with machine learning–derived insights, it provides a pathway toward equitable, AI-augmented care. While prospective multicenter validation remains essential, this work highlights the promise of interpretable, accessible, and open clinical decision support systems—not only for pediatric appendicitis, but as a framework adaptable to broader healthcare challenges.

Acknowledgement

We would like to thank Marcinkevics R and colleagues for making their anonymized dataset publicly available, which greatly facilitated this research. We are also grateful to our colleague and friend, Amrit Neupane, for his valuable contributions in designing the UI/UX of the Dharma web tool.

Author Contributions

Conceptualization and study design: Anup Thapa, Subash Pahari

Data curation: Anup Thapa, Subash Pahari

Literature review: Anup Thapa

Formal analysis: Anup Thapa

Methodology: Anup Thapa

Software: Anup Thapa (Machine Learning Framework Development), Subash Pahari (web-app development)

Writing – original draft: Anup Thapa, Sashank Timilsina, Binay Chapagain

Writing – review & editing: Anup Thapa

Writing – Manuscript Formating: Subash Pahari

References

[1] Hijaz NM, Friesen CA. Managing acute abdominal pain in pediatric patients: current perspectives. Pediatric Health Med Ther. 2017 June 29;8:83-91. doi: 10.2147/PHMT.S120156. PMID: 29388612; PMCID: PMC5774593.

[2] Tseng YC, Lee MS, Chang YJ, Wu HP. Acute abdomen in pediatric patients admitted to the pediatric emergency department. Pediatr Neonatol. 2008 Aug;49(4):126-34. doi: 10.1016/S1875-9572(08)60027-3. PMID: 19054918.

[3] Schneider C, Kharbanda A, Bachur R. Evaluating appendicitis scoring systems using a prospective pediatric cohort. Ann Emerg Med. 2007 Jun;49(6):778-84, 784.e1. doi: 10.1016/j.annemergmed.2006.12.016. Epub 2007 Mar 26. PMID: 17383771.

[4] Sencan A, Aksoy N, Yıldız M, Okur Ö, Demircan Y, Karaca I. The evaluation of the validity of Alvarado, Eskelinen, Lintula and Ohmann scoring systems in diagnosing acute appendicitis in children. Pediatr Surg Int. 2014 Mar;30(3):317-21. doi: 10.1007/s00383-014-3467-0. PMID: 24448910.

[5] Pogorelić Z, Rak S, Mrklić I, Jurić I. Prospective validation of Alvarado score and Pediatric Appendicitis Score for the diagnosis of acute appendicitis in children. Pediatr Emerg Care. 2015 Mar;31(3):164-8. doi: 10.1097/PEC.0000000000000375. PMID: 25706925.

[6] Sağ S, Basar D, Yurdadoğan F, Pehlivan Y, Elemen L. Comparison of Appendicitis Scoring Systems in Childhood Appendicitis. Turk Arch Pediatr. 2022 Sep;57(5):532-537. doi: 10.5152/TurkArchPediatr.2022.22076. PMID: 36062441; PMCID: PMC9524470.

[7] Jukić M, Nizeteo P, Matas J, Pogorelić Z. Trends and Predictors of Pediatric Negative Appendectomy Rates: A Single-Centre Retrospective Study. Children (Basel). 2023 May 15;10(5):887. doi: 10.3390/children10050887. PMID: 37238435; PMCID: PMC10217643.

[8] Kundiona I, Chihaka OB, Muguti GI. Negative appendicectomy: evaluation of ultrasonography and Alvarado score. Cent Afr J Med. 2015 Sep-Dec;61(9-12):66-73. PMID: 29144064.

[9] Oyetunji TA, Ong'uti SK, Bolorunduro OB, Cornwell EE 3rd, Nwomeh BC. Pediatric negative appendectomy rate: trend, predictors, and differentials. J Surg Res. 2012 Mar;173(1):16-20. doi: 10.1016/j.jss.2011.04.046. Epub 2011 May 19. PMID: 21696768.

[10] Patel H, Kamel M, Cooper E, Bowen C, Jester I. The Variable Definition of "Negative Appendicitis" Remains a Surgical Challenge. Pediatr Dev Pathol. 2024 Nov-Dec;27(6):552-558. doi: 10.1177/10935266241255281. Epub 2024 June 6. PMID: 38845117.

[11] Marcinkevics R, Reis Wolfertstetter P, Wellmann S, Knorr C, Vogt JE. Using Machine Learning to Predict the Diagnosis, Management and Severity of Pediatric Appendicitis. Front Pediatr. 2021 Apr 29;9:662183. doi: 10.3389/fped.2021.662183. PMID: 33996697; PMCID: PMC8116489.

[12] Bălănescu L, Băetu AE, Cardoneanu AM, Moga AA, Bălănescu RN. Predictors of Complicated Appendicitis with Evolution to Appendicular Peritonitis in Pediatric Patients. Medicina (Kaunas). 2022 Dec 22;59(1):21. doi: 10.3390/medicina59010021. PMID: 36676645; PMCID: PMC9866196.

[13] Callahan MJ, Rodriguez DP, Taylor GA. CT of appendicitis in children. Radiology. 2002 Aug;224(2):325-32. doi: 10.1148/radiol.2242010998. PMID: 12147823.

[14] Kim DW, Yoon HM, Lee JY, Kim JH, Jung AY, Lee JS, Cho YA. Diagnostic performance of CT for pediatric patients with suspected appendicitis in various clinical settings: a systematic review and meta-analysis. Emerg Radiol. 2018 Dec;25(6):627-637. doi: 10.1007/s10140-018-1624-9. Epub 2018 Jul 12. PMID: 30003463.

[15] Sivit CJ, Applegate KE, Stallion A, Dudgeon DL, Salvator A, Schluchter M, Berlin SC, Myers MT, Borisa VJ, Weinert DM, Morrison SC, Grisoni ER. Imaging evaluation of suspected appendicitis in a pediatric population: effectiveness of sonography versus CT. AJR Am J Roentgenol. 2000 Oct;175(4):977-80. doi: 10.2214/ajr.175.4.1750977. PMID: 11000147.

[16] Stephen AE, Segev DL, Ryan DP, Mullins ME, Kim SH, Schnitzer JJ, Doody DP. The diagnosis of acute appendicitis in a pediatric population: to CT or not to CT. J Pediatr Surg. 2003 Mar;38(3):367-71; discsussion 367-71. doi: 10.1053/jpsu.2003.50110. PMID: 12632351.

[17] Brennan GD. Pediatric appendicitis: pathophysiology and appropriate use of diagnostic imaging. CJEM. 2006 Nov;8(6):425-32. doi: 10.1017/s1481803500014238. PMID: 17209492.

[18] Mittal MK. Appendicitis: Role of MRI. Pediatr Emerg Care. 2019 Jan;35(1):63-66. doi: 10.1097/PEC.0000000000001710. PMID: 30608328.

[19] Pines JM. Trends in the rates of radiography use and important diagnoses in emergency department patients with abdominal pain. Med Care. 2009 Jul;47(7):782-6. doi: 10.1097/MLR.0b013e31819748e9. PMID: 19536032.

[20]Lastunen K, Leppäniemi A, Mentula P. Perforation rate after a diagnosis of uncomplicated appendicitis on CT. BJS Open. 2021 Jan 8;5(1):zraa034. doi: 10.1093/bjsopen/zraa034. PMID: 33609386; PMCID: PMC7893470.

[21] Almaramhy HH. Acute appendicitis in young children less than 5 years: review article. Ital J Pediatr. 2017 Jan 26;43(1):15. doi: 10.1186/s13052-017-0335-2. PMID: 28257658; PMCID: PMC5347837.

[22] Pogorelić Z, Domjanović J, Jukić M, Poklepović Peričić T. Acute Appendicitis in Children Younger than Five Years of Age: Diagnostic Challenge for Pediatric Surgeons. Surg Infect (Larchmt). 2020 Apr;21(3):239-245. doi: 10.1089/sur.2019.175. Epub 2019 Oct 16. PMID: 31618143.

[23] Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018 Dec;284(6):603-619. doi: 10.1111/joim.12822. Epub 2018 Sep 3. PMID: 30102808.

[24] Harmantepe AT, Dikicier E, Gönüllü E, Ozdemir K, Kamburoğlu MB, Yigit M. A different way to diagnosis acute appendicitis: machine learning. Pol Przegl Chir. 2023 Oct 13;96(2):38-43. doi: 10.5604/01.3001.0053.5994. PMID: 38629278.

[25] Reismann J, Romualdi A, Kiss N, Minderjahn MI, Kallarackal J, Schad M, Reismann M. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. PLoS One. 2019 Sep 25;14(9):e0222030. doi: 10.1371/journal.pone.0222030. PMID: 31553729; PMCID: PMC6760759.

[26] 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 Mar 1;13(5):923. doi: 10.3390/diagnostics13050923. PMID: 36900066; PMCID: PMC100010Forsström JJ, Irjala K, Selén G, Nyström M, Eklund P. Using data preprocessing and single layer perceptron to analyze laboratory data. Scand J Clin Lab Invest Suppl. 1995;222:75-81. doi: 10.3109/00365519509088453. PMID: 7569750.49.

[27] Forsström JJ, Irjala K, Selén G, Nyström M, Eklund P. Using data preprocessing and single layer perceptron to analyze laboratory data. Scand J Clin Lab Invest Suppl. 1995;222:75-81. doi: 10.3109/00365519509088453. PMID: 7569750.

[28] Grigull L, Lechner WM. Supporting diagnostic decisions using hybrid and complementary data mining applications: a pilot study in the pediatric emergency department. Pediatr Res. 2012 Jun;71(6):725-31. doi: 10.1038/pr.2012.34. Epub 2012 Mar 22. PMID: 22441377.

[29] Chadaga K, Khanna V, Prabhu S, Sampathila N, Chadaga R, Umakanth S, Bhat D, Swathi KS, Kamath R. An interpretable and transparent machine learning framework for appendicitis detection in pediatric patients. Sci Rep. 2024 Oct 18;14(1):24454. doi: 10.1038/s41598-024-75896-y. Erratum in: Sci Rep. 2025 Jan 22;15(1):2841. doi: 10.1038/s41598-025-86494-x. PMID: 39424647; PMCID: PMC11489819.

[30] Hua R, O'Brien MK, Carter M, Pitt JB, Kwon S, Ghomrawi HMK, Jayaraman A, Abdullah F. Improving Early Prediction of Abnormal Recovery after Appendectomy in Children using Real-world Data from Wearables. Annu Int Conf IEEE Eng Med Biol Soc. 2024 Jul;2024:1-4. doi: 10.1109/EMBC53108.2024.10782031. PMID: 40039430.

[31] Ghomrawi HMK, O'Brien MK, Carter M, Macaluso R, Khazanchi R, Fanton M, DeBoer C, Linton SC, Zeineddin S, Pitt JB, Bouchard M, Figueroa A, Kwon S, Holl JL, Jayaraman A, Abdullah F. Applying machine learning to consumer wearable data for the early detection of complications after pediatric appendectomy. NPJ Digit Med. 2023 Aug 16;6(1):148. doi: 10.1038/s41746-023-00890-z. PMID: 37587211; PMCID: PMC10432429.

[32] Pop GN, Costea FO, Lungeanu D, Iacob ER, Popoiu CM. Ultrasonographic findings of child acute appendicitis incorporated into a scoring system. Singapore Med J. 2022 Jan;63(1):35-41. doi: 10.11622/smedj.2020102. Epub 2020 Jul 15. PMID: 32668829; PMCID: PMC9251212.

[33] Adir A, Braester A, Natalia P, Najib D, Akria L, Suriu C, Masad B, Igor W. The role of blood inflammatory markers in the preoperative diagnosis of acute appendicitis. Int J Lab Hematol. 2024 Feb;46(1):58-62. doi: 10.1111/ijlh.14163. Epub 2023 Aug 29. PMID: 37644670.

[34] Moris D, Paulson EK, Pappas TN. Diagnosis and Management of Acute Appendicitis in Adults: A Review. JAMA. 2021 Dec 14;326(22):2299-2311. doi: 10.1001/jama.2021.20502. PMID: 34905026.

[35] Yajima H, Fujita T, Yanaga K. Profile of signs and symptoms in mild and advanced acute appendicitis. Int Surg. 2010 Jan-Mar;95(1):63-6. PMID: 20480844.

[36] 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 Jan;219(1):154-163. doi: 10.1016/j.amjsurg.2019.04.018. Epub 2019 Apr 27. PMID: 31056211.

[37] Biehl CM, Elliver M, Gudjonsdottir J, Salö M. Utility of Urine Dipstick Testing in Pediatric Appendicitis: Assessing its Role in Identifying Complicated Cases and Retrocecal Appendicitis. Eur J Pediatr Surg. 2024 Dec 19. doi: 10.1055/a-2490-1156. Epub ahead of print. PMID: 39701137.

[38] Singh, C. Machine Learning in Pattern Recognition. European Journal of Engineering and Technology Research. 2023 Apr;8(2):63–68. doi:10.24018/ejeng.2023.8.2.3025.

[39] Issaiy M, Zarei D, Saghazadeh A. Artificial Intelligence and Acute Appendicitis: A Systematic Review of Diagnostic and Prognostic Models. World J Emerg Surg. 2023 Dec 19;18(1):59. doi: 10.1186/s13017-023-00527-2. PMID: 38114983; PMCID: PMC10729387.

[40] Scrimgeour DS, Driver CP, Stoner RS, King SK, Beasley SW. When does ultrasonography influence management in suspected appendicitis? ANZ J Surg. 2014 May;84(5):331-4. doi: 10.1111/ans.12415. Epub 2014 Jan 9. PMID: 24405944.

[41] Held JM, McEvoy CS, Auten JD, Foster SL, Ricca RL. The non-visualized appendix and secondary signs on ultrasound for pediatric appendicitis in the community hospital setting. Pediatr Surg Int. 2018 Dec;34(12):1287-1292. doi: 10.1007/s00383-018-4350-1. Epub 2018 Oct 6. PMID: 30293146.

[42] Nijssen DJ, van Amstel P, van Schuppen J, Eeftinck Schattenkerk LD, Gorter RR, Bakx R. Accuracy of ultrasonography for differentiating between simple and complex appendicitis in children. Pediatr Surg Int. 2021 Jul;37(7):843-849. doi: 10.1007/s00383-021-04872-8. Epub 2021 Mar 7. PMID: 33677613; PMCID: PMC8172400.

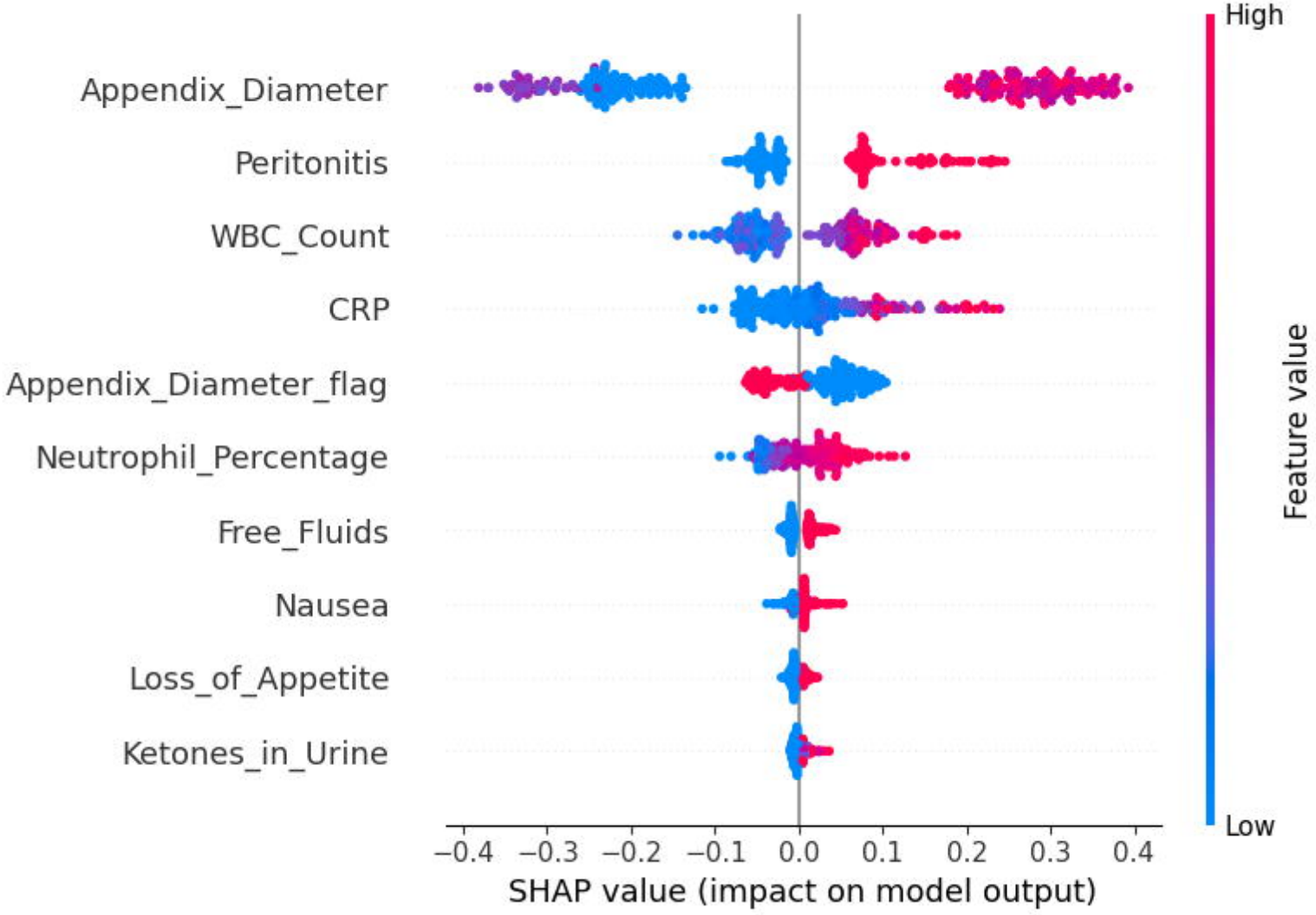
[43] Mills LD, Mills T, Foster B. Association of clinical and laboratory variables with ultrasound findings in right upper quadrant abdominal pain. South Med J. 2005 Feb;98(2):155-61. doi: 10.1097/01.SMJ.0000129927.88863.65. PMID: 15759944.

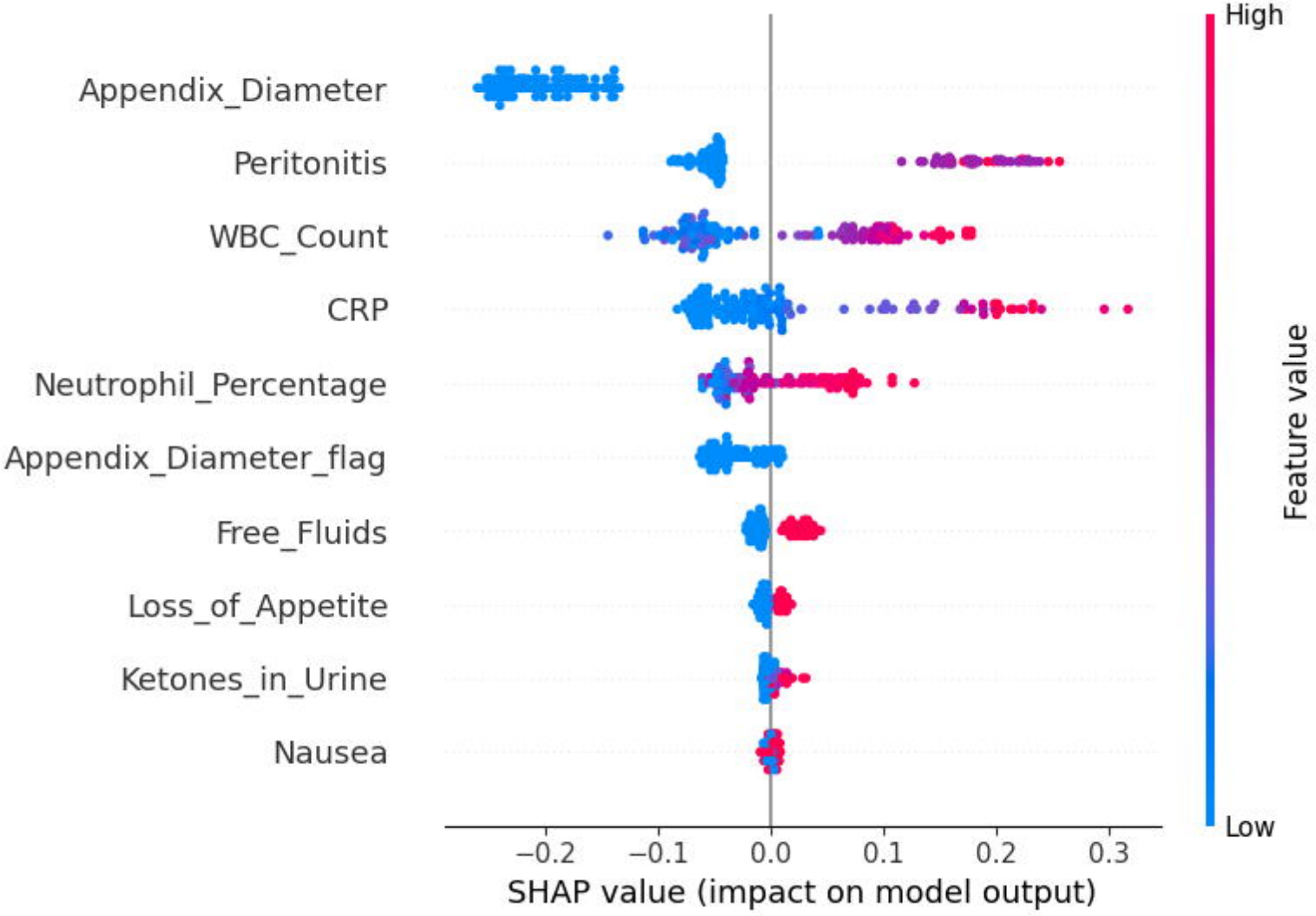
[44] Marcinkevičs R, Reis Wolfertstetter P, Klimiene U, Ozkan E, Chin-Cheong K, Paschke A, et al. Regensburg Pediatric Appendicitis Dataset. Zenodo; 2023. https://zenodo.org/records/7669442

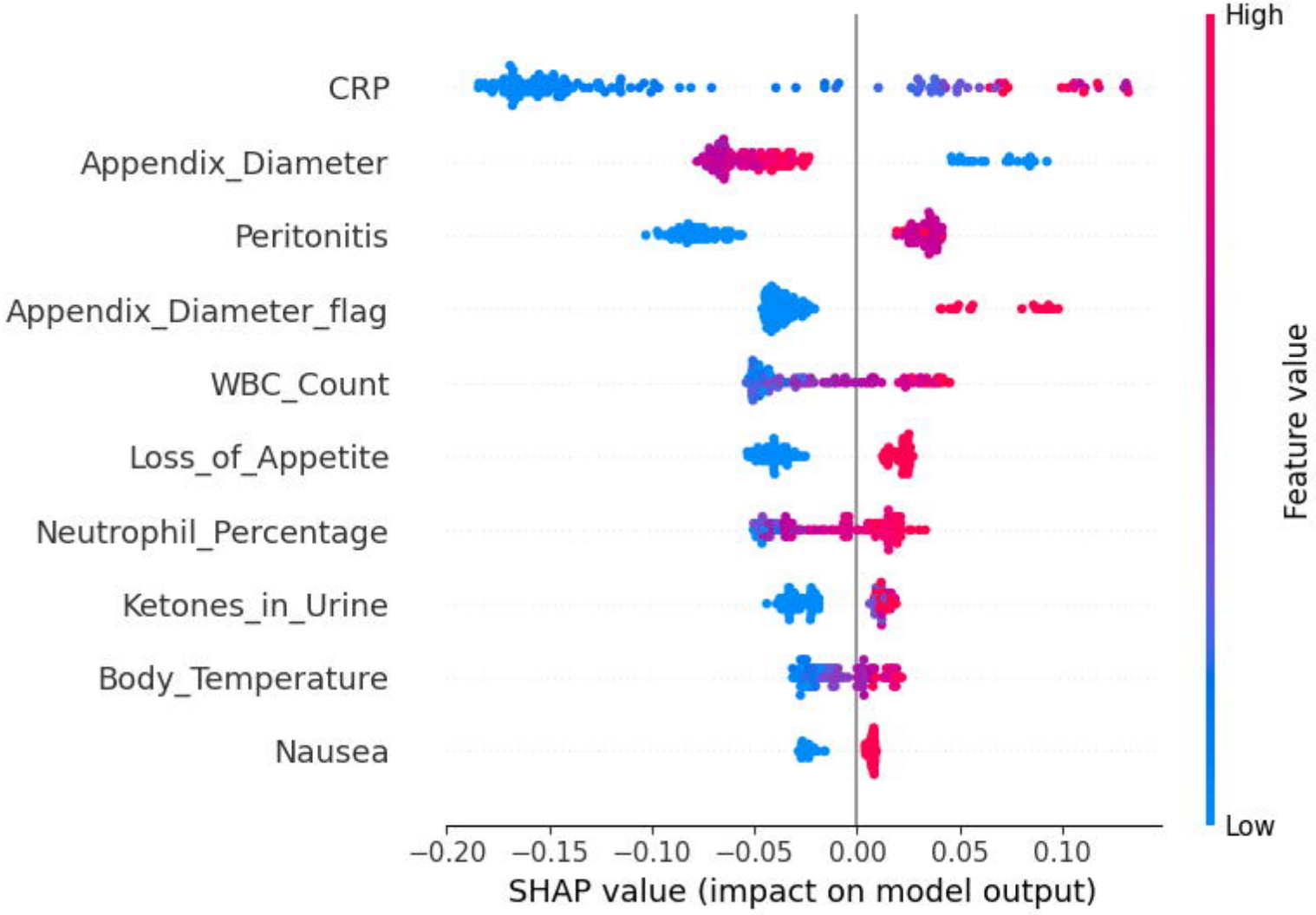
[45] Steiner Z, Buklan G, Stackievicz R, Gutermacher M, Litmanovitz I, Golani G, Arnon S. Conservative treatment in uncomplicated acute appendicitis: reassessment of practice safety. Eur J Pediatr. 2017 Apr;176(4):521-527. doi: 10.1007/s00431-017-2867-2. Epub 2017 Feb 16. PMID: 28210834

[46] Thapa A. Dharma-AppendicitisModel [Internet]. 2025 [cited 2025 May 26]. Available from: https://github.com/ajung17/Dharma-AppendicitisModel

[47] Marcinkevičs R, Sokol K, Paulraj A, Hilbert MA, Rimili V, Wellmann S, Knorr C, Reingruber B, Vogt JE, Wolfertstetter PR. External Validation of Predictive Models for Diagnosis, Management and Severity of Pediatric Appendicitis. *medRxiv*. 2024 Oct 28;2024.10.28.24316300. doi: 10.1101/2024.10.28.24316300.


Appendix_Diameter
Peritonitis
WBC_Count
CRP
Appendix_Diameter_flag
Neutrophil_Percentage
Free_Fluids
Nausea
Loss_of_Appetite
Ketones_in_Urine
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
SHAP value (impact on model output)
High
Feature value
Low


Appendix_Diameter
Peritonitis
WBC_Count
CRP
Neutrophil_Percentage
Appendix_Diameter_flag
Free_Fluids
Loss_of_Appetite
Ketones_in_Urine
Nausea
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low


CRP
Appendix_Diameter
Peritonitis
Appendix_Diameter_flag
WBC_Count
Loss_of_Appetite
Neutrophil_Percentage
Ketones_in_Urine
Body_Temperature
Nausea
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
SHAP value (impact on model output)
High
Feature value
Low

DharmaAI

dharma-ai.org

DHARMA
APPENDICITIS MODEL

NAUSEA

Select option

LOSS OF APPETITE

Select option

PERITONITIS

Select option

WBC COUNT × 10^3 (per μL)

10.0

BODY TEMPERATURE (°C)

38.0

NEUTROPHIL PERCENTAGE (%)

75.0

CRP (mg/L)

0.0

KETONES IN URINE

Select option

APPENDIX DIAMETER (MM)

CRP (mg/L)

55

KETONES IN URINE

Trace

APPENDIX DIAMETER (MM)

7

FREE FLUIDS

No

Get Prediction

Explaination

Dharma Score: 95%

High likelihood of acute appendicitis.
95% Confidence Interval: 93% - 96%
Threshold: 64%
Diagnostic Certainty: High confidence
Clinical Note: Manage in line with appendicitis protocol.

Severity Assessment:

Risk of Complications: 80%
95% Confidence Interval: 76% - 83%
High Risk of Complications

Prediction Explanation

Diagnosis Explanation

Base Value: 0.50

Feature	SHAP Value
Appendix_Diameter	0.3058
Peritonitis	0.0900
CRP	0.0501
Neutrophil_Percentage	0.0396
Nausea	0.0117
Loss_of_Appetite	0.0016
Ketones_in_Urine	0.0010
Free_Fluids	-0.0074
WBC_Count	-0.0438
Final Prediction	0.95

Complication Explanation

Base Value: 0.65

Feature	SHAP Value
CRP	0.2197
Loss_of_Appetite	0.0509
Peritonitis	0.0457
Body_Temperature	0.0241
Nausea	0.0059
Free_Fluids	-0.0014
Ketones_in_Urine	-0.0047
WBC_Count	-0.0923
Appendix_Diameter	-0.0992
Final Prediction	0.80

Download Full Report