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:

E-mail: ajungthapa11@gmail.com (AT)

Abstract

Background: Acute appendicitis remains a challenging diagnosis in pediatric populations, with high rates of misdiagnosis and negative appendectomies despite advances in imaging modalities. Current diagnostic tools, including clinical scoring systems like Alvarado and Pediatric Appendicitis Score (PAS), lack sufficient sensitivity and specificity, while reliance on CT scans raises concerns about radiation exposure, contrast hazards and sedation in children. Moreover, no established tool effectively predicts progression from uncomplicated to complicated appendicitis, creating a critical gap in clinical decision-making.

Objective: To develop and evaluate a machine learning model that integrates clinical, laboratory, and radiological findings for accurate diagnosis and complication prediction in pediatric appendicitis and to deploy this model as an interpretable web-based tool for clinical decision support.

Methods: We analyzed data from 780 pediatric patients (ages 0-18) with suspected appendicitis admitted to Children's Hospital St. Hedwig, Regensburg, between 2016 and 2021. For severity prediction, our dataset was augmented with 430 additional cases from published literature and only the confirmed cases of acute appendicitis(n=602) were used. After feature selection using statistical methods and recursive feature elimination, we developed a Random Forest model named

Dharma, optimized through hyperparameter tuning and cross-validation. Model performance was evaluated on independent test sets and compared with conventional diagnostic tools.

Results: Dharma demonstrated superior diagnostic performance with an AUC-ROC of 0.96 (±0.02 SD) in cross-validation and 0.97-0.98 on independent test sets. At an optimal threshold of 64%, the model achieved specificity of 88%-98%, sensitivity of 89%-95%, and positive predictive value of 93%-99%. For complication prediction, Dharma attained a sensitivity of 93% (±0.05 SD) in cross-validation and 96% on the test set, with a negative predictive value of 98%. The model maintained strong performance even in cases where the appendix could not be visualized on ultrasonography (AUC-ROC 0.95, sensitivity 89%, specificity 87% at the threshold of 30%).

Conclusion: Dharma is a novel, interpretable machine learning based clinical decision support tool designed to address the diagnostic challenges of pediatric appendicitis by integrating easily obtainable clinical, laboratory, and radiological data into a unified, real-time predictive framework. Unlike traditional scoring systems and imaging modalities, which may lack specificity or raise safety concerns in children, Dharma demonstrates high accuracy in diagnosing appendicitis and predicting progression from uncomplicated to complicated cases, potentially reducing unnecessary surgeries and CT scans. Its robust performance, even with incomplete imaging data, underscores its utility in resource-limited settings. Delivered through an intuitive, transparent, and interpretable web application, Dharma supports frontline providers—particularly in low- and middle-income settings—in making timely, evidence-based decisions, streamlining patient referrals, and improving clinical outcomes. By bridging critical gaps in current diagnostic and prognostic tools,

Dharma offers a practical and accessible 21st-century solution tailored to real-world pediatric surgical care across diverse healthcare contexts. Furthermore, the underlying framework and concepts of Dharma may be adaptable to other clinical challenges beyond pediatric appendicitis, providing a foundation for broader applications of machine learning in healthcare.

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

Author Summary

Accurate diagnosis of pediatric appendicitis remains challenging, with current clinical scores and imaging tests limited by sensitivity, specificity, predictive values, and safety concerns. We developed Dharma, an interpretable machine learning model that integrates clinical, laboratory, and radiological data to assist in diagnosing appendicitis and predicting its severity in children. Evaluated on a large dataset supplemented by published cases, Dharma demonstrated strong diagnostic and prognostic performance, including in cases with incomplete imaging—making it potentially especially useful in resource-limited settings for early decision-making and streamlined referrals. Available as a web-based tool, it provides real-time support to healthcare providers in making evidence-based decisions that could reduce negative appendectomies while avoiding hazards associated with advanced imaging modalities such as sedation, contrast, or radiation exposure.

Furthermore, the open-access concepts and framework underlying Dharma have the potential to address diverse healthcare challenges beyond pediatric appendicitis.

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), 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

Data Curation

This study utilizes data from a cohort of pediatric patients with suspected appendicitis who were admitted with abdominal pain to Children's Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021. The original data collection was approved by the Ethics Committee of the University of Regensburg (approval numbers: 18-1063-101, 18-1063_1-101, and 18-1063_2-101). The dataset has since been anonymized and is now publicly available for secondary use [44].

Data Description And Preprocessing

This study utilizes a dataset comprising 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)

The dependent variables with more than 30% missing data were excluded from the study, except for appendix diameter. Despite a substantial amount of missing data, this variable was retained owing to the known challenges associated with appendix visualization through ultrasonography (USG). Moreover, the diameter of the appendix holds well-established clinical significance in the pathophysiology and diagnosis of acute appendicitis.

Binary categorical variables were mapped as 0 and 1 while the ordinal categorical variables were mapped as 0,1 ,2 and 3. The missing values were imputed as 0 to simulate the real-world difficulty in obtaining complete medical data, especially in resource- limited settings.

For the primary classification task (diagnosis), the full dataset comprising 780 cases was used. For the complication prediction task, the dataset was augmented by integrating 430 suspected appendicitis cases from the dataset published by Marcinkevics et al. [11]. After removing the duplicates, the combined dataset consisted of 1,015 unique patient records. Only cases with a confirmed diagnosis of appendicitis (n = 602) were included for training and testing the severity assessment model. Treatment modality was not analyzed in the study directly, as we believe, it should be determined by the operating surgeon based on the diagnosis and the anticipated risk of complications in the diagnosed cases.

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.

For continuous variables, the Mann-Whitney U test was used to compare distributions between positive and negative classes, and rank-biserial correlation measured the strength of association. Features with a significant difference ($p < 0.001$) and a correlation coefficient > 0.40 were included in the model.

All statistical analyses of features with the targets were conducted using Python, employing libraries: Pandas, NumPy, and SciPy, on the original dataset comprising 780 cases of suspected appendicitis.

RFECV was used to identify the parameters with the greatest influence over the decision-making process. Different subsets of selected parameters were then tested independently to select the optimum combination and arrangement of features for the model.

Data mining was done independently for both the targets.

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 cross-validation was performed on 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.

We named the final random forest model **Dharma** after selecting the optimal hyperparameters and decision threshold. The model's performance was subsequently evaluated on the independent test

set using sensitivity, specificity, AUC-ROC, and other relevant metrics to ensure an unbiased assessment.

Additionally, we curated an independent test set of 218 cases by removing duplicates from the dataset of 430 cases used by Marcinkevics et al [11]. Dharma's performance was further evaluated on this fresh set. The performance of commonly used pediatric appendicitis diagnostic tools was also evaluated on this dataset.

For severity prediction, we combined our dataset with a cohort of 430 suspected appendicitis cases from Marcinkevics et al [11] and removed the duplicates, resulting in a combined dataset (n = 1015). Only the confirmed appendicitis cases (n = 602) were included for model development and testing.

The complication prediction dataset was highly imbalanced, with a positive-to-negative class ratio of 148:454. To address this, we applied a combination of oversampling and undersampling techniques—using SMOTE to increase the positive instances and NearMiss (n=1) to reduce the negative class—after splitting the dataset into a temporary set (adjusted from 122:359 to 359:179) and a test set (26:95). The temporary set was subsequently divided into a training set (262:141) and a validation set (97:38). The test set remained unaltered to preserve the original class distribution and ensure an unbiased performance evaluation. We prioritized optimizing Dharma's

sensitivity at the expense of specificity to quantify the risk of possible complications. This trade-off was considered appropriate, as missing true positives could result in severe, life-threatening outcomes like gangrenous or perforated appendicitis, whereas false positives are more acceptable given the patient is already a diagnosed case of acute appendicitis and can be managed surgically.

Model development was conducted using Scikit-learn in Python.

Explainability Analysis

To understand the decision-making process of Dharma, we analyzed the model using SHAP's Tree Explainer (SHapley Additive exPlanations Tree Explainer), a game theory-based approach that assigns contribution values (SHAP values) to each feature by comparing them to a baseline prediction before any data is introduced (base SHAP value). Each feature's SHAP value quantifies its influence on shifting the probability of a positive outcome either towards or away from this baseline.

We used SHAP's summary plot method to visualize these contributions and provide a global interpretation of how Dharma assessed suspected appendicitis cases for both diagnosis and severity prediction tasks in our test set. Additionally, we integrated SHAP values to explain individual predictions in our web app, along with the 95% confidence interval of predictions from each decision tree in our random forest providing an additional layer of uncertainty quantification. This integration is intended to assist attending physicians in exploring the reasoning behind the model's predictions thus enhancing transparency, interpretability and 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, 602 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 1 and 2..

Table 1. 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

272

273

274

Table 2. 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 3 and 4. 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 3. Diagnosis and Continuous variables (n=780)

Feature	U_statistic	P_value	Rank-Biserial correlation	Mean (Positive Class)	Mean(Negative Class)	Samples
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

287

288

289

290

Table 4. 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

Performance metrics

The Alvarado Score (AS) and Pediatric Appendicitis Score (PAS) were significantly associated with appendicitis diagnosis ($p < 0.001$), demonstrating AUC-ROC values of 0.78 and 0.72, respectively. The median scores for positive and negative cases were 7 vs. 5 for AS, and 6 vs. 4 for PAS, indicating higher scoring trends among the confirmed cases. Ultrasonography (USG) exhibited the highest diagnostic performance among the available tools, with an AUC-ROC of 0.93. However, appendix diameter measurements were unavailable in 34% of suspected cases, likely due to the inherent challenges in appendix visualization. A summary of diagnostic associations and performance metrics of these conventional tools are presented in Table 5 and 6.

Table 5. Association of Diagnostic Tools

Scoring Systems	U Statistic	P-value	Rank-Biserial Correlation	Median(Positive Class)	Median(Negative Class)
Alvarado Score	33163.5	<0.001	0.48	7	5

Pediatric Appendicitis Score	38445.5	<0.001	0.4	6	4

304

Table 6. Performance of Diagnostic tools						
Tools	**AUC-ROC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**
Alvarado Score	0.78	91%	45%	71%	77%	72%
Pediatric Appendicitis Score	0.72	48%	76%	74%	50%	59%
Ultrasonography	0.93	93%	94%	98%	79%	93%

305

306 Dharma demonstrated high diagnostic and prognostic performance in predicting appendicitis and
307 its complications. The model achieved a diagnostic AUC-ROC of 0.96 (±0.02 SD) in 10-fold
308 cross-validation. On the split and curated test sets, the AUC-ROC, specificity, positive predictive
309 value (PPV), sensitivity, and negative predictive value (NPV) of |0.97-0.98|88% -98%|93%- 99%
310 |89%- 95%| and |86%-91%| was achieved.

311 For complication prediction, Dharma achieved a sensitivity of 0.93 (±0.05 SD) in 10-fold cross-
312 validation. On the unaltered test set, the sensitivity and NPV were 96% and 98%, respectively.

313

314 The optimal hyperparameters and decision thresholds for both diagnosis and complication
315 prediction tasks are detailed in Table 7 and 8, respectively.

Table 7. Hyperparameters and Threshold for Diagnosis	
n_estimators	555
min_samples_split	12
min_samples_leaf	1
max_depth	35
class_weight	"balanced"
random_state	17
Threshold	0.64
Threshold(appendix not visualized)	0.3

Table 8. Hyperparameters and Threshold for Complication Prediction	
n_estimators	555
min_samples_split	3
min_samples_leaf	1
max_depth	33
random_state	17
Threshold	0.52

320 Performance metrics and comparative evaluations for diagnostic and severity prediction are
321 summarized in Table 9 and 10. The list of features and their arrangement is listed in Table 11.

322

323

Table 9. Dharma's Diagnostic Performance						
Dharma	**AUC-ROC**	**Specificity**	**PPV**	**Sensitivity**	**NPV**	**Accuracy**
10-fold cross- validation	0.96±0.02	-	-	92%±0.02	-	91%±0.03
Split Test Set(n=156)	0.98	88%	93%	95%	91%	93%
Curated Test Set(n=216)	0.97	98%	99%	89%	86%	93%
Appendix not visualized(n=127)	0.95	87%	74%	90%	95%	87%

324

Table 10. Dharma's Performance for Severity Prediction					
Dharma	**Sensitivity**	**NPV**	**AUC-ROC**	**Specificity**	**PPV**
10-fold cross-validation	93±0.05	-	-	-	-
Unaltered Test Set	96%	98%	0.88	64%	42%

325

326

Table 11.List of features and their arrangement	
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

SHAP Analysis

The SHAP (Shapley Additive Explanations) analysis revealed the following insights about the Dharma model's predictions:

Baseline (Base) Values

For the diagnosis task, the base value (i.e., the average model output before any feature input) was 0.499 and for the severity assessment task, the base value was 0.649.

Feature Importance

In the diagnostic predictions, the appendix diameter had the largest impact, followed by the white blood cell (WBC) count, presence of peritonitis, and C-reactive protein (CRP) levels.

In the complications prediction task, CRP exhibited the highest influence, followed by WBC count, appendix diameter, body temperature, and appetite loss. Ineterestingly, high values of appendix diameter showed both positive and negative effects on the model's complications risk assessment, indicating a complex and context-dependent relationship. This observation warrants further exploration in future studies.

Absence of Appendix diameter measurement

Notably, the absence of appendix diameter data decreased the model's probability of a positive diagnosis prediction from the base value. Subsequent inputs of other features then pushed the prediction upward to arrive at the final output.

Visual Summary

The SHAP summary dot plots illustrating the global feature importance for the diagnosis and severity prediction tasks are presented in S1 and S2 Figs. for the diagnosis task and in S3 Fig. for the severity assessment task.

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 our study, the Alvarado Score (AS) and Pediatric Appendicitis Score (PAS) demonstrated extremely weak abilities to differentiate between positive and negative cases, with AUC-ROC values of 0.75 and 0.71 respectively. As expected, the median AS and PAS scores were significantly higher among confirmed positive cases compared to negative ones (7 vs 5 for AS and 6 vs 4 for PAS) [$p<0.001$]. The extremely low specificity of these systems, 45% and 76% for AS and PAS respectively, as also reported by Sefa Sağ et al. [6] (14.3%-57.1%), contributes to a high rate of false-positive diagnoses, ultimately leading to unnecessary negative appendectomies (NA).

Ultrasonography (USG) demonstrated strong diagnostic performance in , with an AUC-ROC of 0.94 and using a 6 mm appendix diameter cutoff, sensitivity and specificity of 93%, and 94% respectively. However, the appendix was not visualized in 35% of the cases. This limitation aligns with findings by Duncan et al. [40] and Jenny et al. [41], who reported an inability to visualize the appendix in 71%-76% of the suspected appendicitis cases. The failure to visualize the appendix may be attributed to variable anatomical positioning, patient-specific anthropometric differences,

372 or the operator-dependent nature of USG [13]. Moreover, advanced findings, other than the
373 diameter of the appendix like peri-appendiceal fluid collection (PALC), peri-appendiceal fat
374 stranding, target sign, hyperemia on color doppler, free fluids can only be reliably estimated by
375 an experienced radiologist, which may not be easily available in resource-limited primary care
376 settings. Additionally, USG findings should also be correlated with clinical and laboratory findings
377 to formulate a complete diagnosis [43].

378

379 Although advanced imaging modalities like CT and MRI provide high diagnostic accuracy for
380 appendicitis [14-16,18], the risks associated with sedation, contrast use, and high-dose ionizing
381 radiation [17] make these modalities less favorable for pediatric populations. Moreover, these
382 modern imaging tools are often inaccessible in the primary and secondary healthcare settings,
383 particularly in resource-limited regions like ours. Furthermore, like other diagnostic tools, they do
384 not effectively distinguish between uncomplicated and complicated appendicitis, which is crucial
385 for guiding optimal management [20].

386

387 Significant progress has been made in applying machine learning to the diagnosis, severity
388 stratification, and management of pediatric appendicitis [11,24–27]. However, many of these
389 studies are technically complex and AI-centric, making their outputs difficult for clinicians to
390 interpret and integrate into practice. Unlike traditional scoring systems, these machine learning
391 models also require a dedicated, easily accessible interface to be of use in real-life clinical
392 scenarios. For instance, the image-based approach by Byun et al. [26] is computationally intensive

393 and difficult to implement in real-time clinical settings. While Marcinkevics et al. [11] attempted
394 to improve accessibility through an online tool, their model relies heavily on complex
395 ultrasonographic features that require expert radiological interpretation. Moreover, it includes
396 variables that are not directly involved in the model's decision-making process, limiting its utility
397 in primary care or resource-constrained emergency departments. Consequently, this tool remains
398 better suited for academic research than for frontline clinical use.

399

400 To address the diagnostic challenges associated with pediatric appendicitis, we present a novel
401 Machine Learning framework, Dharma, designed to guide attending physicians in making optimal
402 management decisions based on diagnosis and anticipated risk of complications. Dharma assesses
403 suspected cases of acute appendicitis by analyzing a comprehensive pattern of clinical features—
404 mirroring the approach of an attending pediatric surgeon. These include symptoms such as nausea
405 or vomiting and loss of appetite; signs like peritonitis; physiological parameters such as body
406 temperature; laboratory findings including white blood cell (WBC) count, percentage of
407 neutrophils, C-reactive protein (CRP), and urinary ketones; as well as radiological findings such
408 as appendiceal diameter and the presence of free fluid on ultrasonography. These features are
409 already well-established in existing literature for the diagnosis and risk stratification of acute
410 appendicitis [32-37].

411

412 In our study, Dharma achieved consistently high diagnostic performance, with AUC-ROC values
413 of 0.96 ± 0.02, 0.98, and 0.97 across 10-fold cross-validation and two independent test sets. This

414 demonstrates a robust discriminatory ability to differentiate between positive and negative
415 appendicitis cases, regardless of threshold. Furthermore, this consistent performance of Dharma
416 across varied clinical presentations underscores its generalizability and strong pattern recognition
417 capabilities.

418

419 At a threshold of 64%, Dharma achieved a high specificity and positive predictive value (PPV) of
420 88%-98% and 93%-99%, respectively, alongside an equally strong sensitivity of 89%-95%. This
421 balance, also supported by the base SHAP value of 0.499, makes Dharma an ideal diagnostic tool,
422 minimizing false positives and reducing the rate of negative appendectomies (NA) while
423 maintaining an extremely low risk of false negatives. Such performance characteristics allow
424 Dharma to provide reliable objective evidence to support clinical suspicion of acute appendicitis,
425 ideally following ultrasonography and before exposing pediatric patients to the potential hazards
426 of CT imaging and sedation. This model offers a viable strategy for reducing CT utilization,
427 particularly in settings where its use is limited by accessibility or concerns about radiation sedation,
428 contrast or radiation hazards.

429

430 Even in the test cases where the appendix measurements were unavailable, Dharma demonstrated
431 strong diagnostic performance, achieving an AUC-ROC of 0.95, misclassifying only 16 out of 127
432 cases (12.5%). High sensitivity and specificity of 89% and 87% respectively, at a threshold of
433 30%, makes it a reliable tool for affirming clinical suspicion of a positive diseased state even with
434 missing appendix diameter value. Additionally, the excellent negative predictive value (NPV) of

95% underscores the model's ability in ruling-out appendicitis and prompting consideration of alternative diagnoses. In this context, Dharma may be particularly beneficial in resource-limited rural and semi-urban primary-to-secondary care settings of low- and middle-income countries(LMICs), supporting early decision-making and guiding referrals. It could also prove valuable in oversaturated tertiary-care emergency departments (EDs) of LMICs, offering timely support to clinical assessments even when USG is unavailable or yields inconclusive results. Additionally, even in well-equipped centers, Dharma could assist in evaluating probable cases with ambiguous imaging findings. This usability of Dharma in diverse clinical scenarios makes it a reliable clinical decision-support tool, in both resource-limited and resource-sufficient settings.

In addition to diagnosis, Dharma is trained to estimate the risk of progression from uncomplicated to complicated appendicitis. With a sensitivity of 0.93 (±0.05 SD) and 96% in cross-validation and test sets, along with a high negative predictive value (NPV) of 98%, Dharma can reliably exclude complications in confirmed appendicitis cases. This feature equips attending surgeons with evidence-based guidance for deciding between surgical and conservative management, optimizing patient outcomes and potentially reducing unnecessary interventions in uncomplicated cases [45]. Similarly, in centers with established surgical management guidelines, Dharma's risk stratification ability through severity assessment may prove useful in scheduling emergency operation theatres as per the urgency of need of surgical care.

The severity assessment dataset (n=602) exhibited significant class imbalance, with a heavy predominance of uncomplicated cases (75%). Training the model on this asymmetry introduced a bias towards negative predictions, increasing the risk of false negatives—i.e., missing possible

complications. To address this, we applied a combination of oversampling using the Synthetic Minority Oversampling Technique (SMOTE) and undersampling via the NearMiss (n=1) algorithm, adjusting the class distribution of the combined training and validation sets from 122:359 to 359:179 in favor of the positive (complicated) class. Training Dharma on this synthetically generated imbalance towards the positive class, reversed its bias, improving the model's ability to identify almost all the possible complicated cases. The base SHAP value of 0.65 for positive-class confirms this bias adjustment. By prioritizing sensitivity, the model minimizes false negative instances and effectively flags almost all of the potentially complicated cases, making it an appropriate screening tool for severity grading in confirmed appendicitis patients.

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.

Dharma addresses the limitations of similar machine learning models in existing literature by providing not only superior diagnostic and prognostic performance [11,24–31], but also a practical,

479 interpretable, and easily accessible web-based decision support tool tailored for real-world clinical
480 environments. Our web app takes patient features, analyzes the case, and provides a Dharma
481 Score—calculated as the mean probability of a positive disease state across all 555 decision trees
482 in Dharma's diagnostic random forest. For severity assessment, the risk of complications is
483 similarly estimated by the trained 555 prognostic trees. In addition, the app provides the probability
484 range within a 95% confidence interval, along with SHAP values for each feature to explain the
485 prediction confidence and the reasoning behind the model's decisions. This prevents the models
486 from being mere "black boxes" and ensures that their results are explained in a way that is easy to
487 understand and interpret by any attending physician.

488

489 Nevertheless, our study had certain limitations that should be acknowledged. A major limitation
490 was the lack of access to reliable, high-quality medical datasets, which significantly hampers
491 progress in leveraging statistical and machine-learning tools—tools we view as critical for
492 advancing 21st-century medicine [23]. We strongly advocate for the open sharing of anonymized
493 medical datasets to support broader, more impactful research. Additionally, as our model was
494 developed using single-center data, its adaptability to other populations is unknown. The absence
495 of exclusively histologically confirmed appendicitis cases may have constrained the model's
496 ability to optimize pattern recognition for both diagnostic and prognostic tasks. Future work will
497 prioritize multicenter validation—both prospective and retrospective—including studies
498 conducted in low- and middle-income countries (LMICs) to rigorously evaluate the real-world
499 clinical utility and generalizability of Dharma across diverse healthcare settings. We also

encourage fellow researchers and med-tech enthusiasts to conduct independent validation studies on Dharma.

Conclusion

In this study, we developed Dharma, a novel machine learning–based clinical decision support tool that integrates readily available clinical, laboratory, and radiological findings into a unified predictive framework. By offering real-time, evidence-based diagnostic and prognostic assistance through an easily accessible and interpretable web-based tool, Dharma aims to reduce diagnostic uncertainty and facilitate timely surgical decision-making in pediatric appendicitis. With its high diagnostic specificity and positive predictive value demonstrated during testing, Dharma may also help reduce the rate of negative appendectomies—an important clinical and economic consideration in pediatric surgical care—without the hazards associated with CT or sedation.

Furthermore, Dharma provides a practical, data-driven solution to this common and critical pediatric surgical challenge, particularly in low-resource settings where imaging and subspecialty access may be limited or unavailable. By leveraging readily available clinical and laboratory data, the model can support frontline providers in making early and accurate decisions. In doing so, Dharma may improve pediatric health outcomes by reducing diagnostic errors and guiding timely referrals.

Thus, Dharma offers a reliable, transparent, interpretable, and evidence-based clinical decision support system for the common yet challenging clinical scenario of pediatric appendicitis. We also encourage others to build upon the concepts and framework we have developed, which we believe are adaptable to a wide range of healthcare challenges beyond pediatric appendicitis.

In alignment with open-access principles, we have made the source code publicly available at GitHub Repository Link [46] and the web-based tool accessible at [dharma-ai.org], promoting transparency, reproducibility, and continuous advancement in clinical decision support and medicine as a whole through knowledge-sharing.

Acknowledgement

We thank Marcinkevics R and colleagues [11] for their inspiring work and for publicly providing anonymized data that facilitated our study. We also appreciate the contributions of our colleague Amrit Neupane for designing the user interface and user experience of our web tool.

Author Contributions

Conceptualization and study design: Anup Thapa, Subash Pahari

Data curation: Anup Thapa, Subash Pahari

Literature review: Anup Thapa

537 Formal analysis: Anup Thapa

538 Methodology: Anup Thapa

539 Software: Anup Thapa (model development), Subash Pahari (web-app development)

540 Writing – original draft: Anup Thapa (except introduction), Sashank Timilsina, Binay Chapagain
541 (introduction)

542 Writing – review & editing: Anup Thapa

543 Writing – Manuscript Formating: Subash Pahari

544

545 References

546

547 [1] Hijaz NM, Friesen CA. Managing acute abdominal pain in pediatric patients: current
548 perspectives. Pediatric Health Med Ther. 2017 June 29;8:83-91. doi: 10.2147/PHMT.S120156.
549 PMID: 29388612; PMCID: PMC5774593.

550 [2] Tseng YC, Lee MS, Chang YJ, Wu HP. Acute abdomen in pediatric patients admitted to the
551 pediatric emergency department. Pediatr Neonatol. 2008 Aug;49(4):126-34. doi: 10.1016/S1875-
552 9572(08)60027-3. PMID: 19054918.

553 [3] Schneider C, Kharbanda A, Bachur R. Evaluating appendicitis scoring systems using a
554 prospective pediatric cohort. Ann Emerg Med. 2007 Jun;49(6):778-84, 784.e1. doi:
555 10.1016/j.annemergmed.2006.12.016. Epub 2007 Mar 26. PMID: 17383771.

556 [4] Sencan A, Aksoy N, Yıldız M, Okur Ö, Demircan Y, Karaca I. The evaluation of the validity
557 of Alvarado, Eskelinen, Lintula and Ohmann scoring systems in diagnosing acute appendicitis in
558 children. Pediatr Surg Int. 2014 Mar;30(3):317-21. doi: 10.1007/s00383-014-3467-0. PMID:
559 24448910.

560 [5] Pogorelić Z, Rak S, Mrklić I, Jurić I. Prospective validation of Alvarado score and Pediatric
561 Appendicitis Score for the diagnosis of acute appendicitis in children. Pediatr Emerg Care. 2015
562 Mar;31(3):164-8. doi: 10.1097/PEC.0000000000000375. PMID: 25706925.

563 [6] Sağ S, Basar D, Yurdadoğan F, Pehlivan Y, Elemen L. Comparison of Appendicitis Scoring
564 Systems in Childhood Appendicitis. Turk Arch Pediatr. 2022 Sep;57(5):532-537. doi:
565 10.5152/TurkArchPediatr.2022.22076. PMID: 36062441; PMCID: PMC9524470.

566 [7] Jukić M, Nizeteo P, Matas J, Pogorelić Z. Trends and Predictors of Pediatric Negative
567 Appendectomy Rates: A Single-Centre Retrospective Study. Children (Basel). 2023 May
568 15;10(5):887. doi: 10.3390/children10050887. PMID: 37238435; PMCID: PMC10217643.

569 [8] Kundiona I, Chihaka OB, Muguti GI. Negative appendicectomy: evaluation of ultrasonography
570 and Alvarado score. Cent Afr J Med. 2015 Sep-Dec;61(9-12):66-73. PMID: 29144064.

571 [9] Oyetunji TA, Ong'uti SK, Bolorunduro OB, Cornwell EE 3rd, Nwomeh BC. Pediatric negative
572 appendectomy rate: trend, predictors, and differentials. J Surg Res. 2012 Mar;173(1):16-20. doi:
573 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.

595 [16] Stephen AE, Segev DL, Ryan DP, Mullins ME, Kim SH, Schnitzer JJ, Doody DP. The
596 diagnosis of acute appendicitis in a pediatric population: to CT or not to CT. J Pediatr Surg. 2003
597 Mar;38(3):367-71; discsussion 367-71. doi: 10.1053/jpsu.2003.50110. PMID: 12632351.

598 [17] Brennan GD. Pediatric appendicitis: pathophysiology and appropriate use of diagnostic
599 imaging. CJEM. 2006 Nov;8(6):425-32. doi: 10.1017/s1481803500014238. PMID: 17209492.

600 [18] Mittal MK. Appendicitis: Role of MRI. Pediatr Emerg Care. 2019 Jan;35(1):63-66. doi:
601 10.1097/PEC.0000000000001710. PMID: 30608328.

602 [19] Pines JM. Trends in the rates of radiography use and important diagnoses in emergency
603 department patients with abdominal pain. Med Care. 2009 Jul;47(7):782-6. doi:
604 10.1097/MLR.0b013e31819748e9. PMID: 19536032.

605 [20]Lastunen K, Leppäniemi A, Mentula P. Perforation rate after a diagnosis of uncomplicated
606 appendicitis on CT. BJS Open. 2021 Jan 8;5(1):zraa034. doi: 10.1093/bjsopen/zraa034. PMID:
607 33609386; PMCID: PMC7893470.

608 [21] Almaramhy HH. Acute appendicitis in young children less than 5 years: review article. Ital J
609 Pediatr. 2017 Jan 26;43(1):15. doi: 10.1186/s13052-017-0335-2. PMID: 28257658; PMCID:
610 PMC5347837.

611 [22] Pogorelić Z, Domjanović J, Jukić M, Poklepović Peričić T. Acute Appendicitis in Children
612 Younger than Five Years of Age: Diagnostic Challenge for Pediatric Surgeons. Surg Infect
613 (Larchmt). 2020 Apr;21(3):239-245. doi: 10.1089/sur.2019.175. Epub 2019 Oct 16. PMID:
614 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.

654 [33] Adir A, Braester A, Natalia P, Najib D, Akria L, Suriu C, Masad B, Igor W. The role of blood
655 inflammatory markers in the preoperative diagnosis of acute appendicitis. Int J Lab Hematol. 2024
656 Feb;46(1):58-62. doi: 10.1111/ijlh.14163. Epub 2023 Aug 29. PMID: 37644670.

657 [34] Moris D, Paulson EK, Pappas TN. Diagnosis and Management of Acute Appendicitis in
658 Adults: A Review. JAMA. 2021 Dec 14;326(22):2299-2311. doi: 10.1001/jama.2021.20502.
659 PMID: 34905026.

660 [35] Yajima H, Fujita T, Yanaga K. Profile of signs and symptoms in mild and advanced acute
661 appendicitis. Int Surg. 2010 Jan-Mar;95(1):63-6. PMID: 20480844.

662 [36] Hajibandeh S, Hajibandeh S, Hobbs N, Mansour M. Neutrophil-to-lymphocyte ratio predicts
663 acute appendicitis and distinguishes between complicated and uncomplicated appendicitis: A
664 systematic review and meta-analysis. Am J Surg. 2020 Jan;219(1):154-163. doi:
665 10.1016/j.amjsurg.2019.04.018. Epub 2019 Apr 27. PMID: 31056211.

666 [37] Biehl CM, Elliver M, Gudjonsdottir J, Salö M. Utility of Urine Dipstick Testing in Pediatric
667 Appendicitis: Assessing its Role in Identifying Complicated Cases and Retrocecal Appendicitis.
668 Eur J Pediatr Surg. 2024 Dec 19. doi: 10.1055/a-2490-1156. Epub ahead of print. PMID:
669 39701137.

670 [38] Singh, C. Machine Learning in Pattern Recognition. European Journal of Engineering and
671 Technology Research. 2023 Apr;8(2):63–68. doi:10.24018/ejeng.2023.8.2.3025.

672 [39] Issaiy M, Zarei D, Saghazadeh A. Artificial Intelligence and Acute Appendicitis: A
673 Systematic Review of Diagnostic and Prognostic Models. World J Emerg Surg. 2023 Dec
674 19;18(1):59. doi: 10.1186/s13017-023-00527-2. PMID: 38114983; PMCID: PMC10729387.

675 [40] Scrimgeour DS, Driver CP, Stoner RS, King SK, Beasley SW. When does ultrasonography
676 influence management in suspected appendicitis? ANZ J Surg. 2014 May;84(5):331-4. doi:
677 10.1111/ans.12415. Epub 2014 Jan 9. PMID: 24405944.

678 [41] Held JM, McEvoy CS, Auten JD, Foster SL, Ricca RL. The non-visualized appendix and
679 secondary signs on ultrasound for pediatric appendicitis in the community hospital setting. Pediatr
680 Surg Int. 2018 Dec;34(12):1287-1292. doi: 10.1007/s00383-018-4350-1. Epub 2018 Oct 6. PMID:
681 30293146.

682 [42] Nijssen DJ, van Amstel P, van Schuppen J, Eeftinck Schattenkerk LD, Gorter RR, Bakx R.
683 Accuracy of ultrasonography for differentiating between simple and complex appendicitis in
684 children. Pediatr Surg Int. 2021 Jul;37(7):843-849. doi: 10.1007/s00383-021-04872-8. Epub 2021
685 Mar 7. PMID: 33677613; PMCID: PMC8172400.

686 [43] Mills LD, Mills T, Foster B. Association of clinical and laboratory variables with ultrasound
687 findings in right upper quadrant abdominal pain. South Med J. 2005 Feb;98(2):155-61. doi:
688 10.1097/01.SMJ.0000129927.88863.65. PMID: 15759944.

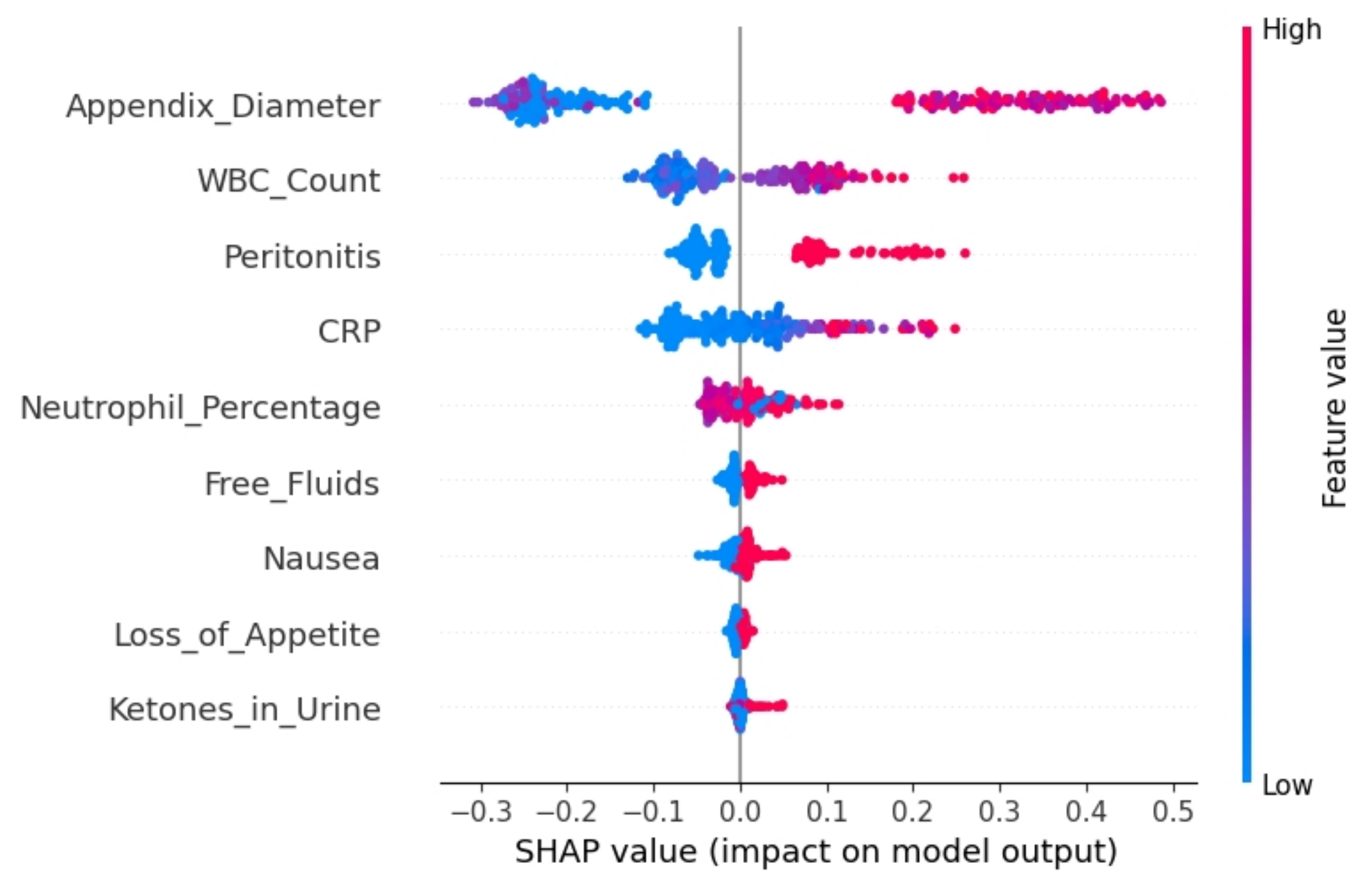
689 [44] Marcinkevičs R, Reis Wolfertstetter P, Klimiene U, Ozkan E, Chin-Cheong K, Paschke A,
690 et al. Regensburg Pediatric Appendicitis Dataset. Zenodo; 2023.

[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

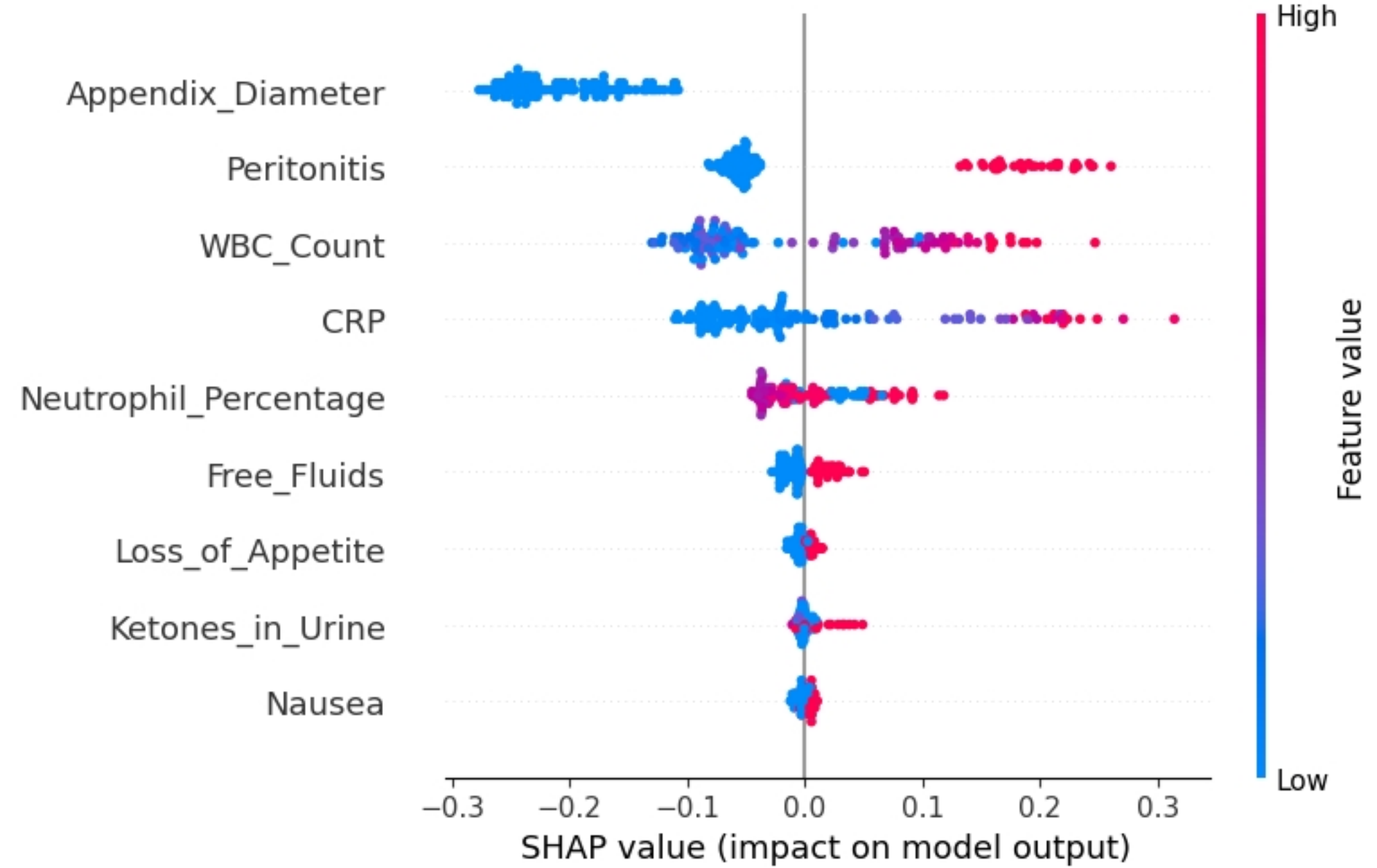
[46] Thapa A. Dharma-AppendicitisModel [Internet]. 2025 [cited 2025 May 26]. Available from: https://github.com/ajung17/Dharma-AppendicitisModel

710

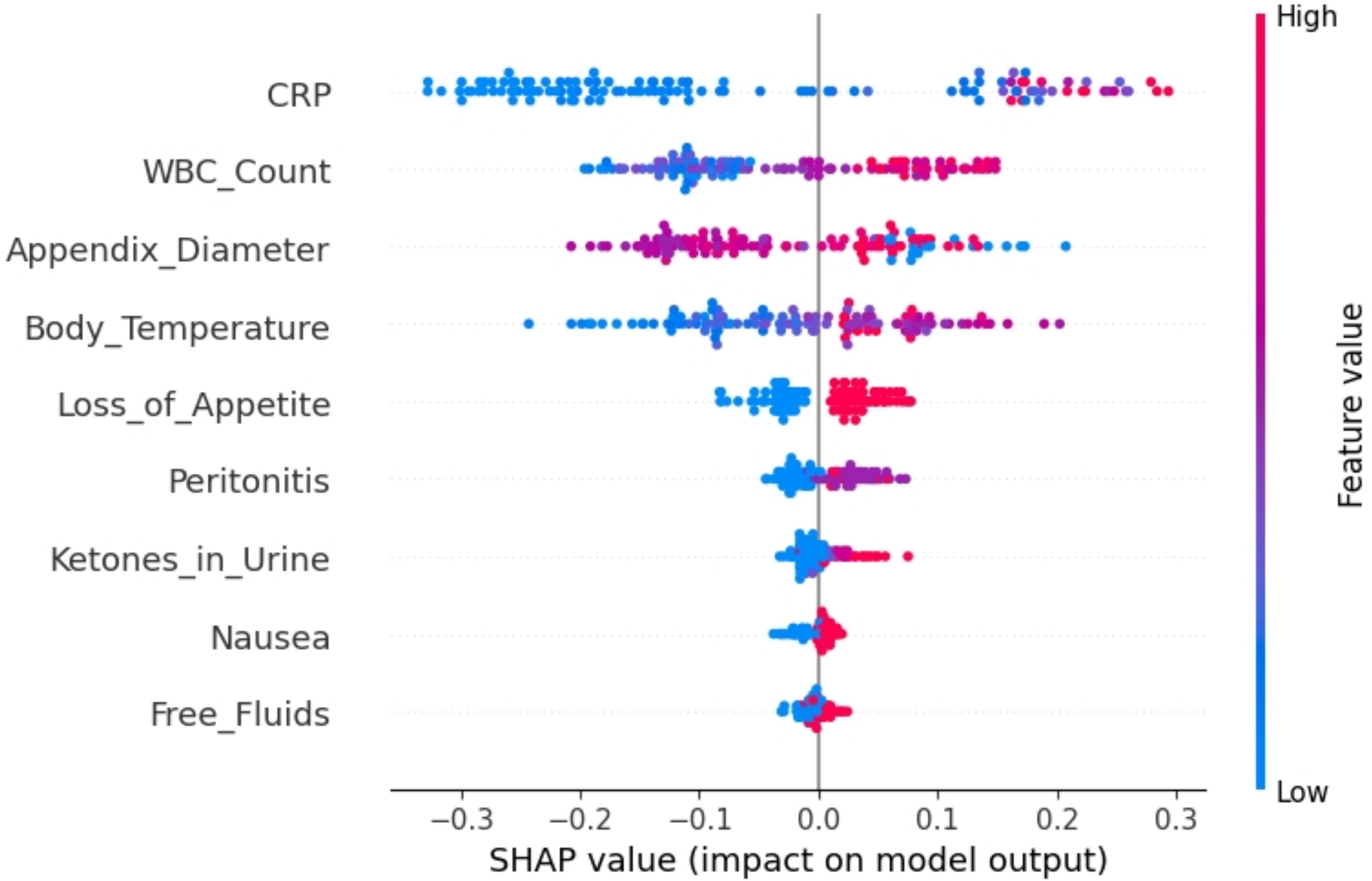
711


Appendix_Diameter
WBC_Count
Peritonitis
CRP
Neutrophil_Percentage
Free_Fluids
Nausea
Loss_of_Appetite
Ketones_in_Urine
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
SHAP value (impact on model output)
High
Feature value
Low

Figure


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

Figure


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

Figure