Frontiers in Pediatrics

TYPE Original Research
PUBLISHED 29 August 2025
DOI 10.3389/fped.2025.1587488

External validation of predictive models for diagnosis, management and severity of pediatric appendicitis

Ričards Marcinkevičs[1†], Kacper Sokol[1*†], Akhil Paulraj[1§], Melinda A. Hilbert[2§], Vivien Rimili[3§], Sven Wellmann[3,4], Christian Knorr[5§], Bertram Reingruber[2], Julia E. Vogt[1‡] and Patricia Reis Wolfertstetter[3,5*‡]

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland, [2]Department of Pediatric Surgery and Pediatric Traumatology, Florence-Nightingale-Hospital Düsseldorf, Düsseldorf, Germany, [3]Faculty of Medicine, University Medical Center Regensburg, Regensburg, Germany, [4]Department of Neonatology, Hospital St. Hedwig of the Order of St. John of God, University Children's Hospital Regensburg (KUNO), Regensburg, Germany, [5]Department of Pediatric Surgery and Pediatric Orthopedics, Hospital St. Hedwig of the Order of St. John of God, University Children's Hospital Regensburg (KUNO), Regensburg, Germany

Background: Appendicitis is a common condition among children and adolescents. Machine learning models can offer much-needed tools for improved diagnosis, severity assessment and management guidance for pediatric appendicitis. However, to be adopted in practice, such systems must be reliable, safe and robust across various medical contexts, e.g., hospitals with distinct clinical practices and patient populations.
Methods: We performed external validation of models predicting the diagnosis, management and severity of pediatric appendicitis. Trained on a cohort of 430 patients admitted to the Children's Hospital St. Hedwig (Regensburg, Germany), the models were validated on an independent cohort of 301 patients from the Florence-Nightingale-Hospital (Düsseldorf, Germany). The data included demographic, clinical, scoring, laboratory and ultrasound parameters. In addition, we explored the benefits of model retraining and inspected variable importance.
Results: The distributions of most parameters differed between the datasets. Consequently, we saw a decrease in predictive performance for diagnosis, management and severity across most metrics. After retraining with a portion of external data, we observed gains in performance, which, nonetheless, remained lower than in the original study. Notably, the most important variables were consistent across the datasets.
Conclusions: While the performance of transferred models was satisfactory, it remained lower than on the original data. This study demonstrates challenges in transferring models between hospitals, especially when clinical practice and demographics differ or in the presence of externalities such as pandemics. We also highlight the limitations of retraining as a potential remedy since it could not restore predictive performance to the initial level.

KEYWORDS

artificial intelligence, machine learning, predictive modeling, medical decision support systems, evaluation

OPEN ACCESS

EDITED BY
Francesco Morini,
Sapienza University of Rome, Italy

REVIEWED BY
Juan A. Tovar,
Hospital Universitario La Paz, Spain
Dinh Tuan Phan Le,
New York City Health and Hospitals Corporation, United States

*CORRESPONDENCE
Kacper Sokol
✉ kacper.sokol@inf.ethz.ch
Patricia Reis Wolfertstetter
✉ patricia.reiswolfertstetter@barmherzige-regensburg.de

[†]These authors have contributed equally to this work and share first authorship

[‡]These authors have contributed equally to this work and share last authorship

[§]PRESENT ADDRESSES
Akhil Paulraj,
Computer Science Department, Princeton University, Princeton, NJ, United States
Melinda A. Hilbert,
Clinic for Trauma, Hand Surgery and Sports Medicine, ViDia Clinics Karlsruhe, Karlsruhe, Germany
Vivien Rimili,
Department of Pediatric Surgery and Pediatric Urology, Children's Hospital of Cologne, Cologne, Germany
Christian Knorr,
Department of Pediatric Surgery, RoMed Klinikum Rosenheim, Rosenheim, Germany

RECEIVED 04 March 2025
ACCEPTED 28 July 2025
PUBLISHED 29 August 2025

CITATION
Marcinkevičs R, Sokol K, Paulraj A, Hilbert MA, Rimili V, Wellmann S, Knorr C, Reingruber B, Vogt JE and Reis Wolfertstetter P (2025) External validation of predictive models for diagnosis, management and severity of pediatric appendicitis.
Front. Pediatr. 13:1587488.
doi: 10.3389/fped.2025.1587488

1 Introduction

Acute appendicitis is a common condition among children and adolescents treated in pediatric surgery departments due to abdominal pain (1, 2). Diagnosis relies on clinical signs and symptoms (in particular, their dynamics and progression under close observation), laboratory tests and imaging, whereas postoperative classification is based on intraoperative findings and histology (3). Scoring systems, such as Alvarado Score (AS) and Pediatric Appendicitis Score (PAS), can facilitate clinical assessment (4, 5). The classical treatment of pediatric appendicitis is surgical, although conservative treatment with antibiotics can be an option in certain cases (6–8). Additionally, spontaneous resolution of uncomplicated appendicitis has been observed and reported, which supports antibiotic-free management based on supportive care for qualifying cases (9–11).

Despite new developments and technologies, early and accurate detection, preoperative classification, and treatment strategy selection are still challenging, especially in young children (1, 12, 13). Widely used clinical and laboratory parameters alone are mostly non-specific at identifying appendicitis (14, 15). Imaging modalities are important tools to guide management and avoid negative appendectomies, but they have limitations, such as operator (investigator) dependency on ultrasonography, radiation exposure for computed tomography, and availability and feasibility of magnetic resonance imaging, not to mention the costs (3, 16).

Recent years have marked impressive progress in Machine Learning (ML) research and the increasing proliferation of tools built upon this technology in medicine. ML algorithms promise to aid in the detection, management and treatment of various diseases, thus improving the overall quality and effectiveness of healthcare. In relation to pediatric appendicitis, ML has been used to diagnose and manage patients suspected of developing this condition (17–27); specifically, such tools were developed to predict diagnosis, management and severity of pediatric appendicitis. These models either rely exclusively on standard clinical and laboratory data (17, 19–21, 27), or additionally utilize imaging modalities (obtained through various methods, e.g., computed tomography or ultrasonography) either directly in their raw format or by extracting hand-crafted annotations (18, 22–26).

Although promising and practical, ML-based tools for pediatric appendicitis are rarely deployed in practice due to the translational barrier inherent to medical machine learning research (28). To overcome this challenge, predictive models need to be validated on external datasets and later go through rigorous clinical trials (which tend to be complex, time-consuming and costly) (29). In this study we make a step in this direction and follow up on our previous work where we developed ML models (23) for predicting diagnosis (*appendicitis* vs. *no appendicitis*), treatment assignment (*surgical* vs. *conservative*) and complications (*complicated appendicitis* vs. *uncomplicated* or *no appendicitis*) of pediatric appendicitis. Specifically, we conduct a principled external validation of the aforementioned ML tools on tabular electronic health records collected in a different hospital, exploring potential challenges associated with the transfer of our predictive models.

The original models (logistic regression, random forest and gradient boosting, all achieving strong performance) were developed on a dataset of 430 patients aged 0 to 18 years admitted with abdominal pain and suspected appendicitis to the Department of Pediatric Surgery at the tertiary Children's Hospital St. Hedwig in Regensburg, Germany, over the period of 2016–2018 (18, 23). The original dataset consists of demographic, clinical, scoring, laboratory and ultrasound (US) predictor variables (see Table 1 for their list).[1] The external validation dataset was acquired at the Department of Pediatric Surgery and Pediatric Traumatology, Florence-Nightingale-Hospital, Düsseldorf, Germany. This cohort consists of 301 pediatric patients hospitalized between 2015 and 2022, and the dataset format and predictor variables adhere to the format of the Regensburg dataset. The study design is summarized schematically in Figure 1.

In this retrospective study, we present an external validation of the aforementioned models on a new and independent cohort of patients. To this end, we:

1. compare the datasets to better understand their differences (Section 3.1);
2. evaluate the models without any adaptation to test their *external validity* under real-world distribution shift (Section 3.2);
3. retrain the models, and then evaluate and compare them again to explore possible gains in performance (Section 3.2); and
4. study feature importance across the models to elucidate their functioning (Section 3.3).

Our study demonstrates the transferability of the models across hospitals and outlines the steps necessary to facilitate such a safe adaptation.

2 Material and methods

2.1 External data acquisition and description

To facilitate external validation, we collected and reviewed retrospective data from children and adolescents aged 0–17 years who were admitted to the Department of Pediatric Surgery and Pediatric Traumatology at Florence-Nightingale-Hospital in Düsseldorf with abdominal pain and suspected appendicitis from January 1st, 2015 to February 1st, 2022. Patients who had undergone an appendectomy before their admission were excluded. Similarly, we did not include subjects with chronic intestinal diseases or current antibiotic treatment if therapy was conservative. In total, 301 patients met the inclusion criteria. The study, including data acquisition and transfer, was approved by the Ethics Committee of the University of Regensburg

[1] This dataset is available at https://github.com/i6092467/pediatric-appendicitis-ml.

TABLE 1 Description of the Regensburg and Düsseldorf datasets containing summary statistics for each variable. For numerical variables, we report medians alongside interquartile ranges; categorical variables are binarized and summarized as frequencies. Additionally, we report adjusted p-values from the unpaired two-sample t-test and chi-squared test for proportions.

Feature		Regensburg $n = 430$		Düsseldorf $n = 301$		p-value
Demographic	Age [years]	11.5	[9.3, 13.9]	10.1	[7.7, 11.7]	≤ **0.001**
	Male sex [%]	53.7		58.1		0.260
	Height [cm]	150.5	[138.0, 162.9]	140.0	[128.3, 150.0]	≤ **0.001**
	Weight [kg]	42.0	[31.1, 55.0]	35.0	[26.0, 43.0]	≤ **0.001**
	Body mass index [kg/m^2]	18.1	[15.85, 21.2]	17.9	[15.8, 20.1]	≤ **0.050**
Scoring	Alvarado score [points]	6.0	[4.0, 7.0]	6.0	[5.0, 7.0]	0.054
	Pediatric appendicitis score [points]	5.0	[4.0, 6.0]	6.0	[5.0, 7.0]	≤ **0.001**
Clinical	Peritonitis [%]	38.4		64.1		≤ **0.001**
	Migration of pain [%]	25.6		46.6		≤ **0.001**
	Tenderness in right lower quadrant [%]	97.0		94.6		0.129
	Rebound tenderness [%]	34.4		43.0		≤ **0.050**
	Cough tenderness [%]	27.0		48.3		≤ **0.001**
	Psoas sign [%]	30.5		41.8		≤ **0.010**
	Nauseous/vomitting [%]	56.3		68.1		≤ **0.010**
	Anorexia [%]	29.1		68.1		≤ **0.001**
	Body temperature [°C]	37.4	[37.0, 38.2]	37.0	[36.5, 37.7]	≤ **0.001**
	Dysuria [%]	5.4		4.0		0.415
	Abnormal stool [%]	27.8		19.5		≤ **0.050**
Laboratory	White blood cell count [10^3/l]	11.9	[8.4, 15.8]	14.9	[10.3, 19.3]	≤ **0.001**
	Neutrophils [%]	74.9	[59.1, 82.9]	72.7	[59.8, 82.1]	0.293
	C-reactive protein [mg/l]	7.0	[1.0, 31.3]	19.0	[5.0, 58.0]	≤ **0.010**
	Ketones in urine [%]	38.4		53.7		≤ **0.001**
	Erythrocytes in urine [%]	22.1		33.9		≤ **0.010**
	White blood cells in urine [%]	12.4		17.0		0.153
Ultrasound	Visibility of appendix [%]	64.5		24.3		≤ **0.001**
	Appendix diameter [mm]	7.3	[6.0, 9.1]	9.0	[7.0, 12.0]	≤ **0.001**
	Free intraperitoneal fluid [%]	43.6		25.2		≤ **0.001**
	Irregular appendix layers [%]	35.9		7.2		≤ **0.001**
	Target sign [%]	46.0		30.8		≤ **0.010**
	Appendix perfusion [%]	65.5		–		–
	Surrounding tissue reaction [%]	71.7		16.4		≤ **0.001**
	Pathological lymph nodes [%]	68.5		2.7		≤ **0.001**
	Mesenteric lymphadenitis [%]	80.4		6.6		≤ **0.001**
	Thickening of the bowel wall [%]	40.9		11.9		≤ **0.001**
	Ileus [%]	14.5		0.0		≤ **0.001**
	Coprostasis [%]	37.8		4.8		≤ **0.001**
	Meteorism [%]	72.9		26.4		≤ **0.001**
	Enteritis [%]	46.3		0.0		≤ **0.001**
Response	Appendicitis [%]	57.2		76.2		≤ **0.001**
	Surgical management [%]	38.4		80.5		≤ **0.001**
	Complicated appendicitis [%]	11.9		16.4		0.091

For significant differences, p-values are given in bold.

(18-1063-101, 18-1063-3-101) and was performed in accordance with the relevant guidelines and regulations.

In terms of management, the cohort was divided into conservative and operative groups. Patients admitted and receiving supporting therapy, e.g., intravenous fluids, enemas and analgesics, with clinical improvement without surgery were classified as conservative. Otherwise, having undergone an appendectomy, subjects were labeled as operative. For the surgical group, histological findings were recorded. In the case of negative appendectomy (i.e., histology presenting a normal appendix without inflammation), the corresponding data were used to predict the diagnosis and severity but not the management.

As in the prior study (23), diagnosis (*appendicitis* vs. *no appendicitis*) was assessed for all included patients. For patients treated surgically, appendicitis diagnosis was based on histology. In the nonoperative group, patients were classified retrospectively as having appendicitis if their AS or PAS were at least 4, combined with an appendix diameter of $\geq$6 mm. Conservatively treated patients classified as having appendicitis were followed up after discharge for recurrences. Patients who had a recurrence and underwent secondary operation were relabeled as surgical in the analysis. The follow-up telephone interview was performed at least one year after discharge, between January 2023 and February 2024. Informed consent was obtained from the parents or legal representatives of the patients who underwent the follow-up.

Furthermore, appendicitis severity was assessed. Patients treated non-operatively, both with and without appendicitis, with no recurrence during the follow-up period were classified as

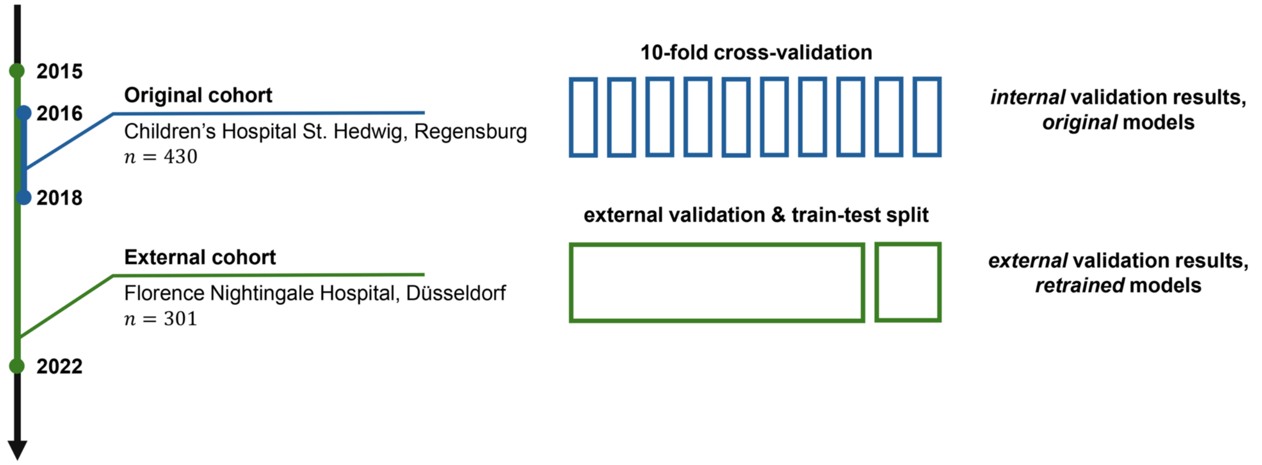
2015
2016
2018
2022
Original cohort
Children's Hospital St. Hedwig, Regensburg
$n = 430$
External cohort
Florence Nightingale Hospital, Düsseldorf
$n = 301$
10-fold cross-validation
external validation & train-test split
internal validation results, *original* models
external validation results, *retrained* models

FIGURE 1
An overview of the study design. The original predictive models were trained and validated on the cohort of patients ($n = 430$) from Regensburg, Germany (23). This article presents the results of the external validation on another cohort ($n = 301$) from Düsseldorf, Germany. In particular, in this study, we validate the original models on the external data and retrain them to assess the potential for improvement.

uncomplicated. For patients treated operatively, classification was based on the histology: *simple/uncomplicated* (subacute/catharral/chronic, phlegmonous) or *complicated* (abscess, perforation).

During the exploratory analysis presented below, we compute summary statistics across both datasets—the original from Regensburg and the external from Düsseldorf—and perform hypothesis tests for the differences between internal and external data. Specifically, we report medians and interquartile ranges (IQR) for numerical attributes and frequencies for categorical features. For statistical analysis, we utilize the unpaired two-sample t-test and Pearson's chi-squared test for the equality of proportions. We adjust the resulting p-values for multiple comparisons to control the false discovery rate using the Benjamini-Hochberg procedure (30) at the level $q = 0.05$.

2.2 Original predictive models

We leverage the dataset from the Florence-Nightingale-Hospital, Düsseldorf, for the external validation of the predictive models developed on the Regensburg cohort (23). The original analysis (23) was concerned with predicting three response variables: (i) diagnosis (*appendicitis* vs. *no appendicitis*), (ii) management (*surgical* vs. *conservative*), and (iii) severity (*complicated appendicitis* vs. *uncomplicated* or *no appendicitis*). In particular, logistic regression (LR), random forest (RF) (31) and gradient boosting (GB) (32) models were trained on the dataset of 430 patients with 38 predictor variables.

In the current study, we train these models on the *full* Regensburg cohort, replicating the original R programming language code (23, 33) in the Python programming language (v3.11.9) using the scikit-learn library (v1.4.2). We use hyperparameter configurations and perform preprocessing steps similar to those described in the original study (23), imputing missing values with the k-nearest neighbors algorithm (with $k = 5$). Note that we limit our analysis to models trained on the full set of features and we do not consider ablations with feature selection or without the US-related variables.

2.3 Model retraining

In addition to the purely external validation, we retrain the predictive models on a combination of the Regensburg and Düsseldorf cohorts, building the models on the 100% of the Regensburg and 80% of the Düsseldorf data. In this setting, we test the models on the remaining, withheld 20% of the external dataset (the data were split at random). We conduct this experiment to gauge the possible benefits of a multicenter cohort approach and to better understand if the predictive performance improves with the inclusion of external data points in the training set.

2.4 Evaluation

For both original and retrained model evaluation, we report the area under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves. Additionally, we investigate the tradeoffs among sensitivity, specificity as well as positive (PPV) and negative (NPV) predictive values by varying the threshold applied to the classifiers' output. Lastly, to better understand the models' predictions, we compute the permutation feature importance (31) of predictor variables using the test set.

3 Results

3.1 External dataset

Both of the datasets investigated in this study are overviewed in Table 1. Therein we report summary statistics for all the variables observed across the Regensburg ($n = 430$) and Düsseldorf ($n = 301$) cohorts. Additionally, we provide the adjusted results of statistical hypothesis tests for the differences in means and proportions of values respectively for numerical and categorical variables.

We observe significant differences across the distribution of most variables. Generally, subjects from the Düsseldorf cohort are younger and exhibit a higher frequency of clinical examination findings. Similarly, the external data exhibit overall higher laboratory parameter values for the variables correlated with appendicitis. Despite this, we observe no statistically significant difference in the neutrophil percentage, likely due to the high rate of missing values for this predictor in the external dataset (see Figure 4).

Furthermore, the Düsseldorf cohort has a lower frequency of positive US findings. We attribute this trend to the higher rate of missing values for relevant variables in the Regensburg dataset (refer to Figure 4) and the fact that the summary statistics shown in Table 1 have been calculated only across the non-missing entries without imputation. By contrast, the reported appendix diameter is significantly larger for the external dataset subjects. Lastly, it is worth noting that the information about the appendix perfusion is entirely missing in the Düsseldorf dataset.

The datasets also differ considerably in two of the response variables: diagnosis and management. The Düsseldorf cohort has a significantly higher prevalence of appendicitis cases (76.2% vs. 57.2%) with, consequently, more patients managed surgically (80.5% vs. 38.4%). While the external dataset has a higher prevalence of complicated appendicitis cases (16.4% vs. 11.9%), this difference is not statistically significant.

In summary, the external dataset from Düsseldorf and the original dataset from Regensburg exhibit statistically significant differences with regard to the distribution of the majority of the observed variables (consult Table 1), including the response variables. Moreover, the frequency of missing values also varies across the cohorts (refer to Figure 4). These dissimilarities potentially pose challenges for the generalization of predictive models across institutions.

3.2 Predictive performance

We now turn to the external validation of the ML models. Table 2 contains AUROC and AUPR measurements for predicting the diagnosis, management and severity of appendicitis on the Regensburg and Düsseldorf datasets. The results for the Regensburg cohort are taken from the original work (23) and were obtained by 10-fold cross-validation. When validating on the Düsseldorf data, we assess the variability in performance using bootstrapping. For reference, we additionally include the expected metric values for a fair coin flip (*random*), which serve as our baselines.

For the models trained exclusively on the Regensburg data (*original*), we observe a sizable decrease in the average AUROC for the diagnosis and management when evaluating on the external dataset. For example, the AUROC of the random forest model decreases from 96% to 85% for the diagnosis and from 94% to 85% for the management. In contrast the external AUPR is comparable to the one from the internal validation for these response variables. For the severity, we observe a larger overall decrease in both metrics. For instance, for the random forest, the AUROC decreases from 90% to 75%, and the AUPR drops from 70% to 45%.

Additionally, we explore the tradeoff between the sensitivity, specificity, PPV and NPV while varying the value of the threshold applied to the classifiers' output. We focus our analysis exclusively on the random forest model as it exhibits the most balanced performance across all the response variables for both datasets. These findings are summarized in Figure 2. For the diagnosis and management targets, using the threshold value of 0.50 explored in the original analysis (23), we observe a deterioration in the classifiers' sensitivity, specificity and NPV. For the severity target, by contrast, there is a decline in sensitivity and PPV. Arguably, these changes may be related to the prevalence shift (34) described in Section 3.1 and suggest the necessity for the threshold and model recalibration.

To verify if the models' performance improves after including a portion of the Düsseldorf data in the training set, we retrain all the

TABLE 2 Validation results for the logistic regression (LR), random forest (RF) and gradient boosting (GB) models predicting the diagnosis, management and severity of appendicitis. The results on the Regensburg dataset are copied from the original study (23), which conducted 10-fold cross-validation. For the Düsseldorf data, we report averages and standard deviations obtained by bootstrapping for the models trained exclusively on the Regensburg cohort (*original*) and retrained on both cohorts (*retrained*). The predictive performance is assessed with the areas under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves.

Dataset	Model	Diagnosis		Management		Severity	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Regensburg	Random (23)	0.50	0.43	0.50	0.38	0.50	0.12
	Original LR (23)	0.91 ± 0.04	0.88 ± 0.07	0.90 ± 0.04	0.88 ± 0.06	0.82 ± 0.13	0.53 ± 0.26
	Original RF (23)	0.96 ± 0.01	0.94 ± 0.03	0.94 ± 0.02	0.92 ± 0.05	0.90 ± 0.08	0.70 ± 0.17
	Original GBM (23)	0.96 ± 0.01	0.94 ± 0.03	0.94 ± 0.02	0.93 ± 0.04	0.90 ± 0.07	0.64 ± 0.21
Düsseldorf	Random	0.50	0.76	0.50	0.81	0.50	0.16
	Original LR	0.80 ± 0.04	0.92 ± 0.02	0.80 ± 0.04	0.94 ± 0.02	0.70 ± 0.06	0.34 ± 0.08
	Original RF	0.85 ± 0.03	0.95 ± 0.01	0.85 ± 0.03	0.96 ± 0.01	0.75 ± 0.04	0.45 ± 0.07
	Original GBM	0.83 ± 0.03	0.94 ± 0.02	0.82 ± 0.03	0.95 ± 0.01	0.72 ± 0.04	0.40 ± 0.07
	Retrained LR	0.84 ± 0.08	0.94 ± 0.04	0.83 ± 0.08	0.95 ± 0.03	0.74 ± 0.12	0.45 ± 0.17
	Retrained RF	0.87 ± 0.06	0.96 ± 0.02	0.83 ± 0.08	0.95 ± 0.03	0.75 ± 0.11	0.49 ± 0.17
	Retrained GBM	0.86 ± 0.07	0.95 ± 0.03	0.82 ± 0.09	0.95 ± 0.03	0.75 ± 0.11	0.47 ± 0.18

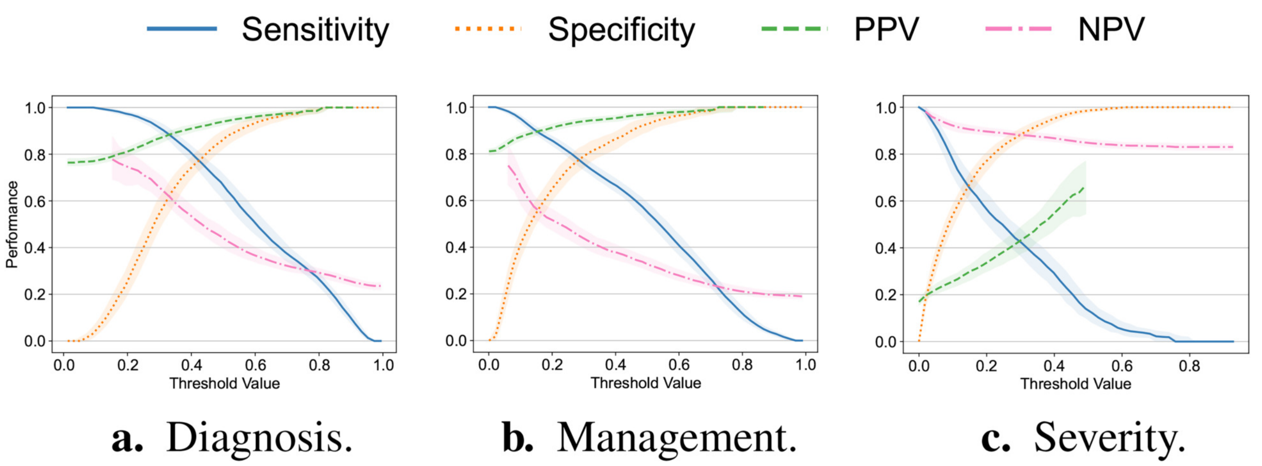

FIGURE 2
Sensitivity, specificity as well as positive (PPV) and negative (NPV) predictive values plotted against the value of the threshold applied to the output of the random forest model trained exclusively on the original Regensburg dataset for the **(a)** diagnosis, **(b)** management and **(c)** severity of appendicitis. All the metrics were assessed on the external (Düsseldorf) dataset. Bold lines correspond to the medians with the confidence bounds given by the interquartile ranges.

models on the aforementioned mixture of the Regensburg and Düsseldorf subjects (see the *retrained* models in Table 2), assessing them on the withheld portion of the external dataset. For all three classifiers, the average AUROC and AUPR metric values attained on the Düsseldorf data increase moderately after retraining. However, the resulting level of performance is still substantially lower than that of the original models on the Regensburg dataset. The lack of bigger improvement in predictive performance may be due to distribution shift, in particular the discrepancies in the missingness patterns and reporting across the two datasets (see Figure 4).

3.3 Feature importance

To elucidate the predictions made by our models on the external dataset, we calculate *permutation feature importance* on the test set. Specifically, we assess the importance of individual predictors by permuting (i.e., shuffling) their values and then quantifying the resulting change in the AUROC metric. The outcomes of this analysis are summarized in Figure 3. We limit our investigation to the diagnosis response variable and the random forest model given that it attained the best well-balanced performance across all the settings (refer back to Table 2).

Similar to the original findings on the Regensburg data (23), the three most important features are the diameter and visibility of the appendix as well as peritonitis. Likewise, the surrounding tissue reaction, target sign, WBC count and neutrophil percentage have an importance score, on average, above 0. Generally, the variable importance on the Düsseldorf data follows a pattern comparable to the results obtained previously on the Regensburg cohort. However, the variability across bootstrap resamples is considerably higher. Nonetheless, these results are not indicative of any concerning trends or spurious associations and fall well within our expectations. Notably, these observations hold for the other two response variables; for treatment the three most important features are peritonitis, appendix diameter and WBC count, and for complications these are CRP, peritonitis and appendix diameter, which is consistent with the results reported in the Regensburg study (23). In the original analysis, the most important predictors were appendix diameter, peritonitis and CRP respectively for diagnosis, treatment and complications.

4 Discussion

In this article, we performed a comprehensive external validation of ML models for predicting the diagnosis, management and severity in pediatric patients with suspected appendicitis (see Figure 1). Specifically, we have focused on the models initially trained on the dataset from the tertiary care hospital in Regensburg, Germany (23). To conduct the analysis, we have acquired an external dataset at the Florence-Nightingale-Hospital in Düsseldorf, Germany.

We observed that the external Düsseldorf dataset presents a statistically significant shift in the distribution of the covariates (captured in Table 1), including the response variables. Furthermore, the rates of missing values differ considerably across the two hospitals (as shown in Figure 4), especially for US-related variables and the percentage of neutrophils. Such discrepancies pose substantial challenges to the transferability of ML models to settings different from those considered at the training time (35).

In assessing the models' predictive performance (reported in Table 2), we observed the patients' diagnoses and treatment assignments could be predicted on the external Düsseldorf data by the models trained solely on the Regensburg cohort. In particular, compared to the original analysis (23), there was no decrease in AUPR and a moderate 10 percentage point decrease in

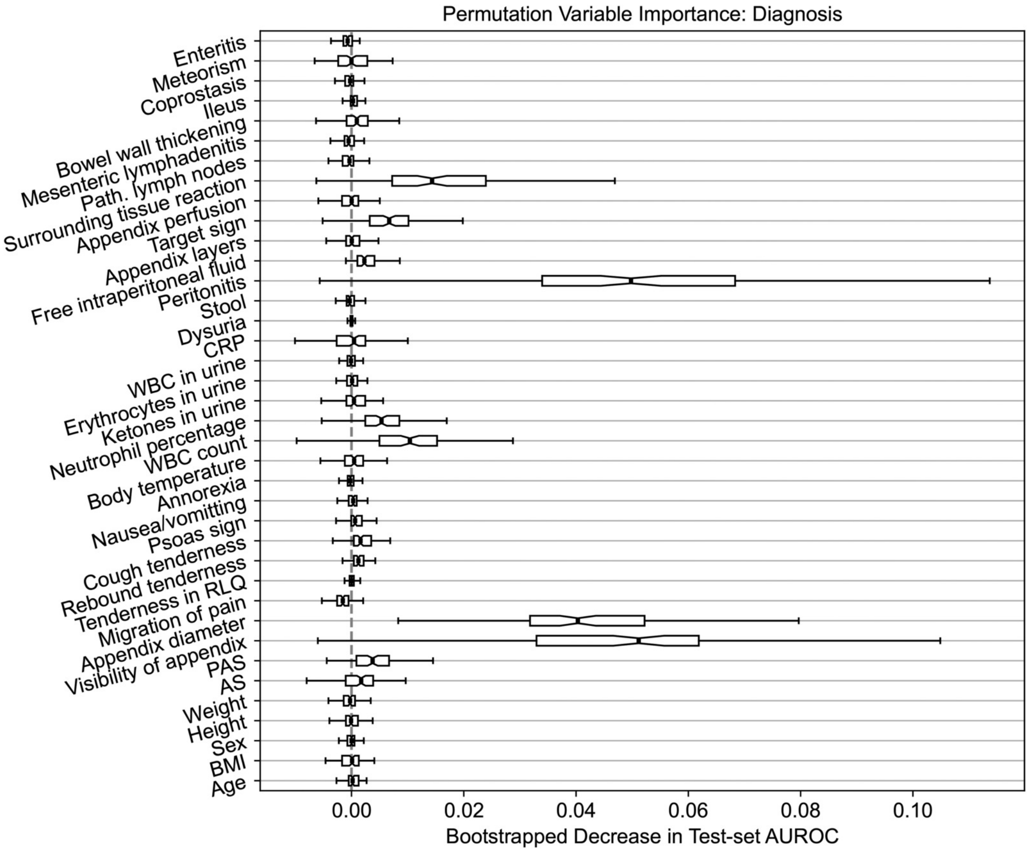

FIGURE 3
Permutation feature importance for the random forest model predicting the diagnosis of appendicitis. The importance is quantified by the decrease in the AUROC predictive performance metric after permuting the values of the predictor variable of interest. The variability in importance is assessed using bootstrapping and it is visualized using box plots.

AUROC. These performance levels are close to the AUROC of 90% reported as the baseline in a recent systematic review assessing the accuracy of artificial intelligence-based tools used in pediatric appendicitis diagnosis (36). In contrast, the predictive performance for the severity decreased more substantially. In addition to AUROC and AUPR measurements, we examined the tradeoff among the sensitivity, specificity, PPV and NPV (visualized in Figure 2). Furthermore, the feature importance analysis on the external dataset (shown in Figure 3) exhibited no concerning patterns.

Several risk prediction models for appendicitis have been built, evaluated and validated, including for varied patient cohorts (adults vs. children), in- and out-patient treatments, and outcomes. A recent prospective multicenter study also evaluated the performance of risk scores to identify appendicitis among children brought to the hospital emergency department (37). It identified that low appendicitis scores (≤ 2 for Alvarado or PAS, or ≤ 3 for Shera-Score) can be used to preselect children who can be discharged without further evaluation, but was unable to offer guidelines to select children who should proceed directly to surgery, indicating that patients with medium and high risk scores should undergo routine imaging examination (37). Our original and current study relied on a wide selection of variables, including the aforementioned risk scores (Alvarado score and PAS) and imaging parameters (ultrasound) from pediatric patients who were suspected of appendicitis, not only to exclude children with a low probability of appendicitis, but also to predict the diagnosis, management and severity of appendicitis (11).

To explore the potential of model updating (35), we retrained the classifiers on a mixture of the two datasets. This led to a moderate increase in AUROC and AUPR across all the target variables (refer to Table 2), suggesting that model updating, indeed, helps to tackle cross-hospital distribution shifts. More generally, our empirical findings indicate some degree of transferability of the considered predictive models across the two hospitals. Nonetheless, the decrease in predictive performance across several evaluation metrics is noticeable and could not be fully mitigated by retraining alone (as demonstrated by Table 2 and Figure 2). We hypothesize that this decrease in performance may be attributed to the shift in the prevalence of appendicitis cases, different missing value and data recording patterns, and variability in patient management routines. Below, we discuss these challenges in more detail.

As stated in Section 3.1, the distribution of most parameters differed across the two datasets. Unique regional and internal hospital practices can, at least partially, explain the observed differences. Notably, the dataset from Regensburg was acquired

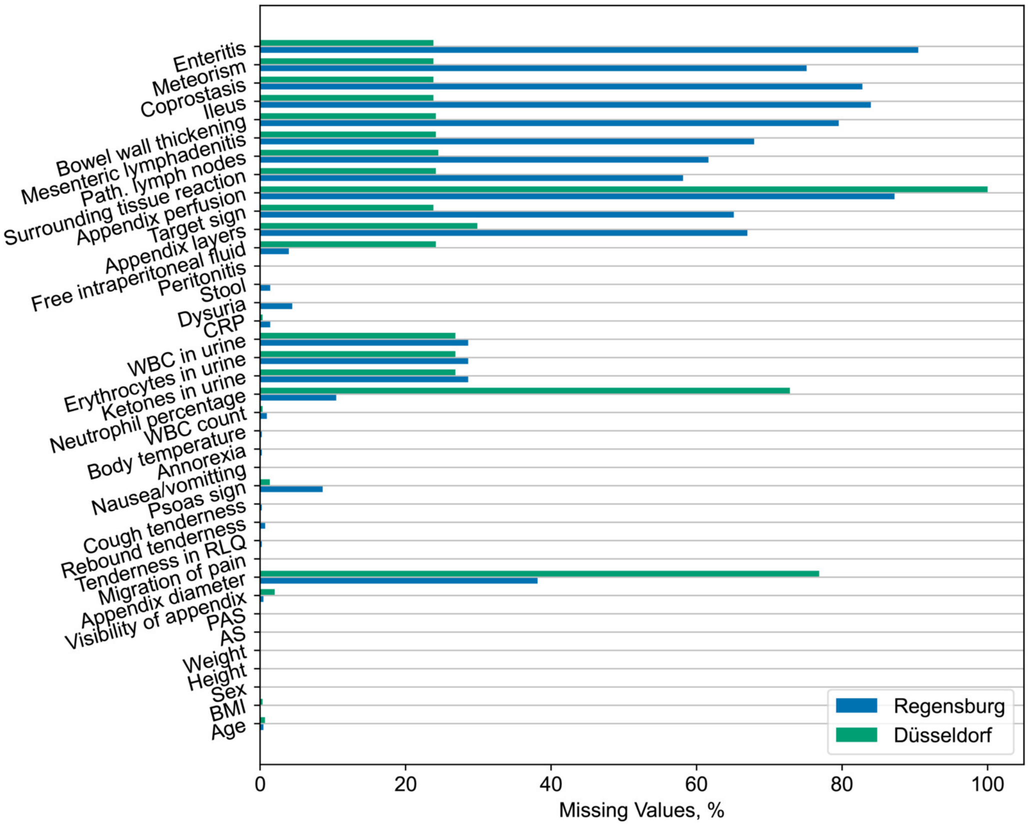

FIGURE 4
Percentages of missing values across all features for the original Regensburg and external Düsseldorf data. We observe considerable differences in the rates of missing values, especially for the ultrasonographic findings and neutrophil percentage.

from in-hospital patients admitted to a pediatric surgery department of a *specialized* pediatric hospital. The Düsseldorf dataset, on the other hand, was acquired from a pediatric surgery department of a *general* hospital with other surgical disciplines, such as general and orthopedic surgery. As a consequence, children aged 14 years or older were treated by general surgeons in Düsseldorf, and only those younger than 14 were seen and managed by pediatric surgeons. Consequently, only the latter group of patients was included in the study, which explains why the children from the Florence-Nightingale-Hospital were younger (median and IQR in years: 10.1 [7.7, 11.7]) than in Regensburg (11.5 [9.3, 13.9]).

The differences in the frequency of variable documentation reflect the internal organizational habits of the hospitals and departments, including variations in standardized admission reports and internal emergency department standards. Additionally, in Regensburg, children and adolescents were admitted by pediatric surgeons or residents in pediatric surgery or pediatrics, whereas in Düsseldorf, the admission was performed by both pediatric surgeons or residents and residents in general or orthopedic surgery working at the emergency department. Consequently, the ultrasound performance and report documentation differs across the two datasets.

Additionally, the clinical pathways and referral practices differed between the two institutions. In Düsseldorf, pediatricians and general practitioners were more likely to refer children to pediatric surgery when the diagnosis of appendicitis was already relatively clear. This preselection process likely contributed to the higher prevalence of confirmed appendicitis and surgical treatment cases in the Düsseldorf cohort. In contrast, the specialized pediatric setting in Regensburg enabled children with less specific abdominal symptoms to be evaluated by multiple specialties, such as pediatric gastroenterology, leading to a broader spectrum of differential diagnoses and, in some cases, more conservative management. Furthermore, the clinical pathway in Regensburg often involved admitting children with tenderness in the right iliac fossa to pediatric surgery for further observation, even in cases with unclear diagnosis. These patients were included in the cohort and may partly explain the higher rate of non-appendicitis cases, lower CRP values and lower frequency of surgical intervention.

Another noteworthy aspect is the time period of data acquisition. While the cohort from Regensburg included patients from January 2016 to December 2018, the Düsseldorf data were acquired from January 2015 to February 2022. Therefore, the latter cohort also included patients admitted during the COVID-19 pandemic and post-pandemic individuals, and negative appendectomy rates were lower during the pandemic (38) as patients might have sought medical care or have been referred to the hospital only if the positive diagnosis had been deemed more probable. This factor, alongside the higher frequency of delayed hospital presentations, might have contributed to the higher appendicitis prevalence and the higher rate of complicated cases observed in Düsseldorf (38, 39).

From the medical perspective, the limitations of the current study are similar to those reported in the original work that developed ML models on the Regensburg cohort (23). These include absence of confirmed appendicitis diagnosis for patients treated conservatively, limited number of study participants and missing values. Additionally, unique regional and internal hospital practices reduce the comparability of the collected datasets and transferability of the models, which, as we demonstrated, cannot be easily compensated for with model updating. Nonetheless, the observed distributions of parameters from both cohorts are clinically acceptable and display variability that is within expectations. Notably, our study allows to contrast the situatedness of a pediatric hospital against a general hospital where adult surgery and interdisciplinary surgical primary care are dominant. Lastly, the documented clinical, laboratory and ultrasound features are standardized, practical and cost-effective, enabling future analysis and comparison of our models on data from other institutions.

5 Conclusion

In this study, we performed an external validation of machine learning models for predicting the diagnosis, management and severity of pediatric appendicitis. When tested externally, the models exhibited lower predictive performance than on the original data. This was in part due to the shift in the prevalence of appendicitis cases we observed between the original and external datasets. Other possible reasons included intrinsic differences in patient demographics, clinical pathways, variations in referral practices and documentation standards for the two hospitals as well as the downstream effects of the COVID-19 pandemic. Such factors demonstrate the challenges of transferring predictive models between hospitals, which should always be done with care to avoid harmful fallout. As a potential remedy, we investigated model retraining; while it showed promise in restoring predictive performance, further research is necessary to determine the limitations of this approach, which we will explore in our future work. Additionally, we plan to investigate the possible design of the prospective evaluation and deployment of our predictive models. Specifically, we will look into defining the number of necessary blood tests and introducing standardized reporting guidelines for clinical examination and ultrasound findings.

Data availability statement

The analyzed dataset in an anonymized form is available alongside the code in the following GitHub repository: https://github.com/i6092467/pediatric-appendicitis-ml-ext. Further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the University of Regensburg (18-1063-101, 18-1063-3-101). The studies were conducted in accordance with the local legislation and institutional requirements. In accordance with national legislation and institutional requirements, written informed consent was not required from participants or their legal guardians/next of kin for analysis of retrospective data. Informed consent was obtained from participants or their legal representatives for any follow-up investigation.

Author contributions

RM: Conceptualization, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Methodology, Software, Validation, Visualization. KS: Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Methodology, Software, Validation, Visualization. AP: Writing – review & editing, Formal analysis, Investigation, Software, Validation. MAH: Writing – review & editing, Data curation, Investigation. VR: Writing – review & editing, Data curation, Investigation. SW: Writing – review & editing, Supervision. CK: Writing – review & editing, Methodology, Supervision. BR: Writing – review & editing, Methodology, Supervision. JEV: Writing – review & editing, Project administration, Resources, Supervision. PRW: Conceptualization, Writing – original draft, Writing – review & editing, Data curation, Investigation, Methodology, Project administration, Supervision.

Funding

The authors declare that financial support was received for the research and/or publication of this article. The open access funding was provided by ETH Zurich.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Bhangu A, Søreide K, Di Saverio S, Assarsson JH, Drake FT. Acute appendicitis: modern understanding of pathogenesis, diagnosis, and management. *Lancet*. (2015) 386:1278–87. doi: 10.1016/s0140-6736(15)00275-5

2. Rentea RM, Peter SDS, Snyder CL. Pediatric appendicitis: state of the art review. *Pediatr Surg Int*. (2016) 33:269–83. doi: 10.1007/s00383-016-3990-2

3. Di Saverio S, Podda M, De Simone B, Ceresoli M, Augustin G, Gori A, et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. *World J Emerg Surg*. (2020) 15:27. doi: 10.1186/s13017-020-00306-3

4. Dingemann J, Ure B. Imaging and the use of scores for the diagnosis of appendicitis in children. *Eur J Pediatr Surg*. (2012) 22:195–200. doi: 10.1055/s-0032-1320017

5. Escribá A, Gamell AM, Fernández Y, Quintillá JM, Cubells CL. Prospective validation of two systems of classification for the diagnosis of acute appendicitis. *Pediatr Emerg Care*. (2011) 27:165–9. doi: 10.1097/pec.0b013e31820d6460

6. Decker E, Ndzi A, Kenny S, Harwood R. Systematic review and meta-analysis to compare the short- and long-term outcomes of non-operative management with early operative management of simple appendicitis in children after the COVID-19 pandemic. *J Pediatr Surg*. (2024) 59:1050–7. doi: 10.1016/j.jpedsurg.2023.12.021

7. Svensson J, Hall N, Eaton S, Pierro A, Wester T. A review of conservative treatment of acute appendicitis. *Eur J Pediatr Surg*. (2012) 22:185–94. doi: 10.1055/s-0032-1320014

8. Svensson JF, Patkova B, Almström M, Naji H, Hall NJ, Eaton S, et al. Nonoperative treatment with antibiotics versus surgery for acute nonperforated appendicitis in children: a pilot randomized controlled trial. *Ann Surg*. (2015) 261:67–71. doi: 10.1097/sla.0000000000000835

9. Ohba G, Hirobe S, Komori K. The usefulness of combined B mode and Doppler ultrasonography to guide treatment of appendicitis. *Eur J Pediatr Surg*. (2016) 26:533–6. doi: 10.1055/s-0035-1570756

10. Park HC, Kim MJ, Lee BH. Randomized clinical trial of antibiotic therapy for uncomplicated appendicitis. *Br J Surg*. (2017) 104:1785–90. doi: 10.1002/bjs.10660

11. Reis Wolfertstetter P, Ebert JB, Barop J, Denzinger M, Kertai M, Schlitt HJ, et al. Suspected simple appendicitis in children: should we use a nonoperative, antibiotic-free approach? An observational study. *Children*. (2024) 11:340. doi: 10.3390/children11030340

12. Benabbas R, Hanna M, Shah J, Sinert R. Diagnostic accuracy of history, physical examination, laboratory tests, and point-of-care ultrasound for pediatric acute appendicitis in the emergency department: a systematic review and meta-analysis. *Acad Emerg Med*. (2017) 24:523–51. doi: 10.1111/acem.13181

13. Bonadio W, Peloquin P, Brazg J, Scheinbach I, Saunders J, Okpalaji C, et al. Appendicitis in preschool aged children: regression analysis of factors associated with perforation outcome. *J Pediatr Surg*. (2015) 50:1569–73. doi: 10.1016/j.jpedsurg.2015.02.050

14. Acharya A, Markar SR, Ni M, Hanna GB. Biomarkers of acute appendicitis: systematic review and cost—benefit trade-off analysis. *Surg Endosc*. (2016) 31:1022–31. doi: 10.1007/s00464-016-5109-1

15. Zani A, Teague WJ, Clarke SA, Haddad MJ, Khurana S, Tsang T, et al. Can common serum biomarkers predict complicated appendicitis in children? *Pediatr Surg Int*. (2017) 33:799–805. doi: 10.1007/s00383-017-4088-1

16. Koberlein GC, Trout AT, Rigsby CK, Iyer RS, Alazraki AL, Anupindi SA, et al.. ACR appropriateness criteria® suspected appendicitis-child. *J Am Coll Radiol*. (2019) 16:S252–63. doi: 10.1016/j.jacr.2019.02.022

17. Akmese OF, Dogan G, Kor H, Erbay H, Demir E. The use of machine learning approaches for the diagnosis of acute appendicitis. *Emerg Med Int*. (2020) 2020:7306435. doi: 10.1155/2020/7306435

18. Aparicio PR, Marcinkevičs R, Reis Wolfertstetter P, Wellmann S, Knorr C, Vogt JE. Learning medical risk scores for pediatric appendicitis. In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE (2021). p. 1507–12.

19. Aydin E, Türkmen İU, Namli G, Öztürk Ç, Esen AB, Eray YN, et al. A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children. *Pediatr Surg Int*. (2020) 36:735–42. doi: 10.1007/s00383-020-04655-7

20. Deleger L, Brodzinski H, Zhai H, Li Q, Lingren T, Kirkendall ES, et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department. *J Am Med Inform Assoc*. (2013) 20:e212–20. doi: 10.1136/amiajnl-2013-001962

21. Hsieh C-H, Lu R-H, Lee N-H, Chiu W-T, Hsu M-H, Li Y-CJ. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery*. (2011) 149:87–93. doi: 10.1016/j.surg.2010.03.023

22. Marcinkevičs R, Reis Wolfertstetter P, Klimiene U, Chin-Cheong K, Paschke A, Zerres J, et al. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Med Image Anal*. (2024) 91:103042. doi: 10.1016/j.media.2023.103042

23. 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) 9:662183. doi: 10.3389/fped.2021.662183

24. Rajpurkar P, Park A, Irvin J, Chute C, Bereket M, Mastrodicasa D, et al. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci Rep*. (2020) 10:3958. doi: 10.1038/s41598-020-61055-6

25. Reismann J, Romualdi A, Kiss N, Minderjahn MI, Kallarackal J, Schad M, et al. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: an investigator-independent approach. *PLoS One*. (2019) 14:e0222030. doi: 10.1371/journal.pone.0222030

26. Stiel C, Elrod J, Klinke M, Herrmann J, Junge C-M, Ghadban T, et al. The modified Heidelberg and the AI appendicitis score are superior to current scores in predicting appendicitis in children: a two-center cohort study. *Front Pediatr*. (2020) 8:592892. doi: 10.3389/fped.2020.592892

27. Xia J, Wang Z, Yang D, Li R, Liang G, Chen H, et al. Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. *Comput Biol Med*. (2022) 143:105206. doi: 10.1016/j.compbiomed.2021.105206

28. Sokol K, Fackler J, Vogt JE. Artificial intelligence should genuinely support clinical reasoning and decision making to bridge the translational gap. *npj Digit Med*. (2025) 8:1–11. doi: 10.1038/s41746-025-01725-9

29. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat Med*. (2019) 25:1337–40. doi: 10.1038/s41591-019-0548-6

30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

31. Breiman L. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/a:1010933404324

32. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

33. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (2021).

34. Godau P, Kalinowski P, Christodoulou E, Reinke A, Tizabi M, Ferrer L, et al. Deployment of image analysis algorithms under prevalence shifts. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer (2023). p. 389–99.

35. van Royen FS, Moons KG, Geersing G-J, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J*. (2022) 60:2200250. doi: 10.1183/13993003.00250-2022

36. Rey R, Gualtieri R, La Scala G, Posfay Barbe K. Artificial intelligence in the diagnosis and management of appendicitis in pediatric departments: a systematic review. *Eur J Pediatr Surg*. (2024) 34:385–91. doi: 10.1055/a-2257-5122

37. Nepogodiev D, Wilkin RJ, Bradshaw CJ, Skerritt C, Ball A, Moni-Nwinia W, et al. Appendicitis risk prediction models in children presenting with right iliac fossa pain (RIFT study): a prospective, multicentre validation study. *Lancet Child Adolesc Health*. (2020) 4:271–80. doi: 10.1016/S2352-4642(20)30006-7

38. Bethell GS, Gosling T, Rees CM, Sutcliffe J, Hall NJ, CASCADE Study Collaborators and RIFT Study Collaborators. Impact of the COVID-19 pandemic on management and outcomes of children with appendicitis: the children with AppendicitiS during the CoronAvirus panDEmic (CASCADE) study. *J Pediatr Surg*. (2022) 57:380–5. doi: 10.1016/j.jpedsurg.2022.03.029

39. Schäfer F-M, Meyer J, Kellnar S, Warmbrunn J, Schuster T, Simon S, et al. Increased incidence of perforated appendicitis in children during COVID-19 pandemic in a Bavarian multi-center study. *Front Pediatr*. (2021) 9:683607. doi: 10.3389/fped.2021.683607