Contents lists available at ScienceDirect

Journal of Pediatric Surgery

journal homepage: www.sciencedirect.com/journal/journal-of-pediatric-surgery

Machine-learning-assisted Preoperative Prediction of Pediatric Appendicitis Severity

Aylin Erman [a, *], Julia Ferreira [c], Waseem Abu Ashour [c], Elena Guadagno [c], Etienne St-Louis [b, c], Sherif Emil [b, c], Jackie Cheung [a, d], Dan Poenaru [b, c]

[a] Department of Computer Science, McGill University, Montreal, QC, Canada
[b] McGill University Faculty of Medicine and Health Sciences, Canada
[c] Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre, Montreal, Qc, Canada
[d] Canada CIFAR AI Chair, Mila, Canada

ARTICLE INFO

Article history:
Received 30 August 2024
Received in revised form
5 December 2024
Accepted 30 December 2024

Keywords:
Appendicitis
Machine learning
Pediatric

ABSTRACT

Purpose: This study evaluates the effectiveness of machine learning (ML) algorithms for improving the preoperative diagnosis of acute appendicitis in children, focusing on the accurate prediction of the severity of disease.

Methods: An anonymized clinical and operative dataset was retrieved from the medical records of children undergoing emergency appendectomy between 2014 and 2021. We developed an ML pipeline that pre-processed the dataset and developed algorithms to predict 5 appendicitis grades (1 - non-perforated, 2 - localized perforation, 3 - abscess, 4 - generalized peritonitis, and 5 - generalized peritonitis with abscess). Imputation strategies were used for missing values and upsampling techniques for infrequent classes. Standard classifier models were tested. The best combination of imputation strategy, class balancing technique and classification model was chosen based on validation performance. Model explainability was verified by a pediatric surgeon. Our model's performance was compared to another pediatric appendicitis severity prediction tool.

Results: The study used a retrospective cohort including 1980 patients (60.6 % males, average age 10.7 years). Grade of appendicitis in the cohort was as follows: grade 1—70 %; grade 2—8 %; grade 3—7 %; grade 4—7 %; grade 5—8 %. Every combination of 6 imputation strategies, 7 class-balancing techniques, and 5 classification models was tested. The best-performing combined ML pipeline distinguished non-perforated from perforated appendicitis with 82.8 ± 0.2 % NPV and 56.4 ± 0.4 % PPV, and differentiated between severity grades with 70.1 ± 0.2 % accuracy and 0.77 ± 0.00 AUROC. The other pediatric appendicitis severity prediction tool gave an accuracy of 71.4 %, AUROC of 0.54 and NPV/PPV of 71.8/64.7.

Conclusion: Prediction of appendiceal perforation outperforms prediction of the continuum of appendicitis grades. The variables our models primarily rely on to make predictions are consistent with clinical experience and the literature, suggesting that the ML models uncovered useful patterns in the dataset. Our model outperforms the other pediatric appendicitis prediction tools.

The ML model developed for grade prediction is the first of this type, offering a novel approach for assessing appendicitis severity in children preoperatively. Following external validation and silent clinical testing, this ML model has the potential to enable personalized severity-based treatment of pediatric appendicitis and optimize resource allocation for its management.

Level of evidence: 3.

1. Introduction

Acute appendicitis is the most common surgical emergency in children [1]. Despite its frequency, appendicitis presents a significant diagnostic and management challenge in surgical practice, particularly in the preoperative distinction between non-perforated and perforated cases in children. The ability to

This article is part of a special issue entitled: AAP(SoSu) 2024 published in Journal of Pediatric Surgery.

* Corresponding author. Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, Room B04.2022, 1001 Decarie Boulevard, Montreal, Qc H4A 3J1, Canada. Tel.: +1 514 412 4498; fax: +1 514 412 4289

E-mail address: aylin.erman@mail.mcgill.ca (A. Erman).

https://doi.org/10.1016/j.jpedsurg.2024.162151

accurately determine perforated appendicitis preoperatively may influence the therapeutic options [2–8] and predicts clinical outcomes [9]. Moreover, distinguishing between the various grades of perforation is equally vital, enabling healthcare facilities to optimize resource allocation and refine postoperative care strategies [10]. The stratification of perforated appendicitis may also support standardization of treatment in an area with significant variability in care and outcomes [10,11]. Our group has previously developed a perforated appendicitis grading system that correlates with outcomes and resource utilization, and that will guide this present study's stratification of perforated appendicitis [12]. To our knowledge, no methods currently exist that can accurately and preoperatively differentiate between the grades of perforated appendicitis.

Traditional diagnostic tools such as the Alvarado score [13], the Pediatric Appendicitis Score (PAS) [14] and the Pediatric Appendicitis Risk Calculator (pARC) [15] may be helpful in the diagnosis of appendicitis, but fall short in discriminating between perforated and non-perforated cases. Existing methods that endeavor to address this challenge are primarily tailored to adult patients [16–18], failing to account for the nuanced differences in disease presentation and pathophysiology between adults and children [19–21]. Additionally, these methods are either uni-modal (i.e. using a single data source type, e.g. CT scans) [22–24], based on small and thus potentially unrepresentative datasets [8,16,25–27], or lack external validation [8,16,22–24,26,27], rendering them unfit for clinical use [9,28–31].

Machine learning (ML) can identify patterns in large multimodal datasets, and thus presents a promising avenue to enhance diagnostic accuracy and guide clinical decision-making. ML has been increasingly used in healthcare to support diagnostic efforts [32–34], and has already been successfully applied to diagnose appendicitis [35–37], post-appendectomy intra-abdominal abscess [38], and perforated adult appendicitis [18]. Despite the potential of ML in this domain, its application in pediatric appendicitis remains underexplored. We hypothesize that the application of ML for pediatric appendicitis can yield competitive severity prediction results.

The objective of this study is to harness the potential of ML to improve the preoperative diagnosis of appendicitis perforation and grade in children. Our goal is to build a tool that will support clinician decision making rather than replace standard pediatric surgical evaluation. We have built an ML pipeline that cleans, enhances, and analyzes preoperative clinical and imaging data, and accurately categorizes patient profiles by perforation and severity. Following full validation, such an approach promises to significantly enhance the management of appendicitis by enabling more tailored treatment strategies and improving hospital resource management, ultimately leading to better patient outcomes.

2. Methods

2.1. Study design

We retrospectively retrieved anonymized demographic, history, physical, imaging, lab investigation and operative data from the medical charts of children undergoing emergency appendectomy at the Montreal Children's Hospital (MCH) of the McGill University Health Centre (MUHC) between 2014 and 2021.

We developed an ML pipeline that predicts the grade of acute appendicitis given the multi-modal patient data. The pipeline preprocessed the dataset and developed algorithms for multi-class and binary classification tasks. The multi-class classification predicts 5 appendicitis grades (1 - non-perforated, 2 - localized perforation, 3 - abscess, 4 - generalized peritonitis, and 5 - generalized peritonitis with abscess), while the binary classification predicts non-perforated versus perforated appendicitis. The type of appendicitis (non-perforated versus perforated) and the grade of perforation were obtained from the operative report.

To predict the presence and severity of perforation, 2 different computational approaches were tested. The *direct approach* generated 1 classifier algorithm directly predicting the appendicitis grade. The *indirect approach* developed 3 distinct classifiers to predict 1. postoperative intra-abdominal abscess (none, single, or multiple), 2. peritonitis (none, localized, or generalized), and 3. perforation (present or not present) - which were then combined to deterministically identify the appendicitis grade. The indirect approach was tested as it incorporates the perforation grade definitions and may lead to higher prediction performances. Figure 1 illustrates the ML pipeline, which consists of data pre-processing and classification components. Both the direct and indirect approaches were tested in the classification component.

This study was approved by the McGill University Health Center Research Ethics Board (#2021–7255).

The grades of perforation are based on Yousef and colleagues' work [12] and are defined as follows: grade 1 (no perforation), grade 2 (localized or contained perforation), grade 3 (contained abscess with no generalized peritonitis), grade 4 (generalized peritonitis with no dominant abscess), grade 5 (generalized peritonitis with one or more dominant abscesses). A localized or early perforation is diagnosed when the perforation is completely encased by omentum or surrounding structures, or results in free purulence only adjacent to the appendix. An abscess is defined as a discrete and distinct collection of contained pus. Generalized peritonitis is defined as purulence involving two or more of the 5 regions of the abdomen (pelvis, right lower quadrant, left lower quadrant, right upper quadrant/subdiaphragmatic space, left upper quadrant/subdiaphragmatic space).

2.2. Dataset information

We retrospectively retrieved data from 2056 pediatric patients diagnosed with acute appendicitis at the MCH. Inclusion criteria were children aged 1–18 years who underwent surgical intervention for acute appendicitis with a confirmed diagnosis between 2014 and 2021. Participants formed a consecutive series.

2.3. Instruments and data collection

A web-based case report form was created using the secure Research Electronic Data Capture (REDCap) software. Patient demographics were retrieved through the Outcome and Assessment Information Set (*Oacis*) platform, which is the MUHC electronic health record platform. Variables of interest in unstructured text were extracted by trained medical students and research assistants, and their extraction was verified as accurate using ChatGPT-4 and expert opinion [39]. Specifically, researchers and ChatGPT-4 were both tasked to extract information from reports. Any discrepancy between human and ChatGPT extraction was flagged for adjudication by a pediatric surgeon.

2.4. Data pre-processing

2.4.1. Data cleaning

All patients missing at least one of intra-abdominal abscess, peritonitis, perforation or grade were removed from the dataset. All variables deemed unhelpful to the goal of appendicitis grading prediction (e.g. time of ultrasound) were removed after consultation. Additionally, all post-surgical variables, with the exception of intra-abdominal abscess, peritonitis and perforation, were removed since they would not be used in the preoperative

Please cite this article as: Erman A et al., Machine-learning-assisted Preoperative Prediction of Pediatric Appendicitis Severity, Journal of Pediatric Surgery, https://doi.org/10.1016/j.jpedsurg.2024.162151

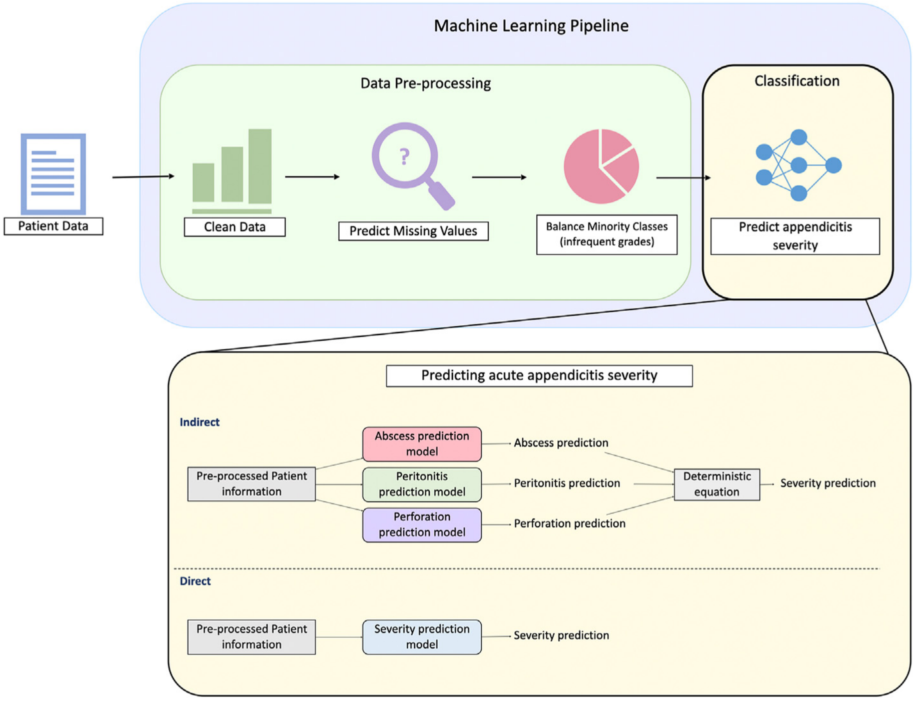
Machine Learning Pipeline
Data Pre-processing
Classification
Patient Data
Clean Data
Predict Missing Values
Balance Minority Classes (infrequent grades)
Predict appendicitis severity
Predicting acute appendicitis severity
Indirect
Pre-processed Patient information
Abscess prediction model
Peritonitis prediction model
Perforation prediction model
Abscess prediction
Peritonitis prediction
Perforation prediction
Deterministic equation
Severity prediction
Direct
Pre-processed Patient information
Severity prediction model
Severity prediction

Fig. 1. Visual summary of the machine learning pipeline.

prediction of appendicitis grade. Finally, patients missing age were removed.

2.4.2. Predicting missing values

Imputations of pre-surgical variables were done on a variable-by-variable basis, in 6 different ways (Table 1). This resulted in 6 differently imputed datasets, one for each strategy.

For each of the 6 datasets with distinct imputation methods, we divided the data into subsets for development (model training), hyperparameter tuning (i.e. optimization of values that control some aspect of a ML model's learning and prediction process), validation (choosing the best model) and final testing. We allocated 70 %, 10 %, 10 %, and 10 % of the data to each respective subset in preparation for the classification stage of the pipeline.

2.5. Balancing minority classes

Before moving onto the classification step, we implemented class-balancing methods on each of the 6 datasets' development subset to prevent majority class bias (in our case for the most common grade 1, i.e. non-perforated appendicitis). This entailed up-sampling minority classes to at least 10 %, 25 %, 50 %, 75 % and 100 % of the majority class, using the Synthetic Minority Over-sampling Technique (SMOTE) [40]. We also kept a copy of the original datasets with no class balancing method for comparison. Thus, 7 datasets with different class balances were produced from each of the 6 datasets, giving 42 final datasets each with its own class balance and imputation method. An overview visual of dataset generation can be found in Figure A1 in the Appendix.

2.6. Classification

2.6.1. Overview

The goal of the classification step was to predict presence of appendicitis perforation and grade. For the former, the direct and indirect approach were both tested, while only the direct approach was used for predicting perforation grade. The best ML pipeline (i.e. best combination of imputation strategy, class-balancing technique and classification model) for these 3 setups were chosen based on a given metric.

2.6.2. Indirect approach

We generated 3 models each predicting one of postoperative intra-abdominal abscess, peritonitis or perforation. To find the best prediction model for each of these 3 targets, we developed several different candidate ML models which varied in terms of learning

Table 1
The strategies used for imputing variables.

Strategy	Method if variable was continuous	Method if variable was discrete
1	Imputed with mean of given variable's non-missing values	Assigned a novel category
2	Imputed with mean of given variable's non-missing values	Imputed with mode of given variable's non-missing values
3	Predicted with linear regression. Input was other variables with no missingness.	Predicted with logistic regression. Input was other variables with no missingness.
4	Predicted with decision tree regressor. Input was other variables with no missingness.	Predicted with decision tree classifier. Input was other variables with no missingness.
5	Predicted with KNeighbour regressor. Input was other variables with no missingness.	Predicted with KNeighbour classifier. Input was other variables with no missingness.
6	Ensemble of strategies 2-5	Ensemble of strategies 2-5

Please cite this article as: Erman A et al., Machine-learning-assisted Preoperative Prediction of Pediatric Appendicitis Severity, Journal of Pediatric Surgery, https://doi.org/10.1016/j.jpedsurg.2024.162151

Table 2
All categorical features used in the model and their relative prevalence in the dataset.

Feature	Values	Prevalence (%)
Sex at birth	Female/Male	39.6/60.4
Pain migration to RLQ	Yes/No/Not recorded	72.3/5.7/22.0
Fever	Yes/No/Not recorded	35.3/53.7/11.0
Anorexia	Yes/No/Not recorded	43.8/13.3/42.9
Nausea	Yes/No/Not recorded	31.6/18.1/50.3
Vomiting	Yes/No/Not recorded	59.6/32.0/8.4
Diarrhea	Yes/No/Not recorded	16.5/64.4/18.9
Abdominal tenderness	Localized/Generalized/Not recorded	57.2/17.0/25.7
Guarding	Yes/No/Not recorded	44.3/14.0/41.7
Rebound tenderness	Yes/No/Not recorded	32.9/21.9/45.1
Appendix identified on ultrasound	Yes/Partial/No/Not recorded	70.8/13.5/14.5/1.2
Fat stranding on ultrasound	Yes/No/Not recorded	63.3/6.2/30.5
Fluid around the appendix on ultrasound	Yes/No/Not recorded	35.9/25.3/38.8
Phlegmon or inflammatory mass on ultrasound	Yes/No/Not recorded	6.2/17.3/76.5
Presence of fecalith on ultrasound	Yes/No/Not recorded	26.5/22.4/51.0
Intra-abdominal abscess on ultrasound	None/Single/Multiple/Not recorded	52.9/7.3/1.4/38.4
Intra-abdominal abscess from operative report	None/Single/Multiple/Not recorded	80.5/11.9/6.0/1.6
Peritonitis from operative report	None/Localized/Generalized/Not recorded	74.0/12.0/11.1/2.9
Presence of perforation from operative report	Yes/No/Not recorded	30.9/68.5/0.5

Other features not listed in the table include age (0.2 % missing, median: 11.0), duration of symptoms (3.6 % missing, median: 24), temperature (3.3 % missing, median: 37.1), WBC count (5.5 % missing, median: 14.7), neutrophils percent (7.0 % missing, median: 81), and appendix max diameter in mm (45.5 % missing, median: 9).

algorithm used and dataset trained on. The learning algorithms tested included logistic regression, K-nearest neighbors (KNN), random forest and decision tree models, as well as an ensemble of all these models. Each of these algorithms was tested across all 42 final datasets. The combination with the best prediction performance on the validation subset was chosen for each target. Accuracy was used as the only metric to optimize, for the indirect approach.

For the best performing abscess, peritonitis and perforation models, SHapley Additive exPlanations (SHAP) values [41] were determined to better understand how the models made their predictions. They explain what variables the model is most dependent on to make their predictions. These variables were reviewed by a pediatric surgeon.

We then applied the 3 models on each patient in the held-out test set and used a deterministic equation that resolved the grade of appendicitis from the predicted perforation, peritonitis and intra-abdominal abscess values. This equation is based on the perforated appendicitis grading scheme proposed previously by our group [12]. Figure 2 is a visual representation of the deterministic equation used to resolve grade of appendicitis from perforation, peritonitis and intra-abdominal abscess values. Based on these results, the final metrics were recorded. We used a bootstrap method on the test set to estimate a 95 % confidence interval. We also ran the entire method 10 times to observe stability of the ML pipeline.

2.6.3. *Direct approach*

We built 2 models, one that directly predicts the perforation grade, and the other that predicts the presence of perforation. Similarly to the indirect approach, candidate ML models varying in learning algorithm and dataset choice were built for each prediction target, and the model with the best performance on the respective validation set was selected. If the prediction target was perforation grade, the metrics tested included accuracy and the Area Under the Receiver Operator Curve (AUROC). AUROC is a common metric used in computer science to describe how well the ML model can distinguish between prediction categories. If the end goal prediction was presence of appendicitis perforation, then accuracy, AUROC, a utility score (*described below*), and negative and positive predictive values (NPV and PPV) were tested.

For the utility score, true negatives (TN), true positives (TP), false negatives (FN) and false positives (FP) were all given a score, depending on how heavily we wanted to penalize or reward each of these outcomes, based on the real-life consequences of each. Negative scores were given to penalize and positive scores to reward, score magnitude representing the severity of the outcome. Each patient is thus given a score based on the predicted and actual states, then all scores summed to generate the utility metric.

Within our clinical context, FN was identified as the worst outcome. While perforated appendicitis has several possible treatment strategies, the gold-standard treatment for non-perforated appendicitis is prompt appendectomy [42]. Moreover, the postoperative complication rate, length of hospital stay and resource utilization is higher with perforated appendicitis [11,43]. Thus, to best support clinicians in treatment selection and postoperative care decision-making, it is critical to correctly identify perforated cases. Therefore, missed perforated cases (i.e. FNs) were more heavily penalized than missing non-perforated cases (FPs). Both TPs and TNs were rewarded, with TNs being more highly rewarded for this same reason.

The final chosen models were then applied to the respective held-aside test sets to get the final performance metrics. A95 % confidence interval was calculated using the bootstrap method on the test set. We ran the ML pipeline 10 times for each metric to observe stability.

2.7. *Prediction model comparison*

For comparison, we used the predictive model for appendicitis perforation developed by Feng and colleagues [25] for children younger than 5 years. All patients with missing values in WBC count or duration of symptoms were discarded, as these variables are required to test the predictive model. This method was then applied to the remaining patients and the accuracy, AUROC, NPV and PPV values were recorded.

3. Results

From our patient data, 6 were treated non-operatively and were excluded from our dataset.The study included 2056 patients that underwent emergency appendectomy. Of these 76 were excluded from analysis due to missing information (appendicitis grade in 72, and age in 4). Of the remaining 1980 patients, 1200 (60.6 %) were male and the average and median age was 10.7 and 11.0, respectively. The grade of appendicitis in the cohort was distributed as

Please cite this article as: Erman A et al., Machine-learning-assisted Preoperative Prediction of Pediatric Appendicitis Severity, Journal of Pediatric Surgery, https://doi.org/10.1016/j.jpedsurg.2024.162151

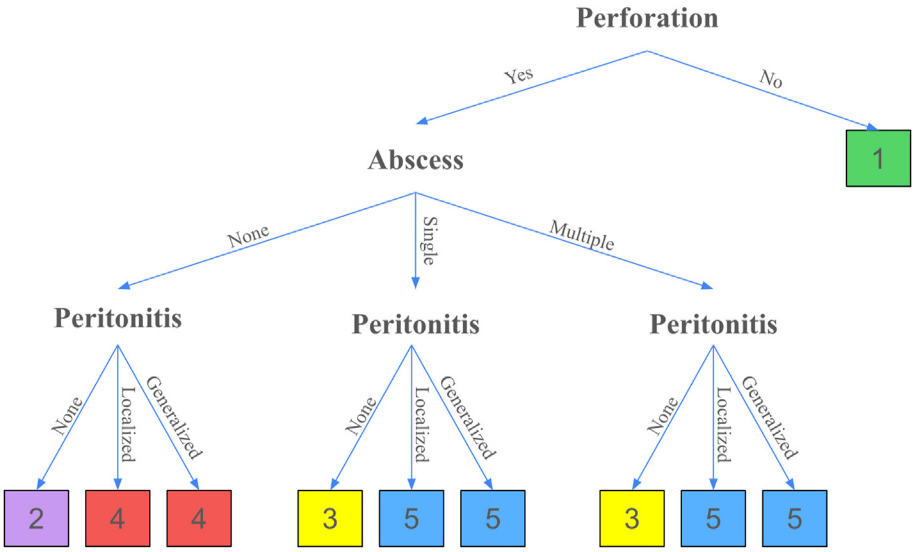

Fig. 2. Visual representation of the deterministic decision tree. The highlighted boxes indicate the final determined appendicitis grade. The colors used include green (grade 1), purple (grade 2), yellow (grade 3), red (grade 4) and blue (grade 5). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

follows: grade 1 (1378), grade 2 (161), grade 3 (142), grade 4 (132), and grade 5 (167). There were therefore 1378 non-unperforated appendicitis cases and 602 perforated.

We included 22 pre-operative variables, including demographic information, patient history, physical signs, laboratory results and ultrasound results. Table 2 describes the 19 categorical pre-operative variables used.

Continuous pre-operative variables include age (0.2% missing, median: 11.0), duration of symptoms (3.6% missing, median: 24), temperature (3.3% missing, median: 37.1), WBC count (5.5% missing, median: 14.7), neutrophils percent (7.0% missing, median: 81), and appendix max diameter in mm (45.5% missing, median: 9).

3.1. ML pipeline results

3.1.1. Predicting grade of perforation

When optimizing for accuracy, the *indirect* grade prediction pipeline resulted in an accuracy of 63.4 ± 0.2 %. The best performing models included logistic/linear regression imputation, no upsampling and random forest classifier for abscess prediction, ensemble imputation, 10 % upsampling and ensemble classifier for peritonitis prediction and logistic/linear regression imputation, SMOTE balancing and random forest classifier for presence of perforation prediction, For comparison, the direct grade prediction pipeline optimized on accuracy achieved an accuracy of 70.1 ± 0.2 %. The model that achieved this accuracy used logistic/linear regression imputation, 10 % upsampling and a random forest classifier. The *direct* grade prediction optimized on AUROC gave an AUROC of 0.77 ± 0.2. All results are reported with a 95 % confidence interval.

3.1.2. Predicting presence of perforation (using the direct architecture)

The pipelines optimized on accuracy and AUROC resulted in a 76.4 ± 0.2 % prediction accuracy and 0.79 ± 0.00 AUROC, respectively. The best performing models used decision tree and mean imputation, 25 % and 50 % upsampling, and random forest classifiers, respectively. The pipeline optimized on NPV and PPV produced an NPV of 82.8 ± 0.2 and PPV of 56.4 ± 0.4. The model in this setting used decision tree imputation, 75 % upsampling and random forest classifer. All results are reported with a 95 % confidence interval.

From running the ML pipeline 10 times for each setting (i.e. metric, indirect or direct approach and prediction target), we observed that the random forest classifier consistently performed best. In fact, 85 % of all best-performing ML pipelines used this classifier. The remaining 15 % of ML pipelines used the ensemble method. Between the imputation strategies and upsampling techniques, no single method consistently performed better than the rest.

Finally, we tested several utility metric weights. Since utility scores are based on binary predictions, this was only tested for the binary classification (presence of perforation) setting, rather than the grade prediction setting. From a sampling of possible utility metrics for our clinical setting, the relative utility scores in a 0–100 range, all fell between 71.4 and 75.5. Results can be found in Table A1 in the Appendix.

3.2. Variable importance analysis

Figures 3 and 4 illustrate the SHAP values for the best-performing ML pipelines, optimized on various metrics and using the direct approach and indirect approaches, respectively.

For the *direct* approach, the predictions generally weighed most heavily symptom duration, temperature, neutrophil number, fever and appendiceal diameter.

For the *indirect* approach, the perforation prediction model gave most importance to symptom duration, neutrophil number, fever, age, and appendiceal diameter. The abscess prediction model gave most importance to intra-abdominal abscess on ultrasound, fever, symptom duration, WBC count and appendiceal diameter. The peritonitis prediction model gave most importance to fever, neutrophil number, fluid around the appendix, temperature, and symptom duration.

3.3. Prediction model comparison

Applying Feng and colleagues' prediction model [25] to the 1718 patients in our dataset, it achieved an accuracy of 71.4 %, AUROC of 0.54, NPV of 71.8 and PPV of 64.7. To compare, our model achieved an accuracy of 76.4 %, AUROC of 0.79, NPV of 82.8 and PPV of 56.4.

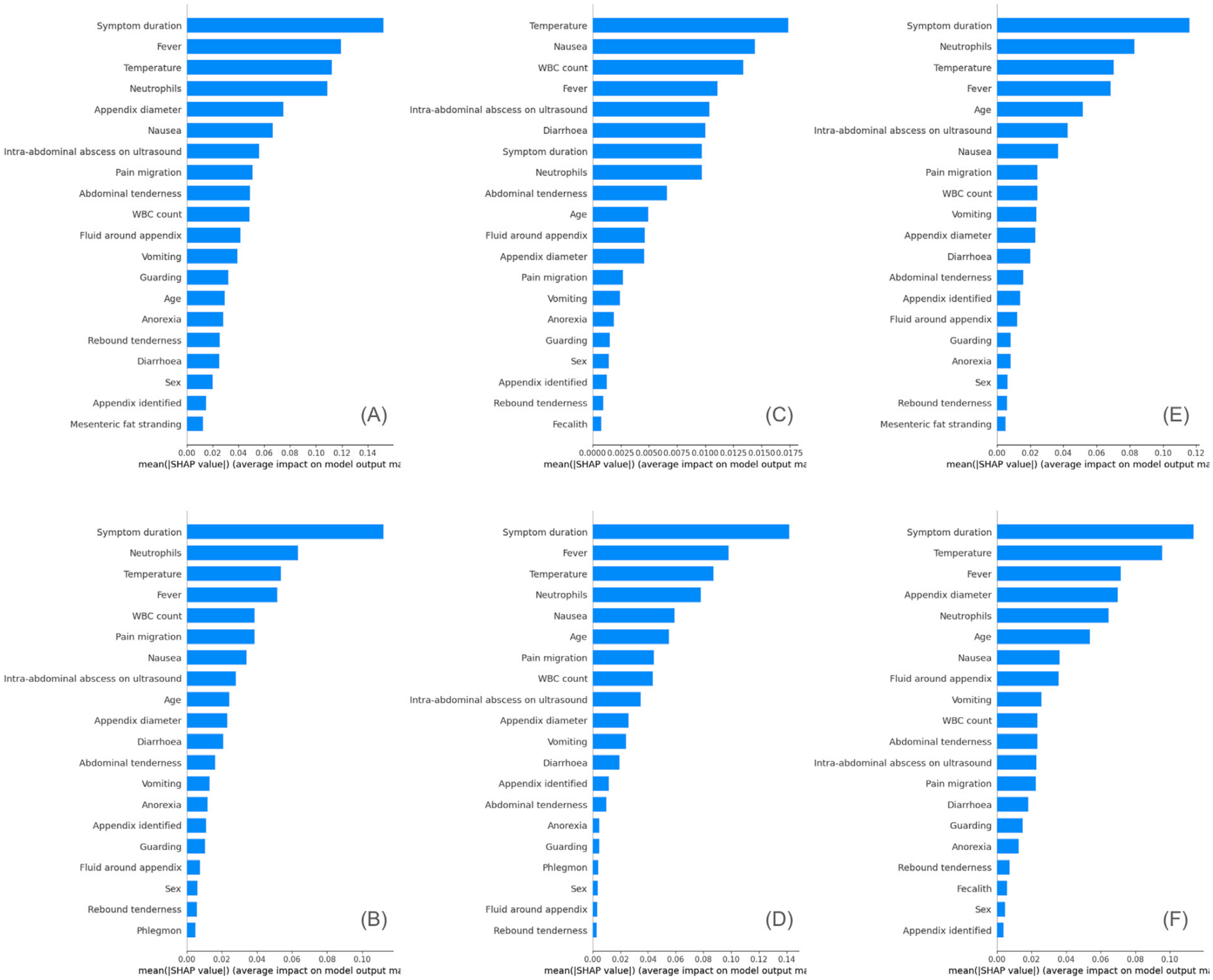

Fig. 3. SHAP values of all the final direct ML pipelines.These are the feature importance rankings for all the direct models resulting from optimizing for different metrics and prediction targets: (A) model that predicts presence of perforation, derived from optimizing for accuracy, (B) model that predicts grade of perforation, derived from optimizing for accuracy, (C) model that predicts presence of perforation, derived from optimizing for AUROC, (D) model that predicts grade of perforation, derived from optimizing for AUROC, (E) model that predicts presence of perforation, derived from optimizing for NPV and PPV, and (F) model that predicts presence of perforation, derived from optimizing for utility score (TN: 2, TP: 1, FN: −2, FP: 0).

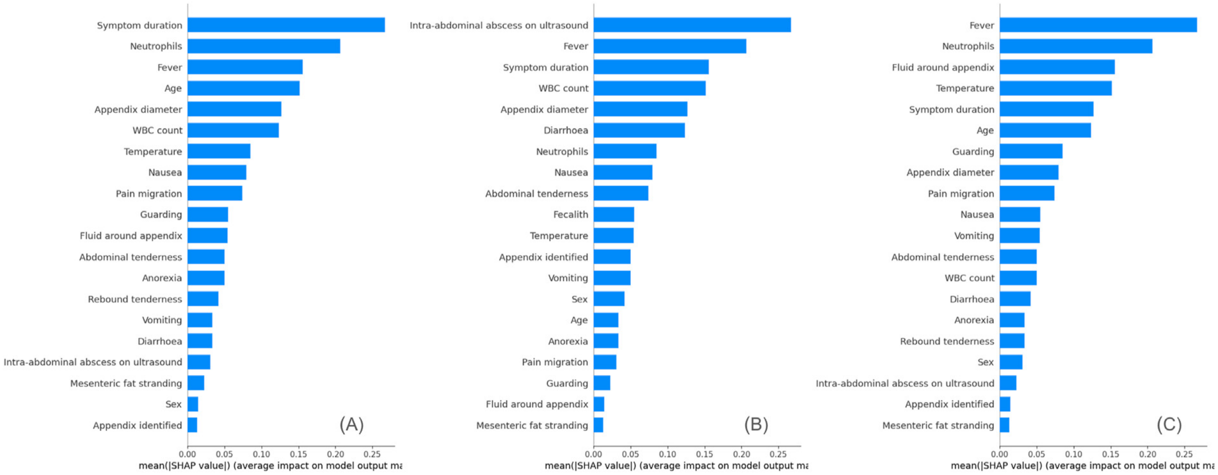

Fig. 4. SHAP values of ML models used in the indirect prediction of grades, optimized on accuracy. (A) perforation model, (B) abscess model, (C) peritonitis model.

4. Discussion

The goal of this study was to build an ML pipeline that can accurately predict appendicitis severity in children using a large single-institutional retrospective dataset. This research builds upon work our group previously conducted, including the definition of perforation grades, based on correlation with outcomes and resource utilization [12], and the accurate extraction of these grades from operative reports, using human and ChatGPT-4 opinion [39]. We tested direct and indirect perforation presence and grade prediction approaches, and optimized models on multiple outcome metrics (accuracy, AUROC, utility metric, and NPV/PPV). Each metric used resulted in a different selected ML pipeline. The development of these multiple pipelines represents the most thorough implementation of ML for appendicitis severity prediction in the literature, and the only implementation of ML for appendicitis grade prediction that we are aware of.

The *indirect* approach optimized on accuracy and the *direct* approach optimized on accuracy and AUROC predicted perforation grade with 63.4 ± 0.2 % accuracy, 70.1 ± 0.2 % accuracy and 0.77 ± 0.00 AUROC, respectively. The direct approach therefore had a better performance in this setting.

The rest of the pipelines used the direct approach to predict appendicitis perforation. The pipelines optimized on accuracy, AUROC, and NPV/PPV predicted presence of perforation with 76.4 ± 0.2 % accuracy, 0.79 ± 0.00 AUROC, 82.8 ± 0.2 NPV and 56.4 ± 0.4 PPV, respectively. The pipeline optimized on the utility metric spanned 71.4—75.5 in relative position, depending on the utility weights chosen.

Overall, we found that prediction based on perforation grade was less accurate than if based on perforation presence. This is at least partly due to multi-class classification (based on 5 grades) generally performing worse than binary classification (based on presence/absence of perforation). It is also reasonable to expect that distinguishing perforated from non-perforated appendicitis is easier than distinguishing between presence or absence of peritonitis, or between no abscess, single abscess, or multiple abscesses.

A separate observation is that almost all ML pipelines have selected the *random forest classifier* among various AI algorithms. This learning algorithm is indeed very popular for medical diagnosis prediction[44—46], due to its robustness against outliers, noisy data, and class imbalance.

Explainability is one of the most important values in healthcare AI [47]. It allows the clinician to see what the otherwise “black box” AI model has chosen to make its decisions on, and determine if this makes clinical sense. We examined our models’ explainability by identifying which specific clinical variables they primarily used (prioritized) in making their predictions - and found these to be consistent with clinician experience and the literature. For direct perforation presence and grade prediction, our most important features were predictably symptom duration, temperature, neutrophil number, fever, and appendiceal diameter - which are also supported by the literature [16,23—25,27,48]. Notably, WBC and C-reactive protein (CRP) levels, generally good indicators of perforated appendicitis [49,50], are less dependable in children [51]. While the WBC count is a marker of perforated appendicitis on its own, it has a stronger predictive performance when combined with the CRP - possibly explaining why it consistently appears in the importance rankings, yet is not a top feature [49,52]. We were not able to use CRP in our pipelines due to its high missingness values, as the test has only been used clinically in recent years.

As expected, the perforation prediction model in the indirect approach placed importance on similar features to the models directly predicting presence of perforation. The main difference was the addition of age as a key predictor of perforation — a logical finding, given that age is inversely correlated with perforation rate [53]. The important features for abscess and for peritonitis prediction were also consistent with the literature. For abscess prediction, there is a known correlation with fever and symptom duration [51], and the self-evident correlation with abscess identification on ultrasound. For peritonitis prediction, the correlation with symptom duration, fever, and neutrophil number are well-established in the literature [51,54,55]. Ultimately, the consistency of feature importance values across our models suggests that these models, regardless of the chosen metrics, have generally identified useful trends in the data.

Our study was not designed to identify the optimal model performance metrics for making clinical predictions in children with appendicitis. The accuracy metric, commonly used in AI studies, is less than ideal in clinical medicine, where the datasets are often imbalanced (as in the grades of appendicitis). Using only the accuracy metric can lead to a model that simply predicts the majority class (grade 1 appendicitis in our case) to achieve high accuracy, at the expense of sensitivity or specificity. Class imbalance also greatly biases NPV and PPV values [56]. In contrast, AUROC performs consistently regardless of class imbalance. In an effort to assist clinicians in their real-life decision-making process, we have also explored a “utility metric”, representing the relative value of either over-classifying or under-classifying their patients on the appendicitis grade or presence of perforations. Unfortunately, utility metrics depend on the therapeutic consequences of each misclassification and correct classification, and are therefore very complex, and dependent on local human and facility resources as well as clinician preferences. In the particular case of appendicitis, the weights assigned to each potential misclassification (FPs and FNs) and correct classification (TPs and TNs) will depend on the specific treatment alternatives (such as offering antibiotics alone instead of appendectomy, or chosing abscess drainage by interventional radiology) which are customary in each setting. These therapeutic choices remain clinically non-validated in pediatric appendicitis, and must therefore be the urgent focus of future prospective trials.

There are several published predictive models for distinguishing perforated and non-perforated appendicitis. However, comparing them to ours is not meaningful as none are tailored to the pediatric population. There are nuanced differences in disease presentation, pathophysiology and diagnosis norms (e.g. use of CT) between adults and children [19—21]. The existing predictive variables and patterns available in the different populations may vary in predictive strength and thus significantly impact predictive performance. For instance, CT scans are known to be more accurate than ultrasounds for diagnosing appendicitis perforation [57,58].

Nonetheless, we offer a short analysis of these methods. One important note is that we determined our final metrics on a set of patients that have not yet been seen during model development. This approach provides a more accurate measure of the model's ability to generalize to new, unseen data. In contrast, some of these published models appear to have reported final metrics on the same set of patients used to develop the model [8,16]. Others determined their final metrics by resampling, with replacement, patient records from the dataset the models were trained on [17]. These practices risk inflating performance metrics.

Phan-Mai and colleagues are the only group to use a fully ML-based approach to predicting appendicitis perforation [18]. They have high performance (AUROC: 0.894) on a larger held-aside set. ML is generally better suited at finding useful patterns in high dimensional big data or when working with complex non-linear relationships [59]. ML also does not need explicit variable selection which allows it to better explore the data and potentially find surprising predictors.

There exists a model developed for children younger than 5 years of age [25]. To properly compare our model to this one, we implemented their perforation prediction model onto our dataset. Our results outperformed their pediatric perforation prediction model. The accuracy and NPV/PPV values were comparable - our model achieved an accuracy of 76.4 % and NPV/PPV of 82.8/56.4 for perforation prediction, while theirs achieved an accuracy of 71.4 % and NPV/PPV of 72/65. However, our model demonstrated a superior AUROC of 0.79, compared to 0.54 for their model. AUROC is a measure of a model's discriminatory power, indicating that our ML pipeline more effectively distinguishes between perforated and non-perforated appendicitis. The small dataset size they used for development (156 patients) could have affected the model's generalizability to larger datasets (1718 patients). As well, their population of interest (less than 5 years old) is different from ours (entire pediatric population) so comparison at all may not be valid.

4.1. *Limitations*

Our study has several limitations. Firstly, we use a retrospective dataset, resulting in high missingness for certain variables, caused by unrecorded values, incomplete emergency department presentation data and incomplete or poorly written operative dictations. Imputation methods were used to address missingness, but they have the potential to alter patterns in the data and reduce the true variability of the imputed variable, both factors potentially affecting model performance. Addressing the class imbalances in the dataset may have had a similar effect. Upsampling of minority classes may have over- or under-emphasized important patterns in the data, and SMOTE, which creates plausible synthetic patient records, may have introduced otherwise non-existent patterns or obscured existing ones.

Moreover, data retrieval included patient records over a 7-year period (2014—2021), during which clinical practice could have potentially changed.

Other limitations include the use of a currently unvalidated grading score - though our group is in the process of validating it. The size of the dataset we used is considered small for typical ML implementations. With larger dataset size, other ML models with high data requirements could be trained, such as the generative adversarial imputation networks (GAINs) for imputations. As well, explainable AI is able to produce descriptions of how an AI system makes predictions generally, but is unreliable for individual predictions [60].

Our models are currently purely *in-silico*. To be useful locally, we first need to run *silent prospective trials* in which the ML model will be deployed for patients in real-time, but the predictions will be unavailable to clinicians so as to not bias therapeutic decisions [61]. Before potential deployment we must bring our validation beyond digital metrics, proving that our model can actually improve clinical outcomes, and we must analyze model outputs to ensure there is no bias towards vulnerable populations [62]. To be useful in other institutions, further *external validation* on datasets from these institutions should be conducted, ensuring that our model is generalizable to other health centers and not overly optimized for our local institutional population. Continuous model *validation over time* is also necessary to ensure that potential changes in patients and practices don't affect our pipeline's performance [63].

4.2. *Future directions*

The utility metric shows strong potential as a future standard for evaluation, due to its ability to reflect the practical benefit of using clinical ML models. For this to occur, advances in the standardization and validation of the treatment of pediatric appendicitis must first occur. Therapeutic standardization would allow the accurate quantification of harm and benefit of true and false positives and negatives needed for the utility metric to be determined.

The next steps for this current model include external validation and silent prospective trials to establish both our pipeline's generalizability and local applicability. Further ways to improve our model include using a validated appendicitis perforation score, and the inclusion of CRP as a variable to the dataset.

5. Conclusion

We have successfully built ML pipelines capable of predicting appendicitis perforation and grade in children, which performed equally or better than existing non-AI tools and were clinically explainable. Our models highlight the potential of AI in the accurate preoperative classification of appendicitis in children, but will require external validation and prospective testing before becoming useable in a clinical setting. Such a tool could influence therapeutic options, help optimize resource allocation, refine postoperative care strategies and support standardization of treatment for pediatric appendicitis.

Previous communication

N/A.

Financial support statement

This research received funding through the Canadian Institute of Health Research Project Grant #465962.

Conflicts of interest

None.

APPENDIX

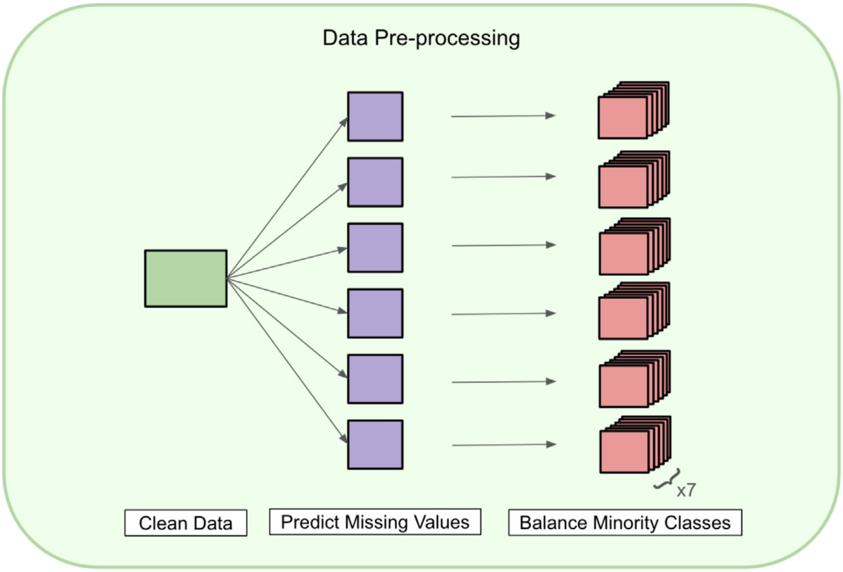

Figure A1. Visual summary of the dataset generation during data pre-processing. The green box represents the single dataset after data cleaning. The purple boxes represent the 6 datasets generated after imputing the missing values. The red boxes represent the 42 final datasets generated after applying 7 class balancing methods to each of the 6 datasets from the previous step.

Table A1
Results from optimizing for utility score while predicting on perforation, reported with a 95 % confidence interval. The relative position of the score is the utility score relative to a 0–100 range.

Utility metric weights				Minimum possible utility score	Maximum possible utility score	Direct approach utility score	Relative position of the score
TN	TP	FN	FP				
2	1	−2	0	−112	314	202 ± 1	73.7 ± 0.0
2	1	−1	0	−56	314	222 ± 1	75.1 ± 0.0
1	1	−2	0	−112	185	100 ± 1	71.4 ± 0.0
1	1	−1	−1	−185	185	91 ± 1	74.6 ± 0.0
3	1	−3	0	−168	443	293 ± 1	75.5 ± 0.0
3	2	−3	−1	−297	499	291 ± 2	73.9 ± 0.0

References

[1] Addiss DG, Shaffer N, Fowler BS, et al. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol 1990;132:910–25. https://doi.org/10.1093/oxfordjournals.aje.a115734.

[2] Cobben LPJ, De Van Otterloo AM, Puylaert JBCM. Spontaneously resolving appendicitis: frequency and natural history in 60 patients. Radiology 2000;215:349–52. https://doi.org/10.1148/radiology.215.2.r00ma08349.

[3] Andersson RE. The natural history and traditional management of appendicitis revisited: spontaneous resolution and predominance of prehospital perforations imply that a correct diagnosis is more important than an early diagnosis. World J Surg 2007;31:86–92. https://doi.org/10.1007/s00268-006-0056-y.

[4] The CODA Collaborative. A randomized trial comparing antibiotics with appendectomy for appendicitis. N Engl J Med 2020;383:1907–19. https://doi.org/10.1056/NEJMoa2014320.

[5] Lee SL, Islam S, Cassidy LD, et al. Antibiotics and appendicitis in the pediatric population: an American pediatric surgical association outcomes and clinical trials committee systematic review. J Pediatr Surg 2010;45:2181–5. https://doi.org/10.1016/j.jpedsurg.2010.06.038.

[6] Rollins KE, Varadhan KK, Neal KR, et al. Antibiotics versus appendicectomy for the treatment of uncomplicated acute appendicitis: an updated meta-analysis of randomised controlled trials. World J Surg 2016;40:2305–18. https://doi.org/10.1007/s00268-016-3561-7.

[7] Sallinen V, Akl EA, You JJ, et al. Meta-analysis of antibiotics *versus* appendicectomy for non-perforated acute appendicitis. Br J Surg 2016;103:656–67. https://doi.org/10.1002/bjs.10147.

[8] Kang C-B, Li W-Q, Zheng J-W, et al. Preoperative assessment of complicated appendicitis through stress reaction and clinical manifestations. Medicine (Baltim) 2019;98:e15768. https://doi.org/10.1097/MD.0000000000015768.

[9] Bom WJ, Scheijmans JCG, Salminen P, et al. Diagnosis of uncomplicated and complicated appendicitis in adults. Scand J Surg 2021;110:170–9. https://doi.org/10.1177/14574969211008330.

[10] Yousef Y, Youssef F, Homsy M, et al. Standardization of care for pediatric perforated appendicitis improves outcomes. J Pediatr Surg 2017;52:1916–20. https://doi.org/10.1016/j.jpedsurg.2017.08.054.

[11] Emil S. Clinical pediatric surgery: a case-based interactive approach. Boca Raton, FL: CRC Press; 2020.

[12] Yousef Y, Youssef F, Dinh T, et al. Risk stratification in pediatric perforated appendicitis: prospective correlation with outcomes and resource utilization. J Pediatr Surg 2018;53:250–5. https://doi.org/10.1016/j.jpedsurg.2017.11.023.

[13] Alvarado A. A practical score for the early diagnosis of acute appendicitis. Ann Emerg Med 1986;15:557–64. https://doi.org/10.1016/S0196-0644(86)80993-3.

[14] Samuel M. Pediatric appendicitis score. J Pediatr Surg 2002;37:877–81. https://doi.org/10.1053/jpsu.2002.32893.

[15] Kharbanda AB, Vazquez-Benitez G, Ballard DW, et al. Development and validation of a novel pediatric appendicitis risk calculator (pARC). Pediatrics 2018;141:e20172699. https://doi.org/10.1542/peds.2017-2699.

[16] Avanesov M, Wiese NJ, Karul M, et al. Diagnostic prediction of complicated appendicitis by combined clinical and radiological appendicitis severity index (APSI). Eur Radiol 2018;28:3601–10. https://doi.org/10.1007/s00330-018-5339-9.

[17] Atema JJ, Van Rossem CC, Leeuwenburgh MM, et al. Scoring system to distinguish uncomplicated from complicated acute appendicitis. Br J Surg 2015;102:979–90. https://doi.org/10.1002/bjs.9835.

[18] Phan-Mai T-A, Thai TT, Mai TQ, et al. Validity of machine learning in detecting complicated appendicitis in a resource-limited setting: findings from Vietnam. BioMed Res Int 2023;2023:1–8. https://doi.org/10.1155/2023/5013812.

[19] Dahal GR. Acute appendicitis in children: how is it different than in adults? Gd. Med J 2019;1:35–40. https://doi.org/10.3126/gmj.v1i1.22404.

[20] Pogorelić Z, Domjanović J, Jukić M, et al. Acute appendicitis in children younger than five years of age: diagnostic challenge for pediatric surgeons. Surg Infect 2020;21:239–45. https://doi.org/10.1089/sur.2019.175.

[21] Lee SL, Ho HS. Acute appendicitis: is there a difference between children and adults? Am Surg 2006;72:409–13. https://doi.org/10.1177/000313480607200509.

[22] Foley TA, Earnest F, Nathan MA, et al. Differentiation of nonperforated from perforated appendicitis: accuracy of CT diagnosis and relationship of CT findings to length of hospital stay. Radiology 2005;235:89–96. https://doi.org/10.1148/radiol.2351040310.

[23] Lietzén E, Mällinen J, Grönroos JM, et al. Is preoperative distinction between complicated and uncomplicated acute appendicitis feasible without imaging? Surgery 2016;160:789–95. https://doi.org/10.1016/j.surg.2016.04.021.

[24] Bixby SD, Lucey BC, Soto JA, et al. Perforated versus nonperforated acute appendicitis: accuracy of multidetector CT detection. Radiology 2006;241:780–6. https://doi.org/10.1148/radiol.2413051896.

[25] Feng W, Zhao X-F, Li M-M, et al. A clinical prediction model for complicated appendicitis in children younger than five years of age. BMC Pediatr 2020;20:401. https://doi.org/10.1186/s12887-020-02286-4.

[26] Kang C-B, Li X-W, Hou S-Y, et al. Preoperatively predicting the pathological types of acute appendicitis using machine learning based on peripheral blood biomarkers and clinical features: a retrospective study. Ann Transl Med 2021;9:835. https://doi.org/10.21037/atm-20-7883. 835.

[27] Imaoka Y, Itamoto T, Takakura Y, et al. Validity of predictive factors of acute complicated appendicitis. World J Emerg Surg 2016;11:48. https://doi.org/10.1186/s13017-016-0107-0.

[28] Huang Y, Du C, Xue Z, et al. What makes multi-modal learning better than single (provably. https://doi.org/10.48550/ARXIV.2106.04538; 2021.

[29] Ajiboye AR, Abdullah-Arshah R, Qin H, et al. Evaluating the effect of dataset size on predictive model using supervised learning technique. Int J Comput Syst Sci Eng 2015;1:75–84. https://doi.org/10.15282/ijsecs.1.2015.6.0006.

[30] Althnian A, AlSaeed D, Al-Baity H, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci 2021;11:796. https://doi.org/10.3390/app11020796.

[31] Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Progr Biomed 2021;208:106288. https://doi.org/10.1016/j.cmpb.2021.106288.

[32] Habehh H, Gohel S. Machine learning in healthcare. Curr Genom 2021;22:291–300. https://doi.org/10.2174/1389202922666210705124359.

[33] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–8. https://doi.org/10.7861/futurehosp.6-2-94.

[34] Park DJ, Park MW, Lee H, et al. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep 2021;11:7567. https://doi.org/10.1038/s41598-021-87171-5.

[35] Rajpurkar P, Park A, Irvin J, et al. AppendiXNet: deep learning for diagnosis of appendicitis from A small dataset of CT exams using video pretraining. Sci Rep 2020;10:3958. https://doi.org/10.1038/s41598-020-61055-6.

[36] Issaiy M, Zarei D, Saghazadeh A. Artificial intelligence and acute appendicitis: a systematic review of diagnostic and prognostic models. World J Emerg Surg 2023;18:59. https://doi.org/10.1186/s13017-023-00527-2.

[37] Mijwil MM, Aggarwal K. A diagnostic testing for people with appendicitis using machine learning techniques. Multimed Tool Appl 2022;81:7011–23. https://doi.org/10.1007/s11042-022-11939-8.

[38] Alramadhan MM, Al Khatib HS, Murphy JR, et al. Using artificial neural networks to predict intra-abdominal abscess risk post-appendectomy. Ann Surg Open 2022;3:e168. https://doi.org/10.1097/AS9.0000000000000168.

[39] Abu-Ashour W, Emil S, Poenaru D. Using artificial intelligence to label free-text operative and ultrasound reports for grading pediatric appendicitis. J Pediatr Surg 2024;59:783–90. https://doi.org/10.1016/j.jpedsurg.2024.01.033.

[40] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.

[41] Lundberg S, Lee S-I. A unified approach to interpreting model predictions 2017. https://doi.org/10.48550/ARXIV.1705.07874.

[42] Holcomb GW, Murphy PJ. Ashcraft's pediatric surgery. 5th ed. Philadelphia: Saunders/Elsevier; 2010.

[43] Bolmers MDM, De Jonge J, Bom WJ, et al. In-hospital delay of appendectomy in acute, complicated appendicitis. J Gastrointest Surg 2022;26:1063–9. https://doi.org/10.1007/s11605-021-05220-w.

[44] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inf Decis Making 2011;11:51. https://doi.org/10.1186/1472-6947-11-51.

[45] Dai B, Chen R-C, Zhu S-Z, et al. Using random forest algorithm for breast cancer diagnosis. In: 2018 int. Symp. Comput. Consum. Control IS3C, Taichung. Taiwan: IEEE; 2018. p. 449–52. https://doi.org/10.1109/IS3C.2018.00119.

[46] Ooka T, Johno H, Nakamoto K, et al. Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutr Prev Health 2021;4:140–8. https://doi.org/10.1136/bmjnph-2020-000200.

[47] Rasheed K, Qayyum A, Ghaly M, et al. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. Comput Biol Med 2022;149:106043. https://doi.org/10.1016/j.compbiomed.2022.106043.

[48] Hajibandeh S, Hajibandeh S, Hobbs N, et al. Neutrophil-to-lymphocyte ratio predicts acute appendicitis and distinguishes between complicated and uncomplicated appendicitis: a systematic review and meta-analysis. Am J Surg 2020;219:154–63. https://doi.org/10.1016/j.amjsurg.2019.04.018.

[49] Panagiotopoulou IG, Parashar D, Lin R, et al. The diagnostic value of white cell count, C-reactive protein and bilirubin in acute appendicitis and its complications. Ann R Coll Surg Engl 2013;95:215–21. https://doi.org/10.1308/003588413X13511609957371.

[50] Patmano M. Laboratory markers used in the prediction of perforation in acute appendicitis. Turkish J Trauma Emerg Surg 2021. https://doi.org/10.14744/tjtes.2021.83364.

[51] Mattei P, editor. Fundamentals of pediatric surgery. New York, NY: Springer New York; 2011. https://doi.org/10.1007/978-1-4419-6643-8.

[52] Yang J, Liu C, He Y, et al. Laboratory markers in the prediction of acute perforated appendicitis in children. Emerg Med Int 2019;2019:1–4. https://doi.org/10.1155/2019/4608053.

[53] Bonadio W, Peloquin P, Brazg J, et al. Appendicitis in preschool aged children: regression analysis of factors associated with perforation outcome. J Pediatr Surg 2015;50:1569–73. https://doi.org/10.1016/j.jpedsurg.2015.02.050.

[54] Buscher K, Wang H, Zhang X, et al. Protection from septic peritonitis by rapid neutrophil recruitment through omental high endothelial venules. Nat Commun 2016;7:10828. https://doi.org/10.1038/ncomms10828.

[55] Catar RA, Chen L, Cuff SM, et al. Control of neutrophil influx during peritonitis by transcriptional cross-regulation of chemokine CXCL1 by IL -17 and IFN -γ. J Pathol 2020;251:175–86. https://doi.org/10.1002/path.5438.

[56] Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genom 2012;13:S2. https://doi.org/10.1186/1471-2164-13-S4-S2.

[57] Crocker C, Akl M, Abdolell M, et al. Ultrasound and CT in the diagnosis of appendicitis: accuracy with consideration of indeterminate examinations according to STARD guidelines. Am J Roentgenol 2020;215:639–44. https://doi.org/10.2214/AJR.19.22370.

[58] Reich B, Zalut T, Weiner SG. An international evaluation of ultrasound vs. computed tomography in the diagnosis of appendicitis. Int J Emerg Med 2011;4:68. https://doi.org/10.1186/1865-1380-4-68.

[59] Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018;15:233.

[60] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3:e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.

[61] Kwong JCC, Erdman L, Khondker A, et al. The silent trial - the bridge between bench-to-bedside clinical AI applications. Front Digit Health 2022;4:929508. https://doi.org/10.3389/fdgth.2022.929508.

[62] Amoei M, Poenaru D. Patient-centered data science: an integrative framework for evaluating and predicting clinical outcomes in the digital health era 2024. 2024. https://doi.org/10.48550/ARXIV.2408.02677.

[63] Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2:e489–92. https://doi.org/10.1016/S2589-7500(20)30186-2.

Please cite this article as: Erman A et al., Machine-learning-assisted Preoperative Prediction of Pediatric Appendicitis Severity, Journal of Pediatric Surgery, https://doi.org/10.1016/j.jpedsurg.2024.162151