Learning Medical Risk Scores for Pediatric Appendicitis

Pedro Roig Aparicio
ETH Zürich

Ričards Marcinkevičs
ETH Zürich
ricards.marcinkevics@inf.ethz.ch

Patricia Reis Wolfertstetter
University Children's Hospital Regensburg

Sven Wellmann
University Children's Hospital Regensburg

Christian Knorr
University Children's Hospital Regensburg

Julia E. Vogt
ETH Zürich

***Abstract*—Appendicitis is a common childhood disease, the management of which still lacks consolidated international criteria. In clinical practice, heuristic scoring systems are often used to assess the urgency of patients with suspected appendicitis. Previous work on machine learning for appendicitis has focused on conventional classification models, such as logistic regression and tree-based ensembles. In this study, we investigate the use of risk supersparse linear integer models (risk SLIM) for learning data-driven risk scores to predict the diagnosis, management, and complications in pediatric patients with suspected appendicitis on a dataset consisting of 430 children from a tertiary care hospital. We demonstrate the efficacy of our approach and compare the performance of learnt risk scores to previous analyses with random forests. Risk SLIM is able to detect medically meaningful features and outperforms the traditional appendicitis scores, while at the same time is better suited for the clinical setting than tree-based ensembles.**

***Index Terms*—interpretable machine learning, pediatric appendicitis, diagnosis, treatment, decision support**

I. Introduction

More than one in every fourteen people suffers acute appendicitis during a lifetime, with the highest incidence rate at an age between 10 and 19 years [1]. The cause for this spontaneous disorder is still poorly understood and there are no consolidated management guidelines. The most common treatment is the appendectomy, a surgical removal of the appendix. However, there is increasing evidence of similar efficacy of nonsurgical interventions with antibiotics [2], [3]. Appendicitis requires prompt treatment, since increased time to intervention is associated with a higher risk of developing life threatening conditions [4]. For this reason, hospital practitioners often use risk scoring systems to assess patient urgency. These include the Alvarado (AS) and pediatric appendicitis scores (PAS) [5], [6] (see Table I). These scores were developed heuristically by experts and their predictive performance is limited to discarding extreme cases of appendicitis.

TABLE I
ALVARADO (*top*) AND PEDIATRIC APPENDICITIS (*bottom*) SCORES [5], [6]. RLQ STANDS FOR RIGHT LOWER ABDOMINAL QUADRANT; AND WBC – FOR WHITE BLOOD CELL COUNT.

Alvarado Score:

Condition	**Score**
RLQ tenderness	2
Elevated temperature: > 37.3 °C	1
Rebound tenderness	1
Migration of pain to the RLQ	1
Anorexia	1
Nausea or vomiting	1
Leukocytosis: WBC $> 10000/\mu l^{-1}$	2
Leukocyte left shift: $> 75\%$ neutrophils	1
Total	10

Pediatric Appendicitis Score:

Condition	**Score**
RLQ tenderness at cough, percussion, or hopping	2
Anorexia	1
Fever: ≥ 38.0 ℃	1
Nausea or vomiting	1
Tenderness over right iliac fossa	2
Leukocytosis: WBC $> 10000/\mu l^{-1}$	1
Neutrophilia: $> 75\%$ neutrophils	1
Migration of pain to RLQ	1
Total	10

A. Machine Learning and Appendicitis

There exists a rich body of literature on using machine learning (ML) models for predicting appendicitis in both general [7]–[9] and pediatric populations [10]–[14]. Most of these studies focus on relatively simple tabular datasets and leverage conventional classification models, such as logistic regression, decision trees, support vector machines, neural networks, and tree-based ensembles.

Published version: P. R. Aparicio, R. Marcinkevičs, P. Reis Wolfertstetter, S. Wellmann, C. Knorr and J. E. Vogt, "Learning Medical Risk Scores for Pediatric Appendicitis," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021, pp. 1507-1512, doi: 10.1109/ICMLA52953.2021.00243.

This raises the question, whether a more interpretable machine learning model which is amenable to the clinical setting could be utilized to achieve predictive performances comparable to previous results. In our work, we address this need by learning risk scores reminiscent of the classical AS and PAS in a data-driven manner and demonstrate that these simple and interpretable classifiers are on par with opaque techniques, such as random forests [15]. In particular, we apply risk supersparse linear integer models [16] to a dataset of 430 pediatric patients with suspected appendicitis to derive risk scores for diagnosis, necessity of surgical intervention, and developing complications.

B. Interpretable Machine Learning

Recently, there has been a significant interest in interpretable machine learning models, *i.e.* models that are designed to be human-understandable [17]. This interest is driven by practitioners applying machine learning in high-stakes decisions and new regulations postulating a "right to explanation" from algorithmic decision-making tools [18]. Despite the perceived trade-off between interpretability and predictive performance, many studies with legal, financial, and healthcare applications have demonstrated the efficacy of interpretable ML [19]–[23].

While classical machine learning models previously applied to predict appendicitis, *e.g.* logistic regression, could be deemed interpretable by some audiences, a user-centered perspective is important, and we must pay attention to the target user of our models [24]. In the context of this study, an integer risk score is more familiar and recognisable to a clinical practitioner than a simple logistic model with real-valued coefficients; and therefore, it is more likely to be accepted and deployed in daily clinical practice.

II. Methods

A. Data Acquisition and Preprocessing

Our analysis focuses on a publicly available dataset[1] from the Deparment of Pediatric Surgery at the tertiary Children's Hospital St. Hedwig in Regensburg, Germany, described in detail by Marcinkevics *et al.* [14]. It contains records for 430 patients aged between 0 and 18 years, admitted with abdominal pain and suspected appendicitis. The records include 38 variables, among which demographic, clinical, scoring, laboratory, and ultrasound (US) predictors. We consider three target variables: (*i*) *diagnosis* (appendicitis vs. no appendicitis), (*ii*) *treatment* (conservative vs. surgical), and (*iii*) *complications* (present vs. absent). Since patients who did not undergo surgery had no confirmed diagnosis, those who were treated with a conservative approach and had AS/PAS ≥ 4 and an appendix diameter ≥ 6 mm were assumed to have appendicitis. Most of the considered categories are balanced, however, only 12% of patients had a case of complicated appendicitis.

During exploratory analysis, non-sensible values and typos, such as negative body temperature, were removed manually and subsequently imputed. Ultrasound variables had many missing values (over 50%), which were imputed manually assuming that missingness was associated with absence of ultrasonographic observations. Continuous variables were encoded using 4-quantile-based discretization. Categorical variables were represented using either ordinal or one-hot encoding. Remaining missing values were imputed using k-nearest neighbors, as implemented in the Python scikit-learn package [25].

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

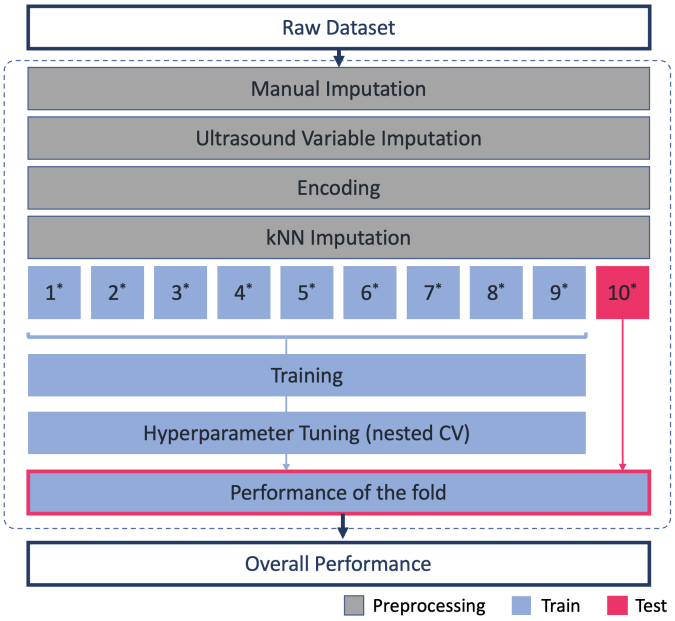

Fig. 1. End-to-end pipeline: the preprocessing consists of 4 steps. Hyperparameter grid search was conducted using cross-validation on the *training* set. The overall performance refers to the average across the 10 withheld test folds.

The whole data analysis pipeline is shown in Figure 1.

B. Risk SLIM Model

Risk SLIM [26] is a sparse linear model with integral coefficients designed for risk assessment. It is learnt by optimizing the logistic loss and outputs the risk of a patient, rather than a binary label. On the other hand, the original SLIM model [16] optimizes the 0-1 loss and is designed for decision making, outputting only a binary-valued label. In the clinical setting, it is crucial to prioritize the treatment of patients based on their risk. For this reason, risk SLIM was chosen over SLIM.

Risk SLIM solves the following optimization problem:

$$\min_{\boldsymbol{\lambda} \in \mathcal{L}} \; l(\boldsymbol{\lambda}) + C_0 \|\boldsymbol{\lambda}\|_0, \tag{1}$$

where the logistic loss is defined as

$$l(\boldsymbol{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{-y_i \boldsymbol{\lambda}^\top \mathbf{x}_i}\right), \tag{2}$$

and the ℓ_0-*norm* is given by $\|\boldsymbol{\lambda}\|_0 = \sum_{j=1}^{d} \mathbf{1}_{\{\lambda_j \neq 0\}}$. By minimizing the logistic loss we expect to achieve high AUROC

TABLE II
TUNED HYPERPARAMETER VALUES FOR RISK SLIM MODELS PREDICTING DIAGNOSIS, BASED ON THE FULL SET OF VARIABLES AND WITHOUT ULTRASOUND, TREATMENT, AND COMPLICATIONS.

	# var-s	$\mathcal{L}$	C_0	Time lim., s
Diagnosis (full)	6	$\{-6,...,6\}$	10^{-6}	500
Diagnosis (w/o US)	6	$\{-5,...,5\}$	10^{-6}	300
Treatment	4	$\{-10,...,10\}$	10^{-6}	60
Complications	10	$\{-2,...,2\}$	10^{-6}	300

and well-calibrated risk predictions. C_0 is the parameter controlling the trade-off between the logistic loss and the sparsity of the coefficient vector $\boldsymbol{\lambda}$. Finally, $\mathcal{L} \subset \mathbb{Z}^{d+1}$ denotes a set of feasible coefficients as defined by the user, *e.g.* $\mathcal{L} = \{-5, -4, ..., 0..., 4, 5\}^{d+1}$.

Finding integer-valued sparse coefficients $\boldsymbol{\lambda}$ is NP-hard, since Equation 1 is a mixed-integer linear program (MILP). Ustun and Rudin [26] propose a new lattice cutting plane algorithm to solve the problem described above. The running time scales linearly to the number of data points and allows applying risk SLIM to large datasets.

III. RESULTS

In our experiments, we apply risk SLIM to the mentioned dataset of 430 pediatric patients to learn risk scores for the three target variables. In addition, we compare the performance of risk SLIM to random forests (RF) [15] in terms of areas under receiver operating characteristic (AUROC) and precision-recall (AUPR) curves that were reported before for the same dataset by Marcinkevics *et al.* [14]. The experiments were implemented in Python; and the original implementation of risk SLIM was used.[2]

A nested 10-fold cross validation (CV) with a stratified split was used for model hyperparameter tuning and comparison. The hyperparameters of risk SLIM models corresponding to the final results are reported in Table II. For random forests, we replicated the results by Marcinkevics *et al.* [14] ourselves with the scikit-learn implementation of RFs [25].

A. Comparison with Random Forests

We first evaluate the predictive performance of scores learnt with risk SLIM and compare it with random forests. Figure 2 shows 10-fold CV results for random forests and risk SLIM across all three target variables. We observe that risk SLIM achieves average AUROCs and AUPRs comparable to random forest's performance for diagnosis and treatment. For complications, the average AUPR for risk SLIM is slightly lower than for RFs, however, this difference is not significant given a large standard deviation across the folds. We attribute relatively low AUPRs for complications attained by all models to the extremely low prevalence of complicated appendicitis cases (12%). Overall, these results suggest that for the considered simple tabular dataset, a linear regularized model, such as risk SLIM, does not perform worse than highly flexible random forests and thus, is a viable and more interpretable alternative.

[2]https://github.com/ustunb/risk-slim

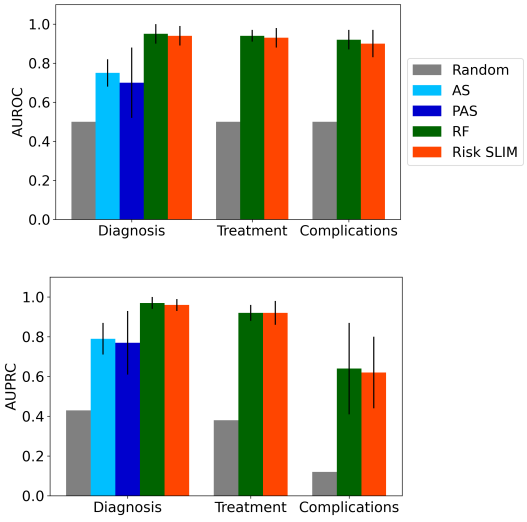

Fig. 2. Performance of AS (■) and PAS (■) scores, random forests (■), and risk SLIM (■) at predicting diagnosis, treatment assignment, and complications. We tuned the number of trees in RF, using 80, 200, and 400 estimators for diagnosis, treatment, and complications, respectively. In addition, we provide the expected performance of a random guess (■) as a naïve baseline. Averages and standard deviations (error bars) are reported across 10 folds in stratified cross-validation.

B. Comparison with Traditional Appendicitis Scores

We now examine the performance of traditional appendicitis scores. As mentioned in Section I, AS and PAS are commonly used to assess patients with suspected appendicitis (see Table I). Originally, Alvarado and PAS are intended for diagnosing extreme cases of appendicitis. Therefore, it makes no sense to use these scores to predict treatment and complications, so we discarded these targets from the evaluation, focusing only on diagnosis. Figure 2 shows the performance of appendicitis scores at predicting the diagnosis across the ten folds of CV.

The overall performance of both scores is considerably worse than for machine learning approaches. Their low predictive performance is not surprising, since these scores are only used in practice to discard extreme cases: patients with a score ≤ 4 can be discarded and patients with a score ≥ 7 are likely to have the disease [27]. Moreover, AS and PAS are purely based on laboratory and clinical findings and do not consider ultrasonographic information. To facilitate a fairer comparison, we fit a risk SLIM score without ultrasound predictor variables. The ultrasound-free model achieves an AUROC and AUPR of 0.85 ± 0.04 and 0.90 ± 0.03, respectively, performing worse than the full risk SLIM score, but considerably better than AS and PAS, both on average and in terms of variability in performance.

To conclude, we can see that data-driven approaches which incorporate ultrasound results offer a considerable improve-

ment in predictive performance over the traditional appendicitis scores, such as AS and PAS.

C. Qualitative Results

Risk SLIM outputs scoring tables from which the score for each patient can be easily calculated by adding small integers if certain conditions apply. In our case, these conditions are based on symptoms and medical findings. After calculating the score, the associated risk can be evaluated according to the formula, following the cumulative logistic distribution:

$$\text{Patient Risk} = \frac{1}{1 + e^{\beta_0 - \boldsymbol{\lambda}^\top \mathbf{x}_i}}, \tag{3}$$

wherein $\boldsymbol{\lambda}^\top \mathbf{x}_i$ corresponds to the total score of the i-th patient and can be easily calculated by adding integer coefficients in $\boldsymbol{\lambda}$ for the symptoms that the patient has. β_0 is a parameter learnt by risk SLIM and is used as a threshold in patient classification. For example, for diagnosis (see Table III), the model learns $\beta_0 = 7$. Since having to use a calculator can be inconvenient, especially when the health of a patient is at stake, this step can be made easier by providing a table that converts integral scores to risks, such as Table IV for diagnosis.

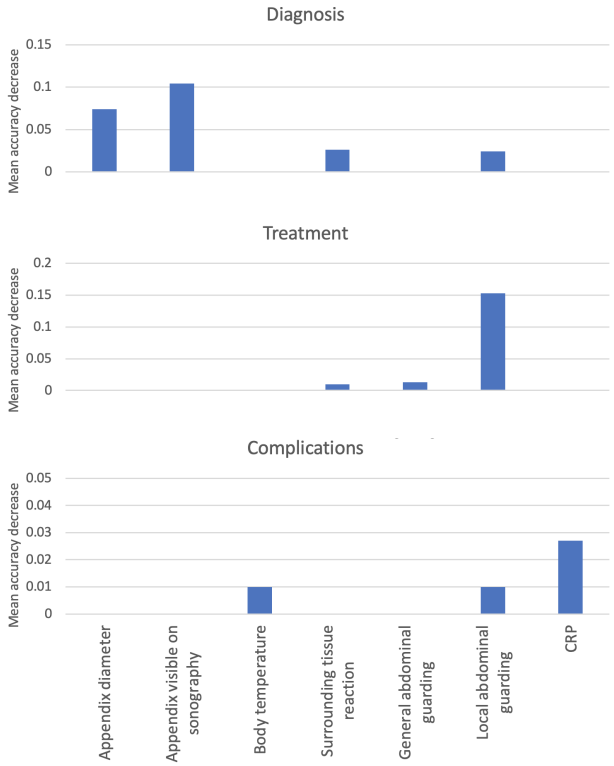

Fig. 3. RF variable importance values, given by mean decrease in accuracy, for a few most important predictors across three responses. For predicting diagnosis (*top*), the most important features are the visibility of the appendix in the US, appendix diameter, local abdominal guarding and inflammation signs in the tissue surrounding the appendix. For the treatment (*middle*), the RF model mostly focuses on local abdominal guarding. In the case of complications (*bottom*), the most relevant variable is the level of C-reactive protein, followed by local abdominal guarding and body temperature.

The rules learnt by risk SLIM (see Table III) are in line with the medical literature [6], [7], [28], [29], assigning highest

TABLE III
SCORING SYSTEMS OBTAINED BY RISK SLIM FOR DIAGNOSIS (*top*), TREATMENT (*middle*), AND COMPLICATIONS (*bottom*).

Diagnosis (full):

Rule	**Score**
Peritonitis generalized	6
Appendix diameter 9–17 mm	6
Appendix diameter 5.9–9 mm	5
Appendix on sonography	4
Peritonitis local	2
β_0	7

Diagnosis (w/o US):

Rule	**Score**
Peritonitis generalized	2
Peritonitis local	2
WBC count > 15.8 10^3/µl	2
Migratory pain	1
Weight 31.1–42.0 kg	1
Body temperature 37.0–37.4 °C	1
WBC count 11.8–15.8 10^3/µl	1
Neutrophil percentage 73.6–82.0%	1
CRP at entry 7.0–31.75 mg/l	1
Dysuria	-1
β_0	2

Treatment:

Rule	**Score**
Peritonitis generalized	10
Peritonitis local	4
Appendix wall layers irregular	2
WBC count 15.8–33.6 10^3/µl	2
Appendix diameter 9–17 mm	1
Pathological lymph nodes	-1
β_0	3

Complications:

Rule	**Score**
Peritonitis generalized	4
Peritonitis local	4
WBC in urine	3
Migratory pain	2
Body temperature 38.2–40.3 °C	2
CRP at entry 31.75–365.10 mg/l	2
Nausea	1
Height 138.0–150.5 cm	-1
Neutrophil percentage 73.55–82.0%	-1
Weight 55–98 kg	-2
β_0	8

TABLE IV
SCORE-TO-RISK CONVERSION TABLE FOR DIAGNOSIS (*with ultrasound*). THE DOCTORS CAN EASILY CONVERT TOTAL PATIENT SCORES OBTAINED FROM RISK SLIM CONDITIONS TO THE RISK OF HAVING APPENDICITIS.

Score	0–4	5	6	7	8	9	10–23
Risk	0.0	0.1	0.3	0.5	0.7	0.9	1.0

importance to abdominal guarding and ultrasound findings, such as visibility and diameter of appendix. For diagnosis, risk SLIM focuses only on abdominal guarding and ultrasound variables in contrast to the traditional scores, which are based on clinical findings, such as nausea, anorexia, and fever. When predicting treatment assignment, the model associates the abdominal guarding to a risk of surgical intervention. High white blood cell (WBC) count, which is associated with inflammation, is also relevant when deciding if the patient should be treated conservatively or surgically. Interestingly, a negative score is assigned for the presence of pathological lymph nodes. For example, mesenteric lymphadenitis mimics the symptoms of appendicitis [30] and if differentiated should be accounted for in treatment assignment. For complications, alongside with peritonitis, risk SLIM includes elevated body temperature and high concentration of C-reactive protein at entry. The association between the latter and complicated appendicitis has been explored in the medical literature as well [29]. Interestingly, the model assigns negative coefficients to the height and body weight. This might be due to overfitting or correlations with other clinically relevant variables, such as BMI and age.

Figure 3 shows random forest feature importance, given by the average decrease in accuracy upon randomly permuting the considered feature, for all three response variables. Similar to risk SLIM, RF relies heavily on ultrasonography for predicting diagnosis, abdominal guarding – for treatment, and CRP levels – for complications. While it appears that both RF and risk SLIM utilise similar features and achieve comparable predictive performance, scores learnt by risk SLIM are less opaque, have a format familiar to physicians, and can be easily computed by hand.

Comparing the full risk SLIM model to AS and PAS (*cf.* Table I), the only feature common across all scores is abdominal guarding. Whereas risk SLIM relies mostly on ultrasonographic findings, traditional appendicitis scores were designed for quick screening of patients and do not require ultrasound imaging. On the other hand, the risk SLIM score trained without US predictors is quite similar to the medical appendicitis scores, assigning high importance to abdominal guarding, WBC levels, fever, and neutrophilia.

IV. CONCLUSIONS AND OUTLOOK

In this paper, we investigated the utility of interpretable machine learning models for predicting the diagnosis, management, and complications of appendicitis in pediatric patients. We demonstrated that sensible and medically relevant risk scores can be learnt using risk SLIM in a purely data-driven manner that perform as well as random forests used by previous researchers [14]. This resonates well with the previous claims that interpretability does not harm predictive performance in simple datasets and in the presence of meaningful features [19]–[23].

Our models offer a significant improvement over traditional medical scores based on expertise and hold a promise of supporting clinical practitioners in their decisions. An important advantage of our classifiers over the previous attempts to leverage machine learning for appendicitis is that the simple format of a risk score is more amenable to an average medical doctor and is consistent with the frequently used scoring systems like Alvarado and PAS.

The obtained scores should be validated externally and on a more diverse set of patients from multiple clinical centers. Note, that all of the patients in the dataset were suspected to have appendicitis, therefore, the scoring system is limited to the described setting and should be adapted or retrained if applied to a wider patient population.

In the future work, we plan to investigate other interpretable classification approaches, particularly, born-again tree ensembles [31] and generalized additive models [32]. We also plan to extend the current dataset with raw US images to leverage recent advancements in deep learning for medical ultrasound [33], [34].

ACKNOWLEDGEMENT

The authors thank Ece Özkan Elsen for valuable discussion and comments. Ričards Marcinkevičs is supported by the SNSF grant #320038189096.

REFERENCES

[1] D. G. Addiss, N. Shaffer, B. S. Fowler, and R. V. Tauxe, "The epidemiology of appendicitis and appendectomy in the United States," *American Journal of Epidemiology*, vol. 132, no. 5, pp. 910–925, 1990.

[2] J. F. Svensson, B. Patkova, M. Almström, H. Naji, N. J. Hall, S. Eaton, A. Pierro, and T. Wester, "Nonoperative treatment with antibiotics versus surgery for acute nonperforated appendicitis in children," *Annals of Surgery*, vol. 261, no. 1, pp. 67–71, 2015.

[3] J. Svensson, N. Hall, S. Eaton, A. Pierro, and T. Wester, "A review of conservative treatment of acute appendicitis," *European Journal of Pediatric Surgery*, vol. 22, no. 3, pp. 185–194, 2012.

[4] N. A. Bickell, A. H. Aufses, M. Rojas, and C. Bodian, "How time affects the risk of rupture in appendicitis," *Journal of the American College of Surgeons*, vol. 202, no. 3, pp. 401–406, 2006.

[5] A. Alvarado, "A practical score for the early diagnosis of acute appendicitis," *Annals of Emergency Medicine*, vol. 15, no. 5, pp. 557–564, 1986.

[6] M. Samuel, "Pediatric appendicitis score," *Journal of Pediatric Surgery*, vol. 37, no. 6, pp. 877–881, 2002.

[7] C.-H. Hsieh, R.-H. Lu, N.-H. Lee, W.-T. Chiu, M.-H. Hsu, and Y.-C. J. Li, "Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks," *Surgery*, vol. 149, no. 1, pp. 87–93, 2011.

[8] L. Deleger, H. Brodzinski, H. Zhai, Q. Li, T. Lingren, E. S. Kirkendall, E. Alessandrini, and I. Solti, "Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department," *Journal of the American Medical Informatics Association*, vol. 20, no. e2, pp. e212–e220, 2013.

[9] P. Rajpurkar, A. Park, J. Irvin, C. Chute, M. Bereket, D. Mastrodicasa, C. P. Langlotz, M. P. Lungren, A. Y. Ng, and B. N. Patel, "AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining," *Scientific Reports*, vol. 10, no. 1, 2020.

[10] J. Reismann, A. Romualdi, N. Kiss, M. I. Minderjahn, J. Kallarackal, M. Schad, and M. Reismann, "Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach," *PLOS ONE*, vol. 14, no. 9, p. e0222030, 2019.

[11] E. Aydin, İ. U. Türkmen, G. Namli, Ç. Öztürk, A. B. Esen, Y. N. Eray, E. Eroğlu, and F. Akova, "A novel and simple machine learning algorithm for preoperative diagnosis of acute appendicitis in children," *Pediatric Surgery International*, vol. 36, no. 6, pp. 735–742, 2020.

[12] O. F. Akmese, G. Dogan, H. Kor, H. Erbay, and E. Demir, "The use of machine learning approaches for the diagnosis of acute appendicitis," *Emergency Medicine International*, vol. 2020, pp. 1–8, 2020.

[13] C. Stiel, J. Elrod, M. Klinke, J. Herrmann, C.-M. Junge, T. Ghadban, K. Reinshagen, and M. Boettcher, "The modified heidelberg and the AI appendicitis score are superior to current scores in predicting appendicitis in children: A two-center cohort study," *Frontiers in Pediatrics*, vol. 8, 2020.

[14] R. Marcinkevics, P. R. Wolfertstetter, S. Wellmann, C. Knorr, and J. E. Vogt, "Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis," *Frontiers in Pediatrics*, vol. 9, 2021.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2015.

[17] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[18] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[19] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for HealthCare," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[20] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.

[21] J. Zeng, B. Ustun, and C. Rudin, "Interpretable classification models for recidivism prediction," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 3, pp. 689–722, 2016.

[22] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning certifiably optimal rule lists for categorical data," 2018, arXiv:1704.01701.

[23] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018.

[24] M. Ribera and À. Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI," in *IUI Workshops*, 2019.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] B. Ustun and C. Rudin, "Optimized risk scores," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

[27] C. D. Douglas, "Randomised controlled trial of ultrasonography in diagnosis of acute appendicitis, incorporating the alvarado score," *BMJ*, vol. 321, no. 7266, pp. 919–919, 2000.

[28] T. D. Owen, H. Williams, G. Stiff, L. R. Jenkinson, and B. I. Rees, "Evaluation of the Alvarado score in acute appendicitis," *Journal of the Royal Society of Medicine*, vol. 85, no. 2, pp. 87–88, 1992.

[29] H.-P. Wu, C.-Y. Lin, C.-F. Chang, Y.-J. Chang, and C.-Y. Huang, "Predictive value of C-reactive protein at different cutoff levels in acute appendicitis," *The American Journal of Emergency Medicine*, vol. 23, no. 4, pp. 449–453, 2005.

[30] B. Toorenvliet, A. Vellekoop, R. Bakker, F. Wiersma, B. Mertens, J. Merkus, P. Breslau, and J. Hamming, "Clinical differentiation between acute appendicitis and acute mesenteric lymphadenitis in children," *European Journal of Pediatric Surgery*, vol. 21, no. 2, pp. 120–123, 2011.

[31] T. Vidal and M. Schiffer, "Born-again tree ensembles," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9743–9753.

[32] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 150–158.

[33] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[34] R. J. G. van Sloun, R. Cohen, and Y. C. Eldar, "Deep learning in ultrasound imaging," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 11–29, 2019.