Machine learning for detecting negative appendectomies in pediatric patients, a bilirubin subset analysis

Berković, Karlotta Sofie

UNIVERSITY OF SPLIT
SCHOOL OF MEDICINE

Karlotta Sofie Berković

MACHINE LEARNING FOR DETECTING NEGATIVE APPENDECTOMIES IN PEDIATRIC PATIENTS, A BILIRUBIN SUBSET ANALYSIS

Diploma Thesis

Academic year:
2023/2024

Mentor:
Assoc. Prof. Zenon Pogorelić, MD, PhD

Split, July 2024

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First, I would like to thank my mentor, Prof. Zenon Pogorelić, MD, PhD, and Dr. Ivan Maleš and Dr. Josip Vrdoljak, for helping and supporting me during the development of my diploma thesis.

Since this chapter of my life marks an important period of my personal growth, I want to take this opportunity to express my appreciation to my parents, who made this entire journey possible, and my siblings, who reminded me never to lose sight of my fun.

I am deeply thankful for my closest friends and my partner for inspiring me, giving me the strength and support to never give up on myself and to always get back up on my feet even stronger than I was before.

I am very grateful for your unwavering support.

Furthermore, I want to extend my heartfelt thanks to all my friends and colleagues for making these past six years one of the most joyful and memorable times of my life.

Split, you will always be my home!

List of Abbreviations:

AA – Acute Appendicitis

AIR score – Appendicitis Inflammatory Response score

AUS – Abdominal Ultrasound

BMI – Body Mass Index

CBC – Complete Blood Count

CI – Confidence Interval

CRP – C-reactive Protein

CT – Computed Tomography

IL-6 – interleukin-6

LA – Laparoscopic Appendectomy

LGR-1 – leucine-Rich α-2-glycoprotein 1

LLQ – Left Lower Quadrant

MCHC – Mean Corpuscular Hemoglobin Concentration

MDI – Mean Decrease Impurity

ML – Machine Learning

MMT – Maximal Mural Thickness

MOD – Maximal Outer Diameter

MPV – Mean Platelet Volume

MRI – Magnetic Resonance Imaging

NGAL – neutrophil gelatinase-associated lipocalin

NLR – Neutrophil/Lymphocyte ratio

NOM – Non-operative Management

OA – Open Appendectomy

pARC score – Pediatric Appendicitis Risk Calculator

PAS – Pediatric Appendicitis Score

PHD – Pathohistological Diagnosis

PMN leukocytes – Polymorphonuclear Leukocytes

PTX-3 – Pentraxin-3

RDW – Red Cell Distribution Width

RLQ – Right Lower Quadrant

ROC – Receiver Operating Characteristic

SD – Standard Deviation

SHAP – Shapley Additive Explanations

TLR – Thrombocyte/Lymphocyte ratio

WBC count – White Blood Cell count

XGBoost – Extreme Gradient Boosting

1. INTRODUCTION

1.1. THE APPENDIX

1.1.1. Embryology

Understanding the embryologic development of the midgut is crucial for understanding the intimately related development of the appendix.

The midgut herniates into the umbilical cord at 4 weeks, while the foregut and hindgut are fixed due to retention bands. The gut rotates counterclockwise at 5 weeks leading to the return of the prearterial segment of the midgut into the abdomen. Following this rotation, at 12 weeks, the cecum is positioned in the upper abdomen, having undergone a 270° rotation, once the postarterial segment has decreased. As the gut elongates, the duodenum, ascending and descending colon become fixed as parts of the primitive mesentery fuse to the posterior abdominal wall. The appendix emerges as a bud from the cecum, and as it is pushed ahead of the cecum, it adopts various anatomical positions, which are further discussed in the Anatomy part (1).

1.1.2. Anatomy

The appendix is an approximate 9 cm long pencil-shaped structure, but short and long forms also exist. The typical origin of the appendix is on average 1.7 to 2.5 cm below the terminal part of the ileum near the ileocecal valve (2).

The most common location of the appendix is retrocecal, but other normal anatomic variations can be found subcecal, pre-ileal, post-ileal, and in pelvic positions, which can complicate diagnosing appendiceal pathologies (3, 4) (Figure 1). The findings on clinical examination including the site of pain, will be influenced the anatomic position of the appendix.

The appendix is a true diverticulum at the posteromedial border of the cecum (4). Its wall consists of the hypoechoic mucosal layer, the echogenic submucosal layer, the hypoechoic muscularis propria layer (longitudinal and circular), and the outermost echogenic serosal covering (3, 5). The ileocolic artery, which terminates in the appendiceal artery, is the main blood supply of the appendix. The vessel crosses the length of the mesoappendix and terminates at the tip of the organ (4). Lymph drainage from both the appendix and parts of the cecum is via the ileocolic lymph nodes, which proceeds to the superior mesenteric lymph node (6).

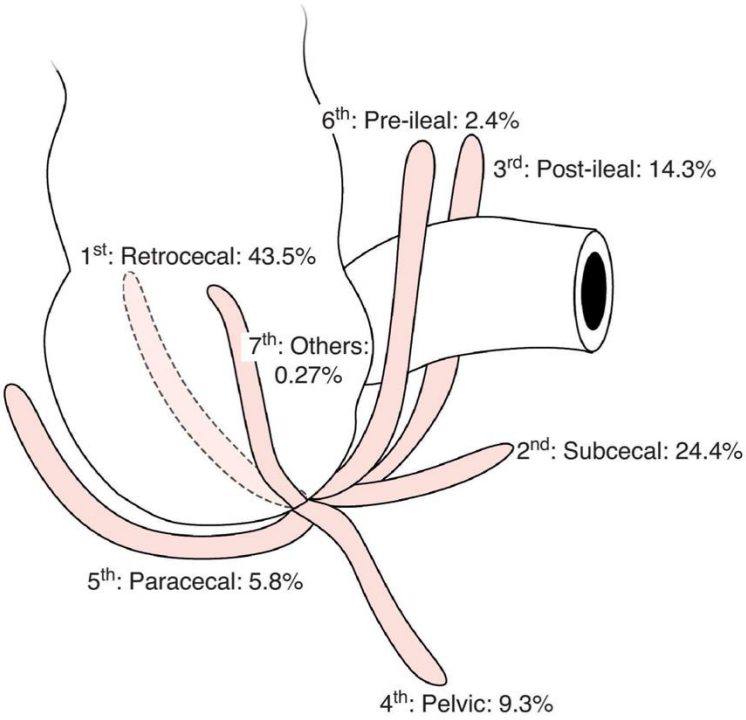

Figure 1. Positions of the vermiform appendix according to their probability;
Source: Randal Bollinger R, Barbas AS, Bush EL, Lin SS, Parker W. Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. J Theor Biol. 2007;249:826-31.

1.1.3. Physiology and Function

The appendix mainly has two different types of tissues: lymphoid tissue and neuroendocrine cells. The lymphoid tissue aids in maturation of B lymphocytes and IgA antibodies, while the neuroendocrine cells assist with various biological control mechanisms, through production of amines and hormones. Since there is no clear evidence for its function in humans, the appendix is seen as a vestigial organ (6). New research suggests that the appendix has some immune function, because of its association with lymphatic tissue, although the specific nature of that function is still unknown. Theories propose that the appendix is a “safe house” for commensal bacteria, which provides support for bacterial growth and the ability of re-inoculation after exposure with a pathogen to the colonic mucosa and the following purging of intestinal tract contents. This theory is based on new understandings of immune-mediated biofilm formations, biofilm distribution in the larger bowel, the association of lymphoid tissue with the appendix, and the ability of biofilms to protect and support colonization (7).

1.2. ACUTE APPENDICITIS

1.2.1. Definition

The definition of acute appendicitis (AA) is an acute inflammation of the vermiform appendix. This condition can be classified into two main categories: uncomplicated appendicitis and complicated appendicitis. In the absence of complications, an inflamed appendix is termed `simple appendicitis'. On the other hand, complicated appendicitis is characterized by appendicitis accompanied by various complications, including perforation, death of tissue due to reduced blood supply (gangrene), the formation of a periappendicular abscess filled with pus, the presence of an appendiceal fecalith, or even the existence of a tumor.

In summary, in contrast to simple appendicitis, complicated appendicitis involves additional complications or factors that make the condition more severe and may require different management approaches (8). In pediatric cases AA is more difficult to recognize because most children cannot articulate or localize their abdominal pain accurately (5).

1.2.2. Epidemiology

Although no age is exempt, the incidence of AA is highest between the ages of 10 and 20 years, with a male to female ratio of 1.4:1.In pediatric population it is most common between the ages 12 and 14 (9). In the United States the lifetime overall risk amounts to 8.6% for males and 6.7% in females, therefore exhibiting slight male predisposition (8).

1.2.3. Pathogenesis and Pathophysiology

AA is the most common etiology of acute abdomen in the pediatric population, requiring surgery (5). The Etiology of AA is most likely multifactorial, with all mechanisms leading to obstruction of the appendiceal orifice (3, 5). This can be evoked by many different mechanical etiologies like from an appendicolith, appendiceal tumors, intestinal parasites, and hypertrophied lymphatic tissue or foreign bodies (11, 12).

Mechanical obstruction leads to bacteria buildup of both aerobic and anaerobic bacteria including *Escherichia coli*, *Peptostreptococcus*, *Bacteroides*, and *Pseudomonas* (3). At the same time, it can be found combined with lymphoid hyperplasia which results in further inflammation with an intraluminal and intramural pressure buildup, resulting in small vessel occlusion and lymphatic stasis. This vascular and lymphatic compromise is aggravated by the appendix itself still producing mucus. The culmination of these mechanisms will ultimately result in perforation (3).

1.2.4. Histopathology

The greater the inflammation, both in terms of its degree and extent, the more severe and prolonged the condition of AA becomes. Proliferation of neutrophils can be seen under the microscope in muscularis propria layer, while in later stages other tissues like periappendicular fat and surroundings become inflamed as well (3).

1.2.5. Signs and Symptoms

The primary presenting complaint of AA is colicky central abdominal and in 50% of patients pain followed by vomiting with migration of pain to the right iliac fossa or McBurney's point can be found. The pain usually intensifies in the first 24 hours becoming more constant and sharper until it migrates, usually accompanied by loss of appetite. Movement will often exacerbate the pain, while the position with maximal comfort will be in the right lateral decubitus position. Low-grade fever is usually seen in AA, but whenever the temperature exceeds 38.3°C perforation should be suspected (12). Typical or atypical presentations are influenced by the anatomical position of the appendix in each patient as well as the patients age (13).

Establishing a proper diagnosis in children younger than the age of six is a challenge and therefore often delayed, because they frequently have an unusual clinical presentation. The younger the child, the more advanced the stage of disease and the greater is the risk of perforation (14). Every child that seems withdrawn and is showing signs of abdominal pain, fever, and diarrhea, should be considered for appendicitis and even perforation, since these symptoms are present significantly more often in children with perforated AA (9, 15).

1.3. DIAGNOSIS OF ACUTE APPENDICITIS

The diagnosis in AA is mostly based on clinical findings and anamnesis only, without the need of further diagnostic tool (15).

If the diagnosis is not clear after conducting history taking, clinical examination, and blood tests, the diagnostic algorithm prioritizes ultrasound as the primary imaging strategy (16). To prevent unnecessary surgery and to avoid complications history and physical examination have to be as accurate as possible (12). In cases where abdominal ultrasound (AUS) cannot be used for diagnosis or is non-revealing, it is recommended to use a computed tomography (CT) scan or magnetic resonance imaging (MRI), which is especially recommended for pregnant patients instead of CT (16).

1.3.1. Clinical signs

Inspecting specific clinical signs is a major part of physical examination although they occur in less than 40% of patients with a positive AA. Nevertheless, the examiner should be able to establish an accurate diagnosis even in the absence of these signs.

A summary of the most important and well-known signs includes (13, 16, 18):

1. McBurney's point: tenderness in the right lower quadrant (RLQ). The point can be found at two-thirds of the distance from the umbilicus to the right anterior superior iliac spine.
2. Rovsing's sign: RLQ pain while palpating the left lower quadrant (LLQ).
3. Obturator sign: increasing pain in RLQ while the patient is supine and the patient's right leg, flexed at the hip, is internally and externally rotated.
4. Psoas sign: If the patient signals increasing pain while the examiner passively extends the patient's right leg at the hip with knees extended and the patient lying on their left side.
5. Blumberg's sign: rebound tenderness, so pain felt when pressure is released from the RLQ of the abdomen after palpating.

Since these clinical does not have to be positive even though the patient does have AA it is always crucial to have a combination of accurate anamnesis, clinical assessment, laboratory values, and imaging to establish a proper diagnosis (15).

1.3.2. Laboratory findings

The diagnosis of AA is traditionally made by a good history combined with a proper clinical examination. There are no specific laboratory factors that are diagnostic, but leukocytosis with additional elevated factors of inflammation is supportive of the diagnosis. Elevations in C-reactive Protein (CRP) level and white blood cell (WBC) count increase the positive finding of AA five-fold (18).

1.3.2.1. CRP

CRP is synthesized by the liver and is induced by interleukin-6 (IL-6) action during the acute phase of an inflammatory process. Minor elevation of CRP up to 1.0 mg/dL can be found in for example obesity, pregnancy, diabetes, common cold, smoking. Moderate elevation (up to 10.0 mg/dL) can be found in systemic inflammation such as rheumatoid arthritis or other autoimmune diseases and marked elevations of more than 10.0 mg/dL are typically found in acute bacterial infections (19). CRP is a sensitive but non-specific inflammatory marker, which is useful in diagnosing appendicular perforation and abscess formation especially in children (20).

1.3.2.2. WBC count

The normal WBC count in blood varies in between 4 – 10 x 10^9/L (21). Globally, a complete blood count (CBC) is the most recommended laboratory investigation for children suspected of having acute appendicitis. Despite the fact that the WBC count is expected to be elevated in cases of AA, its specificity and sensitivity are limited. An increased WBC count is also observed in other medical conditions such as gastroenteritis, mesenteric lymphadenitis, pelvic inflammatory disease, and various infections. The combination of leukocytosis and an elevated neutrophil count, along with an increased CRP, may achieve a diagnostic sensitivity approaching 98% for acute appendicitis (20).

1.3.2.3. Other inflammatory markers

In addition to the standard biomarkers, there are several new biomarkers for acute appendicitis, such as hyperfibrinogenemia (22), ischemia modified albumin (23), pentraxin-3 (PTX-3) (24), hyperbilirubinemia (25), neutrophil gelatinase-associated lipocalin (NGAL) or IL-6 (26), hyponatremia (27) and leucine-Rich α-2-glycoprotein 1 (LGR-1) (28), have recently been investigated. These biomarkers showed good predictive values for the detection of acute appendicitis and the differentiation between complicated and simple acute appendicitis

1.3.2.4. Bilirubin

Bilirubin is a yellow pigment of bile created by degradation of heme-containing proteins. Plasma bilirubin levels are frequently elevated in patients with liver lesions, which may lead to hyperbilirubinemia, but is neither a specific, nor a sensitive marker of liver function. However, since elevation of bilirubin levels is found frequently, it serves as a laboratory marker that is done routinely and well-established in a variety of patients (29).

Bilirubin serves a positive predictive value. Studies have shown that it is an important indicator for identifying patients at higher risk of appendiceal perforation or gangrene. However, when assessing patients with suspected acute appendicitis, bilirubin levels should be considered alongside clinical examinations and other laboratory tests (30).

1.3.3. Scoring systems

1.3.3.1. Alvarado Score

To decrease morbidity associated with negative findings but concomitantly improve early diagnosis of AA, various scoring systems and algorithms have been introduced. The Alvarado Score, described in 1986, is the most widely recognized and frequently used scoring system, derived from a retrospective analysis of patients undergoing surgery for suspected appendicitis (12, 18). It is simple, effective, user-friendly, and is an accurate and consistent instrument in out ruling AA and detecting patients at higher risk. The scoring system is divided into three categories according to Alvarado: three symptoms (migratory right iliac fossa pain, anorexia, and nausea/vomiting), three physical signs (tenderness/rebound pain and elevation of temperature) and two laboratory findings (leukocytosis and neutrophilic shift to the left).

Each indicator is assigned one number according to their diagnostic weight (Table 1), with a score of 5 or 6 being compatible with the diagnosis of AA, a score of 7 or 8 indicating probable AA, and a score of 9 or 10 being highly likely in the diagnosis of AA. A patient with a score of 7 or more requires surgery. The sensitivity of 71% and specificity of 68% demonstrate that the scoring system does not have 100% diagnostic certainty, and furthermore it elicits limits in female patients (18, 31).

Table 1. The Alvarado Scoring System for Acute Appendicitis.

Alvarado Score	
Signs	***Score***
RLQ tenderness	+2
Temperature ≥ 37.3°C	+1
Rebound tenderness	+1
Symptoms	
Migration of pain to RLQ	+1
Anorexia	+1
Nausea/Vomiting	+1
Laboratory values	
Leukocytosis > $10x10^9$	+2
>75% Neutrophils	+1
Total	**10**

Abbreviations: RLQ – right lower quadrant;

1.3.3.2. The Appendicitis Inflammatory Response Score

The Appendicitis Inflammatory Response Score (AIR) was introduced by Anderson comprising seven indicators. The goal of the AIR score is to discriminate objectively, when there is uncertainty of the diagnosis of AA, to overcome the drawbacks from the Alvarado score. The indicators are graded based on the severity of symptoms and signs. The AIR score presents laboratory variables segmented into intervals, while C-reactive protein has been included additionally due to its discriminatory efficacy in assessing appendicitis. Group 1 (score 0-4) represents patients with a very low probability of suffering from AA, with an outpatient follow-up if there is an unaltered general condition. Patients of group 2 (score 5-8) have a moderate possibility of having AA and will be actively observed in hospital with rescoring or additional measures. Patients with a score of 9-12 (group 3) have a very high probability of suffering of an AA, surgical exploration is proposed (Table 2) (32).

Table 2. The Appendicitis Inflammatory Response (AIR) score.

Appendicitis Inflammatory Response (AIR) Score	
Vomiting	1
Pain in right inferior fossa	1
Rebound tenderness: light	1
medium	2
strong	3
Body temperature ≥ 38°C	1
WBC count: 10.0-14.9 x10^9/L	1
15.0 x10^9/L	2
PMN Leukocytes: 70-84%	1
≥ 85%	2
CRP concentration 10-49 g/L	1
≥ 50 g/L	2
Total score	**0-12**

Abbreviations: WBC – white blood cell, PMN – polymorphonuclear, CRP – C-reactive protein;

1.3.3.3. The Pediatric Appendicitis Score

The Pediatric Appendicitis Score (PAS) focuses specifically on the symptoms and physical signs, shown in Table 3, that are distinctive to children (33). To determine next steps in management, PAS should be connected with AUS or abdominal X-ray (34).

Table 3. Pediatric Appendicitis Scoring system.

The Pediatric Appendicitis Score (PAS)	
RLQ tenderness to cough, percussion, or hopping	+2
Anorexia	+1
Fever: temperature ≥ 38°C	+1
Nausea or vomiting	+1
Tenderness over right iliac fossa	+2
Leukocytosis: WBC count > 10,000	+1
Neutrophilia: ANC > 7,500	+1
Migration of pain to RLQ	+1
Total	**10**

*ANC – absolute neutrophil count, WBC count – white blood cell count, RLQ – right lower quadrant;

Scores of less than 4 show a low likelihood of acute appendicitis, and likely do not need imaging. If there is additional absence of RLQ pain, or pain with walking/jumping, and an ANC of lower than 6,750, the score has a negative predicting value of 95%. A Score of 4-6 should be considered for additional imaging, preferably AUS, together with a surgical consult. In high risk patient with a score of over 6 a surgery is needed, with or without prior imaging (35).

1.3.3.4. Pediatric Appendicitis Risk Calculator

An additional valuable tool, the Pediatric Appendicitis Risk Calculator (pARC), calculates the likelihood of appendicitis based on the following variables (Table 4): age, sex, temperature, nausea and/or vomiting, as well as, pain duration, pain location, pain with walking, pain migration, guarding, WBC and ANC (36).

Table 4. Pediatric Appendicitis Risk Calculator.

Pediatric Appendicitis Risk Calculator (pARC)		
Variable		***Value***
Sex	Female	0
	Male	1.2780
Age (grouped by sex)	Male > 13 years or Female >11 years	0
	Female 3-7 years	0.3810
	Female 8-11 years	0.6513
	Male 3-7 years	-0.6653
	Male 8-13 years	-0.0654
Duration of pain, hours	<24	0
	24 to <48	0.4696
	48 to 96	0.1003
	>96	0
	Unknown (defaults to <24hrs)	0
ANC*, cells x $10^3/\mu L$	<14	$1.7734x\sqrt{ANC}$
	≥14	6.6195
Presence of pain with walking	No	0
	Yes	1.0494
Maximal tenderness in RLQ	No	0
	Yes	1.1435
Abdominal guarding	No	0
	Yes	0.6736
History of migration of pain to RLQ	No	0
	Yes	0.4557

*ANC = (neutrophil, % x WBC, cells x $10^3/\mu L$)/100. If neutrophil count is not available, ANC = $(-0.8783+1.1008 \times \sqrt{WBC, \text{cells x } 10^3/\mu L})^2$;
Abbreviations: RLQ – right lower quadrant, ANC – absolute neutrophil count;

Subsequently one of the following formulas are used: pARC Score (ED) = $e^x/(1+e^x)$ or pARC (Community) = $e^{-0.615+1.1x}/(1+e^{-0.615+1.1x})$, where x = -8.6855 + the addition of the assigned values (Table 4).

The calculator demonstrated a specificity of 99.7% for cases with a pARC score >85%, identified as having the highest risk of appendicitis. Additionally, for cases in the high-intermediate range with a pARC Score of 75–84%, the specificity was 97.5% (Table 5) (36).

Table 5. Interpretation of the Pediatric Appendicitis Risk Calculator (37).

pARC score (SD)	Risk group
≤ 5%	Ultra low
6–15%	Low
16–25%	Low-moderate
26–50%	Moderate
51–75%	
76–90%	Moderate-high
> 90%	High

Abbreviations: pARC – Pediatric Appendicitis Risk Calculator; SD – standard deviation

In cases with a risk of 5% or less, considered as ultra-low, outpatient follow-up is appropriate when primary care provider evaluation is available within 24 hours. Additional imaging is not required. Patients, that are evaluated into the 6-15% risk category need observation for at least 6 hours with additional exams. If improvement is seen, ensure follow-up within 24hours, without further imaging. In instances where the risk is 16–25%, classified as low-moderate, for patients with symptoms lasting less than 24 hours, consider additional observation for additional 12 hours. If there is no improvement, repeat CBC and obtain an AUS. If the pain lasted longer than 24 hours, evaluate the patient with AUS directly (36). For moderate cases with 26–50%, AUS is recommended as first line imaging. If the imaging is equivocal, observation and is suggested. In situations with a risk of 51–75%, AUS is again first line imaging, but consider a CT if AUS results are ambiguous and consult Surgery. In moderate-high and high risk cases of over 75%, consultation with surgery is advised (36).

In summary, each of the existing scoring systems can aid in diagnosing patients and minimize instances of negative appendectomies (38).

1.3.4. Imaging

AUS is considered the golden standard for imaging AA, demonstrating a sensitivity of 55% and a specificity of 95%. The technique offers advantages in pediatric cases, due to the body composition characterized by thinner musculature and less abdominal fat. However, it is essential to note that AUS is highly operator-dependent and as a result, an inconclusive or negative study may not definitively rule out the presences of AA (39). As distinctly depicted in Figure 2 a fecalith is shown within the appendix with findings of AA.

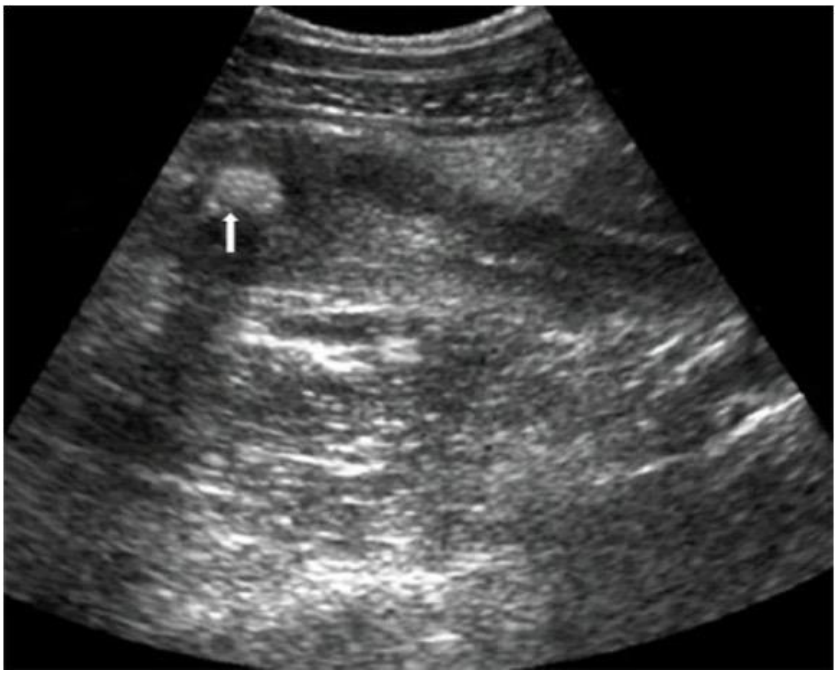

Figure 2. Ultrasound of the right lower quadrant with findings of acute appendicitis. An arrow indicates a fecalith; Source: Mentor's personal archive.

A healthy appendix is a compressible tubular structure that terminates blindly, with a maximal outer diameter (MOD) not exceeding 6 mm. During AUS examinations, the MOD serves as paramount diagnostic criterion to exclude AA. However, it is essential to note that the MOD can be influenced by the presence of intraluminal materials, such as fluids, gas, or feces, potentially exaggerating the measurement. Another critical diagnostic factor is the maximal mural thickness (MMT) of the appendix. MMT is used to reduce the likelihood of false positives based solely on the MOD criterion. Size differences were observed between age groups, including young children, adolescents, and adults, but were only marginally significant. The normal range of the MOD in children was reported to be 0.21–0.64 cm and the MMT to be 0.11–0.27 cm. Additionally, in children less than 6 years of age, an MMT of less than 3 mm is considered within the normal range (5).

Important structures to see in an AUS examination besides the appendix itself are the surrounding peri-cecal fat, peri-appendiceal inflammation, free fluid, presence of reactive lymph nodes, and mural hyperplasia. All these signs of acute inflammation surrounding the appendix are reliable signs of AA, especially in cases where the anatomical position of the appendix is not evaluable, while the lack of their presence is a reliable indicator to rule AA out (39).

Additional imaging tools can be X-ray or CT. In patients with acute abdomen abdominal X-rays are routinely performed, with findings of a soft tissue mass, localized ileus, bowel obstruction, or a fecalith, being suggestive of AA. Since most recent studies show that X-ray can be misleading in cases of AA it is recommended to be used in acute abdomen and not with specific signs of AA as gold standard. In children less than 5 years of age a pre-operative CT scan can reduce the negative appendectomy rate significantly, with a general sensitivity of 0.95 and a specificity of 0.94 (40), but the emitted ionizing radiations emitted while doing a CT scan reveal a higher lifetime risk of developing cancer in children and therefore should be used extremely carefully (14).

1.4. DIFFERENTIAL DIAGNOSIS

There are multiple factors that hinder a proper diagnosis in young children. Most children younger than the age of five present with non-specific clinical presentation, have difficulties to communicate their symptoms, or the symptoms may overlap with other common childhood illnesses. Physical examination can be difficult as well because strong irritability is seen in children with possible AA, delaying diagnosis and leading to a high misdiagnosis rate.

Differential diagnosis in children generally include intussusception, Meckel diverticulum, ectopic pregnancy, testicular torsion, Kidney stones, viral and bacterial gastroenteritis and pelvic inflammatory syndrome (15), upper and lower respiratory tract infections, urinary tract infections, cholecystitis, constipation, blunt abdominal trauma, obstructed hernia, orchitis, right hip septic arthritis, dehydration, sepsis, encephalopathy, and meningitis (20).

1.5. TREATMENT OF ACUTE APPENDICITIS

Standard treatment of AA in the western world is surgical removal, initially via laparotomy (open appendectomy (OA)) but today mostly via laparoscopic appendectomy (LA). Newer studies also show the importance of non-operative management (NOM) as an alternative for special clinical cases (16).

1.5.1. Supportive management

Patients are advised to abstain from oral intake and isotonic crystalloid fluid can be administered intravenously. Antibiotic prophylaxis should be started in coordination with surgeons to ensure that optimal antibiotic levels coincide with the operative procedure. An antibiotic, covering both gram-negative and gram-positive aerobic bacteria, including *Bacteroides fragilis* and *Escherichia coli*, is recommended. Non-perforated AA can be treated with cefoxitin or cefotetan, while in cases with perforated AA, one should consider options like Carbapenem, Ticarcillin-clavulanate, Piperacillin-tazobactam, or Ampicillin-sulbactam. Adequate analgesia should be provided (15).

1.5.2. Surgical management

The preferred treatment for AA remains appendectomy, which can be performed by routine operations - either open surgery or laparoscopy. The success and outcomes of these surgical approaches are primarily determined by the extent of the appendiceal disease, a factor directly linked to morbidity and mortality rates.

Nowadays, the laparoscopic approach is widely regarded as standard of care and is preferred in most cases of AA. The patient is positioned in supine Trendelenburg position combined with a left lateral position, where following a 5 mm supraumbilical incision, a Veress needle is introduced. Through this access point, depending on the patients' age and bodyweight, carbon dioxide is insufflated at pressures of 8–12 mm Hg, creating an artificially induced pneumoperitoneum. A three-port laparoscopic approach is mostly selected, involving a combination of a 5 mm and a 10 mm trocar with a 5 mm scope. The mesoappendix can be dissected using either a harmonic scalpel or thermal fusion technology, while the appendix base is secured using an endoloop, or polymeric clips, and the excised tissue is extracted through the

10 mm trocar (41). The treatment of choice for patients that cannot undergo the laparoscopic approach will be open appendectomy where a transverse incision crossing over the McBurney's point is made and the peritoneum is opened. The mesoappendix is exposed, then tied at the base and removed, while the mucosa revealed afterwards is cauterized and stump inversion is performed by string suture-knot (41).

The laparoscopic approach shows advantages over open surgery in terms of post-operative management. These advantages include lower surgical wound infection rates, fewer adhesive bowel obstructions, less post-operative pain on day one, and earlier hospital discharge. Additionally, it provides aid in inspecting the whole intra-abdominal cavity and can therefore be used as diagnostic and therapeutic tool. On the other hand, there is a lower rate of intraabdominal abscesses, a shorter operative time, and lower costs using open surgery (41). To reduce post-op complications, like wound infections, it is recommended to accompany every single operation by a single dose of antibiotics preoperatively. Antibiotic therapy can be continued for at least 3-5 days, if a more complex AA is found during operation (16).

1.5.3. Additional treatment options

Studies have shown that AA treated with antibiotics can also be a successful option, with an initial course of intravenous antibiotics for 1–3 days, followed by antibiotics given orally for 7 days. Mostly used antibiotics are combinations of either broad-spectrum penicillin with a beta-lactamase inhibitor or cephalosporin combined with tinidazole. During NOM it is crucial to have the first days of treatment in an inpatient setting, with close monitoring of the patient's condition, and the ability to operate in an emergency. Additionally, start of the antibiotic therapy plays a crucial role, showing increased success rates, the earlier the treatment has started after onset of symptoms. If initial treatment fails, surgical management is needed, either laparoscopically or open surgery. Initial treatment with antibiotics is successful in approximately 90% of patients, with the other 10% of patients requiring emergency surgery, and recurrence rates of 20–30% within one year in non-operated patients (16).

1.5.4. Treatment outcomes

While the majority of children experience excellent outcomes after surgery, the incidence of perforation is notably higher in children compared to adults. Mortality rates in surgically treated children with AA are reported to be less than 1%. It has been emphasized through numerous studies that the role of administering antibiotics in children with AA is very crucial, as it reduces perforation rates. AA in neonates displays a higher mortality, primarily because of their inability to verbalize symptoms (15).

1.6. COMPLICATIONS

Complications of AA include gangrene, perforation, appendiceal mass, wound infection, pelvic abscess, wound infections, shock, bowel obstruction, and peritonitis, which all may increase mortality and morbidity and prolong hospital stay (15, 42). These complications are the result of delayed treatment, mostly because of atypical symptoms.

A peri-appendiceal abscess and inflammatory phlegmon can be found in 10% of patients at the time of diagnosis (42). They may arise if the terminal ileum, caecum, and omentum 'wall off' the inflammation. If there is free perforation into the abdominal cavity, Peritonitis will occur (12). As a consequence of that free perforation, approximately 1-3% of children may develop intra-abdominal abscesses and small bowel obstructions (15).

1.6.1. Management of complications

Timing is the most crucial factor in any disease showing signs of complications. Emergency appendectomy will be the choice of treatment in clinical cases with signs of perforated AA or generalized peritonitis. Hemodynamically instable patients that may also show signs of sepsis, will need resuscitation and stabilization, before being transferred to surgery. Stable patients with an appendiceal abscess or a phlegmon will be treated non-operatively initially (16). A percutaneous image-guided drainage is done after initial treatment, where surgeons and interventional radiologists work closely together. If percutaneous drainage is not available or fails as a treatment, surgery is recommended (42).

In general, patients that show longer and more severe duration of disease or extensive complications, early surgery has led to several complications. These include higher rates of postoperative abscesses or enterocutaneous fistulae, as well as higher rates of ileocecal resection rates (16).

1.7. MACHINE LEARNING

Machine learning (ML) differs from traditional programming, because without the need of stepwise programming it directly learns from given data. ML is a computational method that predicts the target variable through fitting a mathematical function to a dataset (43).

1.7.1. Random Forest

Random forest stands out as a widely adopted machine learning technique for constructing predictive models. The model comprises a set of classification and regression trees. These trees employ binary splits on predictor variables to deduce outcome predictions and are simple to apply. Decision trees operate by distinguishing between “high” and “low” values of a predictor linked to the outcome. They are recognized for their practicability and offer an intuitive method for predicting outcomes and besides many advantages, provide poor accuracy for complex datasets. Several classification and regression trees are created by utilizing randomly chosen training datasets and subsets of predictor variables to model outcomes (44).

1.7.2. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that applies a technique known as Gradient Boosting, specifically based on decision trees. In this process, short and basic decision trees are constructed iteratively. Each tree is referred to as a "weak learner" because of its high bias, indicating that it may not perform well on its own. The XGBoost algorithm starts by constructing the first basic tree, which exhibits limited performance initially. Subsequently, it constructs another tree that is trained to predict the errors or shortcomings of the first tree, effectively improving its performance. This sequential process continues, producing a series of weak learners, each correcting the errors of the previous tree, until a stopping condition is met. The stopping condition could be the predefined number of trees (estimators) to be created. XGBoost has additional advantages, including speedy training and the ability to be parallelized or distributed across clusters, making it efficient for large datasets and parallel processing environments (45).

1.7.3. Logistic Regression

Logistic regression is a statistical method used to model the probability of a discrete outcome based on one or more input variables. Typically, logistic regression models are employed for binary outcomes, which involve situations where the result can take only two values, such as true/false, yes/no, or similar.

The primary purpose of logistic regression is to analyze and model the relationship between the input variables and the likelihood of a particular outcome occurring. It is commonly employed in classification problems, where the goal is to determine if a new sample belongs to a specific category or class Logistic regression is extensively employed owing to its simplicity, ease of interpretation, and efficacy in binary classification tasks. Additionally, it can be expanded for multiclass classification (known as multinomial logistic regression) and is applied across diverse domains such as medicine, finance, and social sciences (46).

1.7.4. Machine learning in acute appendicitis

Preoperative prediction of the pathological type of AA aids in not only distinguishing between simple and perforated appendicitis but also preventing negative appendectomy and guiding decisions regarding surgical approach and antibiotic therapy (47). Findings from earlier studies suggest that peripheral blood biomarkers hold promise in predicting the pathological types of acute appendicitis. CRP and WBC count have been the most widely used peripheral blood biomarkers for suspected AA, with an increase in proportion to the severity of infection, because of stimulation of cell-mediating immunity and chemotaxis. Additional diagnostic markers like bilirubin, CRP, and PCT have also been reported significant, as well as the level of Lymphocytes (47).

An alternative study indicated higher association developing AA specifically with being female, and higher levels of PDW, WBC, and MPV. On the other hand, higher levels of neutrophil, RDW, PLT, lymphocyte, and PCT were associated with a lower chances of having appendicitis (48).

2. OBJECTIVES

2.1. AIM OF STUDY

The aim of this study is to evaluate the diagnostic accuracy of a model that includes the total bilirubin level in predicting the instances of negative and positive acute appendicitis in pediatric patients. Specifically, the study seeks to determine whether incorporating bilirubin as a predictor improves the model's performance compared to a prior model that did not include bilirubin levels in their results.

2.2. HYPOTHESIS

We propose the hypothesis that incorporating the total bilirubin level into the diagnostic model will improve the prediction accuracy of acute appendicitis in pediatric patients and aid in diagnostic processes compared to the earlier model that did not include the total bilirubin count as a predictor.

3. PATIENTS AND METHODS

3.1. STUDY DESIGN

This diploma thesis was designed as a single-center retrospective cohort study with information extracted from the patient records at the Department of Pediatric Surgery, University Hospital of Split, involving pediatric patients with suspected acute appendicitis who underwent an appendectomy between January 2019 and July 2023.

The study goal was the assessment of the diagnostic accuracy of a machine learning model that incorporates total bilirubin levels to predict cases of negative and positive acute appendicitis in pediatric patients. Specifically, the study attempts to determine if including bilirubin levels as a predictor improves the model's performance compared to a prior model, derived from the original dataset, which did not explicitly account for bilirubin levels in its results.

All patients underwent surgical treatment, and the diagnoses were confirmed through pathohistological examinations. The selection of surgical approach depended on the preferences of the operating surgeon. The majority of patients underwent three-port laparoscopic appendectomy, with only a small number receiving standard open appendectomy. Both techniques have previously been detailed in paragraph 1.4.2.

3.1.1. Ethical approval

The study protocol received approval from the Ethics Committee of the University Hospital of Split (Approval number 500-03/22-01/188; Date of approval: November 28th, 2022). It adheres to the World Health Organization Declaration of Helsinki from 1975, as revised in 2013, and the International Conference on Harmonization Guidelines on Good Clinical Practice. Rigorous measures were implemented to ensure the strict maintenance of patients' anonymity.

3.1.2. Eligibility criteria

The major inclusion criteria for the original dataset comprised a diagnosis of acute appendicitis in pediatric patients aged 0 to 17 years, with a simultaneous referral for emergent appendectomy. The diagnosis of AA was confirmed via pathohistological diagnosis (PHD). Based on these histopathologic reports, patients were categorized into uncomplicated

appendicitis, including catarrhal or phlegmonous appendicitis, or complicated appendicitis, with gangrenous or gangrenous-perforated types.

Exclusion criteria included age over 17 years, the presence of significant comorbidities such as chronic cardiac, renal, or gastrointestinal conditions and a Body Mass Index (BMI) ≥ 35 kg/m^2. Furthermore, patients with incidental appendectomy during other operations or without histopathology report available were excluded. Lastly patients with a PHD indicating conditions other than appendicitis or a histologically normal appendix, like neuroendocrine tumors or enterobiasis, had to be excluded from this study.

The Subset for this thesis was created after the inclusion and exclusion criteria for the original dataset were applied. Additional inclusion criterium to create the subset of data is the total bilirubin level. Every patient without the total bilirubin level had to be excluded.

3.1.3. Variables and settings

Initially 614 pediatric patients comprised the original dataset out of which 63 were excluded. Patients were excluded if the pathohistological diagnosis (PHD) was unavailable or if there were more than two missing values among key features identified as crucial in previous studies, including neutrophils count, lymphocyte count, WBC count, CRP, and sodium concentration. Following the application of these exclusion criteria, the final analysis included 551 patients. While 47 cases were negative for appendicitis within this group, 252 presented with uncomplicated appendicitis, and 252 with complicated appendicitis, resulting in an imbalanced dataset.

A subset is created from the original dataset, including only the patients that have the total bilirubin count in their laboratory findings, which comprises a total of 297 patients. The initial variable assortment encompassed patient information, data from complete and differential blood counts, biochemical measures, including sodium concentration and CRP. Additional clinical examination findings such as the presence of abdominal pain, rebound tenderness, or guarding were included.

A total of fourteen features were considered for model training and analysis in the comprised subset: weight, height, temperature, leukocyte count, total bilirubin, CRP level, sodium concentration, potassium concentration, chloride concentration, hemoglobin level, hematocrit level, urea and creatinine. The target feature for analysis was the total bilirubin level.

3.2. PREDICTION MODEL TRAINING, OPTIMIZATION AND VALIDATION

Random Forest and Logistic Regression were the machine learning algorithms that were tested. Logistic Regression served as a baseline model, while Random Forest was selected for its recognized effectiveness with tabular data and imbalanced datasets. A nested cross-validation approach, consisting of 5-fold inner and outer cross-validation, was employed to train and validate the models, which was repeated ten times. Subsequently, each outer fold was split into training (80%) and test sets (20%), with stratification on the target variable, to ensure a representative distribution of the target variable. This helps to maintain the balance and integrity of the dataset when splitting it into training and test sets for model evaluation. Inner cross-validation within each outer fold's training set was implemented to improve hyperparameters and conduct threshold adjustment.

3.2.1. Feature Importance

Random Forests are an ensemble learning method primarily used for classification and regression. One of their strengths is the ability to estimate the importance of each feature in the prediction process. After training the model, we extracted the feature importance. The Random Forest model calculates feature importance by averaging the reduction in impurity (Gini impurity or entropy) brought by each feature over all trees in the forest. This method is known as Mean Decrease Impurity (MDI).

3.2.2. Statistical analysis

Statistical analysis involved checking data distribution normality using the Kolmogorov-Smirnov test. The t-test was applied to normally distributed data, while the Mann–Whitney test was used for deviations of normal data. The Chi-squared test was utilized for non-numerical features while a significance threshold of p-value below 0.05 was adopted, when it was considered statistically significant. Statistical analyses and visualizations were conducted using the R programming language.

4. RESULTS

4.1. PATIENT CHARACTERISTICS

The original dataset is comprised of a total of 551 patients that were involved in both model training and evaluation, with patient characteristics and for continuous features provided in detail in Tables 7 and 8. Among the 551 patients, 252 patients were diagnosed with uncomplicated appendicitis, 252 with complicated appendicitis, and 47 had a negative PHD.

Table 7. Patient characteristics for continuous features.

Feature	Negative PHD (n = 47)	Uncomplicated (n = 252)	Complicated (n = 252)	*P**
Age (years)	11.63±3.75	11.73±3.63	11.75±3.92	0.98
Height (cm)	153.32±21.58	153.49±20.34	155.96±21.39	0.42
Weight (kg)	48.76 ± 19.58	46.57 ± 18.25	47.17 ± 20.12	0.46
BMI (kg/m^2)	18.76 ± 4.47	18.91 ± 3.8	19.06 ± 3.8	0.87
Temperature (°C)	36.9 (1.2)	37 (0.9)	37.5 (1.4)	< 0.001
Symptoms duration	28 (31)	24 (15)	30 (24)	< 0.001
CRP (mg/L)	10.3 (47.35)	11.55 (23.3)	46.45 (66.25)	< 0.001
Sodium concentration (mmol/L)	140 (2)	139 (3)	137 (4)	< 0.001
Leukocytes (10^9/L)	11.5±4.61	13.36±4.36	16.58±5.07	< 0.001
Lymphocytes (%)	15.6 (14.65)	13.8 (11.18)	7.95 (5.83)	< 0.001
Neutrophils (%)	76.9 (10.55)	79.3 (12.2)	84.9 (6.62)	< 0.001
Thrombocytes (10^9/L)	276.15±71.86	274.66±67.76	289.93±72.07	0.05
NLR	5.06 (6.22)	5.66 (5.3)	10.64 (8.66)	< 0.001
TLR	1.58 (0.64)	1.5 (0.96)	2.19 (1.42)	< 0.001
RDW (%)	12.7 (0.9)	13 (1)	12.8 (1)	0.07
MCHC (g/L)	343.5 (11.25)	343 (14)	345 (11.25)	0.15
MPV (fL)	8.5 (2.9)	8.1 (2.13)	8.3 (2.05)	0.86

Data presented as mean ± SD or Median (IQR). * one-way ANOVA or Kruskal-Wallis test. Abbreviations: PHD – pathohistological diagnosis; BMI – body mass index; CRP – C-reactive protein; NLR – neutrophil to lymphocyte ratio; TLR – thrombocyte to lymphocyte ratio; RDW – Red blood cell distribution width; MCHC – mean corpuscular hemoglobin concentration; MPV – mean platelet volume

Table 8. Patient characteristics for categorical features.

Feature	Level	Negative PHD (n = 47)	Uncomplicated (n = 252)	Complicated (n = 252)	*P**
Vomiting	0	30	142	75	< 0.001
	1	16	108	172	
Rebound tenderness	0	6	29	20	< 0.001
	1	14	67	36	
	2	15	110	100	
	3	11	42	96	
Nausea	0	15	83	55	< 0.001
	1	30	161	199	
Sex	F	20	167	168	0.005
	M	27	85	84	
Migration	0	27	109	99	0.100
	1	19	140	140	

*Chi-squared test; Abbreviations: PHD – pathohistological diagnosis.

From this original dataset the subset, including only patients with laboratory values of total bilirubin count, was created. A total of 297 patients comprises this subset with fourteen different variables as patients' characteristics provided in Table 9.

Table 9. Patient characteristics of Subset including total bilirubin levels.

Feature	**Male**	**Female**	***P****
Weight (kg)	49.7 (19.6)	43.4 (17.6)	< 0.001
Height (cm)	157.1 (21.5)	149.4 (22.3)	< 0.001
Temperature (°C)	37.4 (0.7)	37.5 (0.6)	0.11
Leukocyte count (10^9/L)	14.9 (5)	15.1 (4.7)	0.82
Total bilirubin	15.9 (10.3)	15.1 (13.6)	0.51
CRP (mg/L)	48.3 (57.7)	50.1 (68.3)	0.77
Sodium	137.9 (3.2)	137.7 (4.3)	0.63
Potassium	4.1 (0.33)	4.09 (0.35)	0.68
Chloride	100.14 (3.5)	101.97 (3.5)	< 0.001
Hemoglobin	136.47 (16.1)	127.51 (15.4)	< 0.001
Hematocrit	0.40 (0.01)	0.38 (0.03)	< 0.001
Glucose	5.49 (1.07)	5.42 (0.83)	0.43
Urea	5.38 (9.3)	3.92 (1.7)	0.30
Creatinine	58.41 (18.9)	45.31 (16.2)	< 0.001

Abbreviations: CRP – C-reactive protein,

The aim of this study was to compare the results applied to our subset of data to a previous model using the original dataset that aimed at decreasing the occurrence of misdiagnosed appendicitis, which resulted in unnecessary appendectomies (negative appendectomies). While surgical removal of the appendix is considered the standard treatment for acute appendicitis due to its low-risk nature, there still is a possibility of complications during or after the procedure, albeit rare. Alternative therapy, although it carries a moderate risk of recurrence, is conservative antibiotic treatment, which typically yields low morbidity and mortality rates. Delaying surgical intervention, however, can increase the likelihood of complications, making surgery the preferred option in most cases.

Considering these factors, the cost associated with false negative diagnoses outweighs that of false positives, therefore it is essential to minimize the occurrence of false negatives, in developing a model to improve the identification of false positives concurrently.

Based on this logic, we chose to refine our model's hyperparameters. A customized metric that employs thresholds on the ROC curve to maximize specificity while preserving maximum sensitivity was utilized. Essentially, a requirement has been imposed that the model must achieve 100% accuracy in diagnosing true appendicitis for patients in the training data. This guarantees that when the model predicts a negative diagnosis, it accurately identifies it as a true negative rather than a false negative.

4.2. MODEL FOR APPENDICITIS PREDICTION

4.2.1. Model characteristics

The Logistic Regression model exhibited the following performance metrics for detecting negative cases: The mean precision for identifying negative cases was 0.26, with a 95% confidence interval (CI) ranging from 0.1 to 0.57. The mean recall was 0.41, with a 95% CI ranging from 0.1 to 0.833. The mean F1-score was 0.31, with a 95% CI ranging from 0.1 to 0.60.

The Random Forest model demonstrated superior performance compared to the Logistic Regression model with the following metrics: The mean precision was 0.416, with a 95% CI ranging from 0.12 to 0.71. The mean recall was 0.697, with a 95% CI ranging from 0.33 to 1.0. The mean F1-score was 0.508, with a 95% CI ranging from 0.18 to 0.75. Mean AUC score for RF was 0.83, with a 95% CI ranging from 0.70 to 0.95 (Figure 3).

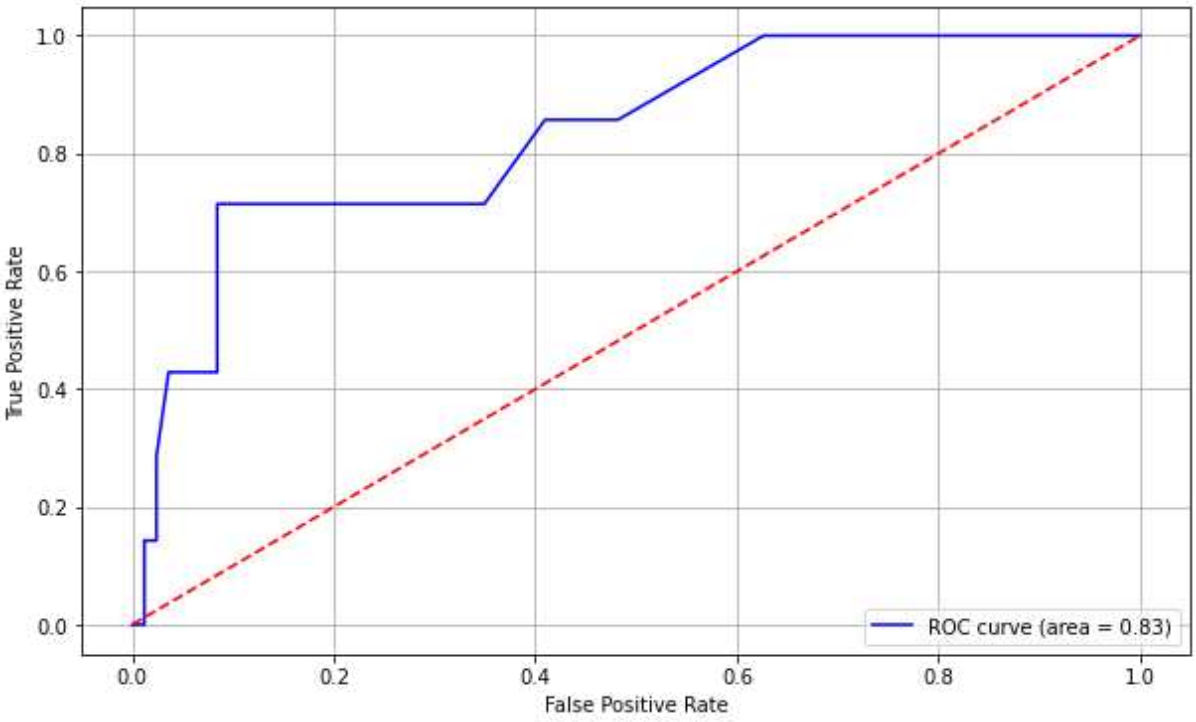

Figure 3. Receiver Operating Characteristic (ROC) for Random Forest classification

The analysis of performance of Logistic Regression versus Random Forest models in detecting negative cases of appendicitis reveals that the Random Forest model significantly outperforms the Logistic Regression model in detecting negative cases of appendicitis. This is evidenced by higher mean values and narrower confidence intervals for all three metrics: precision, recall, and F1-score.

Table 10. Performance of Logistic Regression versus Random Forest models in detecting negative cases of appendicitis.

Metric	Logistic Regression (Mean (95% CI))	Random Forest (Mean (95% CI))
Precision (PPV)	0.26 (0.1 to 0.57)	0.416 (0.12 to 0.71)
Recall (Sensitivity)	0.41 (0.1 to 0.833)	0.697 (0.33 to 1.0)
F1-score	0.31 (0.1 to 0.60)	0.83 (0.70 to 0.95)

Abbreviations: PPV – Positive predictive value, CI – confidence interval,

The Random Forest model's higher precision indicates a lower rate of false positives compared to Logistic Regression. The substantially higher recall of the Random Forest model suggests it is more effective in identifying true negative cases. The F1-score, which balances precision and recall, further corroborates the superior performance of the Random Forest model. The confidence intervals for the Random Forest model are narrower, indicating more reliable estimates compared to those of the Logistic Regression model (Table 10).

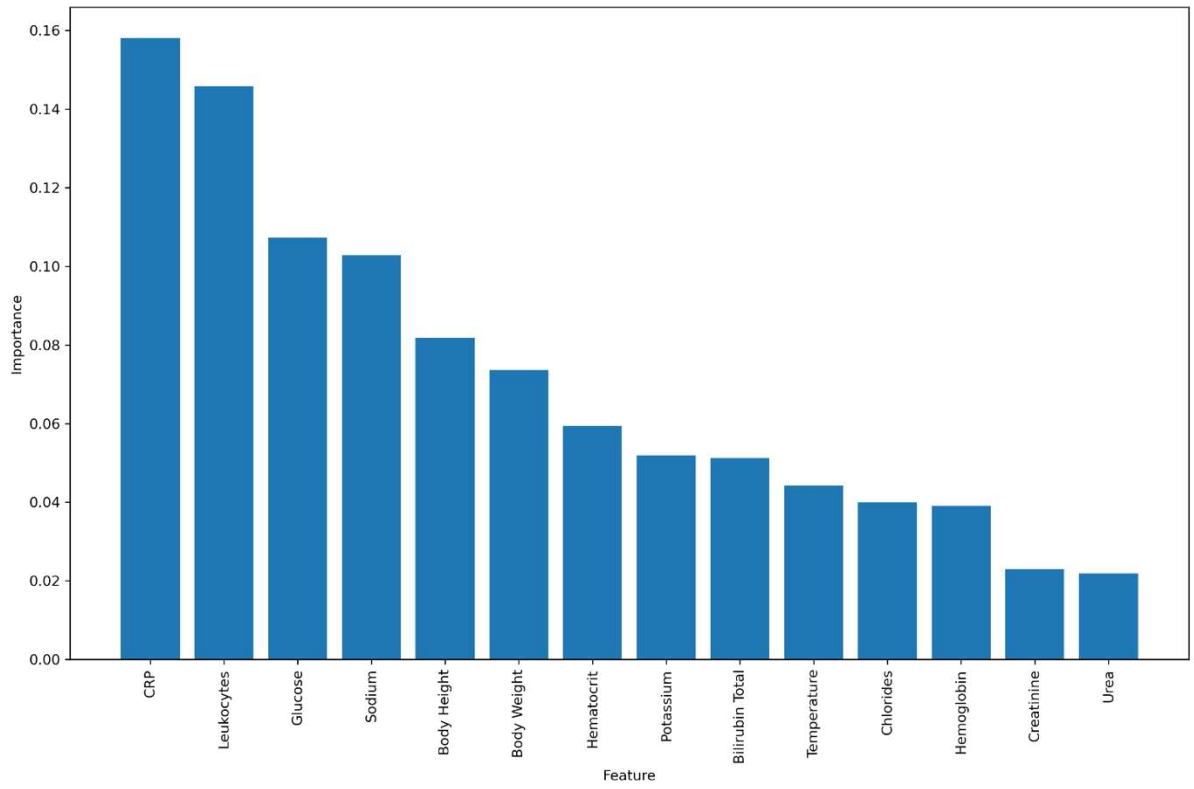

Figure 4. Feature importance for Random Forest. Abbreviations: CRP – C-reactive protein

When analyzing feature importance, total bilirubin was the 9th most important feature out of fourteen in total in the global Random Forest feature importance analysis (Figure 4). The most important feature was CRP, followed by the leukocyte count and the glucose level.

The negative appendicitis group of patients exhibited a mean total bilirubin value of 9.21, with a standard deviation (SD) of 3.8, while the positive appendicitis group of patients had a notably higher mean total bilirubin value of 16.07, accompanied by a larger SD of 11.9. This difference in bilirubin levels between the two groups was statistically significant, as indicated by a p-value of 0.001, derived from the T-test for independent samples (Figure 5).

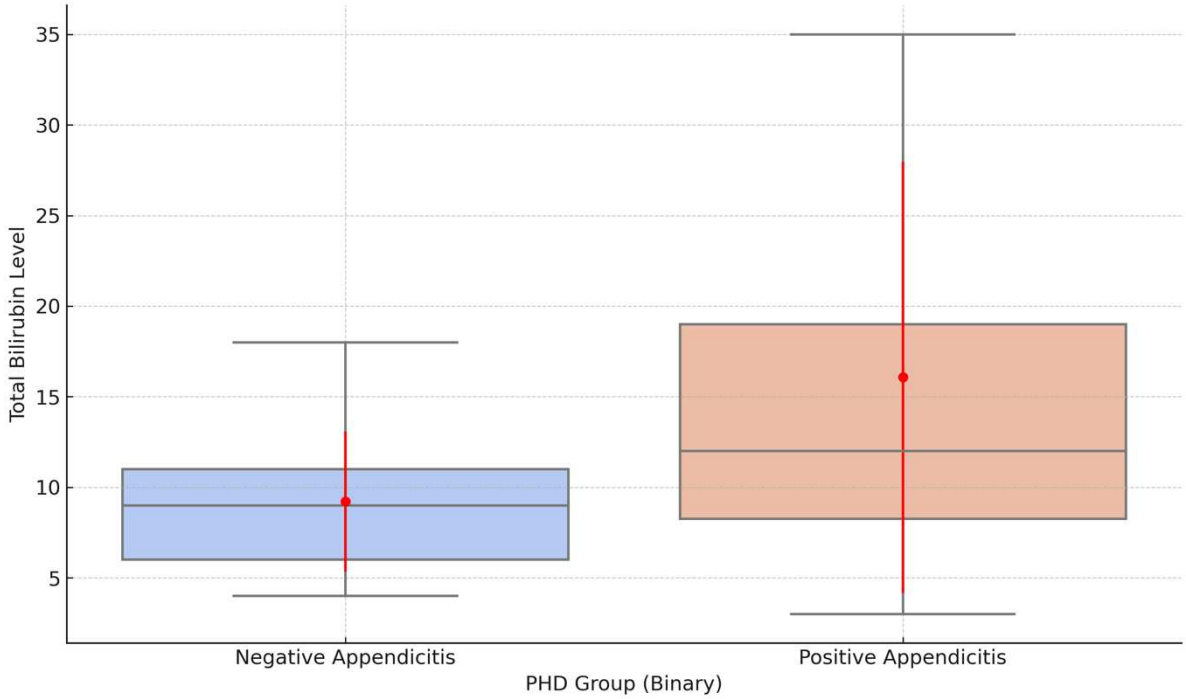

Figure 5. Total bilirubin level in negative (blue) vs. positive (red) appendicitis cases. Data are presented as mean ± standard deviation; Abbreviations: PHD – Pathohistological Diagnostics

5. DISCUSSION

AA is the most common etiology of acute abdomen in the pediatric population, requiring surgery, but despite that, it continues being a diagnostic challenge for clinicians globally (5). Especially children often experience high rates of initial misdiagnosis, leading to delayed treatment and subsequent complications such as perforation (14).

A good history combined with a proper clinical examination is traditionally needed for the diagnosis of AA (15). There are no specific laboratory factors that are diagnostic for AA, but leukocytosis with additional elevation in CRP level increase the positive finding of AA (18). Bilirubin serves as a marker for acute appendicitis, offering a strong positive predictive value, according to previous studies. Additionally, bilirubin counts as an important indicator for identifying patients that are at higher risk of appendiceal perforation or gangrene. However, bilirubin levels should always be considered alongside clinical examinations and other laboratory tests when assessing patients with suspected acute appendicitis (30).

Since total bilirubin level is widely used for its diagnostics value in acute appendicitis, we proposed our hypothesis that incorporating the total bilirubin level into the diagnostic model specifically, will improve the prediction accuracy of acute appendicitis in pediatric patients.

The present study aimed on evaluating the diagnostic accuracy of a model in predicting acute appendicitis in pediatric patients. We created a subset of data including only those patients with the total bilirubin level present in their laboratory work, to compare our results to the current study that did not explicitly exclude the patients without total bilirubin levels. Specifically, this study was conducted to determine whether incorporating bilirubin as a predictor improves the model's performance compared to the prior dataset.

A ML model was developed with the goal of minimizing negative appendectomies in pediatric patients, using clinical and laboratory parameters (49). Many previously developed and similar models prioritize specificity over sensitivity, therefore aiming at diagnosing acute appendicitis, but potentially increasing the risk of misdiagnoses (50, 51). Additionally, some studies fail to address the balance between sensitivity and specificity, which we consider crucial for incorporating tools like this in decision-making within clinical practices (52).

The model was trained on data of pediatric patients who had surgery with previous strong suspecting of AA. All confirmed cases were considered "true positives," while negative AA cases were "false positives." Because the dataset was lacking non-surgery patients, the model was not able to learn "true negatives" or "false negatives." The model is designed for the specific cases, where surgery is highly recommended due to previous examinations (49).

As already evident in the study using the original dataset, the subset reveals that the Random Forest model exhibits higher precision and indicates a lower rate of false positives

compared to Logistic Regression (49). Random Forest model also suggests that it is more effective in identifying true negative cases of AA, highlighted by the substantially higher Recall. The F1-score, balancing precision and recall, further validates the superior performance of the Random Forest model compared to Logistic regression model, as well as the narrower confidence intervals for the Random Forest model, indicating more reliable estimates.

Concluding from our results, Random Forest remains the superior machine learning model in diagnosing AA over the logistic regression model, exhibiting a higher precision and recall. Compared to the original study there is no alternating in comparing the machine learning model's performances on our subset of data.

Does incorporating the total bilirubin level benefit the overall prediction accuracy over the dataset, not explicitly including bilirubin? Contrary to our hypothesis, the findings indicate no apparent significant increase in diagnostic precision compared to the outcome of the previous study's authors (49). A possible reasoning could be, that Bilirubin, even though it is a widely used marker, is only an important indicator for identifying patients that are at higher risk of a severe case of AA. Stated differently, bilirubin has more predictive power in discerning patients with possible perforation or gangrene from a negative AA than is has to a simple AA case. This is also underlined by the feature importance results, where the total bilirubin level is displayed as 9^{th} most important feature out of 14 features in total. As already depicted in other previous studies (53), the CRP and leukocyte count are most important in correctly diagnosing AA in pediatric patients.

The current study has several limitations that must be acknowledged. The first limitation is that the study is retrospective in nature, although much of the data originated from previously conducted prospective studies. Secondly, the dataset is derived from a single hospital, which makes validation on different populations necessary to enhance general applicability. Lastly, the dataset exhibits a fundamental imbalance towards positive appendicitis diagnoses, creating a challenge that cannot be easily resolved given the restraint of clinical decision-making for appendectomy referrals.

On the other hand, the study carries notable strengths: the strongest positive factor is that it relies on definitive pathohistological reports, which evades the need for further radiological techniques and thereby enhancing accessibility, especially in underprivileged regions. Furthermore, unlike many existing ML models, our model maintains an outstanding high sensitivity, ensuring that nearly every patient with AA is correctly identified for surgical treatment.

In summary, incorporating total bilirubin levels into our dataset did not significantly enhance the model's predictive accuracy compared to the previous study, but Random Forest remains the superior machine learning model in diagnosing AA over the logistic regression model, exhibiting a higher precision and recall. Compared to the original study there is no alternating in comparing the machine learning model's performances on our subset of data. The feature importance displays total bilirubin levels as 9th most important feature with the CRP and leukocyte count as most important in correctly diagnosing AA in pediatric patients.

6. CONCLUSION

In conclusion, our findings demonstrate that the Random Forest model outperforms the Logistic Regression model in diagnosing acute appendicitis (AA) in pediatric patients, showing higher precision and recall, like it was already depicted in the original dataset. Bilirubin counts as a useful marker for AA, but it does not rank as highly as CRP and leukocyte count in determining AA diagnosis. Incorporating total bilirubin levels into our dataset did not significantly enhance the model's predictive accuracy compared to the previous study, because bilirubin is mainly important in identification of patients of higher risk of severe AA, such as perforation and gangrene, rather than simple cases. This is supported by the feature importance results, in which bilirubin ranks 9th out of fourteen features, with CRP and leukocyte count highlighted as most crucial for diagnostics. Despite the study's limitations, including its retrospective nature and dataset imbalance, it remains robust due to its reliance on definitive pathohistological reports. The high sensitivity of the ML model ensures accurate identification of patients needing surgical treatment, offering valuable insights for clinical decision-making in pediatric acute appendicitis, with and without including the total bilirubin level into our calculations.

7. REFERENCES

1. Deshmukh S, Verde F, Johnson PT, Fishman EK, Macura KJ. Anatomical variants and pathologies of the vermix. Emerg Radiol. 2014;21:543-52.
2. Schumpelick V, Dreuw B, Ophoff K, Prescher A. Appendix and cecum. Embryology, anatomy, and surgical applications. Surg Clin North Am. 2000;80:295-318.
3. Lotfollahzadeh, Saran, Appendicitis. StatPearls, 12 February 2024.
4. Humes DJ, Simpson J. Acute appendicitis. BMJ. 2006;333:530-4.
5. Park NH, Oh HE, Park HJ, Park JY. Ultrasonography of normal and abnormal appendix in children. World J Radiol. 2011;3(4):85-91.
6. Hodge, Bonnie D., Anatomy, abdomen and pelvis: Appendix. StatPearls, 8 August 2023.
7. Randal Bollinger R, Barbas AS, Bush EL, Lin SS, Parker W. Biofilms in the large bowel suggest an apparent function of the human vermiform appendix. J Theor Biol. 2007;249:826-31.
8. Humes DJ, Simpson J. Acute appendicitis. BMJ. 2006;333:530-4.
9. Pogorelić Z, Janković Marendić I, Čohadžić T, Jukić M. Clinical outcomes of daytime versus nighttime laparoscopic appendectomy in children. Children. 2023;10:750.
10. Pogorelić Z, Ercegović V, Bašković M, Jukić M, Karaman I, Mrklić I. Incidence and management of appendiceal neuroendocrine tumors in pediatric population: a bicentric experience with 6285 appendectomies. Children. 2023;10:1899.
11. Pogorelić Z, Čohadžić T. A bizarre Cause of acute appendicitis in a pediatric patient: An ingested tooth. Children. 2023;10:108.
12. Petroianu A. Diagnosis of acute appendicitis. Int J Surg. 2012;10:115-9.
13. Humes DJ, Simpson J. Acute appendicitis. BMJ. 2006;333:530-4.
14. Pogorelić Z, Domjanović J, Jukić M, Poklepović Peričić T. Acute appendicitis in children younger than five years of age: Diagnostic challenge for pediatric surgeons. Surg Infect. 2020;21:239-45.
15. Gadiparthi R, Waseem M. Pediatric Appendicitis. StatPearls, 3 July 2023.
16. Becker P, Fichtner-Feigl S, Schilling D. Clinical management of appendicitis. Visc Med. 2018;34:453-8.
17. Petroianu A. Diagnosis of acute appendicitis. Int J Surg. 2012;10:115-9.
18. Ohle R, O'Reilly F, O'Brien KK, Fahey T, Dimitrov BD. The Alvarado score for predicting acute appendicitis: a systematic review. BMC Med. 2011;9:139.
19. Nehring, Sara M. C Reactive Protein. StatPearls, 10 July 2023.
20. Almaramhy HH. Acute appendicitis in young children less than 5 years: review article. Ital J Pediatr. 2017;43:15.

21. Blumenreich MS. The white blood cell and differential count. 3rd ed. Boston: Butterworths; 1990.
22. Zhao L, Feng S, Huang S, Tong Y, Chen Z, Wu P, *et al*. Diagnostic value of hyperfibrinogenemia as a predictive factor for appendiceal perforation in acute appendicitis. ANZ J Surg. 2017;87:372-5.
23. Singh A, Pogorelić Z, Agrawal A, Muñoz CML, Kainth D, Verma A, *et al.* Utility of ischemia-modified albumin as a biomarker for acute appendicitis: A systematic review and meta-analysis. J Clin Med. 2023;12:5486.
24. Oztan MO, Aksoy Gokmen A, Ozdemir T, Müderris T, Kaya S, Koyluoglu G. Pentraxin-3: A strong novel biochemical marker for appendicitis in children. Am J Emerg Med. 2019;37:1912-6.
25. Pogorelić Z, Lukšić AM, Mihanović J, Đikić D, Balta V. Hyperbilirubinemia as an indicator of perforated acute appendicitis in pediatric population: A prospective study. Surg Infect. 2021;22:1064-71.
26. Kakar M, Delorme M, Broks R, Asare L, Butnere M, Reinis A, *et al.* Petersons A. Determining acute complicated and uncomplicated appendicitis using serum and urine biomarkers: interleukin-6 and neutrophil gelatinase-associated lipocalin. Pediatr Surg Int. 2020;36:629-36.
27. Anand S, Krishnan N, Birley JR, Tintor G, Bajpai M, Pogorelić Z. Hyponatremia - a new diagnostic marker for complicated acute appendicitis in children: A systematic review and meta-analysis. Children (Basel). 2022;9:1070.
28. Carvalho N, Carolino E, Coelho H, Barreira AL, Moreira L, André M, *et al.* Eosinophil granule proteins involvement in acute appendicitis - An allergic disease? Int J Mol Sci. 2023;24(10):9091.
29. Guerra Ruiz AR, Crespo J, López Martínez RM, Iruzubieta P, Casals Mercadal G, Lalana Garcés M, *et al*. Measurement and clinical usefulness of bilirubin in liver disease. Adv Lab Med. 2021;2:352-72.
30. Emmanuel A, Murchan P, Wilson I, Balfe P. The value of hyperbilirubinaemia in the diagnosis of acute appendicitis. Ann R Coll Surg Engl. 2011;93:213-7.
31. Pogorelić Z, Rak S, Mrklić I, Jurić I. Prospective validation of Alvarado score and Pediatric Appendicitis Score for the diagnosis of acute appendicitis in children. Pediatr Emerg Care. 2015;31:164-8.
32. Pogorelić Z, Mihanović J, Ninčević S, Lukšić B, Elezović Baloević S, Polašek O. Validity of Appendicitis Inflammatory Response Score in distinguishing perforated from non-perforated

appendicitis in children. Children. 2021;8:309.
33. Samuel M. Pediatric appendicitis score. J Pediatr Surg. 2002;37:877-81.
34. Aydin D, Turan C, Yurtseven A, Bayindir P, Toker B, Dokumcu Z, *et al*. Integration of radiology and clinical score in pediatric appendicitis. Pediatr Int. 2018;60:173-8.
35. Goulder F, Simpson T. Pediatric appendicitis score: A retrospective analysis. J Indian Assoc Pediatr Surg. 2008;13:125-7.
36. Tam D, Vazquez H. Calculated decisions: Pediatric appendicitis risk calculator (pARC). Pediatr Emerg Med Pract. 2019;16:CD5–6.
37. Cotton DM, Vinson DR, Vazquez-Benitez G, Margaret Warton E, Reed ME, Chettipally UK, *et al.* Validation of the Pediatric Appendicitis Risk Calculator (pARC) in a community emergency department setting. Ann Emerg Med. 2019;74:471-80.
38. Sağ S, Basar D, Yurdadoğan F, Pehlivan Y, Elemen L. Comparison of appendicitis scoring systems in childhood appendicitis. Turk Arch Pediatr. 2022;57:532-7.
39. Wonski S, Ranzenberger LR, Carter KR. Appendix imaging. StatPearls; 2024.
40. Rud B, Vejborg TS, Rappeport ED, Reitsma JB, Wille-Jørgensen P. Computed tomography for diagnosis of acute appendicitis in adults. Cochrane Database Syst Rev. 2019;2019:CD009977.
41. Pogorelic Z, Buljubasic M, Susnjar T, Jukic M, Pericic TP, Juric I. Comparison of open and laparoscopic appendectomy in children: A 5-year single center experience. Indian Pediatr. 2019;56:299-303.
42. Sartelli M, Chichom-Mefire A, Labricciosa FM, Hardcastle T, Abu-Zidan FM, Adesunkanmi AK, *et al*. The management of intra-abdominal infections from a global perspective: 2017 WSES guidelines for management of intra-abdominal infections. World J Emerg Surg. 2017;12:29.
43. Vrdoljak J, Boban Z, Barić D, Šegvić D, Kumrić M, Avirović M, *et al*. Applying explainable machine learning models for detection of breast cancer lymph node metastasis in patients eligible for neoadjuvant treatment. Cancers. 2023;15:634.
44. Speiser JL, Miller ME, Tooze J, Ip E. A Comparison of Random Forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019;134:93-101.
45. Guillen MD, Aparicio J, Esteve M. Gradient tree boosting and the estimation of production of frontiers. Expert Syst. Appl. 2023;214:119134.
46. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24:12-8.
47. Kang CB, Li XW, Hou SY, Chi XQ, Shan HF, Zhang QJ, *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.
48. Harmantepe AT, Dikicier E, Gönüllü E, Ozdemir K, Kamburoğlu MB, Yigit M. A different way to diagnosis acute appendicitis: machine learning. Pol Przegl Chir. 2023;96:38-43.
49. Males I, Boban Z, Kumric M, Vrdoljak J, Berkovic K, Pogorelic Z, *et al*. Applying an explainable machine learning model might reduce the number of negative appendectomies in pediatric patients with a high probability of acute appendicitis. Sci Rep. 2024;14:12772.
50. Stiel C, Elrod J, Klinke M, Herrmann J, Junge CM, 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.
51. 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.
52. Afzal B, Cirocchi R, Dawani A, Desiderio J, Di Cintio A, Di Nardo D, *et al*. Is it possible to predict the severity of acute appendicitis? Reliability of predictive models based on easily available blood variables. World J Emerg Surg. 2023;18:10.
53. Khan MN, Davie E, Irshad K. The role of white cell count and C-reactive protein in the diagnosis of acute appendicitis. J Ayub Med Coll Abbottabad. 2004;16:17-9.

8. SUMMARY

Objectives: This study aimed on the evaluation of diagnostic accuracy and improvement of model performance of a machine learning model in predicting the instance of AA in pediatric patients, comparing the original dataset to our compressed subset only including patients with a total bilirubin count.

Material and methods: This study was design as a single-center retrospective cohort study. The dataset involves 551 pediatric patients that underwent appendectomy between January 2019 and July 2023, with a subset created including only the patients that have their total bilirubin level in their laboratory findings, comprising 297 patients. Random Forest model and logistic Regression model are the machine learning models used in calculations with a nested cross-validation approach with stratification on the target variable Bilirubin. The feature importance is calculated with Random Forest by averaging the reduction impurity brought by each feature over all trees in the forest, called Mean Decrease Impurity (MDI).

Results: A subset was created, including only patients with laboratory values of total bilirubin count comprised of a total of 297 patients with fourteen different variables as patients' characteristics. For Logistic Regression the mean precision for identifying negative cases was 0.26, with a 95% CI ranging from 0.1 to 0.57 and mean recall was 0.41, with a 95% CI ranging from 0.1 to 0.833. The mean F1-score was 0.31, with a 95% CI ranging from 0.1 to 0.60. The Random Forest model demonstrated superior performance compared to the Logistic Regression model with mean precision of 0.416, with a 95% CI ranging from 0.12 to 0.71 and a mean recall of 0.697, with a 95% CI ranging from 0.33 to 1.0. The mean F1-score was 0.508, with a 95% CI ranging from 0.18 to 0.75. Mean AUC score for Random Forest was 0.83, with a 95% CI ranging from 0.70 to 0.95. The analysis of feature importance reveals that total bilirubin level as the 9th most important feature of fourteen in total, with CRP and leukocyte count being the 2 most important. The negative appendicitis group exhibited a mean total bilirubin value of 9.21, with a SD of 3.8, while the positive appendicitis group had a notably higher mean total bilirubin value of 16.07, accompanied by a larger SD of 11.9. This difference between the two groups was statistically significant, as indicated by a p-value of 0.001.

Conclusions: As already depicted in the study with the original dataset, our findings calculated from our subset of data demonstrated a superiority of Random Forest over Logistic Regression with a higher precision and recall. While bilirubin acts as a useful marker in diagnosing severe cases of AA combined with proper history and examination, incorporating it into our study did not significantly enhance the model's predictive accuracy. This can be attributed to its importance in identifying patients at high risk for perforation and gangrene, rather than simple AA cases.

9. CROATIAN SUMMARY

Naslov: Strojno učenje za otkrivanje negativnih apendektomija u djece, analiza podskupa bilirubina

Ciljevi: Ovo istraživanje ima za cilj procjenu dijagnostičke točnosti i poboljšanje performansi strojnog učenja u predviđanju upale crvuljka u djece, uspoređujući originalni skup podataka s našim komprimiranim podskupom koji uključuje samo djecu s ukupnom razinom bilirubina.
Materijali i metode: Ovo istraživanje je retrospektivna kohortna studija provedena u jednom centru. Uključuje 551 djece koji su podvrgnuti apendektomiji između siječnja 2019. i srpnja 2023. godine, s podskupom od 297 bolesnika koji imaju dostupne podatke o ukupnom bilirubinu. Korišteni su modeli Random Forest i logistička regresija s ugniježđenom unakrsnom validacijom stratificiranom prema varijabli bilirubin. Važnost značajki procijenjena je pomoću Mean Decrease Impurity (MDI) u Random Forest modelu.
Rezultati: Stvoren je podskup koji uključuje samo bolesnike s laboratorijskim vrijednostima ukupnog bilirubina, što čini ukupno 297 djece s četrnaest različitih varijabli kao karakteristikama bolesnika. Za logističku regresiju srednja preciznost za identifikaciju negativnih slučajeva bila je 0,26, s 95% CI od 0,1 do 0,57, dok je srednja osjetljivost iznosila 0,41, s 95% CI od 0,1 do 0,833. Srednja F1-ocjena iznosila je 0,31, s 95% CI od 0,1 do 0,60. Model Random Forest je pokazao superiorniju izvedbu u usporedbi s modelom logističke regresije s prosječnom preciznošću od 0,416, s 95% CI od 0,12 do 0,71, i s prosječnom osjetljivošću od 0,697, s 95% CI od 0,33 do 1,0. Srednja F1-ocjena iznosila je 0,508, s 95% CI od 0,18 do 0,75. Srednje AUC vrijednosti za Random Forest iznosili su 0,83, s 95% CI od 0,70 do 0.95. Analiza važnosti značajki pokazala je da je razina ukupnog bilirubina deveta po važnosti od četrnaest varijabli ukupno, dok su CRP i broj leukocita dvije najvažnije značajke. Skupina s negativnim apendektomijama imala je prosječnu vrijednost ukupnog bilirubina od 9,21, s SD od 3,8, dok je skupina s pozitivnim nalazom akutnog apendicitisa imala značajno višu prosječnu vrijednost ukupnog bilirubina od 16,07, uz veću standardnu devijaciju od 11,9 (p=0,001).
Zaključci: Naši rezultati potvrđuju superiornost Random Forest modela u odnosu na logističku regresiju u predviđanju akutnog apendicitisa. Iako ukupni bilirubin može biti koristan u dijagnostici težih oblika akutnog apendicitisa, njegovo uključivanje nije značajno poboljšalo prediktivnu točnost modela za jednostavne slučajeve upale crvuljka, već za identifikaciju bolesnika s visokim rizikom od komplikacija poput perforacije i gangrene.