International Journal of Medical and All Body Health Research

Comparative Diagnostic Performance of RIPASA, ALVARADO, and AIR Scoring Systems in Acute Appendicitis: A Retrospective Evaluation Using Traditional Metrics and XGBoost Algorithm

Zeynep Kucukakcali [1*], Ipek Balikci Cicek [2]
[1, 2] Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, 44280, Malatya, Turkey

* Corresponding Author: **Zeynep Kucukakcali**

Article Info
ISSN (online): 2582-8940
Volume: 06
Issue: 02
April-June 2025
Received: 11-03-2025
Accepted: 15-04-2025
Page No: 71-76

Abstract
Aim: This study aimed to evaluate the diagnostic performance of three clinical scoring systems—RIPASA, ALVARADO, and AIR—in identifying acute appendicitis. By comparing these methods against histopathological findings, the gold standard for diagnosis, the study further explored whether machine learning integration via the XGBoost algorithm could enhance classification accuracy.
Material and Methods: A publicly available retrospective dataset containing records of 107 patients, including their RIPASA, ALVARADO, AIR scores, and histopathological diagnoses, was analyzed. Diagnostic performance metrics (accuracy, sensitivity, specificity, PPV, NPV, balanced accuracy) were computed for each scoring system. Additionally, an XGBoost classification model was constructed using the three scores as input variables to predict histopathological outcomes. Five-fold cross-validation and grid search optimization were applied to prevent overfitting and enhance generalizability.
Results: RIPASA and AIR scores outperformed the ALVARADO score in both accuracy and sensitivity, with RIPASA demonstrating the most balanced diagnostic profile (accuracy: 86.0%, sensitivity: 87.8%, specificity: 76.5%). AIR achieved the highest sensitivity (97.8%) but had low specificity (29.4%). When combined with XGBoost, the AIR-based model achieved the best overall performance (accuracy: 96.0%, specificity: 100.0%, sensitivity: 90.9%, NPV: 93.3%).
Conclusion: The RIPASA and AIR scoring systems demonstrate superior diagnostic capabilities compared to ALVARADO, particularly when combined with machine learning approaches like XGBoost. The integration of scoring systems with data-driven models holds promise for enhancing diagnostic accuracy and developing robust clinical decision support tools in emergency settings.

DOI: https://doi.org/10.54660/IJMBHR.2025.6.2.71-76

Keywords: Acute appendicitis, RIPASA score, ALVARADO score, AIR score, XGBoost

Introduction

Acute appendicitis is a serious clinical condition caused by inflammation of the appendix, located in the abdominal region, and often requires surgical intervention. It is generally considered one of the most common causes of abdominal pain in emergency department visits. If not treated promptly, it can lead to life-threatening complications such as perforation, peritonitis, and sepsis. Symptoms of this condition include right lower abdominal pain, nausea, vomiting, fever, and elevated white blood cell counts in laboratory tests [1]. These symptoms usually appear progressively and cause patients to seek emergency medical help [2].

During the diagnostic process, it is crucial to thoroughly review the patient's medical history and carefully evaluate the findings of the physical examination. However, since these traditional methods may not always provide a definitive diagnosis, imaging techniques and laboratory analyses are frequently used in modern medicine. In particular, when symptoms are atypical or complex clinical presentations are observed, the use of additional diagnostic tools becomes inevitable [3].

In particular, there are several scoring systems used in the diagnosis of acute appendicitis, including the RIPASA, ALVARADO, and AIR methods. These systems facilitate the systematic evaluation of patients' clinical findings, thereby assisting in determining the need for emergency surgery. Each scoring system is based on different criteria and operates with varying levels of sensitivity and specificity in specific patient groups.

The RIPASA (Royal Infirmary of Edinburgh Appendicitis Score) score is a method that has been accepted, particularly in studies conducted in Asia. This system is calculated based on criteria grouped under 12 headings, including age, gender, duration of symptoms, leukocytosis, nausea, and migration of pain to the right lower quadrant, and demonstrates high reliability [4]. This scoring system makes an important contribution to the clinical decision-making process, particularly in developing countries, by accelerating the diagnostic process and preventing unnecessary imaging or surgical interventions [5].

The Alvarado score is calculated based on 10 items and stands out as a practical tool in the management of patients with suspected appendicitis [6]. This system includes clinical and laboratory findings such as migratory pain, loss of appetite, nausea-vomiting, right lower quadrant tenderness, rebound tenderness, fever, leukocytosis, and neutrophilia. Its ease of application by healthcare professionals ensures its widespread use in emergency department settings [7].

The AIR (Appendicitis Inflammatory Response) score provides a more detailed assessment based on inflammation markers and demographic information [8]. This system, which scores based on C-reactive protein levels, white blood cell count, pain localization, and other physical examination findings, is a suitable option for clinicians who prefer to work with more objective data. This scoring system is not only a helpful tool in the decision-making process for appendicitis diagnosis, but also helps prevent unnecessary surgical procedures [9].

Imaging methods play a critical role in the diagnostic process. Radiological methods such as ultrasonography (USG) and computed tomography (CT) are used in the diagnosis of acute appendicitis with accuracy rates of 90-95% [6]. USG is preferred as the first-line imaging method, especially in pediatric patients and pregnant women, due to its non-invasive nature and lack of radiation. This method can detect findings such as thickening of the appendix, fluid accumulation in the surrounding tissue, or inflammation. On the other hand, CT is frequently preferred in adults due to its ability to provide more detailed anatomical images. With CT, findings such as the diameter of the appendix, wall thickening, and inflammation in the surrounding fat tissue can be evaluated in detail [6, 8]. These findings are particularly valuable in establishing a diagnosis when clinical and laboratory data are insufficient.

White blood cell count and C-reactive protein (CRP) levels are also laboratory tests that aid in the diagnosis of the disease [4]. An increase in these parameters is considered an indicator of an active inflammatory process in the body and plays a supporting role in the diagnosis of appendicitis. To increase diagnostic accuracy, it is recommended that these biomarkers be evaluated in conjunction with scoring systems and imaging results [10].

In conclusion, the diagnosis of acute appendicitis should be made with great care using a series of laboratory tests, imaging methods, and clinical evaluation. Radiological imaging methods and scoring systems play an important role in clinical practice by contributing significantly to the timely treatment of patients. Thanks to such comprehensive and multifaceted approaches, diagnostic accuracy improves, the process of directing treatment accelerates, and hospital resources are utilized more effectively.

In this context, the aim of our study is to compare the diagnostic accuracy of the RIPASA, ALVARADO, and AIR scoring systems, which are commonly used in the diagnosis of acute appendicitis, with the pathology results accepted as the gold standard and to evaluate the accuracy level of each method. This evaluation was conducted using both traditional diagnostic tests (sensitivity, specificity, positive/negative predictive values, and AUC) and machine learning algorithms. Thus, the aim was to determine which of these three systems is more reliable and effective in the diagnosis of appendicitis using both classical statistical approaches and data-driven modern methods. The results obtained may contribute to the clinical decision-making process and serve as a foundation for the development of decision support systems in the future.

Dataset

In this study, a retrospective, open-access dataset comprising 107 patients was utilized to evaluate the diagnostic performance of three different clinical scoring systems (RIPASA, ALVARADO, AIR) developed for the diagnosis of acute appendicitis. The dataset includes four separate variables for each patient: RIPASA score result, ALVARADO score result, AIR score result, and histopathological diagnosis.

The results from each scoring system were coded in binary format:

A value of 1 indicates that the method diagnosed “acute appendicitis,”

while a value of 0 indicates that the method did not diagnose ”acute appendicitis.”

Similarly, the histopathology result is accepted as the gold standard, with patients diagnosed with appendicitis in the pathological evaluation labeled as 1 and those not diagnosed as 0.

The dataset includes information on a total of 107 patients, of whom 87 were histopathologically confirmed to have acute appendicitis. The remaining 20 patients were evaluated as negative by pathology.

Materials and Methods

In this study, the diagnostic efficacy of the RIPASA, ALVARADO, and AIR scoring systems, which are commonly used in the diagnosis of acute appendicitis, was compared with histopathological findings, which are considered the gold standard. For each scoring method, diagnostic metrics including accuracy, balanced accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated based on diagnostic tests. In this context, the agreement of each score with histopathology was evaluated, and its reliability in diagnosis was quantitatively demonstrated. The analyses were performed based on a confusion matrix to evaluate binary classification performance.

Additionally, in addition to diagnostic tests, the discriminative power of these three scoring systems was examined using an Extreme Gradient Boosting (XGBoost) machine learning algorithm. The XGBoost model took RIPASA, ALVARADO, and AIR scores as input variables for each patient and attempted to classify the histopathological diagnosis as the target variable. Model performance was evaluated using the metrics mentioned above and compared with traditional scoring systems. Thus, the classification performance of both classical diagnostic tests and modern machine learning-based approaches was analyzed in a comparative manner.

XGBoost is a gradient-boosted decision tree-based method that stands out for its high accuracy, generalization ability, and processing efficiency, particularly in medical classification problems [11]. During the modeling process, hyperparameter settings were optimized using the grid search method, and 5-fold cross-validation was applied for validation. This was done to eliminate the risk of overfitting and increase the generalizability of the model.

The primary objective of this study is to determine the most effective and reliable method in clinical practice by comparing the performance of these three common scoring systems with both classical diagnostic tests and the XGBoost algorithm. The findings are expected to guide the development of clinical decision support systems and the optimization of emergency department protocols.

Results

1. Findings based on diagnostic tests

The diagnostic accuracy of three different scoring systems used in the diagnosis of acute appendicitis was evaluated according to histopathological results, and the results are presented in Table 1.

Table 1: Analysis results according to diagnostic tests

Methods	Performance metrics	Value	95% CI Lower	95% CI Upper
Ripasa	Accuracy	0.860	0.794	0.926
	Balanced Accuracy	0.821	0.749	0.894
	Sensitivity	0.878	0.792	0.937
	Specificity	0.765	0.593	0.932
	Positive Predictive Value	0.952	0.881	0.967
	Negative Predictive Value	0.542	0.328	0.744
Alvarado	Accuracy	0.636	0.544	0.727
	Balanced Accuracy	0.616	0.524	0.708
	Sensitivity	0.640	0.537	0.743
	Specificity	0.586	0.433	0.738
	Positive Predictive Value	0.892	0.791	0.956
	Negative Predictive Value	0.284	0.121	0.395
Air	Accuracy	0.859	0.805	0.933
	Balanced Accuracy	0.636	0.545	0.727
	Sensitivity	0.978	0.922	0.997
	Specificity	0.294	0.103	0.560
	Positive Predictive Value	0.882	0.800	0.936
	Negative Predictive Value	0.714	0.290	0.963

According to Table 1, when the diagnostic accuracy of three different scoring systems used in the diagnosis of acute appendicitis was evaluated based on histopathological results, it was observed that the RIPASA and AIR scores generally demonstrated higher accuracy and sensitivity compared to the ALVARADO score.

The RIPASA score showed the most balanced performance with an accuracy rate of 86.0% (Balanced Accuracy = 0.821). Its sensitivity was found to be 87.8%, and its specificity was 76.5%. The positive predictive value was quite high at 95.2%, while the negative predictive value was 54.2%.

The AIR score achieved 85.9% accuracy and demonstrated remarkable sensitivity of 97.8%. However, its specificity was quite low at 29.4%. This indicates that the method is strong in identifying positive cases but weak in distinguishing negative ones.

The ALVARADO score had an accuracy rate of 63.6%, with a balanced accuracy value of 61.6%. Its sensitivity was 64.0%, and specificity was 58.6%. The positive predictive value was 89.2%, while the negative predictive value was quite low at 28.4%.

The results indicate that the RIPASA and AIR scores have higher diagnostic accuracy than the ALVARADO score. In particular, RIPASA was found to have a more balanced structure in terms of specificity and balance, while AIR was observed to be stronger in terms of not missing positive cases.

2. Findings obtained with the XGBoost Model

The results of the XGBoost algorithm applied to examine the integration of scoring systems with machine learning are presented in Table 2.

Table 2: Results obtained with XGBoost

Methods	Performance metrics	XGBoost
Ripasa	Accuracy	0.905
	Balanced Accuracy	0.806
	Sensitivity	0.667
	Specificity	0.944
	Positive Predictive Value	0.667
	Negative Predictive Value	0.944
Alvarado	Accuracy	0.812
	Balanced Accuracy	0.500
	Sensitivity	1.000
	Specificity	0.000
	Positive Predictive Value	0.812
	Negative Predictive Value	NaN
Air	Accuracy	0.960
	Balanced Accuracy	0.955
	Sensitivity	0.909
	Specificity	1.000
	Positive Predictive Value	1.000
	Negative Predictive Value	0.933

According to Table 2, the XGBoost algorithm applied to examine the integration of scoring systems with machine learning has demonstrated superior performance in terms of diagnostic accuracy compared to the results obtained with classical methods in some aspects.

The XGBoost model based on the AIR score showed the highest classification success with 96.0% accuracy and 95.5% balanced accuracy. The sensitivity of this model was determined to be 90.9%, and its specificity was 100.0%. The positive predictive value was 100.0%, and the negative predictive value was 93.3%, indicating that it is quite successful in terms of both correct diagnosis and exclusion power.

The XGBoost model created with the RIPASA score achieved 90.5% accuracy and 80.6% balanced accuracy. While its specificity was high at 94.4%, its sensitivity was found to be 66.7%. These results indicate that the model is particularly effective in excluding negative cases.

The ALVARADO-based XGBoost model achieved 81.2% accuracy but the balanced accuracy value remained at 50.0%. The model identified all positive cases (sensitivity = 1.000), but specificity was calculated as 0.0%. Therefore, the model was found to be ineffective in distinguishing negative cases.

Based on these findings, the most balanced and robust classification performance was achieved using the AIR score for both classical diagnostic tests and the XGBoost model. Specifically, when the AIR score was used in combination with XGBoost, both positive and negative cases were correctly classified, and this combination was found to be effective for use in future clinical decision support systems.Discussion

The present study evaluated the diagnostic effectiveness of three widely utilized scoring systems (RIPASA, ALVARADO, and AIR) in diagnosing acute appendicitis, comparing their accuracy with histopathological results, the established gold standard. Diagnostic accuracy in appendicitis is critical due to the frequent nature of the disease and the associated complications that may arise from delays or inaccuracies in diagnosis. The effective differentiation between acute appendicitis and other causes of abdominal pain remains a clinical challenge, often necessitating the utilization of reliable scoring systems to aid decision-making processes in emergency departments globally. Misdiagnosis can lead to either unnecessary surgeries or delayed interventions, resulting in complications such as perforation, peritonitis, and prolonged hospital stays, thus emphasizing the significance of accurate diagnostic tools.

Acute appendicitis remains one of the most frequent surgical emergencies encountered worldwide, with a lifetime incidence of approximately 7-8%, representing a significant clinical and economic burden. Although predominantly affecting adolescents and young adults, appendicitis can manifest at any age, presenting variable clinical scenarios that complicate diagnosis. Classical symptoms include migratory abdominal pain, typically beginning in the peri-umbilical area and moving to the right lower quadrant, accompanied by nausea, vomiting, anorexia, and fever. However, atypical presentations are common, especially in pediatric, elderly, and pregnant patients, significantly complicating clinical assessment and diagnosis [3, 12, 13]. The high variability in clinical presentation necessitates reliance on objective diagnostic scoring systems and imaging studies to ensure timely and accurate diagnosis, minimizing complications and optimizing patient outcomes [2, 3].

Our results demonstrated clear differences in diagnostic performance among the evaluated scoring systems. The RIPASA scoring system exhibited superior performance, with an accuracy rate of 86.0%, sensitivity of 87.8%, and specificity of 76.5%. These findings are consistent with previous studies emphasizing the reliability of the RIPASA score, particularly in Asian populations, due to its comprehensive and regionally relevant criteria [5, 14, 15]. The high positive predictive value (95.2%) underscores the potential of RIPASA to minimize unnecessary surgeries, enhance patient safety, and optimize the utilization of healthcare resources [16].

Conversely, the ALVARADO scoring system, despite its widespread recognition for simplicity and ease of use, showed comparatively lower diagnostic accuracy (63.6%) and specificity (58.6%). The limitations observed in ALVARADO might lead to diagnostic ambiguity and increase the likelihood of either unnecessary interventions or missed diagnoses, especially in atypical clinical scenarios. Prior systematic reviews have similarly reported inconsistent diagnostic accuracy with the ALVARADO score, suggesting it may require supplementation with other diagnostic modalities such as ultrasonography or computed tomography [7, 16, 17].

The AIR scoring system showed outstanding sensitivity (97.8%) but significantly lower specificity (29.4%). This high sensitivity indicates strong performance in identifying true cases of acute appendicitis, though the low specificity suggests substantial false positives, increasing the potential for unnecessary surgical explorations. To mitigate this risk, additional diagnostic evaluations, including imaging techniques such as ultrasound and CT scans, are advisable, particularly for ambiguous cases [8, 9, 18].

Integrating advanced machine learning methods, specifically Extreme Gradient Boosting (XGBoost), substantially improved diagnostic performance. Machine learning techniques, particularly XGBoost, leverage complex algorithms to analyze and interpret vast datasets effectively, identifying subtle clinical patterns that traditional scoring methods may overlook. Notably, the AIR-based XGBoost model significantly outperformed others, demonstrating an accuracy of 96.0%, balanced accuracy of 95.5%, sensitivity of 90.9%, and a specificity of 100%. These superior metrics highlight the algorithm's capacity for high-precision classification and its substantial contribution to enhancing diagnostic accuracy beyond traditional methods [11, 19].

Furthermore, the RIPASA-based XGBoost model displayed excellent specificity (94.4%), highlighting its robustness in correctly excluding non-appendicitis cases, thus preventing unnecessary surgical procedures and optimizing resource utilization. These results are consistent with previous studies advocating for machine learning applications in clinical settings to improve diagnostic efficiency, patient management, and healthcare resource allocation [20, 21].

However, the ALVARADO-based XGBoost model demonstrated inadequate specificity (0%), highlighting the critical need for careful consideration and validation of scoring systems when used in conjunction with machine learning algorithms. This underscores the necessity for meticulous dataset curation and algorithmic tuning to achieve reliable and clinically meaningful outcomes.

Overall, this study illustrates the considerable potential of integrating clinical scoring systems with advanced machine learning techniques. Such integrated approaches promise significant improvements in diagnostic accuracy, reductions in unnecessary surgical interventions, optimized healthcare resource use, and enhanced patient outcomes. Future research should focus on prospective, multicenter validations involving diverse patient populations and healthcare settings to ensure the broad applicability and generalizability of these advanced diagnostic approaches. Additionally, future studies should incorporate a broader range of clinical, biochemical, genetic, and radiologic parameters, facilitating a comprehensive understanding of acute appendicitis. The development of standardized protocols and the validation of these predictive models across various demographic and clinical subsets will further enhance their robustness and reliability. Furthermore, investigations should explore real-time integration into electronic health records and clinical decision support systems to streamline clinical workflows and decision-making processes, ultimately contributing significantly to the advancement of personalized medicine and precision diagnostics in acute appendicitis management.

References

1. Köksal M, Şam AD, Köksal F, Özdemir Hİ, Çelikbaş B, Özcan BG. The comparison of histopathological diagnosis and preoperative ultrasonography and WBC results of cases operated on with a prediagnosis of childhood acute appendicitis. Turkish Journal of Pediatric Disease. 2015. [DOI: 10.12956/tjpd.2015.122]
2. Di Saverio S, Birindelli A, Kelly MD, Catena F, Weber DG, Sartelli M, *et al*. WSES Jerusalem guidelines for diagnosis and treatment of acute appendicitis. World Journal of Emergency Surgery. 2016;11:34. [PMID: 27437029 PMCID: PMC4949879 DOI: 10.1186/s13017-016-0090-5]
3. Bhangu A, Søreide K, Di Saverio S, Assarsson JH, Drake FT. Acute appendicitis: modern understanding of pathogenesis, diagnosis, and management. Lancet (London, England). 2015;386(10000):1278-1287. [PMID: 26460662 DOI: 10.1016/s0140-6736(15)00275-5]
4. Peksöz R, Albayrak Y, Şebin E, Dişçi E, Aşkın S, Baştürk İ. Akut apandisit hastalarında ghrelin hormon düzeyi bakılması ve ghrelin hormon düzeyinin c-reaktif protein ve beyaz kan hücre düzeyleri ile karşılaştırılması. Genel Tıp Dergisi. 2021;31(3):252-256. [DOI: 10.54005/geneltip.996787]
5. Chong CF, Adi MI, Thien A, Suyoi A, Mackie AJ, Tin AS, *et al*. Development of the RIPASA score: a new appendicitis scoring system for the diagnosis of acute appendicitis. Singapore Medical Journal. 2010;51(3):220-225. [PMID: 20428744]
6. Akgül N, Gündeş E. Neutrophil/lymphocyte ratio in acute appendicitis: a state hospital experience. Turkish Journal of Colorectal Disease. 2016;26(4):121-124. [DOI: 10.4274/tjcd.42243]
7. Ohle R, O'Reilly F, O'Brien KK, Fahey T, Dimitrov BD. The Alvarado score for predicting acute appendicitis: a systematic review. BMC Medicine. 2011;9:139. [PMID: 22204638 PMCID: PMC3299622 DOI: 10.1186/1741-7015-9-139]
8. Dıkıcıer E, Altıntoprak F, Çakmak G, Akbulut G. The use of ultrasound imaging for acute appendicitis. Sakarya Medical Journal. 2011;1(2):64-66. [DOI: 10.5505/sakaryamj.2011.32032]
9. Andersson RE, Hugander A, Thulin AJ. Diagnostic accuracy and perforation rate in appendicitis: association with age and sex of the patient and with appendicectomy rate. The European Journal of Surgery = Acta Chirurgica. 1992;158(1):37-41. [PMID: 1348639]
10. Rezak A, Abbas HM, Ajemian MS, Dudrick SJ, Kwasnik EM. Decreased use of computed tomography with a modified clinical scoring system in diagnosis of pediatric acute appendicitis. Archives of Surgery. 2011;146(1):64-67.
11. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.
12. Snyder MJ, Guthrie M, Cagle S. Acute appendicitis: efficient diagnosis and management. American Family Physician. 2018;98(1):25-33. [PMID: 30215950]
13. Petroianu A. Diagnosis of acute appendicitis. International Journal of Surgery. 2012;10(3):115-119.
14. Butt MQ, Chatha SS, Ghumman AQ, Farooq M. RIPASA score: a new diagnostic score for diagnosis of

acute appendicitis. Journal of the College of Physicians and Surgeons—Pakistan: JCPSP. 2014;24(12):894-897. [PMID: 25523723]
15. Aman T, Yaqoob L, Shah SA, Arif NS, Shaheen H, Shah M, *et al*. Evaluation of RIPASA score for diagnosis of acute appendicitis. KJMS. 2018;11(2):189.
16. Karami MY, Niakan H, Zadebagheri N, Mardani P, Shayan Z, Deilami I. Which one is better? Comparison of the acute inflammatory response, Raja Isteri Pengiran Anak Saleha appendicitis and Alvarado scoring systems. Annals of Coloproctology. 2017;33(6):227-231. [PMID: 29354605 PMCID: PMC5768477 DOI: 10.3393/ac.2017.33.6.227]
17. Kollár D, McCartan D, Bourke M, Cross K, Dowdall J. Predicting acute appendicitis? A comparison of the Alvarado score, the appendicitis inflammatory response score and clinical assessment. World Journal of Surgery. 2015;39(1):104-109.
18. Andersson M, Andersson RE. The appendicitis inflammatory response score: a tool for the diagnosis of acute appendicitis that outperforms the Alvarado score. World Journal of Surgery. 2008;32(8):1843-1849. [PMID: 18553045 DOI: 10.1007/s00268-008-9649-y]
19. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930. [PMID: 26572668 PMCID: PMC5831252 DOI: 10.1161/circulationaha.115.001593]
20. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, *et al*. ChexNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv Preprint arXiv:171105225. 2017.
21. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. Journal of Internal Medicine. 2018;284(6):603-619.