A Comparative Multimodal Machine Learning Approach for Diagnosis and Management of Pediatric Appendicitis

Dr. Ranjan Kumar[1*], Shad Jamil[1,2†], Kunal Sati[1†], Ritu Attri[1,3†]

[1*]Department of Computer Science, Aryabhatta College, University of Delhi, Plot No 5, Benito Juarez Marg, South Campus, Anand Niketan, New Delhi, 110021, Delhi, India.

[2*]Department of Computer Science, University of Delhi, First Floor, Faculty of Mathematical Sciences, University of Delhi, New Delhi, 110007, Delhi, India.

[3*]Department of Computer Science and Information Technology, Central University of Haryana, Jant-Pali, Mahendergarh, 123031, Haryana, India.

*Corresponding author(s). E-mail(s): ranjan301@gmail.com;
Contributing authors: shad.datascience@gmail.com;
kunal.sati07@gmail.com; rituattri735@gmail.com;
†These authors contributed equally to this work.

Abstract

Appendicitis is one of the most common diseases in the pediatric age group. The management of appendicitis in children lacks international agreement, primarily because of the lack of defined global recommendations and the limited availability of data-driven research on pediatric appendicitis. Early diagnosis of appendicitis is very crucial to avoid life threatening conditions like peritonitis, abscess formation, perforation, etc. Conventional methods of diagnosis detection include clinical scoring systems, laboratory reports as well as ultrasound imaging. The data used in this study consist of tabular as well as image data for 782 patients in the 0-18 years age group, collected between the years 2016 and 2021. We have independently analysed the structured (tabular) and unstructured (image) data for prediction of two target variables i.e., diagnosis and management. We have also combined the available data to create a hybrid dataset which further aided the results. Our model for tabular data with XGBoost classifier outperformed all previously proposed models for diagnosis prediction with F1-Score of

0.98 and AUROC of 0.99. The model for hybrid data with LGBM classifier outperformed all other modalities for management prediction with F1-Score of 0.98 and AUC=1.00. Explainable AI (XAI) methods are used to interpret the result of the proposed model. Our proposed model presents a more precise diagnostic approach for appendicitis, offering valuable support to medical practitioners in clinical decision-making.

Keywords: Multimodal, Pediatric Appendicitis, Classification, Ensemble learning, XAI

1 Introduction

Appendicitis is a frequently observed surgical emergency amongst the paediatric age group. It is more common in males with a lifetime risk of 86% as compared to 67% in females[1]. Despite of technological advancements in healthcare and medicine, diagnosis of appendicitis continues to be a challenge owing to a diverse range of observed symptoms in patients which may lead to severe complications. Symptoms which appear commonly in adults such as localized pain in lower abdomen, loss of appetite, nausea, etc. are usually found to be lacking in children whereas atypical symptoms like abdominal pain and diarrhoea are more common. This leads to further complications and could prove to be a potential barrier in treatment.

Children under the age of five, may struggle to articulate their symptoms, adding another layer of complexity in diagnosis. Accurate and timely diagnosis of paediatric appendicitis is crucial in preventing serious infections like perforation, abscess formation, and peritonitis, all of which can be life-threatening. Paediatric appendicitis can be diagnosed through a variety of methods, including physical examination, laboratory tests, imaging techniques, as well as scoring systems. Each method, however, has its own limitations, and the selection of a diagnostic approach should be based on patients' symptoms and clinical representations. Various imaging techniques such as Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI) and ultrasound scans are useful for diagnosing the disease. Ultrasound and CT scan are the commonly used methods. Though CT scans are more accurate, ultrasounds are preferred due to their cost-effectiveness, availability and non-involvement of radiation. After diagnosis, managing appendicitis usually requires surgical intervention. This can be done through either open appendectomy or laparoscopic appendectomy. The choice depends upon several factors such as severity of the disease, complications, surgeon's proficiency, etc. To tackle these difficulties, advancements in artificial intelligence (AI) have started to improve conventional diagnostic approaches [1]. AI encompasses the broader domain of developing intelligent machines. Machine Learning (ML) is a subset of AI concentrated on algorithms that learn from data while Deep Learning (DL) is a subset of ML that utilizes deep neural networks to grasp intricate patterns and representations from extensive datasets. In this paper, ML techniques would be utilized for diagnosis of pediatric appendicitis using relevant clinical data as well as ultrasound images collected from patients. Our proposed model is intended to assist physicians in diagnosing appendicitis.

The main contributions of our paper can be summarized as follows:

- First, Preprocessing techniques are applied on medical data, collected in both tabular and image forms.
- Second, both processed tabular and image data are merged to create a hybrid dataset.
- Third, various feature selection/extraction methods including PCA, Fisher Score, Variance Threshold, Mutual Information, and Boruta algorithm, are applied on all three forms of dataset (tabular, image, and hybrid) subsequently reducing their size without losing valuable information.
- Fourth, ML techniques are applied on the reduced datasets to select the best performing dataset and then ensemble learning techniques are applied on the selected dataset for creating the predictive models.
- Finally, SHAP value is an explainable AI (XAI) technique, used to explain the working of the proposed models.

The rest of the paper is structured as follows: Section 2 provides an overview of the relevant studies on paediatric appendicitis. This section also discusses their limitations if any. Section 3, titled Materials and Methods, includes the description of the dataset, an overview of all the machine learning techniques used for preparation and analysis of the dataset and how they are employed for this dataset. Section 4 presents and discusses the obtained results. Section 5 concludes the paper by highlighting the limitations and implications of our model while touching upon future opportunities for research in this field.

2 Related Work

Several studies have been conducted where ML and AI techniques have been used for diagnosis patients with suspected appendicitis using only ultrasound images of abdominal area. The data of 579 children and adolescents who were hospitalized at the department of Paediatric Surgery and Paediatric Orthopaedics in the tertiary Children's Hospital St. Hedwig in Regensberg, Germany between January 1, 2016, and December 31, 2021, have been processed and analysed using the variants of Concept Bottleneck Model (CBM) and Random Forest (RF) algorithm [2]. ML techniques can also be utilized to aid in diagnosing Acute Appendicitis. A study involved analysing data from 595 patients admitted to Hitit University Training and Research Hospital and revealed that the gradient boosting (GB) tree algorithm demonstrated the highest efficacy, achieving an impressive prediction accuracy of 95.31% [3]. Another analysis conducted using a dataset sourced from a public hospital, spanning the years 2016 to 2019, gathered information from 625 patients, comprising 371 males and 254 females. The findings indicated that the RF algorithm yielded the most favourable outcome, achieving an accuracy of 83.75%, a precision of 84.11%, sensitivity of 81.08%, and specificity of 81.01% [4]. ML techniques such as Support Vector Machine (SVM), Decision Tree (DT), logistic regression (LR), K-nearest Neighbour (KNN), Artificial Neural Network (ANN), and Gradient Boosting (GB) have been incorporated in data collected from 1950 patients, sourced from the Department of Gastrointestinal Surgery

at Gia Dinh People Hospital in Ho Chi Minh City, Vietnam, spanning the period from 2016 to 2020, to predict Appendicitis in a resource-limited setting. Remarkably, the Gradient Boosting algorithm consistently outperformed the others, demonstrating AUC and accuracy values of approximately 0.8 or higher in both adjusted and unadjusted data, surpassing the performance of all other algorithms [5]. To predict paediatric appendicitis, Ensemble Learning (EL) techniques have also been applied on publicly available dataset from the Department of Paediatric Surgery at Regensburg's Children's Hospital St. Hedwig which consists of 430 individuals aged 0 to 18 with suspected appendicitis. 38 predictor variables and 3 target variables as Diagnostic (appendicitis vs. no appendicitis), Treatment (conservative vs. surgical), and Complications (present vs. absent) were considered in the study. EL techniques, namely Weighted Averaging and Majority Voting, have been used and outperformed other classification techniques such as Logistic Regression (LR), Naive Bayes (NB), KNN, SVM, DT, RF, Multi-Layer Perceptron (MLP), AdaBoost etc. Across all cases pertaining to Diagnostic, Treatment, and Complications class values, Majority Voting achieved accuracies of 92.15%, 92.73%, and 94.19% respectively, surpassing the performance of Weighted Averaging and yielding notably higher accuracies [6]. In [7], ML based Explainable Artificial Intelligence (XAI) has been adopted in prediction of Perforated and Non perforated Acute Appendicitis with taken into consideration the data of 1797 patients who diagnosed of AAp by the Department of Surgery of Inonu University Faculty of Medicine between May 2009 and March 2022. A model based on CatBoost classifier was used for classify the perforated and non-perforated AAp patients. The model was interpreted using SHAP Values, which is another XAI method. It achieved an accuracy of 88.2% (with a confidence interval of 85.6% to 90.8%) in discriminating between AAp and non-AAp cases. The model also achieved an accuracy of 92% in distinguishing between perforated and non-perforated AAp cases.

3 Materials and Methods

The workflow for the proposed work is shown in Figure 1. Initially, we prepared three models using three different types of datasets: tabular data, image data, and hybrid data, which combined preprocessed tabular and image data. We applied various feature selection/extraction methods to each dataset, and based on performance, we extracted the best subsets of data for further analysis. We selected the top-performing sub-models from each dataset by evaluating the predicted results using various ensemble learning (EL) techniques. Finally, by comparing these three sub-models, the best-performing model was proposed. Subsequently, SHAP values were used as an explainable artificial intelligence (XAI) technique to help explain the proposed model.

3.1 Dataset Description

The data used for the study is from the Pediatric Appendicitis Database, which is publicly available [https://doi.org/10.5281/zenodo.7711412]. The dataset consists of pediatric patients admitted with abdominal pain, with the suspected cause being appendicitis between 2016 and 2021. Two types of data are available: tabular data

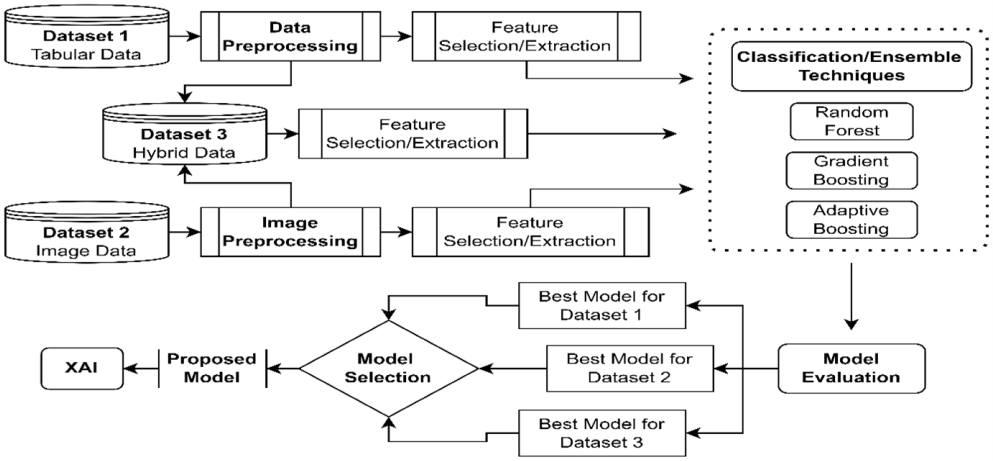

Fig. 1 Workflow of the Proposed System

and image data. The tabular part comprises data from physical examinations, clinical scoring systems, laboratory tests, and manually extracted ultrasonographic findings. The images are B-Mode ultrasound images. Tabular data has 782 instances of patients with 55 predictor variables and 3 target variables, namely diagnosis (appendicitis vs. no-appendicitis), management (conservative vs. primary surgical vs. secondary surgical), and severity (complicated vs. uncomplicated). The image data consists of 2096 abdominal B-Mode ultrasound images of various regions such as the appendix and surrounding tissues, the right lower quadrant (RLQ), adjacent organs (like ileum, cecum, and reproductive organs like the ovaries for females), and lymph nodes. Each patient has somewhere from 1 to 15 images, as various regions of the abdomen are scanned.

3.2 Data Preprocessing

3.2.1 Tabular Data Prerpocessing

Initially, the dataset comprised 782 instances with 55 feature columns and 3 target variables. Columns with missing values and those irrelevant to our analysis, such as 'US_Number', 'US_Performed', 'Length_of_stay' and 'Diagnosis_presumptive', are dropped. All the rows with missing values are also dropped. To make the data machine readable, techniques like one-hot encoding, ordinal encoding, and conversion of string values into numerical are performed accordingly.

Two of the three target variables, diagnosis and management, are chosen for prediction. Diagnosis contains binary values: appendicitis and no appendicitis. Management was initially a multivalued variable, having values as primary surgical, secondary surgical and conservative. It is converted to a binary valued variable by merging 'primary surgical' and 'secondary surgical' values into a single value called 'surgical'. The resultant management variable has binary values 'Conservative' and 'Surgical'. The subset

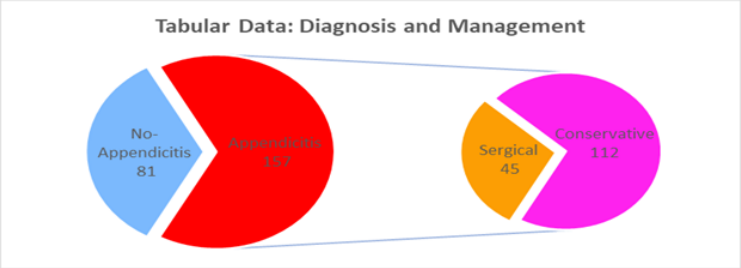

Fig. 2 Distribution of Values for Diagnosis (Left) and Management (Right) for Tabular Data after Preprocessing.

of patients considered for prediction of the management variable are only those who are diagnosed as having appendicitis. Figure 2 depicts the distribution of values for diagnosis (Left) as well as management (Right) target variable.

3.2.2 Image Data Preprocessing

The original ultrasonography images contained graphical user interface elements, distance measurements, markers, as well as other annotations which are commonly witnessed in acquisition of such images in a clinical routine. Therefore, to acquire a cleaner view suited to machine learning purposes, we used DeepFill [8], which is an open-source framework for generative image inpainting task, for image resizing and pixel intensity normalization. All the remaining images are resized to a common size of 400 x 400 px2 with zero padding and the pixel intensities normalized in range (0,1). The images with multiple views are manually removed due to lower quality of available views in the images. After processing the images, radiomics features were extracted from the images using the PyRadiomics package in python. A ROI (Region of Interest) mask of size 400 x 360 px2 is used to exclude zero padded borders that are prevalent in the majority of the resized images. Radiomics refers to the process of extracting many quantitative features from medical images like CT, MRI and PET scans using data characterization algorithms [9]. A total of 744 radiomics features are extracted for each image. The extracted radiomics features can be categorized into 7 types: First Order Statistics Features, Shape Based 2D 3D Features, Gray Level Co-occurrence Matrix (GLCM) Features, Gray Level Run Length Matrix (GLRLM) Features, Gray Level Size Zone Matrix (GLSZM) Features, Gray Level Dependence Matrix (GLDM) Features, and Neighboring Gray Tone Difference Matrix (NGTDM) Features. Figure 3 summarizes this whole process of preprocessing the images. The mentioned features can be extracted from the original image or a derived image using filters such as Laplacian of Gaussian (LoG), Square and Wavelet. Square filter takes squares of pixel intensities whereas LoG filter is an edge enhancement filter [10]. Wavelet filter [10] yields 8 decompositions per level i.e., all possible combinations of both High Pass and Low Pass filters in all 3 dimensions. Out of the 744 extracted radiomics features per image, 93 features each were obtained on the original and square filtered image, 186 on the LoG filtered image and 372 on the wavelet filtered image. Figure 3 summarizes this whole process of preprocessing the images. The mentioned features can be

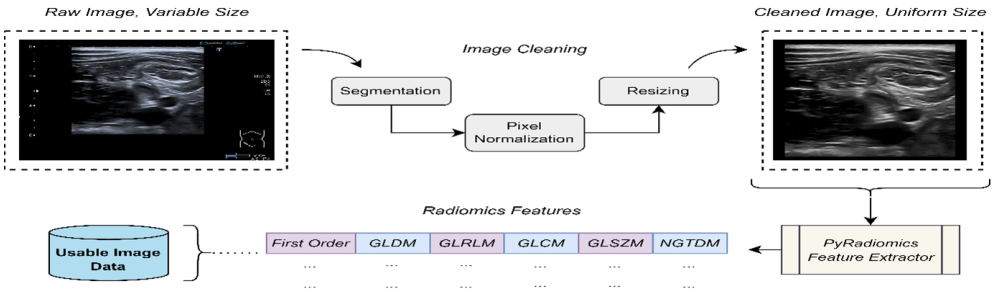

Fig. 3 A comparative view of image before and after preprocessing.

extracted from the original image or a derived image using filters such as Laplacian of Gaussian (LoG), Square and Wavelet. Square filter takes squares of pixel intensities whereas LoG filter is an edge enhancement filter [10]. Wavelet filter [11] yields 8 decompositions per level i.e., all possible combinations of both High Pass and Low Pass filters in all 3 dimensions. Out of the 744 extracted radiomics features per image, 93 features each were obtained on the original and square filtered image, 186 on the LoG filtered image and 372 on the wavelet filtered image.

3.2.3 Hybrid Data

For hybrid data, a subset of Clinical Tabular data is merged with the radiomics features extracted from the images, as shown in Figure 4. This subset of Tabular data only includes Demographics information, clinical scoring system values and laboratory data. Manually extracted ultrasound features are excluded from this subset. Both image and hybrid data contain 808 images each. Within this set, 564 images indicate the presence of appendicitis, while the remaining images do not. Among the 564 cases showing appendicitis, 390 are treated conservatively, while 174 require surgical intervention, as illustrated in Figure 5.

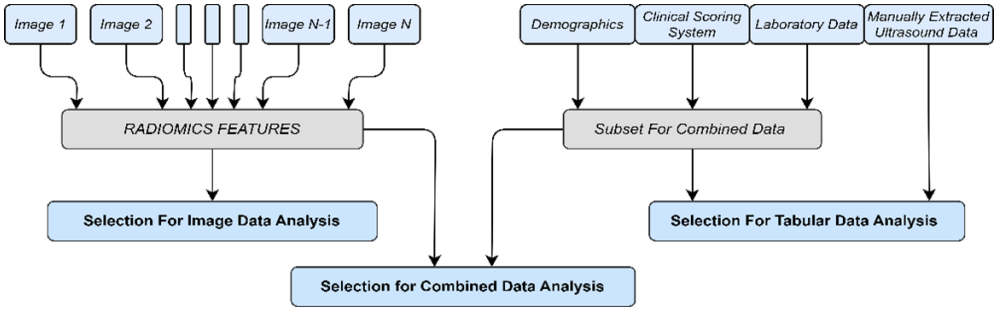

Fig. 4 A descriptive view of the data selected for analysis.

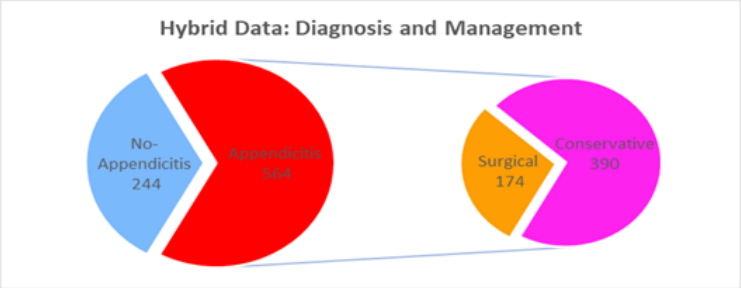

Fig. 5 Distribution of Image and Hybrid data for Diagnosis and Management after preprocessing.

3.3 Feature Selection/Extraction

Feature selection is essential for optimizing the feature set by reducing redundant features and noise, while simultaneously improving model performance and enhances interpretability. We used several feature selection/extraction techniques such as PCA (Principal Component Analysis), Fisher's Scoring, Variance Threshold, Mutual Information, and Boruta Feature Selection Algorithm [12].

3.3.1 Principal Component Analysis

Principal Component Analysis (PCA) takes a large data set with many variables per observation and reduces them to a smaller set of summary indices. These indices retain most of the information in the original set of variables. Analysts refer to these new values as principal components. The principal components themselves are a set of new, uncorrelated variables that are linear combinations of the original variables. Here, the principal components are obtained by preserving 95% of the variance of original datasets.

3.3.2 Fisher's Score

Fisher score is one of the most widely used supervised feature selection methods. The algorithm we had used returns the ranks of the variables based on the fisher's score in descending order. We can then select the variables as per the case. For tabular data only those features were selected which have a value greater than or equal to 5, while for the other two datasets a threshold of 500 was applied.

$$u(\theta) = \nabla_\theta \log p(x \mid \theta) \tag{1}$$

3.3.3 Variance Threshold

The variance threshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. We assume that features with a higher variance may contain more useful information. The threshold value of 0.4 is selected for tabular and image data while a threshold of 0.4 is being used for the hybrid data. These threshold values are selected based on the reduced dataset's

performance.

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{2}$$

3.3.4 Mutual Information

Mutual information of two random variables X and Y is a value that statistically measures the dependency between X and Y. The formula for calculating mutual information between the two random variables X as a feature and Y as a target variable is defined in equation (3). For the tabular data, those features are selected that have mutual information greater than 0.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x,y) \cdot \log \left(\frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)} \right) \tag{3}$$

3.3.5 Boruta Algorithm

Boruta is a wrapper method of feature selection. It is built around the random forest algorithm. It can be used on any classification model. Since it is a Random Forest based method, it works for tree models like Random Forest or XGBoost. Boruta iteratively removes features that are statistically less relevant than a random probe. In each iteration, rejected variables are removed from consideration in the next iteration. It generally ends up with good global optimization for feature selection. Features are selected based on their rank given by the Boruta algorithm.

3.4 Classification Techniques

Simple Machine learning classifiers like K-Nearest Neighbors, Naive Bayes, Logistic Regression, SVM, etc. are applied for the comparison of feature selection techniques. Ensemble learning (EL) techniques like Random Forest and Gradient Boosting using libraries like Light Gradient Boosting, Xtreme Gradient Boosting, and AdaBoost are used for classification purposes [13].

3.5 Evaluation Method and Metrics

For each of the 3 datasets, the basic machine learning classifiers were trained and tested using 5-Fold Cross Validation. Here, the dataset is randomly split into 5 equal-sized disjoint subsets called folds. Then the model is trained on 4 folds and evaluated on the one remaining fold. This step is repeated 5 times, selecting a different combination of 4 folds to train each time. This 5-fold CV method helps to reduce variance and provides better generalization for the model The metrics used for evaluation are F1-Score, Accuracy, Recall, AUROC (Area Under Receiver Operating Characteristic Curve) and AUPR (Area Under Precision-Recall Curve), averaged over 5 folds of cross validation. AUPR is an important evaluation metric in medical data because of several reasons such as rarity of positive class (i.e., when instances of 0-Class are higher than instance of 1-Class) and when cost of false positives are high.

Precision – It is defined as the measure of correct positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall-It is the measure of how well a classifier can predict the positive class.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Accuracy – It defines how well a model can predict the values.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

F1-Score – It is defined as the harmonic mean of precision and recall. It combined two metrics into one that can balance the errors.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

AUC (Area Under Receiver Operating Characteristic Curve) – This metric is helpful in determining how well a classifier can distinguish between the classes.

AUPR (Area Under Precision-Recall Curve) – This metric is useful in determining the model's ability to identify positive instances by considering both precision and recall values.

4 Results and Discussion

4.1 Feature Selection for Appendicitis Diagnosis and Management

We used five feature selection techniques, including PCA, Fisher's Score, variance threshold, mutual information, and Boruta. We applied all five feature selection techniques to each dataset: dataset 1 (tabular data), dataset 2 (image data), and dataset 3 (hybrid data), resulting in five distinct reduced datasets. We further applied ML classifiers to each reduced dataset to select the best one. We evaluated the performance of all ML-based classifiers using accuracy, F1-score, and AUROC. It is evident from the results summarized in Table 1 and Table 2 that the Decision Tree classifier outperformed other classifiers after applying feature selection techniques, which also helps in selecting the final reduced datasets. For the target variable "diagnosis," information gain with a mutual information threshold >= 0.01 proved best for both tabular and hybrid data, whereas the Boruta algorithm gave the best results for image data. For the target variable "management," Fisher's score, Boruta algorithm, and information gain with a mutual information threshold >= 0.02 give the best results for tabular, image, and hybrid data, respectively.

Table 1 Results of Feature Selection/Extraction for Diagnosis

Modality	***Feature Selection Technique***	***Features Selected***	***Evaluation Metric***		
			ACC[1]	***F1***[2]	***AUC***[3]
Dataset 001 (Tabular Data)	*PCA*	3	0.63	0.72	0.58
	Fisher's Score	26	0.90	0.93	0.89
	Variance Threshold	17	0.97	0.97	0.96
	Mutual Information	**26**	**0.98**	**0.99**	**0.99**
	Boruta	18	0.97	0.98	0.97
Dataset 002 (Image Data)	*PCA*	2	0.63	0.74	0.57
	Fisher's Score	244	0.67	0.76	0.61
	Variance Threshold	371	0.67	0.76	0.62
	Mutual Information	356	0.67	0.77	0.61
	Boruta	**314**	**0.68**	**0.76**	**0.62**
Dataset 003 (Hybrid Data)	*PCA*	2	0.63	0.74	0.57
	Fisher's Score	178	0.79	0.85	0.75
	Variance Threshold	416	0.80	0.86	0.76
	Mutual Information	**25**	**0.80**	**0.86**	**0.77**
	Boruta	312	0.79	0.85	0.76

Table 2 Results of Feature Selection/Extraction for Management

Modality	***Feature Selection Technique***	***Features Selected***	***Evaluation Metric***		
			ACC[1]	***F1***[2]	***AUC***[3]
Dataset 004 (Tabular Data)	*PCA*	3	0.64	0.36	0.55
	Fisher's Score	**28**	**0.85**	**0.72**	**0.81**
	Variance Threshold	16	0.70	0.49	0.64
	Mutual Information	22	0.80	0.65	0.76
	Boruta	17	0.78	0.62	0.74
Dataset 005 (Image Data)	*PCA*	2	0.57	0.33	0.51
	Fisher's Score	644	0.68	0.50	0.64
	Variance Threshold	392	0.69	0.50	0.64
	Mutual Information	99	0.67	0.44	0.61
	Boruta	**304**	**0.69**	**0.51**	**0.64**
Dataset 006 (Hybrid Data)	*PCA*	2	0.57	0.69	0.51
	Fisher's Score	278	0.92	0.94	0.91
	Variance Threshold	386	0.89	0.92	0.88
	Mutual Information	**230**	**0.93**	**0.95**	**0.92**
	Boruta	513	0.91	0.93	0.89

[1]Accuracy

[2]F1-Score

[3]AUC-ROC

4.2 Prediction for Appendicitis Diagnosis and Management

Several ML techniques such as KNN (K-Nearest Neighbors), LR (Linear Regression), GNB (Gaussian Naive Bayes), DT (Decision Tree) and various EL like RF (Random Forest), LGBM (Light Gradient Boosting), XGB (Xtreme Gradient Boosting), ADAB (AdaBoost) have been applied for early disease diagnosis and management. The results obtained on all 3 datasets (tabular, image, and hybrid) for both target variables "Diagnosis" and "Management" are depicted in Figure 6. Classifiers based

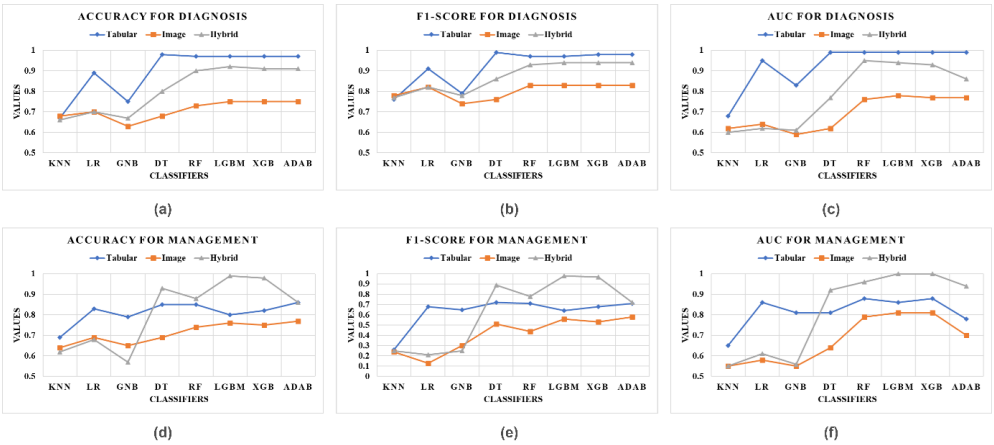

Fig. 6 Line charts comparing performance of various classifiers for each dataset over 3 metrics - Accuracy, F1 and AUC. Graphs 8(a), (b), (c) are plotted for diagnosis target variable whereas 8(d), (e), (f) are plotted for management. KNN, LR, GNB, DT, RF, LGBM, XGB, ADAB refer to K-Nearest Neighbors, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Random Forest, Light Gradient Boosting Model, Xtreme Gradient Boosting and Adaptive Boosting respectively. The blue line represents the tabular data, orange line represents image, and the grey line represents hybrid data.

on EL techniques are proved to be more efficient than classifiers based on simple ML techniques, across all 3 datasets, summarized in Table 3.

According to the results, The EL technique XGB (Xtreme Gradient Boosting) algorithm outperformed other EL techniques on Tabular data whereas LGBM (Light Gradient Boosting Machine) emerged as top performer in case of image and hybrid data for prediction of target variable "Diagnosis". The predictive model using tabular data outperformed with AUROC = 0.99 and AUPR = 1.00 whereas other predictive models based on image and hybrid data having AUROC = 0.78, AUPR = 0.89 and AUROC = 0.94, AUPR = 0.96, respectively. Similarly, RF (Random Forest) classifier performed with AUROC = 0.88 and AUPR = 0.86 on tabular data, which is higher than all other techniques applied for prediction of target variable "Management".

The performance of proposed model for prediction of appendicitis diagnosis "Model 1" is evaluated via ROC which is depicted in Figure 7(a). It shows that the tabular data yielded comparatively superior results but the hybrid model considered to be a more promising approach considering its capacity to reduce human intervention

Table 3 Results of Ensemble Techniques for Diagnosis Management

Modality	*Classifier*	*Diagnosis*				*Management*			
		ACC[1]	***F1***[2]	***AUC***[3]	***AUPR***[4]	***ACC***	***F1***	***AUC***	***AUPR***
Tabular	***RF***	0.97 (±0.01)	0.97 (±0.01)	0.99 (±0.01)	1.00 (±0.00)	**0.85 (±0.05)**	**0.71 (±0.06)**	**0.88 (±0.09)**	**0.86 (±0.08)**
	LGBM	0.97 (±0.01)	0.97 (±0.01)	0.99 (±0.01)	0.99 (±0.00)	0.80 (±0.11)	0.64 (±0.14)	0.86 (±0.07)	0.80 (±0.07)
	XGB	**0.97 (±0.02)**	**0.98 (±0.01)**	**0.99 (±0.01)**	**1.00 (±0.00)**	0.82 (±0.07)	0.68 (±0.10)	0.88 (±0.07)	0.83 (±0.07)
	ADAB-RF	0.97 (±0.02)	0.98 (±0.01)	0.99 (±0.01)	1.00 (±0.00)	0.86 (±0.04)	0.71 (±0.06)	0.78 (±0.04)	0.81 (±0.07)
Image	*RF*	0.73 (±0.03)	0.83 (±0.02)	0.76 (±0.03)	0.88 (±0.03)	0.74 (±0.04)	0.44 (±0.08)	0.79 (±0.05)	0.63 (±0.05)
	LGBM	**0.75 (±0.03)**	**0.83 (±0.02)**	**0.78 (±0.04)**	**0.89 (±0.03)**	**0.76 (±0.04)**	**0.56 (±0.07)**	**0.81 (±0.06)**	**0.68 (±0.05)**
	XGB	0.75 (±0.03)	0.83 (±0.02)	0.77 (±0.04)	0.88 (±0.03)	0.75 (±0.02)	0.53 (±0.04)	0.81 (±0.04)	0.66 (±0.04)
	ADAB-RF	0.75 (±0.04)	0.83 (±0.03)	0.77 (±0.04)	0.87 (±0.02)	0.77 (±0.03)	0.58 (±0.06)	0.70 (±0.04)	0.67 (±0.05)
Image	*RF*	0.90 (±0.02)	0.93 (±0.02)	0.95 (±0.02)	0.97 (±0.01)	0.88 (±0.04)	0.78 (±0.06)	0.96 (±0.02)	0.93 (±0.04)
	LGBM	**0.92 (±0.02)**	**0.94 (±0.01)**	**0.94 (±0.02)**	**0.96 (±0.02)**	**0.99 (±0.01)**	**0.98 (±0.01)**	**1.00 (±0.00)**	**1.00 (±0.00)**
	XGB	0.91 (±0.03)	0.94 (±0.02)	0.93 (±0.03)	0.95 (±0.02)	0.98 (±0.01)	0.97 (±0.01)	1.00 (±0.00)	0.99 (±0.01)
	ADAB-RF	0.91 (±0.02)	0.94 (±0.02)	0.86 (±0.03)	0.95 (±0.01)	0.86 (±0.03)	0.72 (±0.05)	0.94 (±0.02)	0.90 (±0.03)

[1]Accuracy

[2]F1-Score

[3]AUC-ROC

[4]PR-AUC

in the diagnostic process. By leveraging radiomics for feature extraction from the images instead of manual analysis of ultrasound images, the hybrid model significantly improves data quality, thereby reducing the risk of human errors and shortening the diagnosis time. The performance of proposed model for prediction of appendicitis management "Model 3" is evaluated using ROC, as shown in Figure 7(b). It is clear from the results that hybrid data is the more suitable than tabular in predicting appendicitis management. Hence, the performance of models based on hybrid data for predicting both diagnosis and management make it as a worthwhile option to explore.

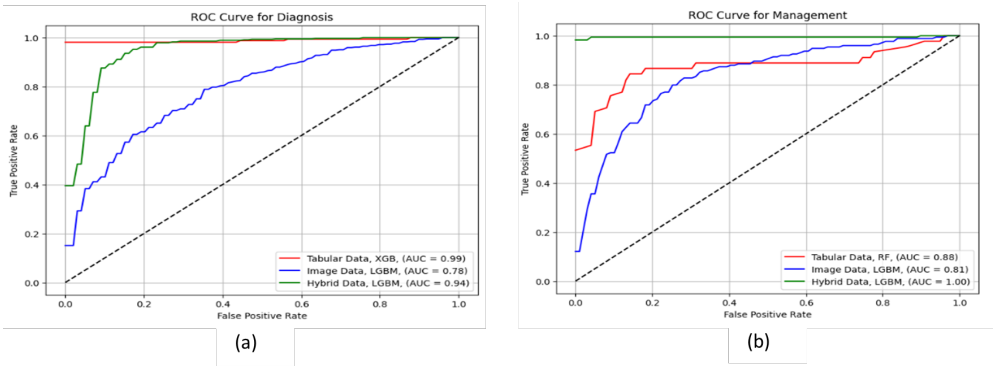

Fig. 7 ROC curves for the best performing classifiers in all 3 datasets. 7(a) is for diagnosis 7(b) for management. The ROC curve observed for tabular data is shown in red colour. Similarly, image and hybrid data are shown in blue and green colour respectively.

4.3 Explainable Artificial Intelligence (XAI) Methods for interpretability of proposed models

Explainable Artificial Intelligence (XAI) methods SHAP Values are utilized to enhance the interpretability of the top performing model for both the target variables "Diagnosis" and "Management". These methods can identify the underlying patterns and relationships in the data, thus providing a transparent look into the whole mechanism of the ML models. It quantifies the contribution of each feature in the predictions made by the model and help the clinicians to understand the reasoning behind the model's outputs. SHAP Value of a particular feature is directly proportional to the feature's impact on the model output. As observed in Figure 10 for prediction of target variable "Diagnosis", Appendix Diameter and Clinical Scoring Parameters i.e., Alvarado score and pediatric appendicitis score had the highest contribution to model output, followed by red blood cell count and neutrophil percentage, both of which are blood test markers. The remaining features such as patient's height, free fluids, white blood cell count, etc. contributed the same on average to the model's output. Similarly, Figure 11 shows the selected Top-30 features for prediction of target variable "Management". The highest contributing variable is Peritonitis which refers to the inflammation of abdomen lining. Other important variables include white blood cell

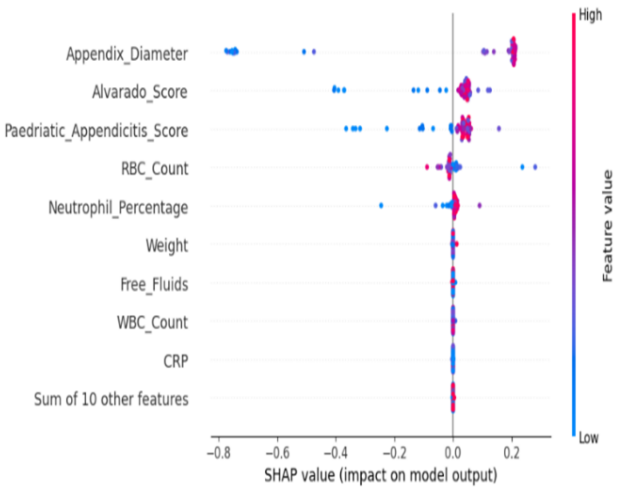

Fig. 8 Bee swarm Plot for SHAP Values of features for tabular data (in diagnosis prediction)

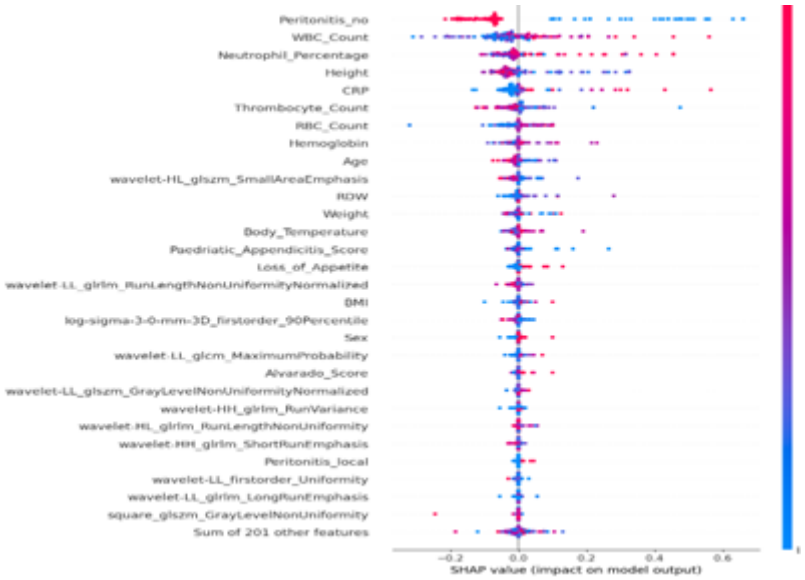

Fig. 9 Bee swarm Plot for SHAP (For Hybrid Management)

count, neutrophil count, c-reactive protein count, thrombocyte count, red blood cell count and hemoglobin. It can be noticed that most of the top contributors are blood test markers. Most of the image learning features have uniform impact on the model's output.

5 Conclusion & Future Scope

As already mentioned, Appendicitis is one of the most common surgical emergencies in children. The aim of this study was to investigate a multi-modal approach for prediction of pediatric appendicitis using the available clinical data on both forms tabular and image. Model for tabular data was proposed due to its superior performance over the other two models. Boosting Classifiers like AdaBoost, LGBM and XGBoost classifiers performed very well with an Average Precision (AP) score of 100% for tabular data, 91% for image data and 98% for the hybrid data. Combination of both data types (tabular and image) does an excellent job at using all the relevant information collected during clinical tests and routine ultrasounds without relying on medical experts to handpick features from ultrasound images, which played a major role in tabular data analysis.

In future research works, there is a possibility to expand our analysis to include patients from additional geographical regions. Models could be trained upon larger datasets to improve their predictive qualities. The field of ML and AI has a huge potential in medical sciences. Our aim would be enhancing the clinical applicability of this study across diverse areas. Moreover, we propose to explore the application of additional ML techniques on different datasets containing ultrasound images. This endeavor is aimed at optimizing our results and potentially achieving improved diagnostic outcomes.

References

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

[2] Marcinkevičs, R., Wolfertstetter, P.R.: Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. Medical Image Analysis **91**, 103042 (2024) https://doi.org/10.1016/j.media.2023.103042

[3] Akmese, O.F., Dogan, G., Kor, H., Erbay, H., Demir, E.: The use of machine learning approaches for the diagnosis of acute appendicitis. Emergency Medicine International **2020**, 73064535 (2020) https://doi.org/10.1155/2020/7306435

[4] Aggarwal, M.M.M.K.: A diagnostic testing for people with appendicitis using machine learning techniques. Multimed Tools Appl **81**, 7011–7023 (2022) https://doi.org/10.1007/s11042-022-11939-8

[5] Phan-Mai, T.-A., Thai, T.T., Mai, T.Q., Vu, K.A., Mai, C.C., Nguyen, D.A.: Validity of machine learning in detecting complicated appendicitis in a resource-limited setting: Findings from vietnam. Biomed Research International **2023**, 5013812 (2023) https://doi.org/10.1155/2023/5013812

[6] Pati, A., Panigrahi, A., Nayak, D.S.K., Sahoo, G., Singh, D.: Predicting pediatric

appendicitis using ensemble learning techniques. Procedia Computer Science **218**, 1166–1175 (2023) https://doi.org/10.1016/j.procs.2023.01.095

[7] Akbulut, S., Yagin, F.H., Cicek, I.B., Koc, C., Colak, C., Yilmaz, S.: Prediction of perforated and nonperforated acute appendicitis using machine learning-based explainable artificial intelligence. Diagnostics **13**, 1173 (2023) https://doi.org/10.3390/diagnostics13061173

[8] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution, 4470–4479 (2019) https://doi.org/10.1109/ICCV.2019.00457

[9] Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. Radiology **278**, 563–577 (2016) https://doi.org/10.1148/radiol.2015151169

[10] Chen, J.S., Huertas, A., Medioni, G.: Fast convolution with laplacian-of-gaussian masks. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-9**, 584–590 (1987) https://doi.org/10.1109/TPAMI.1987.4767946

[11] Othman, G., Zeebaree, D.Q.: The applications of discrete wavelet transform in image processing: A review. Journal of Soft Computing and Data Mining **1(2)**, 31–43 (2020) https://doi.org/https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/7215

[12] Karthikeyan, G.M., Manikandan, R., Gandomi, A.H.: Classification models combined with boruta feature selection for heart disease prediction. Informatics in Medicine Unlocked **44**, 101442 (2024) https://doi.org/10.1016/j.imu.2023.101442

[13] Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN Computer Science **2**, 160 (2021) https://doi.org/10.1007/s42979-021-00592-x

[14] Daunhawer, I., Kasser, S., Koch, G., Sieber, L., Cakal, H., Tütsch, J., al.: Enhanced early prediction of clinically relevant neonatal hyperbilirubinemia with machine learning. Pediatric Research **86**, 122–127 (2019) https://doi.org/10.1038/s41390-019-0384-x

[15] Alvarado, A.: A practical score for the early diagnosis of acute appendicitis. Annals of Emergency Medicine **15(5)**, 557–564 (1986) https://doi.org/10.1016/S0196-0644(86)80993-3

[16] Cheplygin, V.: Cats or cat scans: Transfer learning from natural or medical image source data sets? Current Opinion in Biomedical Engineering **9**, 21–27 (2019) https://doi.org/10.1016/j.cobme.2018.12.005

[17] Jeon, B.G., Kim, H.J., Heo, S.C.: Ct scan findings can predict the safety of delayed

appendectomy for acute appendicitis. Journal of Gastrointestinal Surgery **23(9)**, 1856–1866 (2019) https://doi.org/10.1007/s11605-018-3911-x

[18] Boonstra, P.A., Veen, R.N., Stockmann, H.B.A.C.: Less negative appendectomies due to imaging in patients with suspected appendicitis. Surgical Endoscopy **29**, 2365–2370 (2015) https://doi.org/10.1007/s00464-014-3963-2

[19] ADDISS, D.G., SHAFFER, N., FOWLER, B.S., TAUXE, R.V.: The epidemiology of appendicitis and appendectomy in the united states. American Journal of Epidemiology **132(5)**, 910–925 (1990) https://doi.org/10.1093/oxfordjournals.aje.a115734

[20] Wolff, R.F., Moons, K.G.M., Riley, R.D., Whiting, P.F., Westwood, M., al.: Probast: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine **170**, 51–58 (2019) https://doi.org/10.7326/M18-1376

[21] Malik, A., Paras, Pathania, Monika1, Rathaur, Kumar, V.: Overview of artificial intelligence in medicine. Journal of Family Medicine and Primary Care **8(7)**, 2328–2331 (2019) https://doi.org/10.4103/jfmpc.jfmpc_440_19

[22] Necas, Martin: The clinical ultrasound report: Guideline for sonographers. Australasian Journal of Ultrasound in Medicine **21**, 9–23 (2018) https://doi.org/10.1002/ajum.12075

[23] Castellano, G., Bonilha, L., Li, L.M., Cendes, F.: Texture analysis of medical images. Clinical Radiology **59(12)**, 1061–1069 (2004) https://doi.org/10.1016/j.crad.2004.07.008

[24] Chang, Y.-T., Lin, J.-Y., Huang, Y.-S.: Appendicitis in children younger than 3 years of age: An 18-year experience. The Kaohsiung Journal of Medical Sciences **22**, 432–436 (2006) https://doi.org/10.1016/S1607-551X(09)70334-1

[25] Horwitz, M.J.R., Gursoy, M.M., Jaksic, M.T., Lally, M.K.P.: Importance of diarrhea as a presenting symptom of appendicitis in very young children. The American Journal of Surgery **173(2)**, 80–82 (1997) https://doi.org/10.1016/S0002-9610(96)00417-5

[26] Avanesov, M., Wiese, N.J., Karul, M., Guerreiro, H., Keller, S., Busch, P., etal.: Diagnostic prediction of complicated appendicitis by combined clinical and radiological appendicitis severity index (apsi). European Radiology **28**, 3601–3610 (2018) https://doi.org/10.1007/s00330-018-5339-9

[27] Bonadio, W., Peloquin, P., Brazg, J., Scheinbach, I., Saunders, J., Okpalaji, C., al.: Appendicitis in preschool aged children: Regression analysis of factors associated with perforation outcome. Journal of Pediatric Surgery **50(9)**, 1569–1573 (2015) https://doi.org/10.1016/j.jpedsurg.2015.02.050

[28] Aggarwal, K., Bhamrah, M.S., Ryait, H.S.: Detection of cirrhosis through ultrasound imaging by intensity difference technique. EURASIP Journal on Image and Video Processing volume **2019**, 80 (2019) https://doi.org/10.1186/s13640-019-0482-z

[29] Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., al.: Appendixnet: Deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. Scientific Reports **10**, 3958 (2020) https://doi.org/10.1038/s41598-020-61055-6

[30] Marcinkevics, R., Wolfertstetter, P.R., Wellmann, S., Knorr, C., Vogt, J.E.: Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. Frontiers in Pediatrics **9** (2021) https://doi.org/10.3389/fped.2021.662183

[31] Akbulut, S., Bahçe, Z., Öztaş, T., Gümüş, S., Söğütçü, N., Sakarya, H., al.: Assessment of demographic, clinical and histopathological features of patients who underwent appendectomy due to a presumed diagnosis of acute appendicitis. Ulus Travma Acil Cerrahi Derg **27(3)**, 315–324 (2021) https://doi.org/10.14744/tjtes.2020.73537

[32] Prabhudesai, S.G., Gould, S., Rekhraj, S., Tekkis, P.P., Glazer, G., Ziprin, P.: Artificial neural networks: Useful aid in diagnosing acute appendicitis. World Journal of Surgery **32**, 305–309 (2007) https://doi.org/10.1007/s00268-007-9298-6