Statistical Learning Based Pediatric Appendicitis Prediction

Zitong Wang [*1] **Wen Yuan** [*1] **Jingxing Zhou** [*1]

Abstract

Pediatric appendicitis is one of the most common acute surgical emergencies in childhood, with lifetime risks of 8.6% for males and 6.7% for females. Early and accurate diagnosis is crucial for preventing complications. In this study, we develop machine learning models to predict appendicitis diagnosis, severity, and management decisions using the Regensburg Pediatric Appendicitis Dataset from UCI, comprising 782 patients with 58 features. We focus on a clinically practical scenario where ultrasound imaging is unavailable, using only 45 non-ultrasound features. We conduct comprehensive data preprocessing, exploratory data analysis, and systematic model evaluation including logistic regression, random forest, gradient boosting, and other ensemble methods. Our best basic models achieve test accuracies of 75.64% for diagnosis, 88.46% for management, and 89.10% for severity prediction. We also investigate various missing value imputation strategies (Median, KNN, MICE) and feature selection techniques to handle the inherent data incompleteness in clinical settings. We additionally evaluate advanced tree-based methods with native missing-value handling and find that most variants perform competitively and robustly across tasks. We also study LLM-based prediction under constrained prompting: few-shot in-context examples markedly improve results and can approach strong tree ensembles, whereas injecting LLM-derived signals into tree models does not yield consistent gains beyond the best tree baselines.

[*]Equal contribution [1]School of Mathematical Sciences, Peking University. Correspondence to: Zitong Wang <2300010750@stu.pku.edu.cn>, Wen Yuan <2200010833@stu.pku.edu.cn>, Jingxing Zhou <2300010749@stu.pku.edu.cn>.

Preprint. January 6, 2026.

1. Introduction

Pediatric appendicitis represents one of the most prevalent acute surgical emergencies encountered in childhood medicine (Addis et al., 1990). According to the National Institutes of Health (NIH), the lifetime risk of developing appendicitis is approximately 8.6% for males and 6.7% for females (Afridi et al., 2023). The condition poses significant diagnostic challenges, particularly in pediatric populations where clinical presentation may be atypical and communication with young patients is inherently difficult.

Early and accurate diagnosis of appendicitis is paramount for several reasons. Delayed diagnosis can lead to perforation, peritonitis, and other severe complications that significantly increase morbidity and mortality (Bhangu et al., 2015). Conversely, unnecessary surgical interventions carry their own risks and contribute to healthcare costs. This diagnostic dilemma motivates the development of computational tools that can assist clinicians in making more informed decisions.

In this work, we leverage the Regensburg Pediatric Appendicitis Dataset (Marcinkevičs et al., 2023), sourced from Children's Hospital St. Hedwig in Regensburg, Germany, and collected between 2016 and 2021. This comprehensive dataset contains records from 782 patients with 58 features spanning demographic information, clinical scoring systems, physical examination findings, laboratory tests, and ultrasound imaging results.

Importantly, our primary analysis focuses on a clinically practical scenario where ultrasound imaging is unavailable. Ultrasound, while highly informative for appendicitis diagnosis, is not universally accessible in all clinical settings, particularly in resource-limited environments or emergency situations. By excluding ultrasound-derived features (and the potentially leaky `Length_of_Stay` feature), we develop models using only 45 features after preprocessing that can be readily obtained from demographic data, physical examination, and laboratory tests. This approach ensures our models are applicable in a wider range of clinical contexts.

Our study addresses three prediction tasks of clinical relevance:

- **Diagnosis**: Determining whether a patient has appen-

dicitis

- **Severity**: Classifying appendicitis as complicated or uncomplicated
- **Management**: Predicting whether surgical intervention is required

We employ a systematic machine learning pipeline encompassing data preprocessing, exploratory data analysis, feature selection, model training, and hyperparameter optimization. A particular focus of our work is addressing the challenge of missing data, which is ubiquitous in clinical datasets due to varying clinical protocols and patient conditions. In order to address this issue, we first establish baseline models using median imputation for missing values. Then we make a comprehensive survey of advanced tree-based methods that natively handle missing data, as well as compare different imputation strategies including KNN and MICE (Multiple Imputation by Chained Equations) to evaluate their impact on model performance. Additionally, we investigate the usage of large language models (LLMs) for clinical prediction tasks in the context of appendicitis.

2. Data Preprocessing

2.1. Dataset Overview

The original dataset comprises 782 patient records with 58 features. The features can be categorized into several groups as shown in Table 1.

Table 1. Feature categories in the dataset

Category	Count	Examples
Demographic	6	Age, Sex, Height, Weight, BMI, Length of Stay
Clinical Scores	2	Alvarado Score, Paediatric Appendicitis Score
Clinical Exam	10	Migratory Pain, Lower Right Abdominal Pain, Nausea, Body Temperature
Peritonitis	3	Generalized, Local, No peritonitis
Stool	4	Constipation, Diarrhea, Normal
Laboratory	9	WBC Count, Neutrophil %, CRP, Hemoglobin
Urinalysis	12	Ketones, RBC, WBC in urine
Ultrasound	22	Appendix Diameter, Free Fluids, Perforation, etc.

2.2. Data Cleaning and Transformation

We performed the following preprocessing steps:

Missing Target Removal: Two samples with missing diagnosis labels were removed, resulting in 780 valid samples.

Feature Encoding: Binary categorical variables were transformed using label encoding (0/1), while multi-class categorical variables were converted using one-hot encoding. This process expanded the feature space from the original 58 columns to 78 features.

Dataset Variants: We created two dataset variants:

- **All Features**: 780 samples $\times$ 78 features, including ultrasound-derived measurements
- **No Ultrasound**: 780 samples $\times$ 45 features, excluding 22 ultrasound features and `Length_of_Stay` (identified as a potentially leaky feature since it is determined post-admission)

As emphasized in the introduction, our primary analysis uses the **No Ultrasound** dataset to ensure clinical applicability in settings where imaging is unavailable.

Train-Test Split: Data was partitioned into training (624 samples, 80%) and test (156 samples, 20%) sets using stratified sampling based on the Diagnosis variable to maintain class distribution balance.

2.3. Missing Data Characteristics

Clinical datasets inherently suffer from incompleteness due to varying examination protocols and patient conditions. Table 2 summarizes the missing data rates for key features.

Table 2. Missing data rates for selected features

Feature	Missing %	Category
Segmented Neutrophils	93.1%	Laboratory
Appendix Diameter	36.3%	Ultrasound
RBC in Urine	26.3%	Urinalysis
Ketones in Urine	25.6%	Urinalysis
WBC in Urine	25.4%	Urinalysis
Ipsilateral Rebound	20.8%	Clinical
Neutrophil Percentage	13.2%	Laboratory
Clinical Scores	6.6%	Scoring

The overall missing rates are 36.84% for the all-features dataset and 11.25% for the no-ultrasound dataset. Notably, some features with high predictive value for our targets also exhibit substantial missingness, presenting a fundamental challenge in maximizing data utilization. This motivates our detailed investigation of missing value imputation strategies in Section 5.

3. Exploratory Data Analysis

3.1. Target Variable Distributions

The three target variables exhibit different class distributions, as visualized in Figure 1. The targets are defined as follows:

- **Diagnosis**: Binary classification (appendicitis vs. no appendicitis)
- **Severity**: Binary classification (complicated vs. uncomplicated)
- **Management**: Three-class classification (conservative, primary surgical, secondary surgical)

Figure 1. Distribution of the three target variables: Diagnosis, Severity, and Management. The class imbalance varies across targets, with Severity showing the most pronounced imbalance.

Figure 2 provides a detailed view of the class balance for each target, which is crucial for selecting appropriate evaluation metrics.

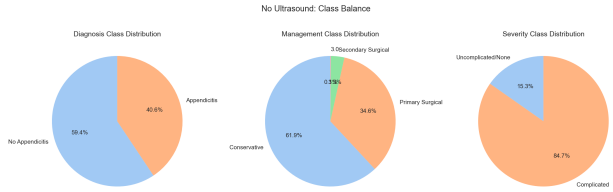

Figure 2. Class balance pie charts for each target variable, showing the proportion of positive and negative cases.

3.2. Feature Correlation Analysis

We conducted comprehensive correlation analysis to understand relationships between features and target variables. Figure 3 shows the correlation heatmap for key features. Observing this heatmap, we can first group the features into three classes: namely basic physical indicators(e.g. Age, BMI), commonly used laboratory markers and clinical scores(e.g. WBC Count, CRP, Alvarado Score) and target variables. The top correlation regions can be identified as the intra-group correlations, and the inter-group correlations between laboratory markers/clinical scores and target variables. This aligns well with clinical understanding, as inflammatory markers and clinical scores are primary diagnostic tools for appendicitis, while basic physical indicators, correlating with themselves, have limited direct correlation with disease status.

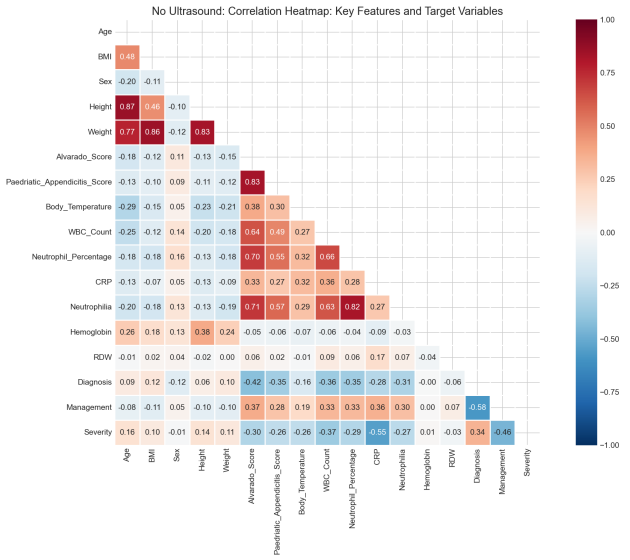

Figure 3. Correlation heatmap showing relationships between features and target variables. Strong correlations are observed between inflammatory markers and between clinical scores.

The analysis revealed several clinically meaningful patterns, as shown in Figure 4:

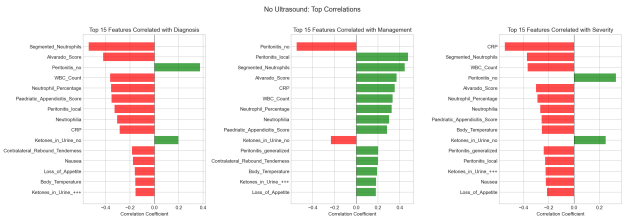

Figure 4. Top 15 features most correlated with each target variable. The correlations align well with established clinical knowledge.

Diagnosis Correlations: The top features correlated with diagnosis include Segmented Neutrophils, Alvarado Score, absence of Peritonitis, and WBC Count. These findings align well with established clinical knowledge, as inflammatory markers and clinical scoring systems are primary diagnostic tools.

Management Correlations: Management decisions show strong correlations with Peritonitis indicators (both local and generalized) and Segmented Neutrophils, reflecting that surgical intervention is often indicated in cases with peritoneal involvement.

Severity Correlations: Severity classification correlates most strongly with CRP (C-Reactive Protein), Segmented Neutrophils, and WBC Count, consistent with the understanding that complicated appendicitis is associated with heightened inflammatory responses.

3.3. Feature Distribution by Target

To understand how features differ between classes, we examined box plots of key features grouped by each target. Figure 5, Figure 6, and Figure 7 show the distribution of numerical features by diagnosis, severity, and management status respectively. Furthermore, The scatter plot matrix of key features against themselves and colored by diagnosis status is provided in Figure 8.

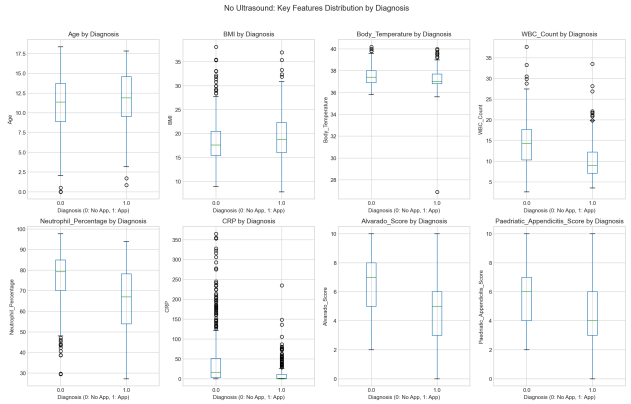

Figure 5. Box plots showing distribution of key features by Diagnosis status. Notable differences are observed in inflammatory markers such as WBC Count and CRP.

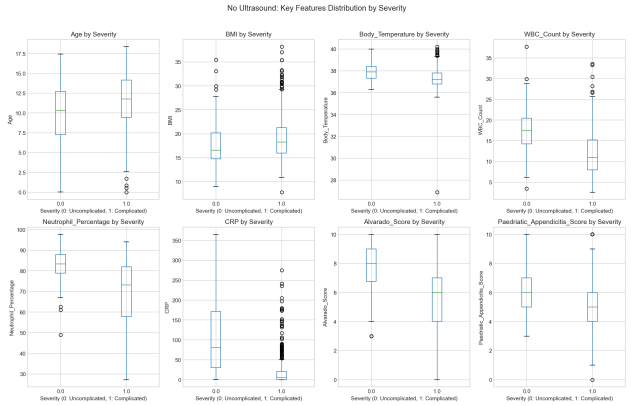

Figure 6. Box plots showing distribution of key features by Severity status.

4. Basic Models

In this section, we present our baseline machine learning models, feature selection methodology, and hyperparameter tuning procedures. All experiments in this section use the No Ultrasound feature set (45 features) with median imputation as the default missing value strategy.

4.1. Model Selection and Formulation

We evaluated seven classification algorithms representing different modeling paradigms:

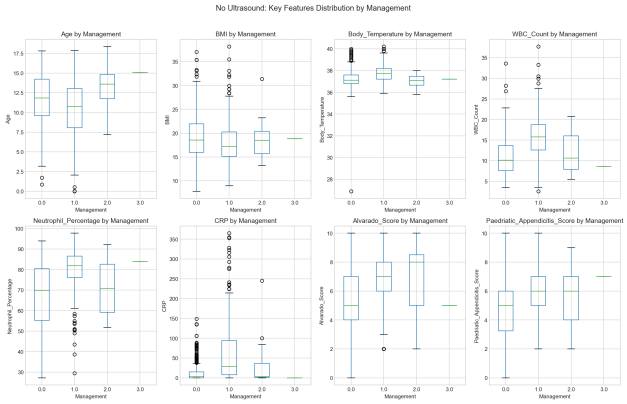

Figure 7. Box plots showing distribution of key features by Management status.

Figure 8. Scatter plot matrix of key features colored by Diagnosis status.

- **Logistic Regression**: A linear model that estimates the probability of class membership using the logistic function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \tag{1}$$

  We use L2 regularization with regularization parameter C.

- **Decision Tree**: A non-parametric model that recursively partitions the feature space based on information gain or Gini impurity.

- **Random Forest** An ensemble of B decision trees, each trained on a bootstrap sample with random feature subsets. Predictions are made by majority voting:

$$\hat{y} = \text{mode}\{h_b(x)\}_{b=1}^{B} \tag{2}$$

- **Gradient Boosting** A sequential ensemble that fits trees to the negative gradient of the loss function. At each iteration m:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{3}$$

where h_m is fitted to the residuals $r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$.

- **HistGradientBoosting**: A variant of gradient boosting that discretizes continuous features into histograms, enabling faster training and **native support for missing values**.

- **Support Vector Machine (SVM)**: A kernel-based classifier using the RBF kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{4}$$

- **K-Nearest Neighbors (KNN)**: An instance-based method that classifies based on the k nearest training samples.

Models that do not natively handle missing values (all except HistGradientBoosting) were paired with median imputation in a scikit-learn pipeline.

4.2. Experimental Results

Table 3 summarizes the best performing models for each target variable on the no-ultrasound feature set.

Table 3. Best model performance on test set (No Ultrasound). Diag. stands for Diagnosis, Man. for Management, Sev. for Severity.

Target	Best Model	Accuracy	F1-Score
Diag.	HistGradientBoosting	75.64%	0.6984
Man.	HistGradientBoosting	88.46%	0.8744
Sev.	Random Forest	89.10%	0.9377

The results demonstrate that severity and management prediction are considerably easier tasks than diagnosis, achieving nearly 90% accuracy. The diagnosis task proves more challenging, likely due to the subtlety of distinguishing appendicitis from other conditions with similar presentations. The detailed results for all models across targets can be found in Appendix A.1.

4.2.1. ROC Curve Analysis

Figure 9 shows the ROC curves for all models on each target variable, providing insight into the trade-off between sensitivity and specificity.

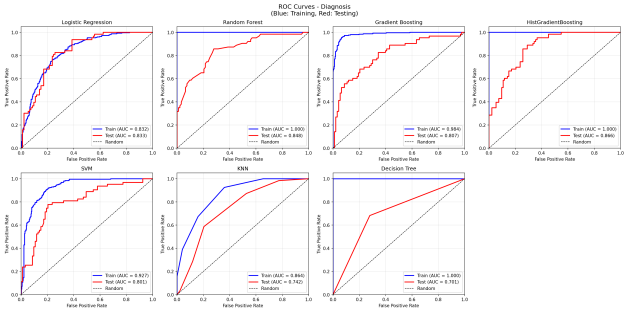

Figure 9. ROC curves for all models on the Diagnosis target. HistGradientBoosting and Random Forest achieve the highest AUC values.

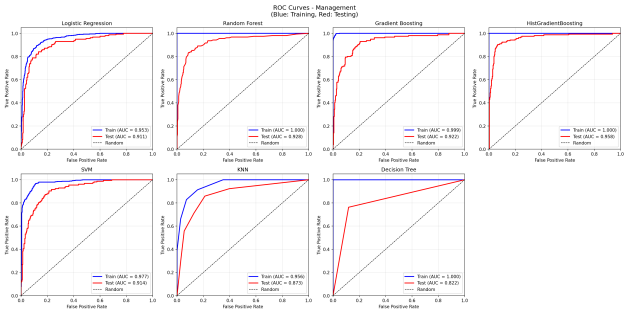

Figure 10. ROC curves for all models on the Management target.

4.2.2. Confusion Matrix Analysis

The confusion matrices in Figure 12 reveal the types of errors made by our best models. However, it is worth noting that due to class imbalance, accuracy alone may not fully capture model performance. For instance, in the Severity task, the model achieves high accuracy partly due to the predominance of uncomplicated cases.

4.2.3. Overfitting Analysis

We monitored the gap between training and test performance to detect overfitting. Table 4 summarizes the overfitting analysis for the Diagnosis target.

While tree-based models show significant overfitting on

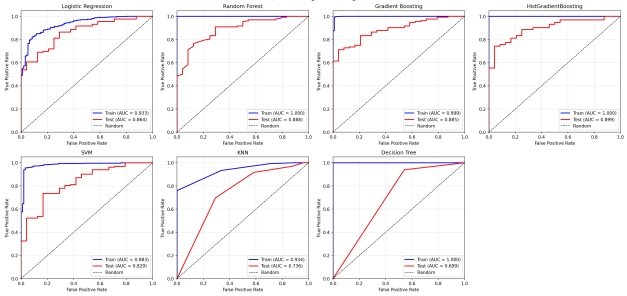

Figure 11. ROC curves for all models on the Severity target.

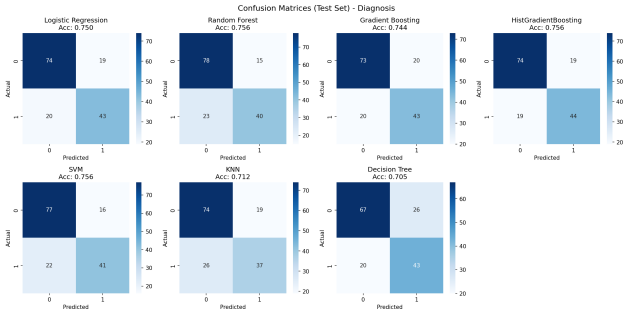

Figure 12. Confusion matrix for the best model on Diagnosis prediction.

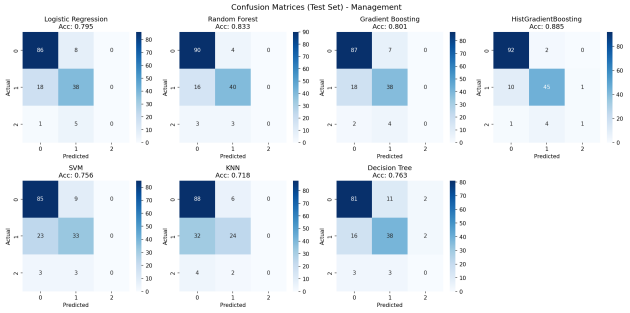

Figure 13. Confusion matrix for the best model on Management prediction.

training data, they still achieve the best test performance, suggesting that their capacity to capture complex patterns outweighs the overfitting risk for this dataset size.

4.3. Feature Selection

To reduce model complexity and potentially improve generalization, we implemented a two-stage stepwise feature selection procedure:

Stage 1 - Forward Selection: Starting from an empty feature set, we iteratively added the feature that most improved cross-validation performance, stopping at 15 features.

Stage 2 - Backward Elimination: From the 15 forward-selected features, we iteratively removed the least important features, arriving at a final set of 10 features per target. Table 5 presents the selected features for each target variable. The full selected feature sets are provided in Appendix A.2.

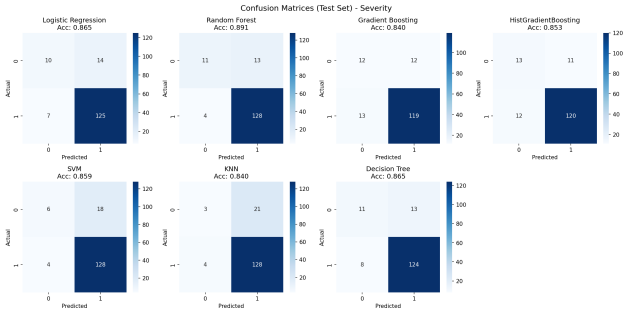

Figure 14. Confusion matrix for the best model on Severity prediction.

Table 4. Train vs Test comparison for Diagnosis (Overfitting Analysis)

Model	Train Acc	Test Acc	Gap
Logistic Reg.	0.753	0.750	+0.003
Random Forest	1.000	0.756	+0.244
Gradient Boost	0.938	0.744	+0.194
HistGradBoost	1.000	0.756	+0.244
SVM	0.853	0.756	+0.096
KNN	0.772	0.712	+0.061
Decision Tree	1.000	0.705	+0.295

Table 5. Selected features after forward-backward selection

Target	Selected Features (10)
Diagnosis	BMI, Sex, Lower Right Abd Pain, Loss of Appetite, WBC Count, RBC Count, Ketones in Urine (+, ++), Peritonitis (generalized, no)
Management	Age, BMI, Coughing Pain, Ketones in Urine (++, +++, no), WBC in Urine (no), CRP, Peritonitis (no), Psoas Sign
Severity	BMI, Sex, Contralateral Rebound Tenderness, Nausea, Segmented Neutrophils, Ketones in Urine (++, +++, no), RBC in Urine (+), CRP

4.3.1. Impact of Feature Selection

Comparing models trained on full features versus selected features revealed mixed results. Figure 15 and Figure 16 show the ROC curves for diagnosis models trained on selected features and full features respectively. One can observe that the performance remains comparable, indicating that the selected features capture most of the predictive signal, while significantly reducing dimensionality.

While feature selection improved interpretability and reduced computational costs, the performance gains were marginal and sometimes negative, suggesting that the gradient boosting methods effectively perform implicit feature selection.

Table 6 provides a detailed comparison of test accuracy with and without feature selection across all models and targets.

The results reveal that feature selection generally degrades performance for Diagnosis and Management tasks, with the exception of Logistic Regression on Diagnosis

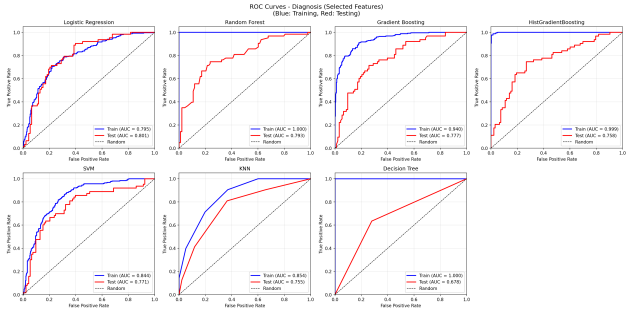

Figure 15. ROC curves for models trained on selected features (Diagnosis). Performance is comparable to full-feature models.

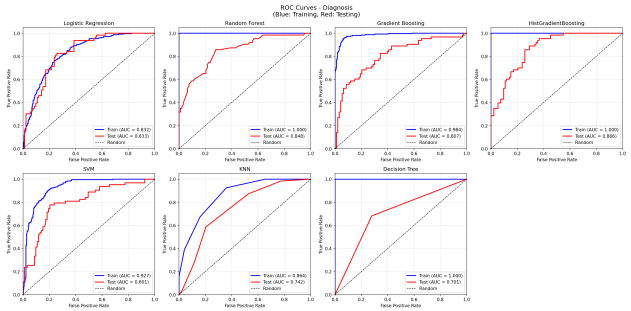

Figure 16. ROC curves for models trained on full features (Diagnosis).

(+0.64%). The Management task shows particularly large drops for HistGradientBoosting (−13.46%) and Decision Tree (−8.97%). Interestingly, the Severity task exhibits more positive outcomes, with Logistic Regression, Gradient Boosting, SVM, and KNN all benefiting from feature selection. This suggests that Severity prediction may rely on a smaller, more focused set of predictive features, while Diagnosis and Management require the full feature space to capture subtle discriminative patterns.

4.4. Hyperparameter Tuning

We performed grid search with 5-fold cross-validation to optimize key hyperparameters for each model. The hyperparameter search spaces were:

- **Random Forest**: $N_{\text{estimators}}$ taking values from 10 to 300.
- **Gradient Boosting**: $N_{\text{estimators}}$ taking values from 25 to 200.
- **KNN**: $N_{\text{neighbors}}$ taking values from 1 to 21.
- **SVM**: C taking values from 0.01 to 10.
- **Decision Tree**: `max_depth` taking values from 2 to 15.

Table 6. Impact of feature selection on test accuracy. Diff is the difference between selected and full feature sets, expressed as a percentage point change.

Target	Model	Full	Selected	Diff
Diag.	Logistic Reg.	75.00	75.64	+0.64
	Random Forest	75.64	75.64	0.00
	Gradient Boost	74.36	71.79	−2.56
	HistGradBoost	75.64	74.36	−1.28
	SVM	75.64	72.44	−3.21
	KNN	71.15	69.87	−1.28
	Decision Tree	70.51	68.59	−1.92
Man.	Logistic Reg.	79.49	78.21	−1.28
	Random Forest	83.33	80.13	−3.21
	Gradient Boost	80.13	78.85	−1.28
	HistGradBoost	88.46	75.00	−13.46
	SVM	75.64	74.36	−1.28
	KNN	71.79	71.15	−0.64
	Decision Tree	76.28	67.31	−8.97
Sev.	Logistic Reg.	86.54	89.10	+2.56
	Random Forest	89.10	85.90	−3.21
	Gradient Boost	83.97	86.54	+2.56
	HistGradBoost	85.26	80.13	−5.13
	SVM	85.90	87.82	+1.92
	KNN	83.97	87.18	+3.21
	Decision Tree	86.54	78.21	−8.33

- **Logistic Regression**: C taking values from 0.001 to 100.

Figure 17, Figure 18, and Figure 19 visualize the grid search results for each target.

5. Handling Missing Values

The integrity of clinical datasets is frequently compromised by missing values, which can introduce substantial bias and reduce statistical power if not addressed rigorously. In our dataset, the missingness rate is approximately 11.25% for the non-ultrasound feature set. Since most standard additive models and boosting algorithms (e.g., standard Gradient Boosting, Logistic Regression) require complete input matrices, the choice of imputation strategy becomes a critical hyperparameter in our modeling pipeline.

In this study, we initially hypothesized that the missing data mechanism might be **Missing at Random (MAR)**. However, to validate this and inform our choice of strategy, we first conducted a visual diagnosis of the missingness patterns.

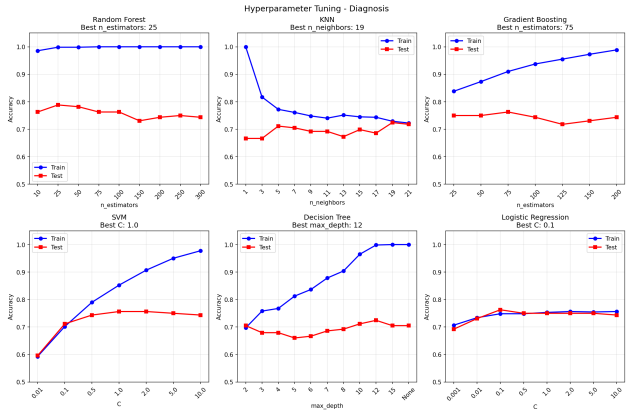

Figure 17. Grid search results for Diagnosis, showing training and validation performance across different hyperparameter values for each model.

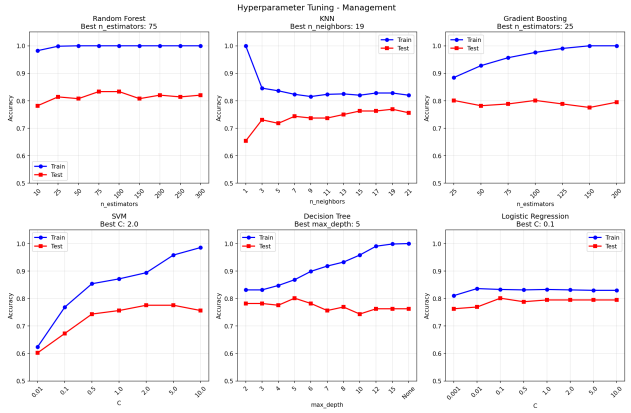

Figure 18. Grid search results for Management, showing training and validation performance across different hyperparameter values for each model.

5.1. Exploratory Analysis and Imputation Strategies

Prior to applying specific imputation algorithms, we utilized visual tools to characterize the nature of the missing data.

Figure 20 presents a vertical heatmap of missingness. The distinct "yellow bands" across the lower features indicate that missingness is structurally clustered rather than sporadic; entire panels of tests (likely secondary markers) are often skipped together for specific patient cohorts.

Furthermore, Figure 21 investigates the **Missing Not at Random (MNAR)** hypothesis by correlating the missingness indicators with clinical targets. Strong correlations (deep red/blue) are observed; for instance, the absence of *Ipsilateral Rebound Tenderness* is negatively correlated with a positive diagnosis. This implies an **informative missingness** mechanism—the very act of not performing a test contains diagnostic signal—which poses a theoretical challenge to imputation methods like MICE that assume MAR.

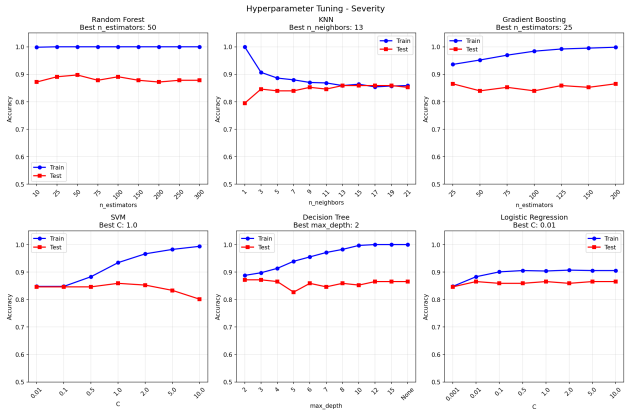

Figure 19. Grid search results for Severity, showing training and validation performance across different hyperparameter values for each model.

Given these observations, we systematically compared three imputation approaches ranging from univariate heuristics to multivariate iterative modeling to determine which best handles this complex mechanism.

5.1.1. Median Imputation (Baseline)

As a baseline, we employ univariate median imputation. This method replaces every missing entry with the median of the observed values for that feature:

$$\hat{x}_{ij}^{\text{miss}} = \text{median}(\{x_{kj} : k \in \mathcal{I}_{obs}(j)\}) \tag{5}$$

where $\mathcal{I}_{obs}(j)$ represents the set of indices for which feature j is observed. While this approach is $O(1)$ in computational complexity and robust to outliers, it is statistically flawed for high-precision modeling:

- **Underestimation of Variance**: By replacing a distribution of missing values with a single point estimate (the median), we artificially reduce the feature's variance. This can lead to narrower confidence intervals and inflated Type I errors (false positives) in subsequent statistical tests.
- **Distortion of Covariance**: Since imputation is performed independently for each feature, the method ignores the inherent correlation structure between variables (e.g., the physiological correlation between white blood cell count and neutrophils). This "decoupling" effect hampers the ability of multivariate models (like GAMs) to learn interaction terms.
- **Bias under MAR**: Median imputation is only unbiased under the strict **Missing Completely at Random (MCAR)** assumption, which rarely holds in clinical data.

Figure 20. **Vertical Heatmap of Missingness.** The y-axis represents clinical features, and the x-axis represents patient samples. Yellow bands indicate missing values. The structural "blocking" suggests that specific test panels are systematically skipped.

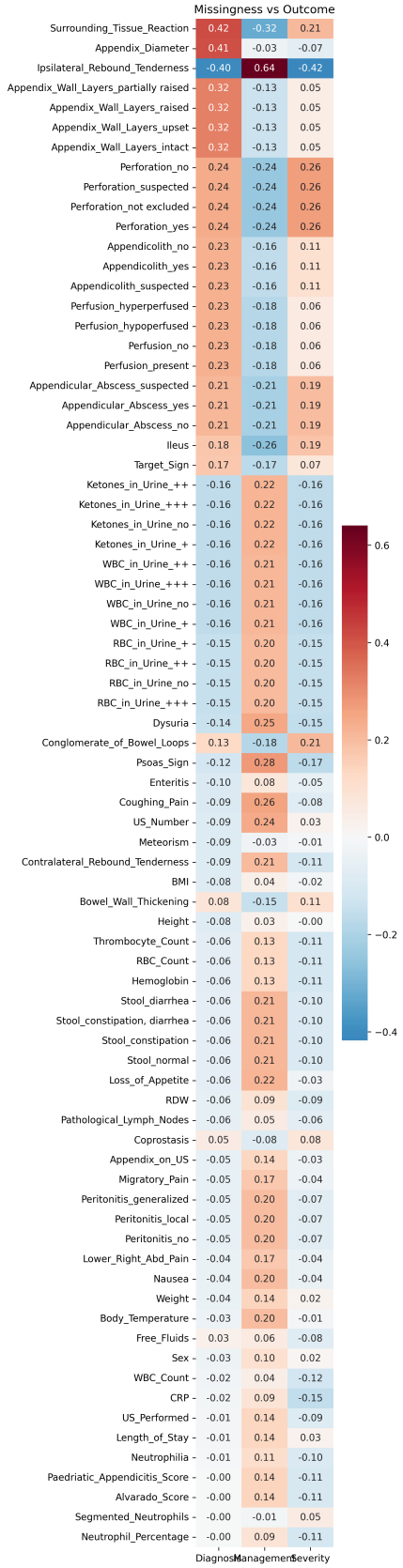

Figure 21. **Correlation between Feature Missingness and Clinical Outcomes.** Red/Blue colors indicate strong correlations between the *absence* of a feature and the target variable. Significant correlations suggest the data mechanism is MNAR rather than MCAR/MAR.

5.1.2. K-NEAREST NEIGHBORS (KNN) IMPUTATION

To preserve local data structures, we utilize KNN imputation, a non-parametric method. For a target sample i with missing values, the algorithm identifies the set of k nearest neighbors, $N_k(i)$, based on the Euclidean distance computed from the observed features. The missing value is then estimated as a weighted average:

$$\hat{x}_{ij}^{\text{miss}} = \frac{\sum_{l \in N_k(i)} w_{il} x_{lj}}{\sum_{l \in N_k(i)} w_{il}} \tag{6}$$

where weights w_{il} are typically proportional to the inverse distance ($1/d(x_i, x_l)$). **Advantages and Limitations:**

- **Local Structure Preservation**: Unlike median imputation, KNN considers the local density of the data, making it capable of capturing non-linear patterns within clusters.
- **Computational Cost**: The algorithm is computationally intensive, scaling with $O(N^2)$ as it requires computing a pairwise distance matrix. This becomes a bottleneck for large datasets.
- **Curse of Dimensionality**: In high-dimensional feature spaces, the distance between all points tends to converge, rendering the concept of "nearest neighbor" less meaningful.
- **Sensitivity to Scaling**: KNN is highly sensitive to feature magnitudes. Pre-processing steps such as standardization (Z-score normalization) are strictly required before applying this method.

5.1.3. MICE (MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS)

Given the limitations of the previous methods, we adopt **MICE** (Van Buuren & Groothuis-Oudshoorn, 2011) as our primary imputation strategy. MICE, also known as Fully Conditional Specification (FCS), avoids defining a rigid joint distribution for the entire dataset. Instead, it decomposes the multivariate missing data problem into a series of univariate regression problems. The algorithm operates in a Round-Robin fashion:

1. **Initialization**: All missing values are initially filled with a simple placeholder (e.g., mean) to allow the regression models to run.
2. **Iteration (Gibbs Sampling Step)**: For each cycle $t = 1, \ldots, T$ and for each variable j:
   - The current variable x_j is set as the target, and all other variables x_{-j} (containing their most current imputed values) are used as predictors.
   - A regression model (e.g., **Bayesian Ridge Regression**) is trained on the observed part of x_j:

$$\phi_j^{(t)} \leftarrow \text{train}(x_j^{\text{obs}} \sim x_{-j}^{(t)}) \tag{7}$$

   - New values for the missing parts of x_j are drawn from the posterior predictive distribution:

$$x_j^{\text{miss}(t)} \sim P(x_j | x_{-j}^{(t)}, \phi_j^{(t)}) \tag{8}$$

   This cycle is repeated until the distribution of imputed parameters stabilizes (convergence).
3. **Uncertainty Modeling**: Unlike single deterministic imputation, MICE can generate multiple datasets to account for the uncertainty of the missing values themselves. For our predictive modeling task, we utilize the single optimal imputation derived from the converged state.

Justification for MICE in Additive Models:

- **Preservation of Multivariate Relations**: By using other variables as predictors, MICE effectively reconstructs the correlation structure of the data. For example, if 'Body Temperature' is missing, MICE can infer it from 'Infection Markers', preserving the physiological consistency required for the Additive Model to function correctly.
- **Flexibility**: The underlying regressor in the chained equations can be adapted. In this study, we employ **Bayesian Ridge Regression**, which introduces regularization to handle potential multicollinearity among clinical features.

5.2. Comparison and Discussion

Table 7 compares imputation strategies across different models and targets.

Table 7. Imputation strategy comparison (Test Accuracy). *Diag.* stands for Diagnosis, *Sev.* stands for Severity.

Target	Model	Median	KNN	MICE	Median(US)
Diag.	Logistic Reg.	0.769	0.788	0.788	0.923
	Gradient Boost	0.795	0.782	0.795	0.948
	HistGradBoost	0.795	0.795	0.821	0.980
Sev.	Logistic Reg.	0.936	0.872	0.885	0.916
	Gradient Boost	0.936	0.929	0.923	0.961
	HistGradBoost	0.942	0.936	0.929	0.942

Several key observations emerge from our imputation comparison:

- **Heterogeneity in Imputation Efficacy**: The optimal imputation strategy proves to be highly target-dependent. For the *Diagnosis* task, multivariate iterative imputation (MICE) combined with HistGradientBoosting achieves superior performance (82.1%), whereas for *Severity*, simple median imputation yields the highest accuracy (94.2%).
- **Resilience of Native Handling**: HistGradientBoosting demonstrates remarkable robustness. Unlike Logistic Regression, which fluctuates significantly depending on the imputation quality, HistGradientBoosting's native ability to handle missing values renders it less sensitive to the choice of pre-imputation, suggesting that sophisticated external imputation may yield diminishing returns for modern tree-based models.
- **Potential Violation of MAR Assumption**: The limited marginal gain from MICE suggests that the missingness may be partly *Missing Not At Random (MNAR)*. In clinical settings, the absence of a test is often informative (e.g., tests skipped for less severe cases), containing a signal that simple imputation or native handling might implicitly preserve better than MICE.
- **Feature Dominance vs. Imputation Utility**: The marginal performance gains observed across imputation methods (Columns 1-3) can be attributed to the **weak predictive power** of the features containing missing values (primarily secondary lab markers). In sharp contrast, the fourth column (Median(US)) demonstrates that simply including the "Appendix Diameter" feature—even with crude median imputation—triggers a massive accuracy surge (e.g., to 98%). This indicates that the **dominant predictive signal** of the ultrasound feature outweighs the benefits of sophisticated imputation strategies applied to weaker clinical features. Essentially, the "informativeness" of the available features matters far more than the recovery of missing secondary data.

6. Advanced Methods

Building upon our basic models, we explore advanced techniques including sophisticated tree-based ensembles and large language model (LLM) approaches.

6.1. Advanced Tree-Based Methods

Beyond the basic models in Section 4, we further evaluate a suite of advanced tree-based methods on the **No Ultrasound** dataset (45 features) under the same fixed train/test split (80/20, stratified by Diagnosis). Unless otherwise stated, models that do not natively support missing values are trained with **median imputation** applied to the training set and then consistently applied to the test set.

Methods. We consider the following advanced tree-based algorithms and ensemble variants:

- **Random Forest (RF).** An ensemble of bagged decision trees trained on bootstrap samples with random feature sub-sampling. RF is robust and nonlinear, but can overfit on small datasets if trees are deep. In our implementation, missing values are handled via median imputation.
- **Gradient Boosting Decision Trees (GBDT).** A sequential additive model that fits weak learners (shallow trees) to residuals. Compared with RF, GBDT often yields stronger performance but can be sensitive to hyperparameters. Missing values are handled via median imputation.
- **AdaBoost.** Adaptive boosting that reweights samples across iterations to focus on misclassified points, typically using decision stumps or shallow trees as base learners. AdaBoost can be competitive on tabular data but may be sensitive to noisy labels. Missing values are handled via median imputation.
- **Histogram-based Gradient Boosting Trees (HistGBDT).** A scalable GBDT variant that bins continuous features into discrete histograms. Importantly, HistGBDT provides **native handling of missing values** by learning default directions for missing-feature splits, which makes it well-suited to clinical datasets with incomplete measurements.
- **LightGBM.** A histogram-based gradient boosting framework with efficient split finding and leaf-wise tree growth. LightGBM supports **native missing-value handling** by assigning missing values to the optimal split direction during training. We use LightGBM as a strong industrial-grade GBDT baseline.
- **CatBoost.** Gradient boosting with symmetric (oblivious) trees and specialized handling of categorical variables. CatBoost also supports **native missing-value handling**. Although our features are already encoded, CatBoost remains a strong contender due to its stable training and regularization effects.
- **Neural Decision Forest (NDF).** A hybrid model combining a neural feature transformation with differentiable decision trees/forests trained end-to-end. NDF aims to capture feature interactions beyond standard trees while maintaining tree-like decision structures. Missing values are handled via the same preprocessing pipeline used for non-native models (median imputation).
- **Tree Ensemble Layer (TEL).** A stacking-style meta-model that aggregates predictions from multiple tree

ensembles using a learned combination layer. Concretely, we train several base tree models to obtain out-of-fold (here: training-set) predicted probabilities/logits, then fit a lightweight combiner (e.g., logistic regression or a shallow tree) on these meta-features. TEL can be viewed as an *ensemble-of-ensembles* strategy to improve robustness.

- **TreeNet.** A two-stage refinement approach: we first train a base tree ensemble (e.g., RF/GBDT) to obtain prediction scores; then we train a second model on the augmented feature vector $(x, \hat{p}(y|x))$ (or concatenated class scores) to refine decisions. This resembles a simple stacking pipeline designed to exploit complementary signal between raw clinical features and a strong tree baseline.

- **MorphBoost.** A heterogeneous-depth boosting ensemble that combines boosting models of different tree depths (e.g., shallow and deeper trees) through weighted averaging or voting. The intuition is to blend low-variance shallow learners with higher-capacity learners to balance bias and variance, which can be beneficial under limited sample size.

- **Meta Tree Boosting.** A meta-ensemble that performs voting/averaging across multiple tree-based predictors (e.g., RF, GBDT, LightGBM, CatBoost, HistGBDT). By aggregating diverse inductive biases, Meta Tree Boosting aims to reduce variance and improve generalization under a single fixed split.

Implementation notes. All models are trained on the same fixed split for fair comparison.

Results. Table 8, 9, 10 reports all exact final results.

6.2. LLM-Based Approaches

We investigate whether Large Language Models (LLMs) can assist clinical prediction on the Regensburg Pediatric Appendicitis Dataset under the **No Ultrasound** setting (45 features). Our goal is *not* to replace statistical learners with free-form text generation, but to evaluate whether LLMs can (i) directly predict clinical outcomes from structured patient features and/or (ii) provide useful high-level clinical abstractions as additional features for downstream models. Total view is in Figure 22.

LLM models. We evaluate two representative LLM families: (1) a GPT-style model (denoted as **GPT**) and (2) **DeepSeek**. We keep temperature and decoding settings fixed across experiments to ensure comparability.

Input formatting. Each patient record is serialized into a compact text block listing all available feature names and values. Missing values are explicitly represented as `Unknown` (or `Not Measured`) rather than being imputed. For categorical features, we present human-readable values (e.g., `Peritonitis = none/local/generalized`) instead of one-hot vectors. This representation is designed to mimic how clinicians read structured triage forms. The prompts are in B.

Evaluation protocol. All LLM experiments are conducted on the same fixed train/test split used in the tree-based methods. For Diagnosis and Severity (binary), we report Accuracy and F1-score of the positive class; for Management (multi-class), we report Accuracy and macro-F1. Unless otherwise specified, LLM predictions are made independently for each test sample without access to true labels.

6.2.1. Direct Prediction

We first test whether LLMs can directly predict clinical outcomes based solely on structured patient features.

- **Direct Ask (Zero-shot).** The LLM receives the patient features and is asked to predict Diagnosis/Severity/Management without examples.

- **Few-shot with Mean Examples.** We prepend several representative examples ("mean" cases summarizing typical positive/negative patterns) to provide context and calibrate the decision boundary.

6.2.2. Iterative Refinement and Hybrid Strategies

We then consider strategies intended to reduce systematic LLM errors:

- **Refine by Error Feedback.** After an initial prediction, we provide the LLM with a short summary of its incorrect predictions on a small held-out calibration subset from the training data (not the test labels) and ask it to update its decision rule and re-predict. This aims to correct recurring mistakes (e.g., over-reliance on a single lab marker).

- **Give Base Model Results.** We provide the LLM with the probability outputs or predicted labels from a strong baseline model (e.g., HistGradientBoosting) as additional evidence. The LLM is instructed to either agree with the base model or override it with an explicit justification (still outputting only final labels).

- **Only Solve Difficult Cases.** We use the base model as the default predictor. The LLM is queried only for

samples near the decision boundary (e.g., low confidence or small probability margin), and its prediction replaces the base model on these difficult cases. This reduces cost and focuses LLM reasoning where it may help most.

6.2.3. LLM AS FEATURE ENGINEER

Rather than asking the LLM to directly output the ground-truth task labels, we leverage its clinical prior knowledge to generate auxiliary labels/features that can be appended to the tabular representation. Concretely, for each patient we ask the LLM to produce three binary indicators: `y_diagnosis`, `y_severity`, `y_management`, *based only on observed features and clinical knowledge, without using any true labels*. These LLM-generated signals are then used as additional features for downstream models (e.g., decision trees, TEL), functioning as high-level clinical abstractions.

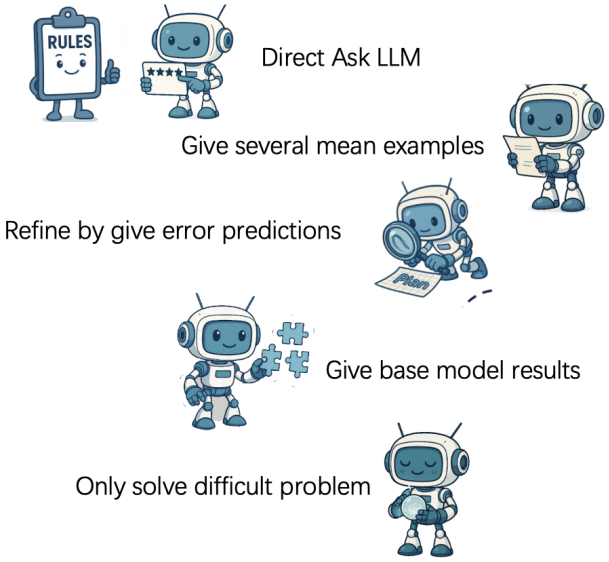

Figure 22. Schematic View for LLM in Modeling

Results. Table 8, 9, 10 reports all exact final results.

Discussion. Overall, direct LLM prediction is consistently weaker than strong tree-based learners trained on the same split, especially on the Diagnosis task. As shown in Figure 23, the efficiency of the LLM can be improved to a level close to that of tree-based models simply by providing a few in-context examples, which is a striking result.

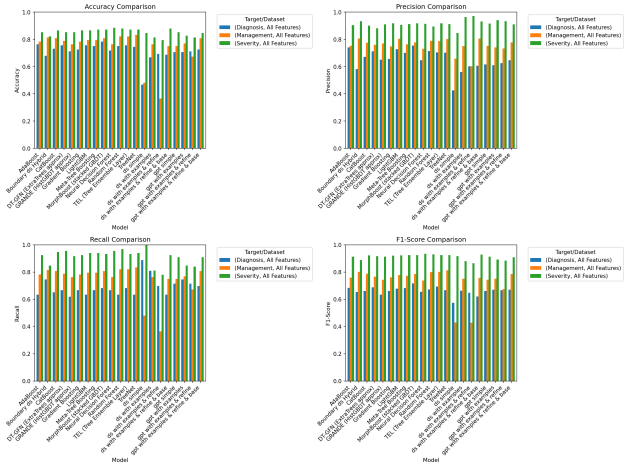

Figure 23. Tree Model vs. LLM

7. Conclusions and Future Work

7.1. Summary

In this study, we developed and evaluated machine learning models for pediatric appendicitis prediction using the Regensburg Pediatric Appendicitis Dataset, with a focus on clinically practical scenarios where ultrasound imaging is unavailable. Our key findings include:

1. **Model Performance**: Using only non-ultrasound features (45 features), HistGradientBoosting and Random Forest emerged as the top performers among basic models, achieving test accuracies of 75.64% (Diagnosis), 88.46% (Management), and 89.10% (Severity).

2. **Clinical Relevance**: The most predictive features align with established clinical knowledge, including inflammatory markers (WBC, CRP, Neutrophils), clinical scores (Alvarado, PAS), and physical examination findings (Peritonitis signs).

3. **Missing Data Handling**: While MICE provides a principled approach to imputation, simpler methods like median imputation can perform comparably when paired with robust models like HistGradientBoosting that handle feature interactions effectively. The optimal imputation strategy is task-dependent.

4. **Feature Selection**: Stepwise selection identified compact feature sets (10 features per target) that maintain competitive performance while improving interpretability.

5. **Clinical Applicability**: By focusing exclusively on non-ultrasound features, our models remain applicable in clinical settings where imaging is unavailable, such as resource-limited environments or time-critical triage situations.

6. **Advanced Tree-Based Methods**: Across the advanced tree family (e.g., gradient-boosted decision trees and modern ensemble variants), performance remained consistently strong and competitive across all three targets, indicating that tree-based learners provide a robust and reliable baseline for this dataset under missingness and mixed feature types.

7. **LLM Competitiveness with In-Context Examples**: Although direct LLM-based prediction underperforms well-tuned tree models in the zero-shot setting, providing a small number of in-context examples substantially narrows the gap, bringing the LLM to near tree-level performance. This suggests that limited task-specific guidance can dramatically improve LLM decision quality even without parameter updates.

8. **Hybridization Limits**: We explored hybrid strategies that inject LLM judgments into tree models (e.g., using LLM outputs as additional signals or refinements). However, these approaches did not yield consistent improvements over the best standalone tree models, suggesting that the tree ensembles already capture most of the predictive structure available in the tabular features, and that LLM-derived signals may be redundant or noisy in this setting.

7.2. Future Directions

Future work could continue to explore enhancing the predictive performance and utilization of machine learning models for pediatric appendicitis, especially in high data missingness scenarios. One can also continue to explore more explainable models and interpretability techniques to facilitate clinical adoption. We have not been able to fully explore the LLM-based methods, and this remains a promising avenue for future research. One can continue to refine the prompt engineering strategies, experiment with extensive LLM providers and architectures, and investigate more sophisticated methods of combining LLMs with traditional ML approaches.

Contributions

The detailed contribution of each team member is as follows:

- **Zitong Wang**: Data preprocessing, exploratory data analysis, implementation and experiment of basic models, and the general write-up of the framework. Specificly, responsible for Section 1, Section 2, Section 3 and Section 4.

- **Wen Yuan**: Implementation of advanced tree-based methods and LLM-based approaches. Specifically Section 6.

- **Jingxing Zhou**: Exploratory Data Analysis, comparative assessment of missing data imputation strategies, and the composition of Chapter 5

Software and Data

The analysis and training of models were conducted using Python with mainly scikit-learn (Pedregosa et al., 2011). The Regensburg Pediatric Appendicitis Dataset is publicly available through the UCI Machine Learning Repository (Marcinkevičs et al., 2023). Code and processed data are available in the supplementary materials.

Acknowledgements

We thank the course instructor Ruibin Xi and teaching assistant Siyi Wu of the 2025 Fall Statistical Learning course at Peking University for their guidance. We also acknowledge the creators of the Regensburg Pediatric Appendicitis Dataset for making this valuable clinical data publicly available.

References

Addis, D. G., Shaffer, N., Fowler, B. S., and Tauxe, R. V. The epidemiology of appendicitis and appendectomy in the united states. *American Journal of Epidemiology*, 132(5):910–925, 11 1990. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a115734. URL https://doi.org/10.1093/oxfordjournals.aje.a115734.

Afridi, M. A., Khan, I., Khalid, M. M., and Ullah, N. Combined clinical accuracy of inflammatory markers and ultrasound for the diagnosis of acute appendicitis. *Ultrasound*, 31(4):266–272, 2023.

Bhangu, A., Søreide, K., Di Saverio, S., Assarsson, J. H., and Drake, F. T. Acute appendicitis: modern understanding of pathogenesis, diagnosis, and management. *The Lancet*, 386(10000):1278–1287, 2015. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(15)00275-5. URL https://www.sciencedirect.com/science/article/pii/S0140673615002755.

Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Ozkan, E., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Knorr, C., and Vogt, J. E. Regensburg pediatric appendicitis dataset, February 2023. URL https://doi.org/10.5281/zenodo.7669442.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

A. Additional Experimental Details

A.1. Complete Model Results

Table 8, 9, 10 presents the complete performance metrics for all model-target combinations on the no-ultrasound dataset with median imputation.

A.2. Feature Selection Results

Table 11 shows the features selected through our two-stage forward-backward selection process.

A.3. Data Processing Summary

B. LLM Prompts

B.1. Direct Prediction Prompt

```
The target is encoded as integer labels. The meaning of each label is:

{mapping_text}

{base_pred_text}
{examples_text}
Now here is a NEW patient (different from all the examples above).

The features and their values of THIS child are:

{case_text}

Possible labels you are allowed to output are: [{class_ids_str}].

Please output ONLY a single integer from [{class_ids_str}] as your prediction,
with no extra words, no explanation.
```

B.2. Iterative Refinement Prompt

```
You are a senior pediatric emergency surgeon.
You will see clinical features of one child and must predict
the label for the target: {target_name}
```

B.3. Prompt for Novel Information Generation

```
You are a senior pediatric emergency surgeon.

You will see the clinical features of one child with suspected appendicitis.
There are three prediction targets:

1) {s_diag}
2) {s_sev}
3) {s_mng}

For each target, you must make a binary decision:
- 1 = the more serious / positive outcome is present
- 0 = the more serious / positive outcome is NOT present

The dataset internally encodes labels as integer codes. Here is the mapping:
{mapping_text}
```

Table 8. Complete model performance on test set (No Ultrasound features, Median Imputation) for Diagnosis

Target	Model	Train Acc	Test Acc	Train F1	Test F1	Train AUC	Test AUC
Diagnosis	Logistic Regression	0.753	0.750	0.691	0.688	0.832	0.833
	Decision Tree	1.000	0.705	1.000	0.652	1.000	0.701
	Random Forest	1.000	0.756	1.000	0.678	1.000	0.848
	Gradient Boosting	0.938	0.744	0.922	0.683	0.984	0.807
	HistGradientBoosting	1.000	0.756	1.000	0.698	1.000	0.866
	SVM	0.853	0.756	0.815	0.683	0.927	0.801
	KNN	0.772	0.712	0.707	0.622	0.864	0.742
	Random Forest		0.750	0.714	0.635		0.672
	Gradient Boosting		0.724	0.656	0.667		0.661
	AdaBoost		0.763	0.741	0.635		0.684
	LightGBM		0.756	0.727	0.635		0.678
	CatBoost		0.731	0.672	0.651		0.661
	Neural Decision Forest		0.718	0.646	0.667		0.656
	GRANDE (HistGBDT approx)		0.712	0.650	0.619		0.634
	DT-GFN (ExtraTrees approx)		0.756	0.712	0.667		0.689
	MorphBoost (stacked GBDT)		0.782	0.754	0.683		0.717
	Meta-Tree Boosting		0.750	0.700	0.667		0.683
	TEL (Tree Ensemble Layer)		0.756	0.705	0.683		0.694
	TreeNet		0.744	0.702	0.635		0.667
	Boundary ds Hybrid		0.679	0.580	0.746		0.653
	ds simple		0.468	0.424	0.889		0.574
	ds with examples		0.667	0.560	0.810		0.662
	ds with examples & refine		0.692	0.603	0.698		0.647
	ds with examples & refine & base		0.686	0.606	0.635		0.620
	gpt simple		0.705	0.616	0.714		0.662
	gpt with examples		0.705	0.610	0.746		0.671
	gpt with examples & refine		0.712	0.625	0.714		0.667
	gpt with examples & refine & base		0.724	0.647	0.698		0.672

Table 9. Complete model performance on test set (No Ultrasound features, Median Imputation) for Management

Target	Model	Train Acc	Test Acc	Train F1	Test F1	Train AUC	Test AUC
Management	Logistic Regression	0.833	0.795	0.820	0.776	0.911	0.848
	Decision Tree	1.000	0.763	1.000	0.756	1.000	0.763
	Random Forest	1.000	0.833	1.000	0.813	1.000	0.899
	Gradient Boosting	0.976	0.801	0.976	0.781	1.000	0.887
	HistGradientBoosting	1.000	0.885	1.000	0.874	1.000	0.943
	SVM	0.872	0.756	0.856	0.734	0.966	0.874
	KNN	0.836	0.718	0.821	0.682	0.926	0.796
	Random Forest		0.821	0.790	0.821		0.799
	Gradient Boosting		0.782	0.749	0.782		0.762
	AdaBoost		0.782	0.751	0.782		0.760
	LightGBM		0.795	0.804	0.795		0.778
	CatBoost		0.808	0.775	0.808		0.788
	Neural Decision Forest		0.763	0.731	0.763		0.739
	GRANDE (HistGBDT approx)		0.763	0.770	0.763		0.745
	DT-GFN (ExtraTrees approx)		0.788	0.759	0.788		0.766
	MorphBoost (stacked GBDT)		0.808	0.776	0.808		0.786
	Meta-Tree Boosting		0.795	0.764	0.795		0.772
	TEL (Tree Ensemble Layer)		0.821	0.787	0.821		0.802
	TreeNet		0.833	0.801	0.833		0.813
	Boundary ds Hybrid		0.814	0.806	0.814		0.802
	ds simple		0.481	0.658	0.481		0.431
	ds with examples		0.763	0.749	0.763		0.752
	ds with examples & refine		0.365	0.602	0.365		0.429
	ds with examples & refine & base		0.750	0.806	0.750		0.757
	gpt simple		0.750	0.751	0.750		0.743
	gpt with examples		0.769	0.740	0.769		0.751
	gpt with examples & refine		0.673	0.733	0.673		0.674
	gpt with examples & refine & base		0.808	0.776	0.808		0.786

Table 10. Complete model performance on test set (No Ultrasound features, Median Imputation) for Severity

Target	Model	Train Acc	Test Acc	Train F1	Test F1	Train AUC	Test AUC
Severity	Logistic Regression	0.904	0.865	0.945	0.923	0.933	0.864
	Decision Tree	1.000	0.865	1.000	0.922	1.000	0.699
	Random Forest	1.000	0.891	1.000	0.938	1.000	0.888
	Gradient Boosting	0.984	0.840	0.991	0.905	0.999	0.885
	HistGradientBoosting	1.000	0.853	1.000	0.913	1.000	0.899
	SVM	0.934	0.859	0.962	0.921	0.983	0.829
	KNN	0.886	0.840	0.937	0.911	0.934	0.736
	Random Forest		0.878	0.895	0.970		0.931
	Gradient Boosting		0.865	0.917	0.924		0.921
	AdaBoost		0.853	0.904	0.924		0.914
	LightGBM		0.865	0.905	0.939		0.922
	CatBoost		0.865	0.899	0.947		0.923
	Neural Decision Forest		0.885	0.913	0.955		0.933
	GRANDE (HistGBDT approx)		0.853	0.910	0.917		0.913
	DT-GFN (ExtraTrees approx)		0.853	0.881	0.955		0.916
	MorphBoost (stacked GBDT)		0.872	0.918	0.932		0.925
	Meta-Tree Boosting		0.872	0.912	0.939		0.925
	TEL (Tree Ensemble Layer)		0.872	0.918	0.932		0.925
	TreeNet		0.872	0.912	0.939		0.925
	Boundary ds Hybrid		0.821	0.933	0.848		0.889
	ds simple		0.846	0.846	1.000		0.917
	ds with examples		0.814	0.964	0.811		0.881
	ds with examples & refine		0.795	0.972	0.780		0.866
	ds with examples & refine & base		0.878	0.931	0.924		0.928
	gpt simple		0.853	0.916	0.909		0.913
	gpt with examples		0.827	0.941	0.848		0.892
	gpt with examples & refine		0.814	0.933	0.841		0.884
	gpt with examples & refine & base		0.846	0.909	0.909		0.909

Table 11. Selected features after forward-backward selection

Target	Stage	Selected Features
Diagnosis	Forward (15)	BMI, Sex, Lower_Right_Abd_Pain, Loss_of_Appetite, WBC_Count, RBC_Count, Ketones_in_Urine_+, Ketones_in_Urine_++, Peritonitis_generalized, Peritonitis_no, Neutrophil_Percentage, Nausea, CRP,
	Backward (10)	BMI, Sex, Lower_Right_Abd_Pain, Loss_of_Appetite, WBC_Count,RBC_Count, Ketones_in_Urine_+, Ketones_in_Urine_++, Peritonitis_generalized, Peritonitis_no
Management	Forward (15)	Age, BMI, Coughing_Pain, Ketones_in_Urine_++, Ketones_in_Urine_+++, Ketones_in_Urine_no, WBC_in_Urine_no, CRP, Peritonitis_no, Psoas_Sign
	Backward (10)	Age, BMI, Coughing_Pain, Ketones_in_Urine_++, Ketones_in_Urine_+++, Ketones_in_Urine_no, WBC_in_Urine_no, CRP, Peritonitis_no, Psoas_Sign
Severity	Forward (15)	BMI, Sex, Contralateral_Rebound_Tenderness, Nausea, Segmented_Neutrophils, Ketones_in_Urine_++, Ketones_in_Urine_+++, Ketones_in_Urine_no, RBC_in_Urine_+, CRP
	Backward (10)	BMI, Sex, Contralateral_Rebound_Tenderness, Nausea, Segmented_Neutrophils, Ketones_in_Urine_++, Ketones_in_Urine_+++, Ketones_in_Urine_no, RBC_in_Urine_+, CRP

Table 12. Data processing summary statistics

Metric	Value
Original samples	782
Samples after removing missing targets	780
Training samples	624 (80%)
Test samples	156 (20%)
All Features count	78
No Ultrasound Features count	45
Ultrasound features excluded	22
Other excluded features	1 (Length_of_Stay)
Missing rate (All Features)	36.84%
Missing rate (No Ultrasound)	11.25%
Missing in training data	3086 values (10.99%)
Missing in test data	787 values (11.21%)
Models evaluated	7
Target variables	3
Imputation strategies compared	3

```
{examples_text}
Now here is a NEW child (different from all examples above).

The features of this child are:
{case_text}

Based only on these features and your clinical knowledge (and not on the true labels),
decide three binary labels:

- y_diagnosis   = 1 if the child truly has appendicitis, else 0.
- y_severity    = 1 if the child has complicated/severe appendicitis, else 0.
- y_management  = 1 if the child requires surgical / operative management, else 0.

Output your answer on a single line in the following exact format:
y_diagnosis=AA; y_severity=BB; y_management=CC

where AA, BB and CC are integers, each either 0 or 1.
Do not output any extra words or explanation.
```