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.

TODO: Add more about advanced tree methods and LLM.

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 appendicitis
- **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, fea-

[*] 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 5, 2026.

ture 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 × 78 features, including ultrasound-derived measurements
- **No Ultrasound**: 780 samples × 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)

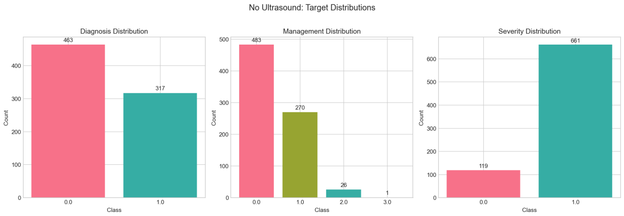

Figure 1. Distribution of the three target variables: Diagnosis, Severity, and Management. The class imbalance varies across targets, with Severity showing the most pronounced imbalance.

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

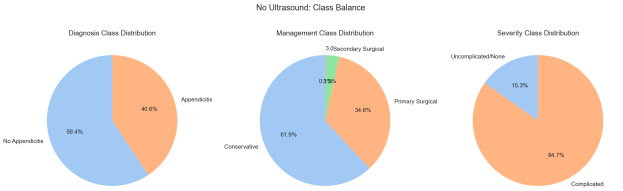

Figure 2. Class balance pie charts for each target variable, showing the proportion of positive and negative cases.

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.

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

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.

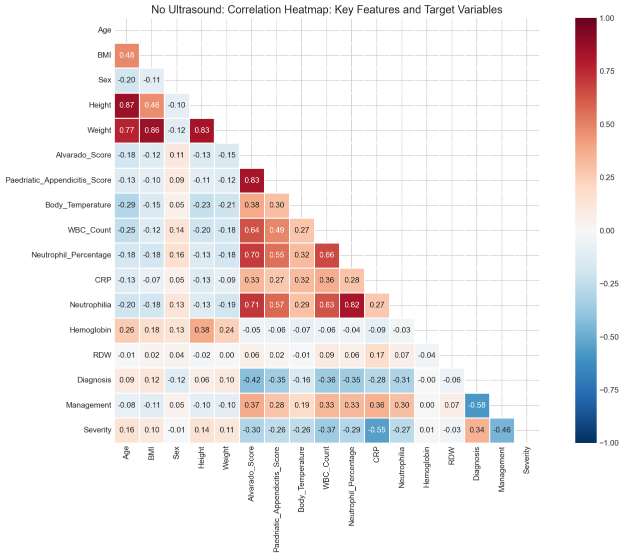

Figure 3. Correlation heatmap showing relationships between features and target variables. Strong correlations are observed between inflammatory markers and between clinical scores.

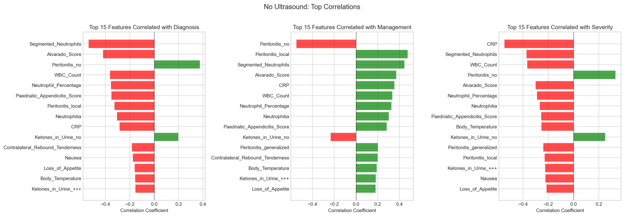

Figure 4. Top 15 features most correlated with each target variable. The correlations align well with established clinical knowledge.

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.

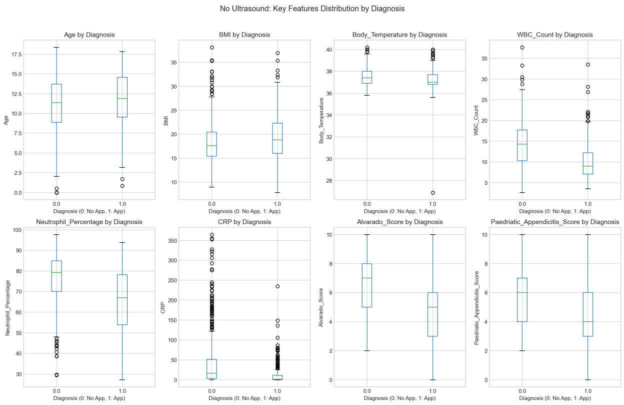

Figure 5. Box plots showing distribution of key features by Diagnosis status. Notable differences are observed in inflammatory markers such as WBC Count and CRP.

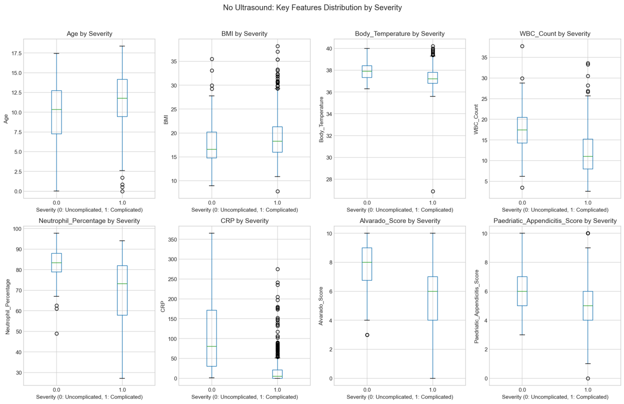

Figure 6. Box plots showing distribution of key features by Severity status.

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:

- **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.

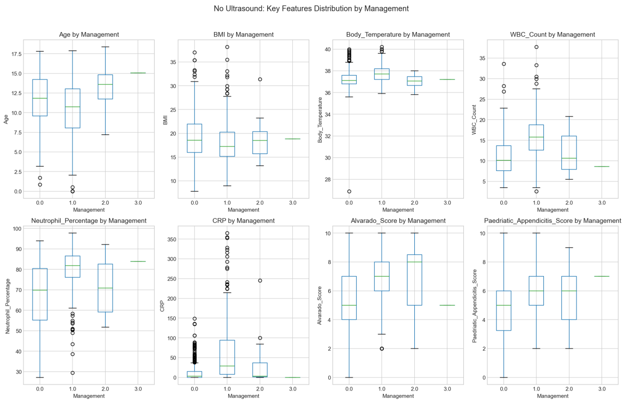

Figure 7. Box plots showing distribution of key features by Management status.

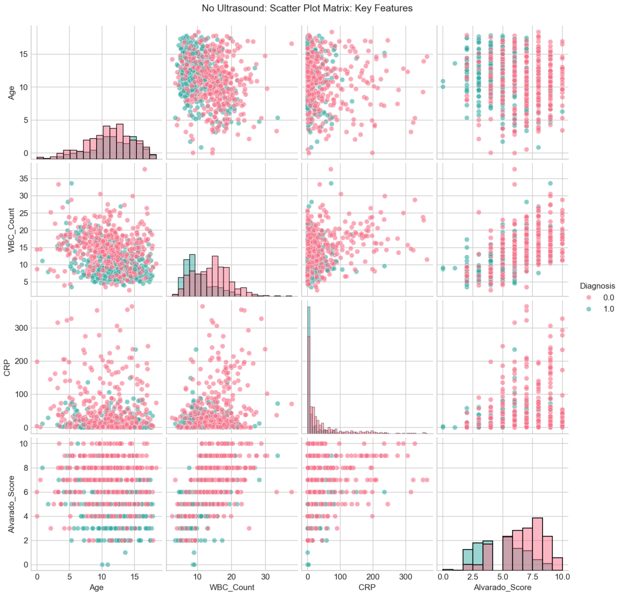

Figure 8. Scatter plot matrix of key features colored by Diagnosis status.

- **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.

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.

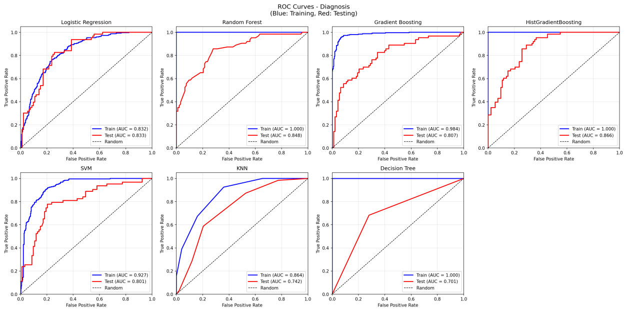

Figure 9. ROC curves for all models on the Diagnosis target. HistGradientBoosting and Random Forest achieve the highest AUC values.

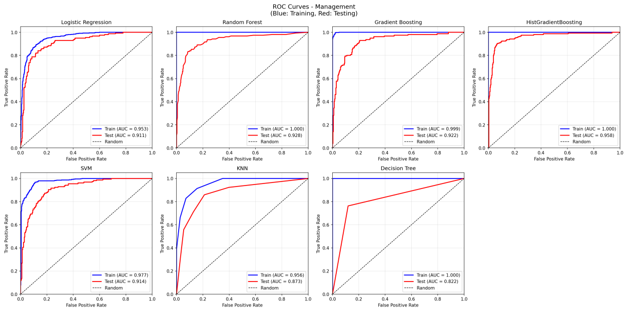

Figure 10. ROC curves for all models on the Management target.

4.2.3. 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 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:

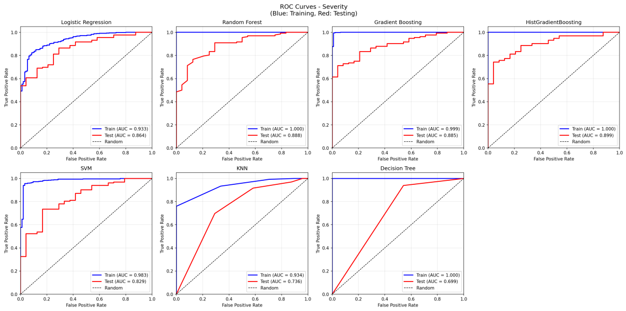

Figure 11. ROC curves for all models on the Severity target.

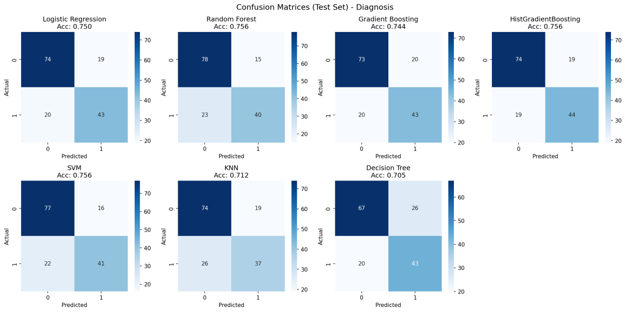

Figure 12. Confusion matrix for the best model on Diagnosis prediction.

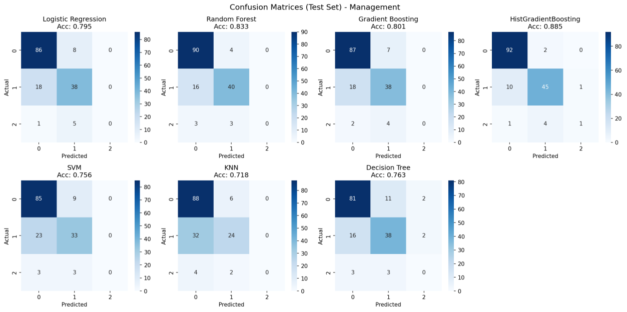

Figure 13. Confusion matrix for the best model on Management prediction.

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.

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.

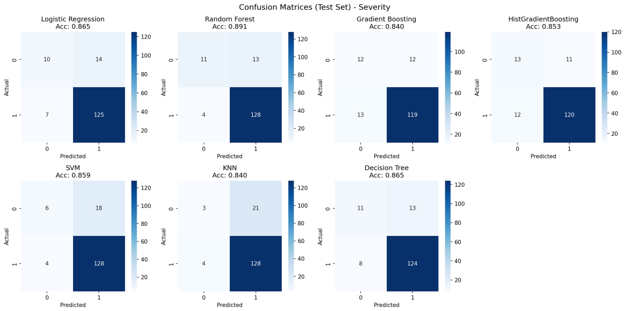

Figure 14. Confusion matrix for the best model on Severity prediction.

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

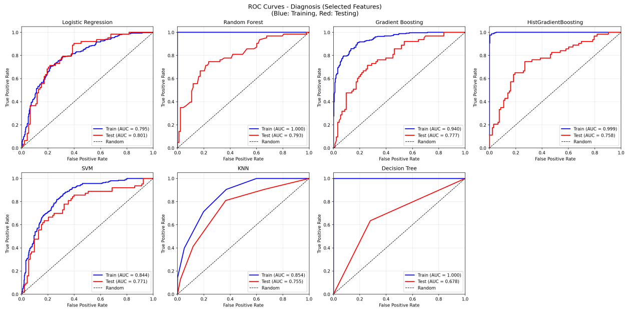

Figure 15. ROC curves for models trained on selected features (Diagnosis). Performance is comparable to full-feature models.

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

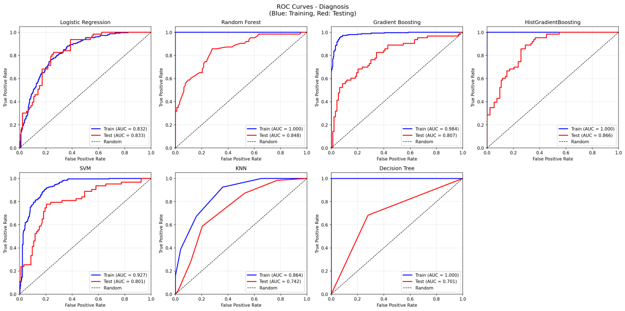

Figure 16. ROC curves for models trained on full features (Diagnosis).

and without feature selection across all models and targets.

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

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

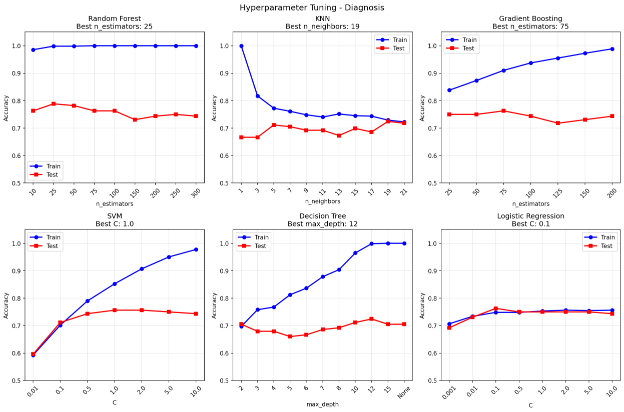

Figure 17. Grid search results for Diagnosis, showing training and validation performance across different hyperparameter values for each model.

5. Handling Missing Values

Given the substantial missing data in our dataset (11.25% overall for the no-ultrasound dataset), the choice of imputation strategy can significantly impact model performance. In this section, we systematically compare three imputation approaches.

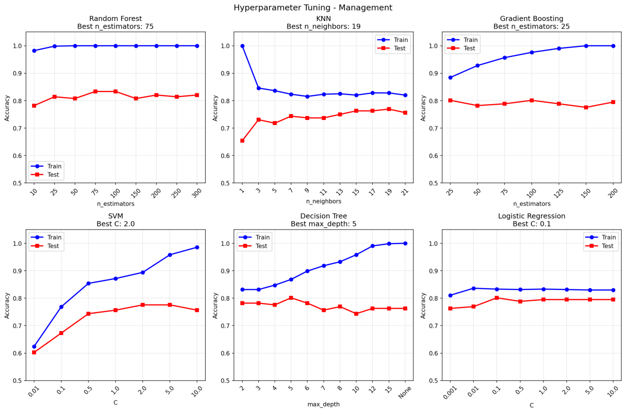

Figure 18. Grid search results for Management, showing training and validation performance across different hyperparameter values for each model.

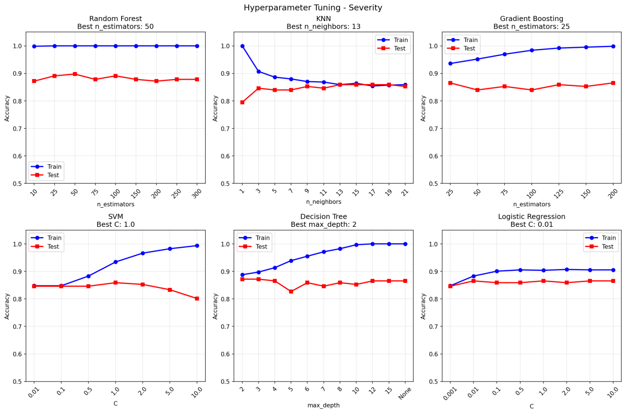

Figure 19. Grid search results for Severity, showing training and validation performance across different hyperparameter values for each model.

5.1. Imputation Methods

5.1.1. Median Imputation

The simplest approach replaces missing values with the feature median:

$$\hat{x}_{ij}^{\text{miss}} = \text{median}(\{x_{kj} : x_{kj} \text{ is observed}\}) \tag{5}$$

While computationally efficient, this method has notable limitations:

- Ignores class-specific patterns in the data
- Distorts covariance and correlation structures
- Can introduce bias if data is not missing completely at random (MCAR)

5.1.2. KNN Imputation

K-Nearest Neighbors imputation estimates missing values based on the k most similar complete samples:

$$\hat{x}_{ij}^{\text{miss}} = \frac{1}{k} \sum_{l \in N_k(i)} x_{lj} \tag{6}$$

where $N_k(i)$ denotes the k nearest neighbors of sample i.

Advantages and limitations:

- Preserves local data structure
- Computationally expensive for large datasets
- Susceptible to the curse of dimensionality
- Sensitive to the choice of k and distance metric

5.1.3. MICE (Multiple Imputation by Chained Equations)

MICE (Van Buuren & Groothuis-Oudshoorn, 2011) is an iterative approach that models each feature with missing values as a function of other features. The algorithm proceeds as follows:

1. **Initialization**: Fill missing values with simple imputation (e.g., mean)
2. **Iteration**: For each feature j with missing values:

$$\theta_j^{(t)} \sim P(\theta_j | X_j^{\text{obs}}, X_{-j}^{(t)}) \tag{7}$$

$$X_j^{\text{miss}(t+1)} \sim P(X_j^{\text{miss}} | X_{-j}^{(t)}, \theta_j^{(t)}) \tag{8}$$

3. **Pooling**: Combine results across multiple imputations:

$$\bar{\beta} = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_k \tag{9}$$

MICE assumes data is **Missing at Random (MAR)**, meaning the probability of missingness depends only on observed data.

5.2. Comparative Analysis

Table 7 compares imputation strategies across different models and targets.

5.3. Discussion

Several key observations emerge from our imputation comparison:

Task Dependence: The optimal imputation strategy varies by task. For Diagnosis, MICE with HistGradientBoosting

Table 7. Imputation strategy comparison (Test Accuracy)

Target	Model	Median	KNN	MICE
Diagnosis	Logistic Reg.	0.769	0.788	0.788
	Gradient Boost	0.795	0.782	0.795
	HistGradBoost	0.795	0.795	**0.821**
Severity	Logistic Reg.	0.936	0.872	0.885
	Gradient Boost	0.936	0.929	0.923
	HistGradBoost	**0.942**	0.936	0.929

achieves the best result (82.1%), while for Severity, simple median imputation performs best (94.2%).

Model Robustness: HistGradientBoosting, which natively handles missing values through its binning strategy, shows consistent performance across imputation methods. This suggests that sophisticated imputation may be less critical when using models with built-in missing value handling.

Computational Trade-offs: MICE is significantly more computationally expensive than median imputation, yet does not consistently outperform simpler methods. For real-time clinical applications, median imputation with HistGradientBoosting offers an attractive balance of simplicity and performance.

MAR Assumption: The limited benefit of MICE may indicate that the MAR assumption is not well-satisfied in our data. Clinical missing data often exhibits patterns related to the severity of the condition (e.g., sicker patients may have more tests performed), violating MAR.

TODO: Create visualization showing missing data patterns across features: (1) Heatmap of missingness across samples and features, (2) Correlation between missingness indicators and target variables.

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

TODO: Implement and evaluate advanced tree-based methods on the no-ultrasound dataset. Document experimental setup, hyperparameters, and training procedures.

6.1.1. Ensemble Variants

Beyond the standard Random Forest and Gradient Boosting, we investigate:

- **AdaBoost**: Adaptive boosting that adjusts sample weights based on classification errors, building a strong learner from weighted weak learners.
- **LightGBM**: Gradient boosting with leaf-wise tree growth strategy, offering faster training and lower memory usage through histogram-based algorithms.
- **CatBoost**: Gradient boosting with oblivious (symmetric) decision trees, providing built-in handling of categorical features and ordered boosting to reduce prediction shift.

6.1.2. Neural and Hybrid Approaches

- **Neural Decision Forest**: Combines deep neural networks with decision forests, using differentiable decision trees that can be trained end-to-end with gradient descent.
- **Tree Ensemble Layer (TEL)**: Aggregates multiple tree ensembles through a learned combination layer, enabling ensemble-of-ensembles architectures.

6.1.3. Meta-Learning Strategies

- **TreeNet**: A two-stage approach where the first random forest rf_1 is trained, then a second model is trained on $(x, rf_1(x))$ to refine predictions.
- **MorphBoost**: Combines gradient boosting models of different depths (e.g., depth 2 and depth 4) through voting.
- **Meta-Tree Boosting**: Ensemble voting across Random Forest, Gradient Boosting, and LightGBM predictions.

TODO: Add experimental results table for advanced tree methods. Expected best results: Diagnosis – MorphBoost (Accuracy: 0.7821, F1: 0.7167); Management – TreeNet (Accuracy: 0.8333, F1: 0.8131); Severity – Neural Decision Forest (Accuracy: 0.8846, F1: 0.9333). Include confusion matrices and comparison with basic models.

6.2. LLM-Based Approaches

TODO: Document LLM experimental setup including: model selection (GPT-4, DeepSeek), prompt engineering strategies, API configurations, and evaluation protocols.

We explore leveraging Large Language Models for clinical prediction through several strategies:

6.2.1. Direct Prediction

- **Zero-shot**: Directly query the LLM with patient features and ask for diagnosis/severity/management predictions without any examples.

- **Few-shot with Examples**: Provide the LLM with several representative examples before making predictions.

6.2.2. Iterative Refinement

- **Error-based Refinement**: After initial predictions, provide the LLM with its errors and ask it to refine its decision-making.
- **Hybrid with Base Model**: Combine LLM predictions with traditional model outputs, potentially using the LLM only for difficult cases near decision boundaries.

6.2.3. LLM as Feature Engineer

Rather than using LLMs for direct prediction, we can leverage their clinical knowledge to generate new features:

> "Based only on these features and your clinical knowledge (and not on the true labels), decide three binary labels: `y_diagnosis`, `y_severity`, `y_management`..."

These LLM-generated features can then be used as additional inputs to traditional ML models.

TODO: Add experimental results for LLM-based methods in two tables: (1) Direct LLM prediction results – Best without base model: Diagnosis (GPT with examples & refine, Acc: 0.7115, F1: 0.6667), Management (DeepSeek with examples & refine & base, Acc: 0.7500, F1: 0.7572), Severity (GPT with examples & refine, Acc: 0.8141, F1: 0.8845). (2) Results with LLM-generated features from DeepSeek: Diagnosis (DT-GFN, Acc: 0.7821, F1: 0.7385), Management (TEL, Acc: 0.8333, F1: 0.8148), Severity (TEL, Acc: 0.8846, F1: 0.9328).

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.

   TODO: Add summary points for advanced methods and LLM approaches once results are finalized.

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 Section 2, Section 3 and Section 4.
- **Wen Yuan**:

  TODO: TBD
- **Jingxing Zhou**:

  TODO: TBD

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.

TODO: reorganize the code and provide a README file with instructions for reproducing experiments. Add the code to the supplementary materials when submitting the report.

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 presents the complete performance metrics for all model-target combinations on the no-ultrasound dataset with median imputation.

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

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
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
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

A.2. Feature Selection Results

Table 9 shows the features selected through our two-stage forward-backward selection process.

A.3. Data Processing Summary

Table 9. 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 10. 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