Predicting Pediatric Appendicitis and Severity Using Machine Learning

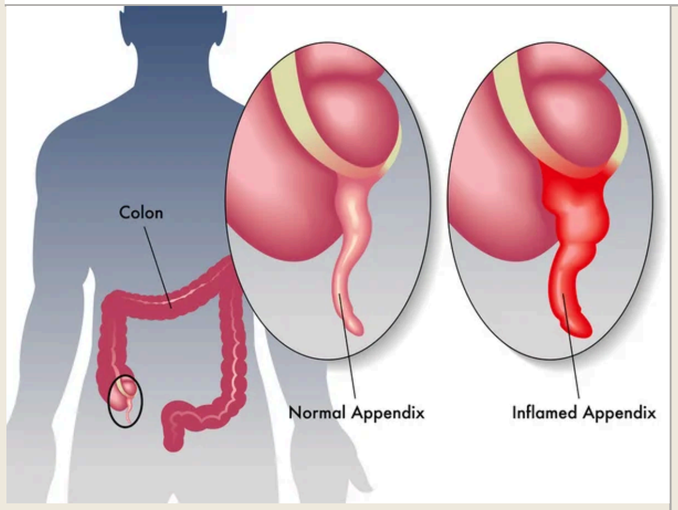
AUTHOR

Eoin Houstoun
ehous001@gold.ac.uk
Department of Computing,
Goldsmiths, University of London

Data Source:
https://archive.ics.uci.edu/dataset/938/regensburg+pediatric+appendicitis

Goldsmiths
UNIVERSITY OF LONDON

An XGBoost machine learning model for diagnosis and severity prediction of pediatric appendicitis using clinical and laboratory features.



01. Introduction

Appendicitis is the most common surgical emergency in children, but diagnosis can be challenging due to overlapping symptoms with other conditions.

This project presents a machine learning-based decision support tool that predicts both the likelihood of appendicitis and its severity, using clinical, laboratory, and ultrasound features. By integrating an XGBoost model with an interactive Gradio web interface, the system offers fast, user-friendly predictions to assist healthcare professionals in improving diagnostic accuracy.

02. Objective

To develop a machine learning-based tool capable of predicting:

1. **The likelihood of appendicitis diagnosis (Binary classification)**
2. **The severity of appendicitis (complicated vs. uncomplicated)**

using structured clinical, laboratory, and imaging data, and to deploy this tool as an intuitive web application for real-time decision support.

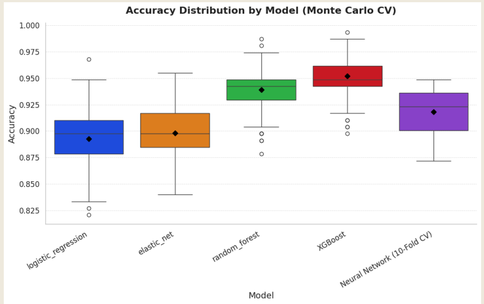
03. Methodology

- **Data Exploration:**
  - Regensburg pediatric appendicitis dataset containing structured clinical, laboratory, and ultrasound-derived features.
- **Preprocessing:**
  - KNN imputation for missing numerical values.
  - Standardisation of numerical features.
  - One-hot encoding of categorical variables.
- **Modeling:**
  - Logistic Regression, Elastic Net, Random Forest, XGBoost, and a MLP (PyTorch) were trained, with hyperparameters optimised through Monte Carlo nested cross-validation.
- **Validation:**
  - Monte Carlo Cross-Validation (MCCV) with 200 iterations for reliable performance estimation (80/20 split).
  - Evaluation metrics included mean & standard deviation of accuracy, recall, precision, F1-score, and AUC.
- **Model Selection & Deployment:**
  - XGBoost was selected as the best model, retrained on the full dataset, and deployed as a Gradio web application.

04. Results

DIAGNOSIS - Appendicitis vs. No Appendicitis (N=780)

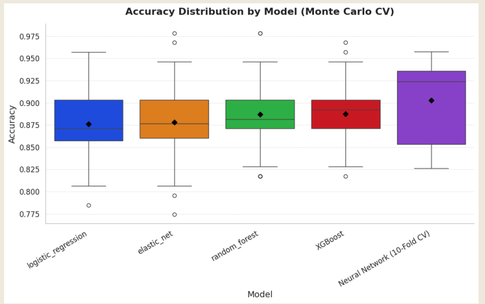
- Best performing model: XGBoost
  - Mean Accuracy: **95%**
  - Mean F1-score **96%**
- Most important features:
  - Detectability of the appendix during ultrasound.
  - Appendix Diameter
  - Length of stay (How long the patient was in hospital)



Only Patients Diagnosed with Appendicitis

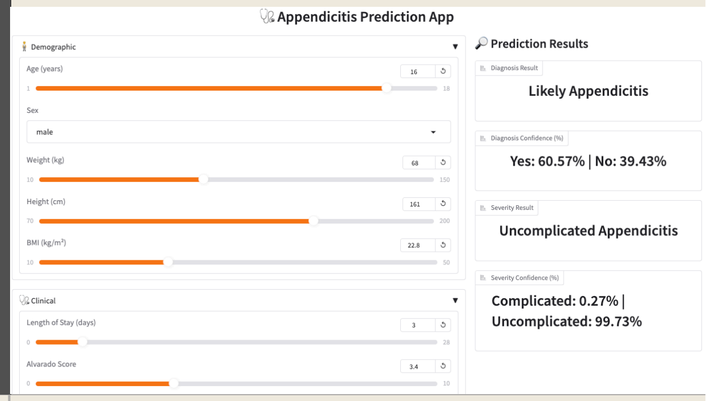
SEVERITY - Complicated vs. Not Complicated (N=486)

- Best performing model: MLP & XGBoost
  - Mean Accuracy: **90%**
  - Mean F1-score **81%**
- Most important features:
  - Length of stay (How long the patient was in hospital)
  - CRP (Protein produced by the liver, elevated in case of inflammation, infection, or injury



05. Web App - Appendicitis Tool

- XGBoost was the most reliable model for both diagnosis and severity.
- Interactive Gradio web application built for real-time appendicitis and severity prediction.
- Users input clinical, laboratory, and imaging features.
- Model predicts likelihood of appendicitis and severity.
- Here is a snippet of the tool in action (further down the app contained the rest of the features.

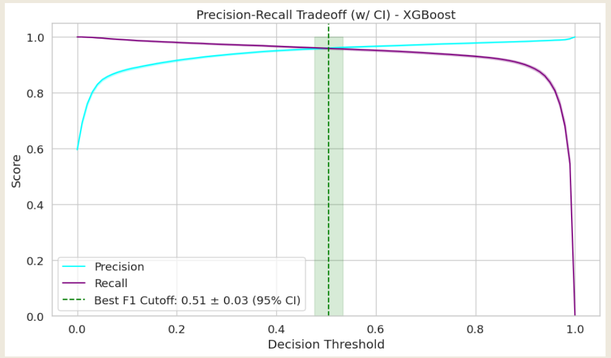


06. Conclusion

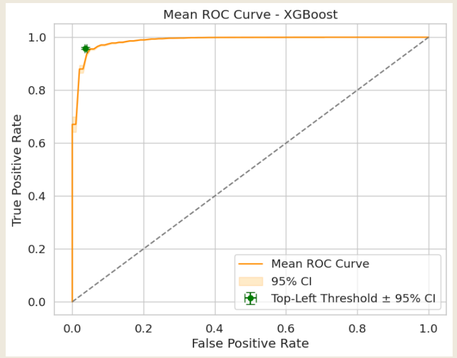
- Logistic Regression, Elastic Net, Random Forest, XGBoost, and MLP Neural Network performed with MCCV used to validate.
- XGBoost achieved the highest and most consistent performance across all metrics for diagnosis classification.
- The Neural Network achieved the highest mean performance for severity prediction but showed greater variance across validation folds. For more stable and reliable results, XGBoost was chosen for deployment.
- Feature importance analysis identified appendix detectability and size (“Appendix on US” & “Appendix Diameter”) as the most critical predictor for diagnosis. “Length of Stay” was a best indicator for the severity of the appendicitis.
- XGBoost deployed in a web application for clinical prediction.
- Findings support machine learning as a useful tool for pediatric appendicitis diagnosis.

Graphs

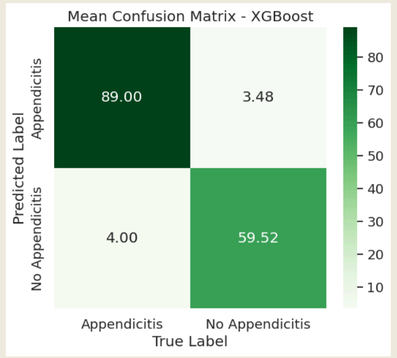
XGBoost Performance on Diagnosis



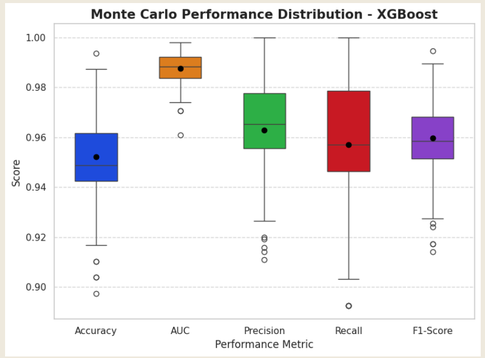
- Mean precision & mean recall across 200 MCCV iterations XGBoost.
- Confidence intervals (shaded blue and purple area) on the precision & recall lines are tight to the mean indicating reliable performance across the iterations.
- Threshold of .51 returns strongest F1-score. (Mean F1=0.96)



- Mean ROC Curve across 200 MCCV iterations (XGBoost)
- C.I. (shaded yellow) on the line is tight to the mean indicating reliable performance across the iterations.
- Curve is close to the top left indicating strong mean AUC results. (Mean AUC=0.988)

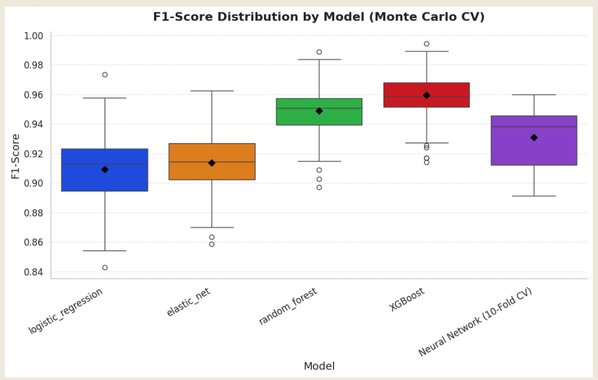


- Mean confusion matrix across 200 MCCV iterations (XGBoost).
- False positives and false negatives are low.



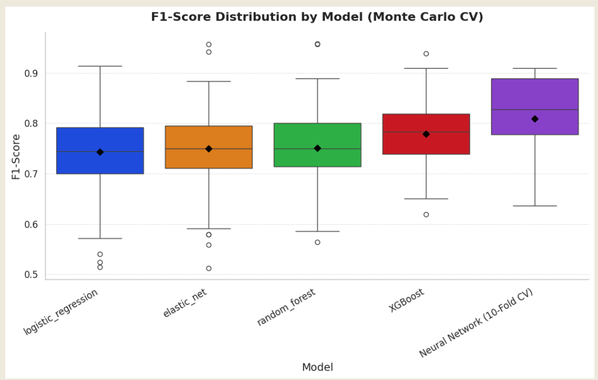
- Mean of the metrics across 200 MCCV iterations (XGBoost).
- All performed well with medians all > .94.
- Variability is low with reasonable consistent results. (box plots range < 2%).

Model Comparison on Diagnosis



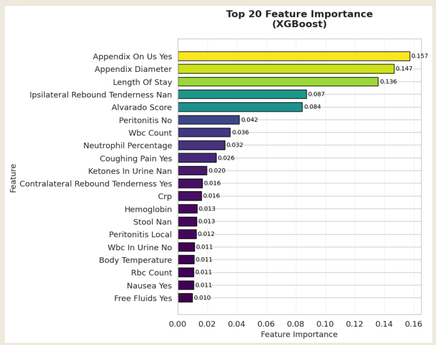
- Mean F1-scores were evaluated across 200 MCCV iterations, prioritising F1-score due to slight class imbalance.
- All models achieved high F1-scores (>90%).
- XGBoost returned the highest mean F1-score with the least variation across iterations (tightest boxplot).
- XGBoost was selected for deployment in the diagnosis prediction tool based on its strong and consistent performance.

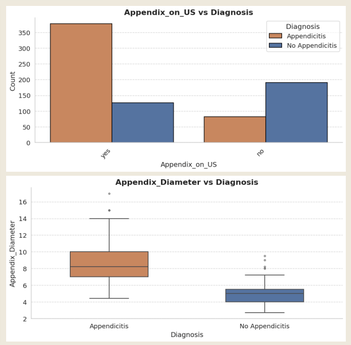
Model Comparison on Severity



- Mean F1-scores were evaluated across 200 MCCV iterations, with F1-score prioritised due to severe class imbalance.
- All models achieved F1-scores above 70%.
- The MLP achieved the highest mean F1-score (81%) but showed higher variability and overestimated accuracy (90%).
- XGBoost delivered a strong F1-score (78%) with lower variance across iterations and lower computational cost.
- XGBoost was selected for deployment in the prediction tool for its stability and efficiency.

Feature Importance on Diagnosis

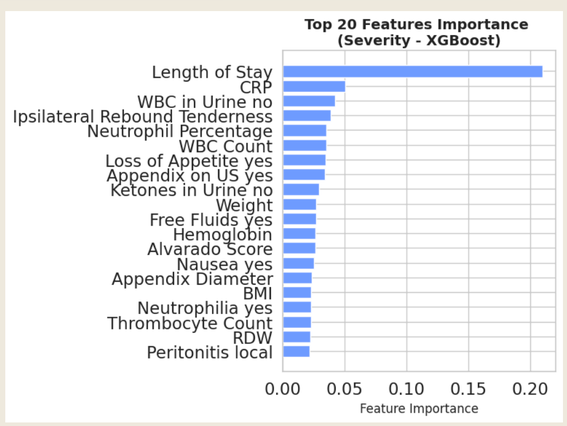


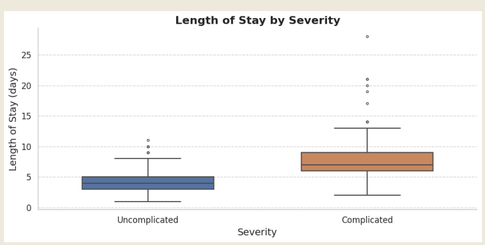


- Appendix visibility on ultrasound: Detection strongly increased appendicitis likelihood.
- In EDA, cases with visible appendix had more appendicitis diagnoses (orange bar); non-visible appendix linked to no appendicitis (blue bar).
- Appendix diameter: Larger appendix diameters were associated with a greater chance of appendicitis (supported by boxplot).

- Length of stay: Longer hospital stays correlated with higher likelihood of appendicitis.
- Ipsilateral rebound tenderness: Absence of tenderness suggested lower appendicitis risk.
- Alvarado Score: Higher scores (based on symptoms, signs, and labs) indicated greater probability of appendicitis.

Feature Importance on Severity





- Length of Stay: Longer hospital stays were strongly associated with complicated appendicitis (supported by boxplot analysis). It was the most important predictor.
- CRP (C-Reactive Protein): Higher CRP levels, a marker of inflammation, were linked to a greater likelihood of a complicated case.