A Two-Stage Patch–Patient Learning Framework for *Helicobacter pylori*–Related Histopathological Diagnosis

Abstract

This study presents a reproducible and interpretable two-stage framework for *Helicobacter pylori*–related histopathological image analysis. Stage I trains a binary classifier on patch-level labels ($Presence \in \{1, -1\}$), while Stage II aggregates patch probabilities and deep embedding features at the patient level to perform three-class classification (`NEGATIVA / BAIXA / ALTA`).

Under strict patient-level grouping and a final HoldOut evaluation protocol, the best-performing patient-level model achieves Accuracy = 0.7759, Macro-F1 = 0.7492, Balanced Accuracy = 0.7368, and Macro OvR ROC-AUC = 0.7978. The intermediate class BAIXA reaches Precision = 0.6250, Recall = 0.5882, and F1 = 0.6061. Grad-CAM visualizations and patient-level prediction reports were generated for interpretability and error analysis.

1 Introduction

The diagnosis of *Helicobacter pylori* infection from histopathological slides requires both local tissue-level evidence and global patient-level assessment. To bridge these two scales, we propose a hierarchical two-stage framework integrating patch-level representation learning with patient-level aggregation.

2 Problem Definition

2.1 Patch-Level Task

Binary classification of patch-level labels:

$$Presence \in \{1, -1\}$$

Samples with $Presence = 0$ are discarded.

2.2 Patient-Level Task

Three-class classification:

$$DENSITAT \in \{NEGATIVA, BAIXA, ALTA\}$$

All splits are performed at the patient level to prevent data leakage.

3 Data and Preprocessing

After cleaning:

- Total annotation rows: 2695
- Retained patches: 2676
- Discarded samples: 19

Patient distribution:

- NEGATIVA: 151
- ALTA: 86
- BAIXA: 72

4 Methodology

4.1 Stage I: Patch-Level Classification

Backbone: ResNet family (best: `resnet18_s42`). Output: positive probability p_{pos}.

4.2 Stage II: Patient-Level Aggregation

Features include:

- Statistical summaries of patch probabilities
- Aggregated deep embeddings (mean + standard deviation)

Final feature dimension: 1033. Classifier selected via cross-validation (`et_baixa_boost`).

5 Experimental Results

5.1 Patch-Level Performance

Validation:

- Accuracy: 0.9804
- F1-score: 0.9853

- ROC-AUC: 0.9947

Test:

- Accuracy: 0.9505
- F1-score: 0.9499
- ROC-AUC: 0.9792

5.2 Patient-Level Performance

Class	Precision	Recall	F1
NEGATIVA	0.8281	0.9138	0.8689
BAIXA	0.6250	0.5882	0.6061
ALTA	0.8500	0.7083	0.7727

Table 1: Per-class performance on HoldOut set

Overall metrics:

- Accuracy: 0.7759
- Macro-F1: 0.7492
- Balanced Accuracy: 0.7368
- Macro OvR ROC-AUC: 0.7978

Confusion matrix:

$$\begin{bmatrix} 53 & 5 & 0 \\ 11 & 20 & 3 \\ 0 & 7 & 17 \end{bmatrix}$$

6 Error Analysis

Out of 116 HoldOut patients, 26 were misclassified (22.4%). Most errors occur between BAIXA and adjacent classes, indicating overlapping intermediate representations.

Grad-CAM visualizations confirm attention on histologically relevant regions.

7 Conclusion

We developed a complete patch-to-patient diagnostic pipeline achieving Accuracy = 0.7759 and Macro-F1 = 0.7492. Future directions include hard-example reweighting, stain normalization, attention-based bag modeling, and external validation.