# Clinical Co-occurrence Priors for Multilabel Classification in Brazilian Gastroscopy Images

**Willgnner Ferreira Santos[1], Paulo Victor dos Santos[3], Daniela Medeiros Milhomem Cardoso[4], Marcella Scoczynski Ribeiro Martins[2] and Ronaldo Martins da Costa[1]**

[1]Institute of Informatics, Federal University of Goiás (INF/UFG), Goiânia, GO, Brazil

[2]Federal University of Technology - Paraná (UTFPR), Brazil

[3]Institute of Innovation in Management, UFG (IIG/UFG), Goiânia, GO, Brazil

[4]General Hospital of Goiânia (HGG) and Institute of the Digestive System (IAD), Goiânia, GO, Brazil

willgnner_santos@discente.ufg.br, {ronaldocosta,paulodos}@ufg.br

marcella@utfpr.edu.br, danielammcardoso@gmail.com

***Abstract.*** *This work evaluates relational formulations for multilabel gastroscopy classification, motivated by finding co-occurrences in upper gastrointestinal endoscopy. On a Brazilian dataset of 1,990 images evaluated across five core clinical labels under stratified five-fold cross-validation with a ResNet50 backbone, we compare independent binary cross-entropy (M0_BCE), Asymmetric Loss (M0_ASL), Classifier Chains (M1), explicit co-occurrence loss regularization (M2), and ML-GCN (M3). Co-occurrence regularization (M2) obtains the highest mean macro-F1 (*$0.537 \pm 0.041$*), whereas ML-GCN underperforms (*$0.323 \pm 0.023$*). However, paired bootstrap resampling shows M2's confidence interval overlaps with plain BCE (*$0.517 \pm 0.045$*), indicating no formulation conclusively outperforms ResNet50+BCE. The primary contribution is a rigorous benchmark and informative null result on relational modeling in small clinical label regimes.*

## 1. Introduction

Upper Gastrointestinal Endoscopy (UGE) is the standard examination for investigating diseases of the esophagus, stomach, and duodenum. The growing volume of images produced in clinical routine has motivated the development of deep learning systems to assist diagnosis [Klang et al. 2023]. Recent works report gains in tasks such as gastric lesion classification, early cancer detection, and identification of findings associated with gastritis and intestinal metaplasia [Zhang et al. 2023, Yang et al. 2023]. However, existing literature predominantly formulates the problem as binary or mutually exclusive multiclass classification, assigning a single label per image. This formulation fails to reflect real-world clinical practice, where multiple findings, such as erythema, erosion, and ulcer, frequently co-exist in the same image and exhibit stable semantic relationships.

Despite the volume of publications, multilabel modeling in UGE remains extremely rare: in a systematic review over 717 articles from IEEE Xplore, Scopus, and

Web of Science, only one study used the term multilabel in the title, and none incorporated clinical co-occurrence as an architectural prior in gastroscopy. The few existing multilabel initiatives focus on anatomical region recognition rather than co-occurrence of pathological findings [Li et al. 2022]. Consequently, standard classifiers with independent heads trained via Binary Cross-Entropy (BCE) treat clinically dense pairs (e.g., erythema with erosion) as statistical coincidences, ignoring relational structures that could provide learning signals in class-imbalanced regimes [Krenzer et al. 2026].

The relevance of this work lies in addressing this gap using a real-world Brazilian UGE dataset of 1,990 images with up to 11 binary labels, where 51.7% of images present two or more positive labels simultaneously. Furthermore, existing literature relies predominantly on Asian public benchmarks, which face domain shifts due to population-specific endemic disease prevalences (e.g., *H. pylori* infection rates and gastric cancer incidence), hardware variations, and local clinical workflows, limiting direct transferability to Brazilian healthcare settings [Silva et al. 2023, Ferreira et al. 2024].

The general objective of this article is to investigate relational formulations for multilabel gastroscopy classification, comparing a baseline with independent heads against strategies incorporating clinical co-occurrence as a prior. We evaluate four formulations: independent heads with Binary Cross-Entropy (BCE, M0), Classifier Chains (M1), explicit co-occurrence regularization added to the loss (M2), and Multi-Label Graph Convolutional Network (ML-GCN, M3) over empirical, Pointwise Mutual Information (PMI), and clinically curated graphs. Rather than claiming a state-of-the-art diagnostic breakthrough, our contribution is a careful empirical evaluation and null finding: simple loss-level priors offer modest mean gains but remain statistically indistinguishable from a plain BCE baseline, whereas complex end-to-end GCNs fail in small clinical label regimes.

## 2. Related Work

Deep learning for UGE image classification has advanced rapidly, focusing on gastric cancer detection, pre-malignant lesion classification, and quality assessment. Prior studies fall into three broad groups: (i) binary or multiclass convolutional networks for tissue differentiation [Zhang et al. 2023]; (ii) Vision Transformer (ViT) and hybrid architectures with transfer learning [Zhang et al. 2023]; and (iii) anatomical region recognition and video quality monitoring pipelines [Bravo et al. 2023]. These axes consolidate the state of the art but treat categories in isolation, ignoring label dependencies.

Studies closest to our work adopt multi-output formulations. Liu et al. [Liu et al. 2024] propose a Kyoto Gastritis Score recognition framework using five parallel GAM-EfficientNet models for atrophy, redness, enlarged folds, metaplasia, and nodularity, representing an implicit multilabel approach without relational priors. Wei et al. [Wei et al. 2025] present a Multi-Granularity Collaborative Network (MGCN) exploring correlations across hierarchical label levels. Cao et al. [Cao et al. 2025] formulate autoimmune gastritis diagnosis as a multitask network with separate anatomical and lesion heads. For limited data, Lopes et al. [Lopes et al. 2024] evaluate self-supervised learning for gastric atrophy screening, while Hossain et al. [Hossain et al. 2023] compare convolutional and Transformer architectures in polyp classification.

Despite these advances, existing studies treat labels independently, hierarchically,

or in a multitask setup without explicitly modeling clinical co-occurrence between distinct pathological findings. Furthermore, results obtained on Asian public datasets face limited transferability to Latin American clinical data. This work addresses this gap by conducting a rigorous multi-fold comparative benchmark of relational priors on Brazilian UGE data.

## 3. Methodology

The methodology encompasses data collection, initial inspection, preprocessing and stratified splitting, label graph construction, model formulations, and statistical evaluation.

### 3.1. Data Collection and Characterization

The dataset comprises 1,990 gastroscopy images collected in clinical routine and approved by the Research Ethics Committee of the General Hospital of Goiânia (Hospital Geral de Goiânia, HGG) via Plataforma Brasil under CAAE 59898122.7.0000.0035, in compliance with Resolution CNS 466/2012. All images originated from anonymized databases of the State Hospital Alberto Rassi (HGG) and the Institute of the Digestive System (IAD) in Goiânia, GO, Brazil, with no access to socio-demographic or patient-identifiable data. Controlled access to the dataset can be requested for research purposes.

Images were manually annotated by experienced gastroenterologists and endoscopists, with consensus review employed to resolve ambiguous cases. Up to 11 binary labels per image were annotated: NORMAL, ALTERED, SALIVA, LIGHT, ERYTHEMA, POLYP, ULCER, EROSION, MICRONODULARITY, VASCULAR ECTASIA, and NEOPLASIA. The task is formulated as multilabel classification: predicting a binary vector $y_n \in \{0,1\}^K$ for image $x_n$. The experimental core considers $K = 5$ clinical labels: ERYTHEMA, POLYP, ULCER, EROSION, and MICRONODULARITY. NEOPLASIA and VASCULAR ECTASIA ($N = 4$ positive cases each) were excluded from primary macro-F1 due to extreme statistical instability; SALIVA and LIGHT were treated as auxiliary artifacts; NORMAL and ALTERED as meta-labels.

### 3.2. Initial Inspection and Exploratory Analysis

An integrity audit identified duplicate image names, orphan records, and missing values. After deterministic cleaning (retaining the row with most positive labels for duplicates and discarding orphans), 1,990 effectively usable images remained. To prevent inter-fold data leakage, visual similarity was computed across all image pairs using ImageNet pretrained ResNet50 feature embeddings; no pair exceeded the cosine similarity threshold of 0.95, confirming the absence of near-duplicate images across patients.

Exploratory analysis quantified joint label distributions: 51.7% of images (1,028/1,990) present two or more positive labels simultaneously. Clinically dense pairs include ERYTHEMA+EROSION (34 images) and ULCER+EROSION (21 images), as well as strong artifact-pathology interactions such as LIGHT with EROSION (87 images).

### 3.3. Data Preprocessing and Splitting

Images were resized to $224 \times 224$ pixels and normalized with ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). In training, only random horizontal flipping was applied; vertical flipping and aggressive color jittering were avoided to prevent

generating anatomically unrealistic orientations or altering mucosal erythema representations.

Data split used multilabel stratified cross-validation ($K = 5$ folds) via second-order Iterative Stratification [Sechidis et al. 2011], preserving joint label distributions. Each fold allocated 1,353 images for training, 239 for validation, and 398 for testing. The split was frozen using a SHA-256 hash, and a programmatic audit confirmed zero train/validation/test image intersection across all folds, ensuring strict experimental reproducibility.

### 3.4. Label Graph Construction

Relational label structure was modeled via a graph $G = (V, E, A)$ with $K = 5$ clinical nodes and adjacency matrix $A \in \mathbb{R}^{K \times K}$. Three variants of $A$ were evaluated. The empirical and PMI variants were estimated separately in each fold on the training set only, while the clinical graph remained fixed:

- $G_{\text{emp}}$ **(empirical):** $A_{ij}$ receives the normalized co-occurrence count $n_{ij} \in [0, 1]$ between labels $i$ and $j$.
- $G_{\text{pmi}}$ **(normalized PMI):** edge weighted by Laplace-smoothed Pointwise Mutual Information ($\alpha = 1$),

$$\text{PMI}(i, j) = \log \frac{P(y_i = 1, y_j = 1)}{P(y_i = 1)\, P(y_j = 1)}, \tag{1}$$

  normalized as $\text{nPMI}(i, j) = \text{PMI}(i, j)/[-\log P(y_i = 1, y_j = 1)]$.
- $G_{\text{clin}}$ **(curated clinical):** weights assigned in $\{0,\ 0.5,\ 1.0\}$ by senior endoscopists based on pathophysiological co-occurrences (1.0 for direct causal/strong co-occurrence like erythema with erosion, 0.5 for indirect associations, 0 for independent findings).

A sparsification threshold $\tau \in \{0,\ 0.1,\ 0.3\}$ evaluated on validation folds was applied to filter weak background noise, keeping edges with weight $\geq \tau$. Graphs were symmetrized, identity matrix added ($A \leftarrow A + I$), and symmetric renormalization $\hat{A} = D^{-1/2}(A + I)D^{-1/2}$ applied [Kipf and Welling 2017]. Figure 1 compares the three graph variants.

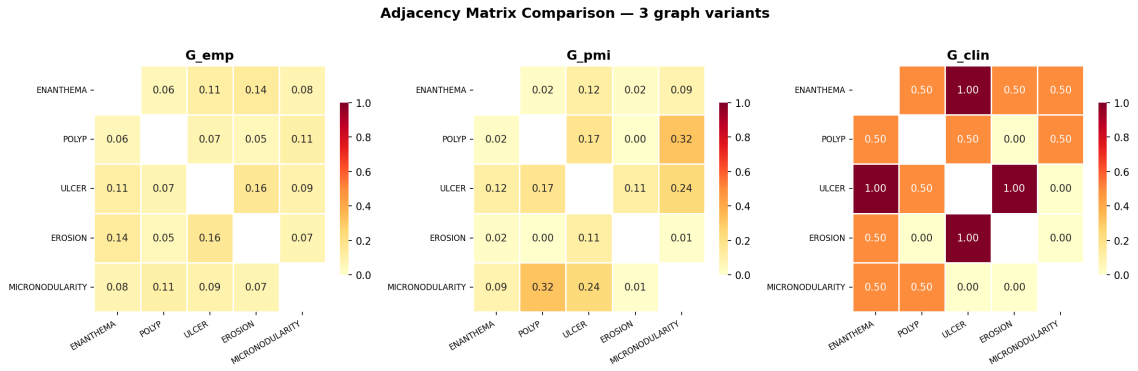



**Figure 1. Comparison of adjacency matrices $G_{\text{emp}}$, $G_{\text{pmi}}$, and $G_{\text{clin}}$ used as relational priors among the five core clinical labels.**

### 3.5. Models and Experimental Configuration

Four model families were compared, including two M0 baseline variants, all sharing a ResNet50 [He et al. 2016] backbone pre-trained on ImageNet-1k (weights `IMAGENET1K_V2`). The backbone extracts a $d_{\text{feat}} = 2048$ visual vector $\phi(x)$ per image. Table 1 summarizes the formulations.

**Table 1. Summary of multilabel classification formulations evaluated in this study.**

| ID | Formulation | Loss function / prior |
|---|---|---|
| M0_BCE | Independent heads | BCEWithLogitsLoss + *pos_weight* |
| M0_ASL | Independent heads | Asymmetric Loss [Ridnik et al. 2021] |
| M1 | Classifier Chains [Read et al. 2011] | Binary chained BCE |
| M2 | BCE + co-occurrence regularization | $\mathcal{L}_{\text{BCE}} + \lambda\,\mathcal{L}_{\text{coo}}$ |
| M3 | ML-GCN [Chen et al. 2019] | GCN over $G \in \{G_{\text{emp}}, G_{\text{pmi}}, G_{\text{clin}}\}$ |

**M0** trains independent linear heads $\hat{z}_i = \mathbf{w}_i^\top \phi(x) + b_i$, using BCEWithLogitsLoss with per-class pos_weight ($w_i = N_{\text{neg},i}/N_{\text{pos},i}$). **M0_ASL** uses Asymmetric Loss [Ridnik et al. 2021]. **M1** implements Classifier Chains [Read et al. 2011], where the $k$-th binary classifier receives predicted probabilities of preceding $k-1$ labels as additional input. **M2** adds an explicit co-occurrence regularization term to BCE:

$$\mathcal{L}_{\text{coo}} = \sum_{(i,j)\in E} A_{ij}^{\text{clin}} \cdot \text{BCE}(\sigma(z_i)\,\sigma(z_j),\, \pi_{ij})\,, \tag{2}$$

where $\pi_{ij} = P(y_i = 1, y_j = 1)$ is estimated exclusively on training folds to prevent leakage. The loss is $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda\,\mathcal{L}_{\text{coo}}$, with $\lambda = 0.3$ pre-fixed based on validation search. **M3** adapts ML-GCN [Chen et al. 2019]: label embeddings of dimension 300 pass through two GCN layers (256 hidden units, LeakyReLU), producing a weight matrix $W_{\text{rel}} \in \mathbb{R}^{K \times d_{\text{feat}}}$ applied to visual features $\phi(x)$, yielding logits $\hat{z} = \phi(x)\, W_{\text{rel}}^\top$. M3 was evaluated combinatorially over three graphs $\{G_{\text{emp}}, G_{\text{pmi}}, G_{\text{clin}}\}$ and three thresholds $\tau \in \{0, 0.1, 0.3\}$.

Training hyper-parameters were kept identical: AdamW optimizer, learning rate $10^{-4}$ for backbone and $10^{-3}$ for heads/GCN, weight decay $10^{-4}$, batch size 32, max 50 epochs, early stopping on validation macro-F1 (patience 10), and ReduceLROnPlateau scheduler (patience 4, factor 0.5). Per-class decision thresholds were optimized on validation folds across five random seeds (seeds 42 to 46) using an NVIDIA A100 GPU.

### 3.6. Evaluation Metrics

The primary metric is macro-F1 over the five core clinical labels:

$$\text{macro-F1} = \frac{1}{K} \sum_{i=1}^{K} \frac{2\,\text{prec}_i\,\text{rec}_i}{\text{prec}_i + \text{rec}_i}. \tag{3}$$

F1-micro, per-class F1, macro-F1 on images with $\geq 2$ positive labels ($F1_{\geq 2}$), Hamming loss, and per-class area under precision-recall curve (PR-AUC, reported as mean Average Precision - mAP computed via numerical integration) are also evaluated. Metrics are reported as mean $\pm$ standard deviation across 5 folds.

### 3.7. Statistical Analysis

Models were compared using stratified paired bootstrap (2,000 pooled resamples across 5 folds), generating 95% confidence intervals (CIs) and paired differences $\Delta\text{F1} = \text{F1}_{\text{model}} - \text{F1}_{\text{M0_BCE}}$. All hyperparameter choices (graph $G$, threshold $\tau$, weight $\lambda$) were selected on validation folds. Multiple metric comparisons were conducted without alpha adjustment, reinforcing conservative interpretations.

### Code Availability

Source code, preprocessing scripts, experiment configurations, graph routines, and statistical evaluation code are available in the public repository[1].

## 4. Results and Discussion

This section presents five-fold cross-validation performance across the five core clinical labels.

### 4.1. Overall Model Performance

Table 2 summarizes aggregated test set metrics for M0_BCE, M0_ASL, M1, M2, and the top validation-selected M3 variant ($G_{\text{clin}}$, $\tau = 0.1$). Figure 2 illustrates macro-F1 distributions.

**Table 2. Aggregated test set metrics across 5 folds (mean ± standard deviation).**

| Model | Macro-F1 | PR-AUC (mAP) | Micro-F1 | Hamming | $\text{F1}_{\geq 2}$ |
|---|---|---|---|---|---|
| M0_BCE | $0.517 \pm 0.045$ | $0.584 \pm 0.031$ | $0.578 \pm 0.017$ | $0.063 \pm 0.005$ | $\mathbf{0.442 \pm 0.102}$ |
| M0_ASL | $0.499 \pm 0.032$ | $0.501 \pm 0.048$ | $0.541 \pm 0.022$ | $0.070 \pm 0.006$ | $0.381 \pm 0.074$ |
| M1_CC | $0.519 \pm 0.028$ | $0.568 \pm 0.039$ | $\mathbf{0.579 \pm 0.026}$ | $\mathbf{0.055 \pm 0.003}$ | $0.367 \pm 0.057$ |
| M2_coo | $\mathbf{0.537 \pm 0.041}$ | $\mathbf{0.611 \pm 0.042}$ | $0.557 \pm 0.050$ | $0.066 \pm 0.008$ | $0.340 \pm 0.064$ |
| M3_clin-$\tau = 0.1$ | $0.323 \pm 0.023$ | $0.243 \pm 0.019$ | $0.349 \pm 0.045$ | $0.162 \pm 0.031$ | $0.410 \pm 0.029$ |

Explicit co-occurrence regularization (M2) obtained the **highest mean macro-F1** (0.537) and **best PR-AUC** (0.611), outperforming M0_BCE by $+2.0$ pp and M1 by $+1.8$ pp. However, as shown in statistical bootstrap analysis, this advantage does not achieve statistical significance. M0_ASL performed worse (0.499), showing that frequency-only reweighting is insufficient. M1 produced the lowest Hamming loss (0.055). Crucially, plain M0_BCE maintained the highest score on the multi-finding subset ($F1_{\geq 2} = 0.442 \pm 0.102$), indicating that M2's co-occurrence prior did not outperform the baseline on images with multiple simultaneous findings. ML-GCN (M3) performed far below the baseline (0.323).

### 4.2. Analysis by Metrics and Evaluation Curves

Per-class analysis reveals trade-offs masked by aggregate metrics. Figure 3 details per-label F1 scores. Relative to M0_BCE, M2 gained $+11.7$ pp in POLYP ($0.506 \rightarrow 0.624$) and $+5.7$ pp in MICRONODULARITY ($0.351 \rightarrow 0.408$), the two lowest-prevalence labels (2.5% and 1.7%). Conversely, M2 lost $-4.2$ pp in EROSION ($0.670 \rightarrow 0.629$) and $-6.8$ pp in ERYTHEMA ($0.538 \rightarrow 0.470$), the most prevalent labels.

[1] https://github.com/Willgnner-Santos/gastroscopy-cooccurrence-priors

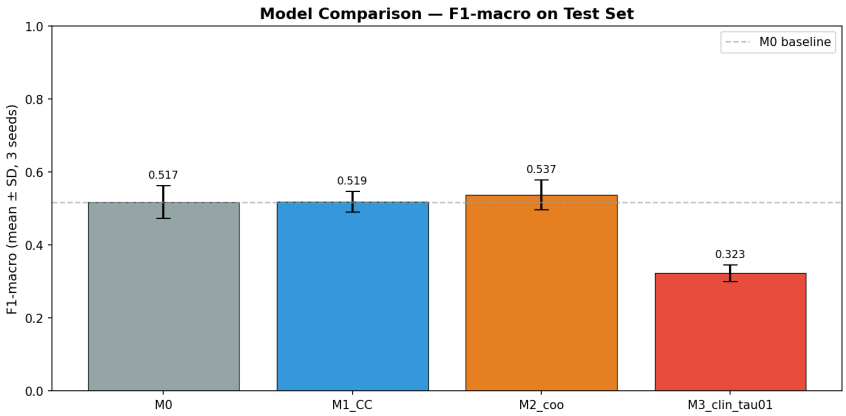



**Figure 2. Macro-F1 comparison among evaluated models (mean and standard deviation across 5 folds).**

*Support and Variance Caution:* In a 5-fold split with 398 test images per fold, POLYP ($2.5\%$) and MICRONODULARITY ($1.7\%$) correspond to roughly 10 and 7 positive test cases per fold. Consequently, per-class F1 shifts of $+11.7$ pp and $+5.7$ pp operate over single-digit positive counts. Without per-class confidence intervals, these swings cannot be cleanly separated from fold-to-fold sample noise and must be interpreted as a capacity reallocation trade-off rather than a uniform improvement. Figure 4 illustrates baseline precision-recall curves across labels.

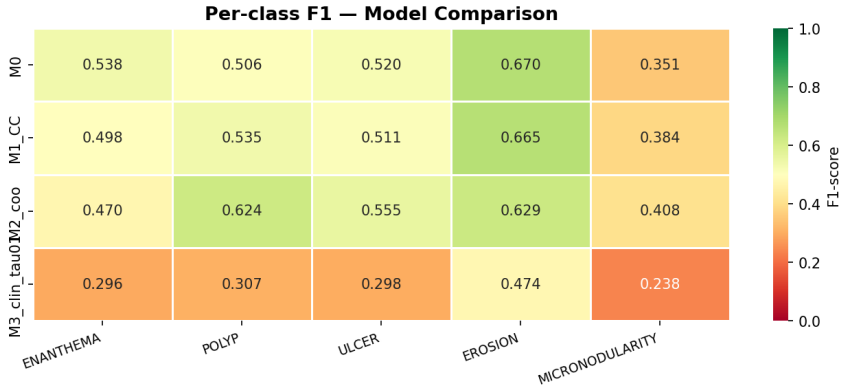



**Figure 3. F1 per label and per model highlighting top-performing formulations per class.**

### 4.3. Effect of Evaluated Strategies

**Co-occurrence regularization (M2):** The $\lambda \cdot \mathcal{L}_{\text{coo}}$ loss term over $G_{\text{clin}}$ improved sensitivity for POLYP. Validation parameter search for $\lambda \in \{0.1, 0.3, 0.5, 1.0\}$ exhibited an inverted-U curve peaking at $0.548$ for $\lambda = 0.5$. The pre-fixed $\lambda = 0.3$ was retained for protocol comparability.

**Post-Logit Smoothing (M0_post):** A heuristic post-hoc coupling ($\hat{Z} = Z + ZA_{\text{clin}}$) degraded macro-F1 to $0.491$, demonstrating that relational priors must dynamically shape features during training rather than act as late filters.

**Classifier Chains (M1):** M1 achieved statistical equivalence in macro-F1 ($0.519$ vs $0.517$) while reducing Hamming loss by $\approx 13\%$ relative, refining predictions pointwise.

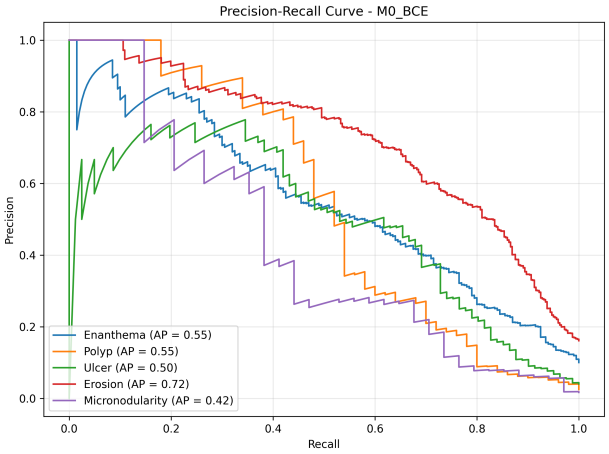



**Figure 4. Per-class precision-recall curves for M0_BCE baseline.**

**ML-GCN (M3):** No ML-GCN configuration beat the baseline (best $\tau = 0.1$ achieved $0.323$; see Figure 5). Initializing label embeddings with GloVe 300D vectors improved macro-F1 to $0.460$, but remained inferior to M0_BCE. This confirms that GCN heads introduce noise and over-parameterization when trained on small clinical label sets.

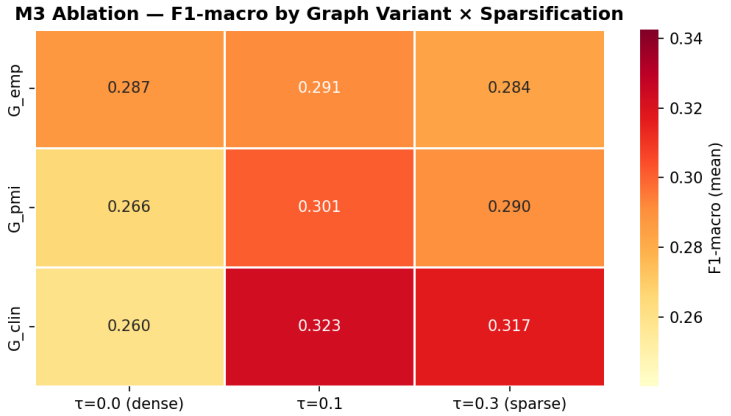



**Figure 5. Ablation of M3 (ML-GCN) across label graphs ($G_{\text{emp}}$, $G_{\text{pmi}}$, $G_{\text{clin}}$) and thresholds $\tau$.**

## 4.4. Statistical Analysis

Table 3 presents paired bootstrap macro-F1 scores and 95% CIs ($B = 2,000$).

**Table 3. Paired macro-F1 and 95% CIs (bootstrap, $B = 2{,}000$).**

| Model | Pooled Macro-F1 | 95% CI |
|---|---|---|
| M0_BCE | 0.514 | [0.474, 0.554] |
| M1_CC | 0.515 | [0.464, 0.562] |
| M2_coo | 0.524 | [0.482, 0.567] |
| M3_clin-$\tau = 0.1$ | 0.292 | [0.266, 0.318] |

Confidence intervals for M0, M1, and M2 **overlap broadly**, confirming that M2's higher mean is not statistically significant at the 5% level. In contrast, M3's CI is disjoint, proving its performance drop is consistent and structural.

## 4.5. Qualitative Error Inspection

We sampled disagreement cases between M0_BCE and M2_coo across test folds (Table 4).

**Table 4. Representative disagreement cases between M0_BCE and M2_coo. MICRONOD. denotes MICRONODULARITY.**

| True label | M0_BCE | M2_coo | Pattern |
|---|---|---|---|
| POLYP | ERYTHEMA, ULCER, EROSION | POLYP, ULCER | Rare-label recovery |
| ULCER | POLYP | ERYTHEMA, POLYP, ULCER, EROSION, MICRONOD. | Coupled false positives |
| ERYTHEMA | ERYTHEMA, POLYP, MICRONOD. | POLYP, ULCER, EROSION | Missed prevalent label |

Qualitative inspection confirms the quantitative trade-off: M2 can recover low-prevalence findings but risks introducing false positives through strong label coupling or missing isolated prevalent findings.

### 4.6. Discussion of Findings

Results show that clinical co-occurrence provides useful prior knowledge for gastroscopy multilabel classification. M2 achieved higher mean macro-F1 and PR-AUC, suggesting loss regularization can complement class reweighting [Krenzer et al. 2026]. However, the advantage is not statistically significant, and M0_BCE retained higher $F1_{\geq 2}$ $(0.442)$. The way priors are integrated is critical: explicit loss regularization shapes feature learning without replacing direct supervision, whereas ML-GCN generates noisy classification weights when trained from scratch on small clinical vocabularies. Architectures designed for large-scale benchmarks (ChestX-ray14, MS-COCO) do not transfer trivially to modest clinical label sets.

## 5. Conclusion

This work evaluated relational formulations for multilabel gastroscopy classification across five core clinical labels on a Brazilian dataset of 1,990 images. Explicit co-occurrence regularization (M2) achieved the highest mean macro-F1 $(0.537 \pm 0.041)$, but bootstrap analysis proved its confidence interval overlaps with the plain BCE baseline $(0.517 \pm 0.045)$. ML-GCN failed to match the baseline. We conclude that simple loss-level priors offer a better cost-benefit ratio than complex end-to-end GCNs in small clinical regimes, though no proposed formulation conclusively outperforms a plain ResNet50+BCE baseline. Future work will investigate medical-domain pre-training, multi-expert clinical graph validation, and adaptive graphs for rare classes under few-shot protocols.

## References

Bravo, J. I., Gomez, A. M., and Rodriguez, C. E. (2023). Spatio-temporal deep learning models for anatomical quality assessment in upper gastrointestinal video endoscopy. *Computers in Biology and Medicine*, 161:107005.

Cao, M., Lin, K., Wang, S., and Zhao, P. (2025). Multitask learning framework for autoimmune gastritis diagnosis and anatomical localization in gastroscopy. *IEEE Transactions on Biomedical Engineering*, 72(3):954–963.

Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5177–5186.

Ferreira, R. A., Souza, F. M., and Martins, M. S. R. (2024). Deep learning frameworks for endoscopic image quality and pathology classification in public healthcare systems. *IEEE Transactions on Medical Imaging*, 43(1):412–422.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hossain, M. A., Rahman, M. M., and Islam, M. S. (2023). Comparative evaluation of convolutional and vision transformer architectures for endoscopic polyp classification and segmentation. *IEEE Access*, 11:94512–94526.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Klang, E., Barash, Y., Soffer, S., Zimlichman, E., Eliakim, R., and Kopylov, U. (2023). Deep learning applications in upper gastrointestinal endoscopy: A systematic review and meta-analysis. *Digestive and Liver Disease*, 55(4):445–453.

Krenzer, A., Shah, Z., Troya, J., and Puppe, F. (2026). Evaluation of loss functions and class weighting schemes for multilabel medical image classification. In *Lecture Notes in Computer Science (LNCS)*, volume 14526, pages 276–288. Springer Nature.

Li, C., Wang, R., Xu, J., and Zhang, Y. (2022). Anatomical landmark and region recognition in upper gastrointestinal endoscopy: A benchmark study. *Medical Image Analysis*, 78:102401.

Liu, X., Chen, H., Zhou, W., and Wang, Y. (2024). Kyoto gastritis score recognition via multi-output efficientnet models in upper endoscopy. *Computer Methods and Programs in Biomedicine*, 245:108021.

Lopes, R. V., Almeida, T. P., and Santos, P. V. (2024). Self-supervised visual representation learning for gastric atrophy and intestinal metaplasia screening. *IEEE Signal Processing Letters*, 31:1405–1409.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Ridnik, E., Ben-Baruch, E., Zamir, N., Noy, A., and Friedman, I. (2021). Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91.

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). On the stratification of multi-label data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 145–158. Springer.

Silva, L. M., Santos, P. H., Oliveira, G. R., and Costa, R. M. (2023). Artificial intelligence in gastrointestinal endoscopy: A brazilian multi-center diagnostic benchmark. *IEEE Journal of Biomedical and Health Informatics*, 27(2):890–899.

Wei, J., Zhang, H., Li, R., and Sun, Y. (2025). Multi-granularity collaborative network for hierarchical pathology classification in upper gastrointestinal endoscopy. *Pattern Recognition*, 158:111045.

Yang, Y., Zhao, Q., Lin, X., and Wang, H. (2023). Automated identification of gastritis and intestinal metaplasia via deep learning in esophagogastroduodenoscopy. *Frontiers in Oncology*, 13:1007237.

Zhang, M., Wang, J., Liu, Y., Chen, X., and Zhao, L. (2023). Gastric lesion classification and early cancer detection in endoscopic images using multi-scale convolutional networks. *IEEE Access*, 11:28450–28461.