QLabelMIL: Inter-Pathology Query Decoding for Multi-Label Gastric Histopathology

Pedro C. Neto[1,3,4][0000−0003−1333−4889], Rita N. Lopes[3,4], and Lígia Prado e Castro[2]

[1] Unilabs.AI, Campus Biotech, Geneva, Switzerland
[2] Unilabs, Campus Biotech, Geneva, Switzerland
[3] Faculdade de Engenharia da Universidade do Porto
[4] Equal contribution

Abstract. Gastric cancer precursor conditions, such as H. pylori infection, intestinal metaplasia, and atrophy, frequently co-occur in patients, as modeled by the Correa Cascade. However, current models for whole-slide image classification are not optimized for multi-label settings, failing to model potentially useful inter-pathology dependencies or provide per-class interpretability. We introduce QLabelMIL, a multiple-instance learning aggregator that uses class-specific queries and a transformer decoder to model patch-to-label relationships, while natively supporting joint multi-label classification and per-class heatmaps. A Graph Convolutional Network variant further leverages the prior distribution of label co-occurrences to score per-label features. It is evaluated on one of the largest gastric histopathology datasets, comprising 5764 WSIs from 3412 patients, against several established MIL frameworks like CLAM, TransMIL, and MambaMIL. QLabelMIL achieved better predictive performance with a macro-averaged AUROC of up to 0.9186 and competitive calibration metrics, offering an efficient and interpretable solution.

Keywords: Digital Pathology · Multiple Instance Learning · Gastric.

1 Introduction

Driven by growing concerns over the impact of gastric diseases [20], efforts increasingly focus on reducing the prevalence of gastric cancer, the fifth most common cancer worldwide and the fourth leading cause of cancer-related deaths [20]. Precursor conditions significantly impact the risk of developing this disease [14, 17, 20]. For instance, *Helicobacter pylori* (H. pylori), a stomach bacterium, is classified as a Group 1 carcinogen by the World Health Organization (WHO) [5] and causes an estimated 76% [14] to 90% [20] of all gastric cancer cases. H. pylori infections are also known to promote the development of other pre-cancerous conditions, such as Chronic/Acute inflammation and Atrophy [8, 15, 23, 24]. The latter is also a main cause for Intestinal Metaplasia [3], which is one of the conditions leading to dysplasia and cancer [3]. This dependency and co-occurrence is known as the Correa Cascade [3]. It has been shown that within the same stomach, several of these conditions can manifest simultaneously [2].

Routine H&E histology of stomach biopsies remains the gold standard diagnosis for most of these conditions [16]. The exponential growth of digital pathology data [12], datasets [13], and models [11] creates an ideal environment for Artificial Intelligence (AI) to improve clinical accuracy and turnaround times. However, research in AI methods for gastric pathology has been rather limited, conducted on rather small datasets [19, 7], and without developing novel strategies to tackle some of its challenges. A notorious flaw is that current methods are not designed for multi-label problems, restricting their ability to model inter-pathology dependencies. Often, literature proposals focus on using one aggregator per condition [21]. This creates an obvious computational burden, does not model class dependencies, and fails to leverage shared information across classes, which might benefit the modeling of less prevalent conditions. Other approaches focus on a single shared representation, which penalizes less frequent conditions and removes interpretability by preventing per-class heatmap creation.

To tackle these gaps, we propose QLabelMIL, inspired by the Query2Label transformer [9]. Our proposed architecture models inter-class dependencies, creates class-specific queries and per-class attention maps, and is specifically designed to improve multi-label classification. Table 1 summarizes our architecture's mechanisms compared to existing literature.

Through this document, we describe two main contributions: 1) the introduction of QLabelMIL, a novel, multi-label friendly multiple-instance learning aggregator; 2) the benchmarking of this aggregator and established literature models on one of the largest gastric datasets documented.

Table 1. Distinctive mechanisms of **QLabelMIL** compared to other methods.

Mechanism	ABMIL	CLAM	ACMIL	MambaMIL	TransMIL	**QLabelMIL**
Per-class attention maps	✗	✓	◑	✗	◑	✓
Class co-occurrence modeling	✗	✗	✗	✗	✗	✓
Label-relationship graph	✗	✗	✗	✗	✗	✓
Joint multi-label	◑	◑	◑	◑	◑	✓
Single multi-label model by design	✗	✗	✗	✗	✗	✓

✓ native ◑ partial ✗ absent.

2 Related Work

Previous works address whole-slide image classification through various Multiple Instance Learning (MIL) architectures.

Baseline architectures may be split into models that process instances independently and those that model spatial context. ABMIL [6] introduces the mechanism of aggregating instances via a trainable attention network. CLAM [10] extends to parallel class-specific attention branches and an instance-level clustering objective. ACMIL [25] utilizes multiple branches with diversity regularization and stochastic masking to capture distinct visual patterns. These architectures are computationally lightweight, but they share the limitation of aggregating

patches independently and do not model spatial or sequential relationships. Furthermore, in a multi-label context, ABMIL and ACMIL are unable to generate class-specific heatmaps, as they calculate global and pattern-specific attention weights, respectively. CLAM resolves this limitation via class-specific branches.

To capture spatial context, TransMIL [18] uses a Transformer with approximated self-attention, while MambaMIL [22] integrates Selective Scan Space State Sequential modules with sequence reordering. However, TransMIL is memory intensive, especially at higher magnifications. MambaMIL achieves linear complexity to resolve this, but its sequence reordering introduces randomness, causing training instability that requires dataset-specific tuning. Furthermore, because TransMIL models patch-to-patch relationships and MambaMIL uses global attention for aggregation, neither can generate class-specific heatmaps in multi-label settings.

In gastric histopathology, GastritisMIL [21] targets the detection and severity grading of chronic and acute gastritis, intestinal metaplasia, and atrophy. Combining a ResNet50 feature extractor with an attention-based aggregation and four downstream tasks, the system reports high internal performance (AUROC: 0.971–0.984). However, their reliance on separate models and removal of large portions of the dataset from use for each pathology may suggest the need to pre-select images for each condition rather than having a continuous workflow for the detection of multiple precursors. Similarly, GasMIL [4] utilizes a semi-supervised, multi-scale MIL framework to diagnose inflammation, acute activity, intestinal metaplasia, and atrophy. While the model had strong performance (AUROC: 0.877–0.981) and improved pathologist performance in observer studies, it shares the same main limitations as GastritisMIL. By relying on independent models for each condition, these systems increase overhead and prevent joint feature learning between naturally interdependent tasks.

3 Methodology

3.1 Problem Definition and Dataset

Initiated primarily by H. pylori, the Correa Cascade models the progression of several gastric conditions [3, 8]. The most frequent sequence is H. pylori $\rightarrow$ chronic inflammation $\rightarrow$ atrophy $\rightarrow$ IM $\rightarrow$ dysplasia $\rightarrow$ cancer, providing insights into the risk of each condition.

To tackle these diagnoses, a large cohort of multi-label data was collected. The dataset comprises 5764 whole-slide images (WSIs) from 3412 patients undergoing gastric biopsies across several institutions, with slide-level labels extracted from clinical reports. All slides were scanned at the same facility.

After scanning, Otsu thresholding extracted the foreground tissue. This tissue was tiled into non-overlapping 40$\times$ patches. Concentric with each 40$\times$ patch, 10$\times$ patches were extracted to form an overlapping grid. Finally, feature embeddings were extracted at both magnifications using pre-trained ResNet18 and ResNet50 backbones. The dataset was then partitioned at the patient level into training (70%), validation (15%), and test (15%) sets.

Table 2. Dataset composition and inter-pathology co-occurrence matrix.

Condition	Total Positives	Co-occurrence Matrix					
		IM	CI	AI	DY	HP	AT
Intestinal Metaplasia (IM)	1372	-	1285	16	8	390	1165
Chronic Inflam. (CI)	4882	1285	-	30	8	2120	1682
Acute Inflam. (AI)	32	16	30	-	0	27	25
Dysplasia (DY)	11	8	8	0	-	0	8
H. pylori (HP)	2145	390	2120	27	0	-	568
Atrophy (AT)	1838	1165	1682	25	8	568	-

As these conditions represent stages of a continuous cascade, they frequently co-occur. Table 2 details the positive cases and co-occurrence frequencies across the dataset. Chronic Inflammation shows a high number of positives without representative negatives, whereas Dysplasia and Acute Inflammation lack positives. While this might occur because dysplasia is rare, for Acute Inflammation we suspect it stems from a lack of standardization in SNOMED-code [1] reporting. Due to these imbalances, models trained exclusively for the IM, H. pylori and Atrophy classes are the focus of the analysis. Additional experiments on 6 classes are still shared for analysis.

3.2 QLabelMIL

The proposed architecture, summarized in Figure 1, is mainly composed of two stages: a shared transformer decoder and a configurable classification head.

Shared query-decoder. Initially, a frozen encoder processes a whole-slide image to produce the raw patch features $\mathbf{H} \in \mathbb{R}^{N \times d_e}$, where N is the number of patches and d_e is the encoder embedding dimension. These features are then projected into a d-dimensional space using a learnable weight matrix $\mathbf{W}_p$ and layer normalization (LN). The resulting patch tokens $\mathbf{P}$ are then processed in the per-label query decoder.

In the decoder, C learnable label queries, where C represents the number of target classes, are initialized as a matrix $\mathbf{E}$. These queries are then processed iteratively by L decoder layers ($\ell = 1, \ldots, L$). Within each layer, the queries undergo multi-head self-attention (MHA) to capture inter-label relationships, followed by cross-attention, where patch features acting as keys and values allow for label-patch relationships to be modeled. The output is passed through a feed-forward network (FFN), with residual connections and layer normalization applied at each step:

$$\mathbf{P} = \mathrm{LN}\big(\mathbf{H}\mathbf{W}_p\big) \in \mathbb{R}^{N \times d}, \qquad \mathbf{Q}^{(0)} = \mathbf{E} \in \mathbb{R}^{C \times d} \tag{1}$$

$$\mathbf{Q}' = \mathbf{Q}^{(\ell-1)} + \mathrm{MHA}\big(\mathrm{LN}\,\mathbf{Q}^{(\ell-1)}, \mathrm{LN}\,\mathbf{Q}^{(\ell-1)}, \mathrm{LN}\,\mathbf{Q}^{(\ell-1)}\big), \tag{2}$$

$$\mathbf{Q}'' = \mathbf{Q}' + \mathrm{MHA}\big(\mathrm{LN}\,\mathbf{Q}', \mathbf{P}, \mathbf{P}\big), \tag{3}$$

$$\mathbf{Q}^{(\ell)} = \mathbf{Q}'' + \mathrm{FFN}\big(\mathrm{LN}\,\mathbf{Q}''\big), \tag{4}$$

This process produces per-label features $\mathbf{Z} = \mathrm{LN}\big(\mathbf{Q}^{(L)}\big) = [\mathbf{z}_1, \ldots, \mathbf{z}_C]^\top \in \mathbb{R}^{C \times d}$, where $\mathbf{z}_c \in \mathbb{R}^d$ is the representation of pathology c.

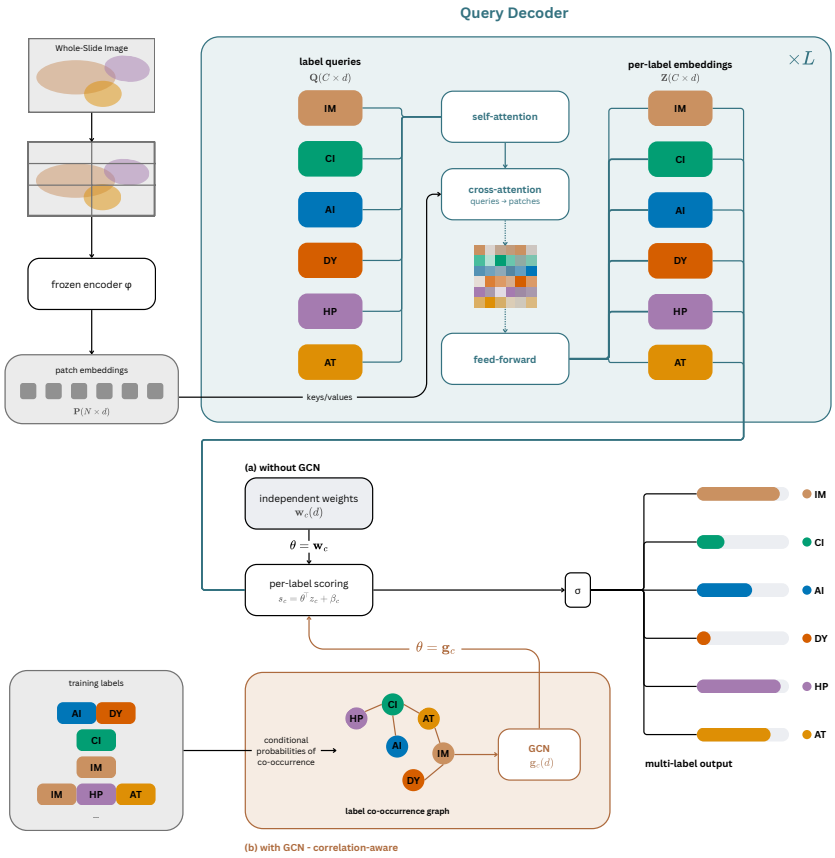

Fig. 1. QLabelMIL architecture. A frozen encoder extracts WSI patch embeddings (**P**) as keys and values. In the query-decoder, learnable label queries (**Q**) undergo self-attention for inter-label relationships, and cross-attention with embeddings for per-label features (**Z**). The classif. head evaluates these via (a) independent weights ($\mathbf{w}_c$), or (b) a GCN processing a label co-occurrence graph for interdependent weights ($\mathbf{g}_c$).

QLabelMIL (without GCN). Each label is scored by an *independent* grouped-linear classifier. The logit s_c for class c is computed via the inner product of the class representation $\mathbf{z}_c$ and a dedicated weight $\mathbf{w}_c$, offset by a bias β_c. A sigmoid activation function (σ) then maps this to predicted probability $\hat{p}_c$:

$$s_c = \mathbf{w}_c^\top \mathbf{z}_c + \beta_c, \qquad \hat{p}_c = \sigma(s_c), \qquad c = 1, \ldots, C, \tag{5}$$

Per-label weights $\{\mathbf{w}_c \in \mathbb{R}^d\}_{c=1}^C$ are learned independently, so no information is shared across labels in this step.

QLabelMIL (with GCN). To leverage the prior distribution of label co-occurrence, in this variant the independent weights are replaced by *correlation-aware* scores, produced by a Graph Convolutional Network (GCN).

First, a label co-occurrence graph is constructed from the training labels $\mathbf{y}^{(n)} \in \{0,1\}^C$. From the marginal co-occurrence count between two classes, defined as $M_{cc'} = \sum_n y_c^{(n)} y_{c'}^{(n)}$, conditional probabilities $P_{cc'} = M_{cc'}/M_{c'c'} = \Pr(c \mid c')$ are calculated. Then, edges in the graph are formed between two classes if $P_{cc'}$ meets a predefined threshold τ, to prevent modeling spurious correlations. Finally, for each label, a weight of p is shared uniformly among connected neighbors, maintaining a weight of $1-p$ on itself. Entries of the resulting adjacency matrix $\widehat{\mathbf{A}}$ are therefore defined as:

$$\widehat{A}_{cc'} = \begin{cases} \dfrac{p}{\sum_{k \neq c} \mathbb{1}[P_{ck} \geq \tau]}, & c \neq c',\ P_{cc'} \geq \tau, \\ 1-p, & c = c', \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Randomly initialized node embeddings $\mathbf{G}^{(0)} \in \mathbb{R}^{C \times d_0}$ are then passed through L_g graph convolution layers. At each layer ℓ, they are multiplied by a learnable weight matrix $\boldsymbol{\Theta}^{(\ell)}$ and by the adjacency matrix $\widehat{\mathbf{A}}$, and passed through a LeakyReLU activation function (δ):

$$\mathbf{G}^{(\ell)} = \delta\big(\widehat{\mathbf{A}}\, \mathbf{G}^{(\ell-1)} \boldsymbol{\Theta}^{(\ell)}\big), \qquad [\mathbf{g}_1, \ldots, \mathbf{g}_C]^\top = \mathbf{G}^{(L_g)} \in \mathbb{R}^{C \times d}, \tag{7}$$

The final embeddings are then used to score the per-label features. Row c of the output matrix, denoted as $\mathbf{g}_c$, acts as the weight for evaluating class c, incorporating a scalar bias β_c:

$$s_c = \mathbf{g}_c^\top \mathbf{z}_c + \beta_c, \qquad \hat{p}_c = \sigma(s_c), \tag{8}$$

Comparison. The two variants differ only in the per-label classifier: (5) uses independent weights $\mathbf{w}_c$, whereas (8) uses $\mathbf{g}_c = [\,\mathbf{G}^{(L_g)}]_c$, which couples the classifiers through the co-occurrence graph (6). Both variants are trained using a Binary Cross Entropy (BCE) loss $\mathcal{L} = \frac{1}{C} \sum_c \mathrm{BCE}(\hat{p}_c, y_c)$.

3.3 Experimental Setup

For each model configuration, independent hyperparameter tuning was performed. Final models were initialized from scratch and trained for up to 200 epochs using the optimal hyperparameters, based on validation AUROC.

Experiments were conducted on a NVIDIA RTX 3080 GPU. Models optimized a binary cross-entropy loss alongside any architecture-specific criteria. Due to the data imbalance, class-specific positive weights were applied to the main criterion. These weights were calculated using the training split data and were defined by the ratio of negative to positive samples for each condition.

Because each slide has a variable number of patches, a batch size of 1 was enforced, but gradient accumulation was included as a tunable option. Optimization was performed using AdamW with a sequential learning rate scheduler that combined a linear warmup with cosine annealing. Finally, to prevent overfitting, early stopping terminated training if the validation AUROC failed to improve upon its best value for 20 epochs.

4 Results

Given the nature of the dataset, we have prepared two sets of experiments. For the first, we trained the models on a 3-class problem where we account exclusively for Atrophy, H. pylori and Intestinal Metaplasia. We then added a second experiment with broader scope on the classification of the 6 classes available in the dataset.

4.1 Training on a 3-Class problem

Table 3. Diagnostic performance (AUROC). Best and second-best values per column within each pathology are **bolded** and underlined, respectively.

Pathology	Architecture	10× Magnification		40× Magnification	
		ResNet18	ResNet50	ResNet18	ResNet50
Intestinal Metaplasia	ABMIL	0.9352	0.9356	0.9326	0.9433
	CLAM	0.9356	0.9377	0.9456	0.9528
	ACMIL	0.9269	0.9380	**0.9489**	0.9513
	MambaMIL	0.9327	0.9469	0.9453	0.9570
	TransMIL	0.9289	0.9377	0.9448	0.9495
	QLabelMIL	**0.9427**	**0.9496**	0.9456	**0.9604**
	QLabelMIL (GCN)	0.9342	0.9459	0.9430	0.9543
H. pylori	ABMIL	0.8986	0.8969	0.9079	0.9051
	CLAM	0.9126	0.9149	0.9148	0.9120
	ACMIL	0.9088	**0.9236**	0.9018	0.9038
	MambaMIL	0.9008	0.9120	0.9034	0.9112
	TransMIL	0.9067	0.9179	0.9049	0.9114
	QLabelMIL	0.9082	0.9204	0.9163	0.9157
	QLabelMIL (GCN)	**0.9147**	0.9117	**0.9194**	**0.9204**
Atrophy	ABMIL	**0.8714**	0.8654	0.8575	0.8660
	CLAM	0.8601	0.8618	0.8560	0.8677
	ACMIL	0.8498	0.8727	0.8600	0.8612
	MambaMIL	0.8563	0.8730	0.8453	0.8714
	TransMIL	0.8606	0.8827	0.8627	**0.8735**
	QLabelMIL	0.8643	**0.8857**	**0.8632**	0.8684
	QLabelMIL (GCN)	0.8642	0.8721	0.8595	0.8711
Macro-Average	**ABMIL**	0.9017	0.8993	0.8993	0.9048
	CLAM	0.9028	0.9048	0.9055	0.9108
	ACMIL	0.8952	0.9114	0.9036	0.9054
	MambaMIL	0.8966	0.9107	0.8980	0.9132
	TransMIL	0.8987	0.9128	0.9041	0.9115
	QLabelMIL	**0.9051**	**0.9186**	**0.9083**	0.9148
	QLabelMIL (GCN)	0.9044	0.9099	0.9073	**0.9153**

QLabelMIL achieved the best macro-averaged performance across all configurations (Table 3). Across the four magnification and backbone combinations, the proposed models outperformed all architectures in every condition in at least two settings. When that was not the case, a QLabelMIL architecture typically got the second-best result, and first and second-best metrics were frequently shared between the two proposed variants. The naive variant proved to have the best diagnostic performance overall, although the GCN brought advantages mostly for H. pylori.

Overall, intestinal metaplasia achieved the best performance across architectures, whereas atrophy consistently performed worst. H. pylori metrics exhibited noticeable gains with QLabelMIL compared to those of most baselines. As expected, ResNet50 generally outperformed ResNet18, though this gap was narrower for atrophy. Atrophy was uniquely advantaged by 10× magnification, whereas metaplasia and H. pylori models performed better at 40×.

Table 4. Calibration (ECE). Best and second-best values per column within each pathology are **bolded** and underlined, respectively.

Pathology	Architecture	10× Magnification		40× Magnification	
		ResNet18	ResNet50	ResNet18	ResNet50
Intestinal Metaplasia	ABMIL	**0.0538**	0.1426	0.2681	**0.0353**
	CLAM	0.1348	0.1093	**0.0445**	0.0386
	ACMIL	0.0721	0.0972	0.1878	0.0994
	MambaMIL	0.1355	0.1256	0.0672	0.0419
	TransMIL	0.1847	0.0649	0.0969	0.1301
	QLabelMIL	0.0728	**0.0432**	0.0585	0.0427
	QLabelMIL (GCN)	0.1703	0.0837	0.0878	0.0735
H. pylori	ABMIL	0.1669	0.0396	0.0766	0.0821
	CLAM	0.0574	0.0404	0.0506	0.0457
	ACMIL	0.0604	**0.0395**	0.0420	0.0743
	MambaMIL	0.0567	0.0686	0.0489	0.0539
	TransMIL	0.0858	0.0652	0.0430	**0.0309**
	QLabelMIL	0.0467	0.0684	0.0723	0.0715
	QLabelMIL (GCN)	**0.0304**	0.0717	**0.0417**	0.0821
Atrophy	ABMIL	0.0544	0.1831	0.3062	**0.0506**
	CLAM	0.1314	0.1659	0.0733	0.0648
	ACMIL	0.0827	**0.0983**	0.2209	0.1531
	MambaMIL	0.1687	0.1210	**0.0650**	0.1132
	TransMIL	0.2146	0.1481	0.1649	0.1590
	QLabelMIL	**0.0530**	0.1112	0.1035	0.1100
	QLabelMIL (GCN)	0.1390	0.2029	0.0913	0.1091
Macro-Average	**ABMIL**	0.0917	0.1217	0.2170	0.0560
	CLAM	0.1079	0.1052	**0.0562**	**0.0497**
	ACMIL	0.0717	0.0784	0.1502	0.1089
	MambaMIL	0.1203	0.1051	0.0604	0.0697
	TransMIL	0.1617	0.0927	0.1016	0.1067
	QLabelMIL	**0.0575**	**0.0743**	0.0781	0.0747
	QLabelMIL (GCN)	0.1133	0.1195	0.0736	0.0882

Regarding calibration (Table 4), QLabelMIL obtained the best macro-averaged ECE in two of the four configurations and achieved the best metrics for every condition in at least one setting. The remaining configurations were mostly competitive with the best-performing architectures. The two proposed variants exhibit similar tendencies to those previously mentioned.

Unlike predictive performance, in calibration, H. pylori was the condition for which models achieved the best metrics, although results appeared less consistent across configurations for all pathologies. The difference between the two proposed backbones was less pronounced, with the smaller backbone achieving the best results in several instances.

Table 5. Comparison of diagnostic performance across architectures, measured by AUROC for six target pathologies. Best values per column within each pathology are **bolded**, and second best values are underlined. AUROC: Area Under the Receiver Operating Characteristic curve; GCN: Graph Convolutional Network.

Pathology	Architecture	10× Magnification		40× Magnification	
		ResNet18	ResNet50	ResNet18	ResNet50
Intestinal Metaplasia	ABMIL	0.9154	0.9180	0.8908	0.9241
	CLAM	**0.9314**	0.9131	0.8849	0.9052
	ACMIL	0.9197	0.9145	0.9361	0.9472
	MambaMIL	0.9190	0.9337	0.8967	0.9258
	TransMIL	0.9043	0.9124	0.9074	0.9321
	QLabelMIL	0.9253	**0.9451**	0.9419	**0.9577**
	QLabelMIL (GCN)	0.9212	0.9348	**0.9462**	0.9539
Chronic Inflammation	ABMIL	**0.8202**	0.8141	**0.8110**	0.7594
	CLAM	0.7915	0.8064	0.7807	0.7869
	ACMIL	0.7986	0.8009	0.8050	0.7821
	MambaMIL	0.7968	0.7748	0.7848	0.7837
	TransMIL	0.7666	0.7783	0.7758	0.7801
	QLabelMIL	0.7894	0.8137	0.7978	0.7776
	QLabelMIL (GCN)	0.8034	**0.8154**	0.7753	**0.7876**
Acute Inflammation	ABMIL	0.7600	0.7517	0.8234	0.8234
	CLAM	0.8106	0.7790	**0.9047**	**0.8906**
	ACMIL	**0.8607**	**0.8247**	0.8220	0.8797
	MambaMIL	0.6700	0.6358	0.6758	0.8011
	TransMIL	0.8054	0.7481	0.8145	0.8597
	QLabelMIL	0.7339	0.7764	0.7811	0.7795
	QLabelMIL (GCN)	0.7632	0.8234	0.7705	0.8512
Dysplasia	ABMIL	0.9694	0.8382	0.9565	0.8974
	CLAM	0.9618	0.9444	**0.9859**	0.9294
	ACMIL	0.9724	0.9741	0.8976	0.9671
	MambaMIL	0.8644	**0.9771**	0.9659	0.9221
	TransMIL	0.8976	0.9082	0.8753	0.9162
	QLabelMIL	**0.9829**	0.9565	0.9500	0.9718
	QLabelMIL (GCN)	0.9588	0.6994	0.9029	**0.9729**
H. pylori	ABMIL	0.8955	0.8927	0.8954	0.8788
	CLAM	0.9001	0.9025	0.8934	0.8993
	ACMIL	**0.9070**	0.8912	0.9101	0.8834
	MambaMIL	0.8948	0.8865	0.8905	0.8886
	TransMIL	0.8912	0.8812	0.8938	0.8937
	QLabelMIL	0.8938	0.9064	0.9048	0.8970
	QLabelMIL (GCN)	0.8977	**0.9121**	**0.9102**	**0.9020**
Atrophy	ABMIL	0.8554	0.8719	0.8388	0.8602
	CLAM	0.8594	0.8668	0.8373	0.8460
	ACMIL	**0.8601**	0.8533	0.8529	0.8541
	MambaMIL	0.8515	**0.8744**	0.8322	0.8512
	TransMIL	0.8405	0.8456	0.8309	0.8566
	QLabelMIL	0.8498	0.8728	**0.8536**	**0.8702**
	QLabelMIL (GCN)	0.8527	0.8603	0.8520	0.8617
Macro-Average	**ABMIL**	0.8693	0.8478	0.8693	0.8572
	CLAM	0.8758	0.8687	**0.8811**	0.8762
	ACMIL	**0.8864**	0.8764	0.8706	0.8856
	MambaMIL	0.8328	0.8470	0.8410	0.8621
	TransMIL	0.8509	0.8456	0.8496	0.8731
	QLabelMIL	0.8625	**0.8785**	0.8715	0.8756
	QLabelMIL (GCN)	0.8662	0.8409	0.8595	**0.8882**

4.2 Training on a 6-Class problem

Differently from a 3-class problem with relatively balanced number of samples, the presence of highly imbalanced classes in a 6-class setting seems to be a harder task. With QLabelMIL variations holding the best performance in three of the four tested configurations when considering their performance on Intestinal Metaplasia, H. Pylori and Atrophy, the remaining classes were shown to be much harder. In fact, if we consider the evaluation of all six classes (Table 5), QLabelMIL is only the best solution in two out of four configurations with a marginal advantage. Hence, in such an adversarial setting, where data is not appropriate to train an artificial intelligence model, the advantages of our method reside mostly on its ability to generate condition-specific heatmaps while being competitive or slightly better than other models in the literature.

In summary, our proposed architecture proved to bring advantages in predictive performance while being competitive in terms of calibration when compared to other commonly used MIL frameworks for pathology. The presence on unbalanced classes seems to also hurt our model, without advantages from the modelling of co-occurrences. Yet, on a 3-class scenario where all classes are to some extent balanced, the performance is, in some cases, far superior to the other methods.

QLabelMIL brings enhanced interpretability without performance trade-offs, achieving top performance in the vast majority of the training configurations.

5 Conclusion

The proposed QLabelMIL architecture natively supports multi-label settings and generation of class-specific heatmaps. Its modeling of label co-occurrences proved to bring advantages to the gastric cancer precursor setting of the Correa cascade, and we expect to validate it on more datasets and problems. The efficiency of the method compared to typical Transformer-based architectures is also promising and should be explored in the future.

Future work should evaluate this aggregator on different pathology specialties, more balanced datasets, and with more aggressive tuning of the proposed method's hyperparameters. From the performance gains brought by the increased backbone capacity, we also expect foundation models to further boost the performance of the aggregator. In addition, a mixed magnification setup could benefit some conditions, such as atrophy. Our next step would therefore be to consider magnification fusion.

References

1. Bos, L., Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Stud Health Technol Inform **121**, 279–290 (2006)
2. Chen, Y., He, L., Zheng, X.: Characteristics of multiple early gastric cancer and gastric high-grade intraepithelial neoplasia. Medicine **102**(49), e36439 (2023)

3. Correa, P., Piazuelo, M.B.: The gastric precancerous cascade. Journal of digestive diseases **13**(1), 2–9 (2012)
4. Fang, S., Liu, Z., Qiu, Q., Tang, Z., Yang, Y., Kuang, Z., Du, X., Xiao, S., Liu, Y., Luo, Y., Gu, L., Tian, L., Liang, X., Fan, G., Zhang, Y., Zhang, P., Zhou, W., Liu, X., Tian, J., Wei, W.: Diagnosing and grading gastric atrophy and intestinal metaplasia using semi-supervised deep learning on pathological images: development and validation study. Gastric Cancer **27**(2), 343–354 (Mar 2024). https://doi.org/10.1007/s10120-023-01451-9, https://doi.org/10.1007/s10120-023-01451-9
5. Forman, D.: Helicobacter pylori and gastric cancer. Scandinavian Journal of Gastroenterology **31**(sup215), 48–51 (1996)
6. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based Deep Multiple Instance Learning. CoRR **abs/1802.04712** (2018), http://arxiv.org/abs/1802.04712, arXiv: 1802.04712
7. Liscia, D.S., D'Andrea, M., Biletta, E., Bellis, D., Demo, K., Ferrero, F., Petti, A., Butinar, R., D'Andrea, E., Davini, G.: Use of digital pathology and artificial intelligence for the diagnosis of Helicobacter pylori in gastric biopsies. Pathologica **114**(4), 295–303 (Aug 2022). https://doi.org/10.32074/1591-951X-751
8. Liu, Q., Tang, J., Chen, S., Hu, S., Shen, C., Xiang, J., Chen, N., Wang, J., Ma, X., Zhang, Y., et al.: Berberine for gastric cancer prevention and treatment: Multi-step actions on the correa's cascade underlie its therapeutic effects. Pharmacological research **184**, 106440 (2022)
9. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)
10. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images (May 2020). https://doi.org/10.48550/arXiv.2004.09666, http://arxiv.org/abs/2004.09666, arXiv:2004.09666 [eess]
11. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
12. Neto, P.C., Montezuma, D., Oliveira, S.P., Oliveira, D., Fraga, J., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Reinhard, S., et al.: An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. NPJ precision oncology **8**(1), 56 (2024)
13. Oliveira, S.P., Montezuma, D., Moreira, A., Oliveira, D., Neto, P.C., Monteiro, A., Monteiro, J., Ribeiro, L., Gonçalves, S., Pinto, I.M., et al.: A cad system for automatic dysplasia grading on h&e cervical whole-slide images. Scientific Reports **13**(1), 3970 (2023)
14. Park, J.Y., Georges, D., Alberts, C.J., Bray, F., Clifford, G., Baussano, I.: Global lifetime estimates of expected and preventable gastric cancers across 185 countries. Nature Medicine **31**(9), 3020–3027 (2025)
15. Pennelli, G., Grillo, F., Galuppini, F., Ingravallo, G., Pilozzi, E., Rugge, M., Fiocca, R., Fassan, M., Mastracci, L.: Gastritis: update on etiological features and histological practical approach. Pathologica **112**(3), 153 (2020)
16. Pimentel-Nunes, P., Libânio, D., Marcos-Pinto, R., Areia, M., Leja, M., Esposito, G., Garrido, M., Kikuste, I., Megraud, F., Matysiak-Budnik, T., Annibale, B., Dumonceau, J.M., Barros, R., Fléjou, J.F., Carneiro, F., van Hooft, J.E., Kuipers, E.J., Dinis-Ribeiro, M.: Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy

(ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy **51**(4), 365–388 (Apr 2019). https://doi.org/10.1055/a-0859-1883
17. Shah, S.C., Piazuelo, M.B., Kuipers, E.J., Li, D.: Aga clinical practice update on the diagnosis and management of atrophic gastritis: expert review. Gastroenterology **161**(4), 1325–1332 (2021)
18. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., others: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
19. Steinbuss, G., Kriegsmann, K., Kriegsmann, M.: Identification of Gastritis Subtypes by Convolutional Neuronal Networks on Histological Images of Antrum and Corpus Biopsies. International Journal of Molecular Sciences **21**(18), 6652 (2020). https://doi.org/10.3390/ijms21186652
20. Thrift, A.P., Wenker, T.N., El-Serag, H.B.: Global burden of gastric cancer: epidemiological trends, risk factors, screening and prevention. Nature reviews Clinical oncology **20**(5), 338–349 (2023)
21. Xia, K., Hu, Y., Cai, S., Lin, M., Lu, M., Lu, H., Ye, Y., Lin, F., Gao, L., Xia, Q., Tian, R., Lin, W., Xie, L., Tan, D., Lu, Y., Lin, X., Yang, X., Zhong, L., Xu, L., Zhang, Z., Wang, L., Ren, J., Xu, H.: GastritisMIL: An interpretable deep learning model for the comprehensive histological assessment of chronic gastritis. Patterns **6**(8) (Aug 2025). https://doi.org/10.1016/j.patter.2025.101286, https://doi.org/10.1016/j.patter.2025.101286, publisher: Elsevier
22. Yang, S., Wang, Y., Chen, H.: MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 296–306. Springer Nature Switzerland, Cham (2024)
23. Yin, Y., Liang, H., Wei, N., Zheng, Z.: Prevalence of chronic atrophic gastritis worldwide from 2010 to 2020: an updated systematic review and meta-analysis. Annals of Palliative Medicine **11**(12), 3697703–3693703 (2022)
24. Zhang, S.L., Lollie, T.K., Chen, Z., Narasimhalu, T., Wang, H.L.: Histopathologic diagnosis of gastritis and gastropathy: A narrative review. Digestive Medicine Research **6** (2023)
25. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification (2024), https://arxiv.org/abs/2311.07125, _eprint: 2311.07125