Scientific Reports

https://doi.org/10.1038/s41598-025-32737-w

Article in Press

Towards deep-learning based detection and quantification of intestinal metaplasia on digitized gastric biopsies: a multi-expert comparative study

Received: 22 July 2025

Accepted: 11 December 2025

Published online: 26 February 2026

Cite this article as: Cano F., Caviedes M., Siabatto A. *et al.* Towards deep-learning based detection and quantification of intestinal metaplasia on digitized gastric biopsies: a multi-expert comparative study. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-32737-w

Fabian Cano, Mauricio Caviedes, Andres Siabatto, Jesus Villarreal, Jose Quijano, Álvaro Bedoya-Urresta, Marino Coral Bedoya, Yomaira Yepez Caicedo, Angel Cruz-Roa, Fabio A. González, Satish E. Viswanath & Eduardo Romero

Towards deep-learning based detection and quantification of intestinal metaplasia on digitized gastric biopsies: a multi-expert comparative study

Fabian Cano[1], Mauricio Caviedes[1], Andres Siabatto[1], Jesus Villarreal[1], Jose Quijano[1], Álvaro Bedoya-Urresta[5], Marino Coral Bedoya[5], Yomaira Yepez Caicedo[5], Angel Cruz-Roa[3], Fabio A. González[2], Satish E. Viswanath[4], and Eduardo Romero[1,*]

[1]Computer Imaging and Medical Applications Laboratory - Cim@lab, Universidad Nacional de Colombia, Bogotá D.C., Colombia
[2]Machine Learning, Perception and Discovery Lab - Mindlab, Universidad Nacional de Colombia, Bogotá D.C., Colombia
[3]GITECX Research Group & Automatic Data-driven Analytics Laboratory - AdaLab, Universidad de los Llanos, Meta, Colombia
[4]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, United States
[5]Fundación Centro de Investigación de Enfermedades Digestivas y Nutricionales - CIEDYN, Pasto, Colombia
[*]Corresponding authors: edromero@unal.edu.co

ABSTRACT

Current gastric cancer (GCa) risk systems are prone to errors since they evaluate a visual estimation of intestinal metaplasia percentages in histopathology images of gastric mucosa to assign a risk. This study presents an automated method to detect and quantify intestinal metaplasia using deep convolutional neural networks as well as a comparative analysis with visual estimations of three pathologists. Gastric samples were collected from two different cohorts: 149 asymptomatic volunteers from a region with a high prevalence of GCa in Colombia and 56 patients from a tertiary hospital. Deep learning models were trained to classify intestinal metaplasia, and predictions were used to estimate a percentage of intestinal metaplasia and to assign an adapted OLGIM stage. Atrophy was not assessed because of the limited reproducibility among pathologists. Results were compared with independent blinded metaplastic assessments performed by three graduated pathologists. The best-performing deep learning architecture classified intestinal metaplasia with F1-Score of 0.80 ± 0.01 and AUC of 0.91 ± 0.01. Among pathologists, inter-observer agreement by a Fleiss's Kappa score ranged from 0.20 to 0.48. In comparison, agreement between the pathologists and the best-performing model ranged from 0.12 to 0.35. Deep learning models show potential to reliably detect and quantify the percentage of intestinal metaplasia, achieving high classification performance. In practice, visual estimation is still the only available method, yet it is marked by considerable inter-observer variability. Deep learning models provide consistent estimates that could help reduce this subjectivity in risk stratification.

Introduction

Gastric cancer (GCa) was the fifth most commonly diagnosed cancer and the fifth leading cause of cancer-related death worldwide in 2022[1]. Approximately 90% of GCa cases were diagnosed as adenocarcinomas, from which the most common type was intestinal adenocarcinoma[2]. The high mortality rate associated with GCa is closely related to asymptomatic progression, however, if diagnosed at early stages, patient survival improves considerably[3]. The way in which this cancer develops and progresses is not completely clear, probably a combination of genetic factors associated with bacterial aggressiveness, the *Helicobacter pylori*, and environmental or lifestyle factors[4]. Given the lack of effective strategies to cure GCa, early diagnosis remains the most promising strategy to reduce both incidence and mortality rates[5].

Currently, the screening protocol is guided by the Sydney System, introduced in 1991[6] and updated in 1994 as the Updated Sydney System (USS)[7]. This system recommends histological evaluation of at least five gastric biopsies obtained from both the antrum and corpus, three from lesser curvature and two from greater curvature. This estimator has been popularized in practice because the incisura angularis has been considered an area with higher risk and therefore has been included as an additional antrum biopsy[8], and some studies have observed premalignant lesions in this region[9]. Overall, *Helicobacter pylori*, mononuclear cells, loss of glands (atrophy), intestinal metaplasia, inflammation or dysplasia can be found in any of the five collected biopsies. Particular attention has been paid

to intestinal metaplasia which has been scored in different categories: 0 (absence), 1 (mild), 2 (moderate), and 3 (severe), a combination of the extent and topographic distribution of microscopic changes[10]. This type of lesion is considered a gastric adaptation to chronic infection with *Helicobacter pylori*[11] and a pivotal event in GCa progression, described as a *"point of no return"*[12]. The final step to establish a progression risk to GCa is to assign a stage[13]. Protocols developed to assign this risk are basically the Operative Link for Gastritis Assessment (OLGA) system[14] and the Operative Link on Gastric Intestinal Metaplasia (OLGIM) system[15] (specifically designed to stage the extent and distribution of intestinal metaplasia). Both systems classify patients from stage 0 (lowest risk) to stage IV (highest risk).

Recently, some concerns have been raised about the uncertainty and clinical implications of current gastric risk stratification[16,17] after the original USS version was introduced. Several publications have reported variability among general pathologists when assigning OLGIM stages[16,18,19] and they have documented that these systems may underestimate or overestimate cancer risk[9,18,20]. Although some studies suggest that OLGIM may stratify GCa risk[15,21], there is currently no robust quantitative evidence to support this statement[18,22]. Therefore, a more precise and reproducible method to quantify these premalignant lesions remains an unmet need[16]. In this context, automatic approaches are particularly appealing as they address many of the limitations associated with visual assessment. Deep learning, in particular, has demonstrated high accuracy for detecting and quantifying patterns in medical images[16], with outstanding results reported in lung cancer[23–25], breast cancer[26–28], prostate cancer[29–31], and gastric cancer[32–34]. Specifically in gastric cancer, deep learning models have shown improved diagnostic accuracy and reproducibility in histopathology images, while significantly reducing both inter- and intra-observer variability[35].

The current OLGIM risk assessment system, which routinely guides in treatment of gastric cancers[12,36], assigns a risk score by heuristically combining visual estimations of intestinal metaplasia from five different gastric biopsy sites. In this context, intestinal metaplasia remains a critical premalignant GCa stage that requires accurate and reproducible quantification[16]. Interestingly, some studies have reported cases classified as OLGIM low-risk who were later diagnosed with gastric cancer[22,37], while others identified as high-risk never developed cancer and, in some cases, reversed the lesion[38,39]. In summary, these investigations indicate that subjectively estimated GCa risk is prone to errors and dependent on expert evaluation[12].

In this study, an automatic deep learning approach is proposed to accurately detect and quantify intestinal metaplasia in hematoxylin and eosin (H&E)-stained images (see Figure 1). A main contribution consists in applying and adapting several state-of-the-art neural networks to compute the percentage of intestinal metaplasia, and their performance was compared with manual estimations made by three pathologists. Evaluation with two complementary cohorts, one from a high-risk population of Colombian volunteers and another from a tertiary hospital, allowed assessment of classification accuracy and robustness across different clinical contexts.

Materials and Methods

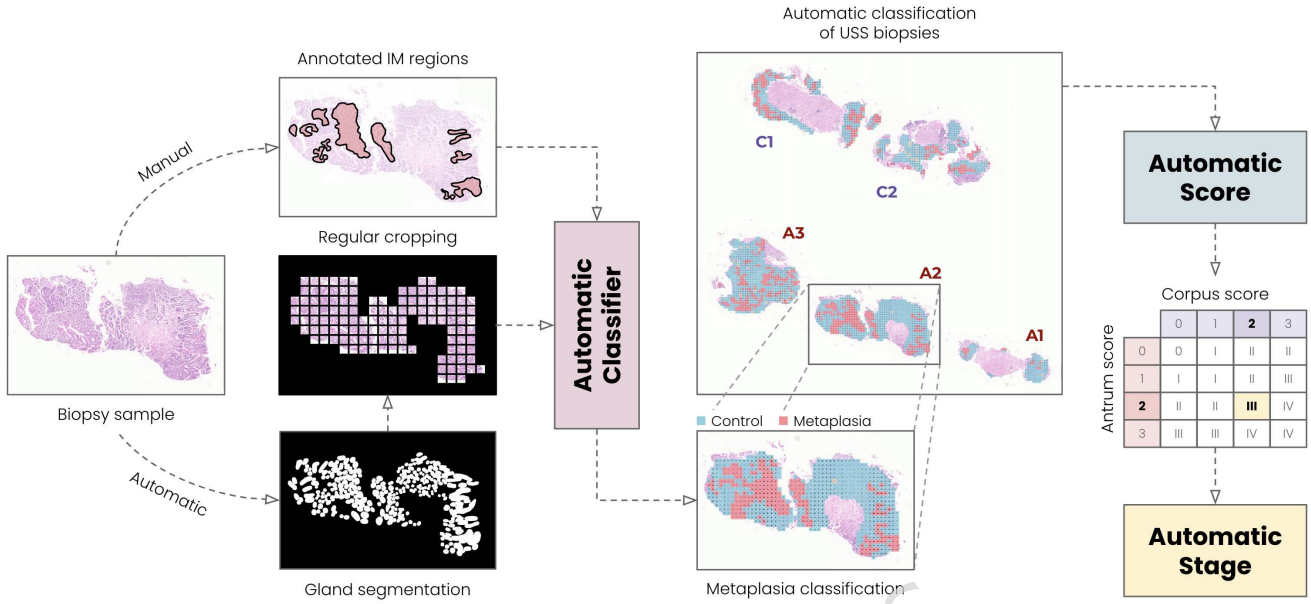

Figure 1. Workflow of the automatic intestinal metaplasia classification with deep learning models and automatic scoring using the proportion of intestinal metaplasia in both antrum and corpus biopsies.

Data Description

Gastric biopsy samples from two independent cohorts were digitized. The first cohort consists of 149 asymptomatic volunteers who were selected from an independent study provided by the *CIEDYN* foundation, a partner of the *Urkunina5000* project[40]. This study performed an endoscopic and pathological characterization of gastric disease in 5.000 volunteers from 55 small villages in *Nariño* Colombia, a region long observed as being GCa prevalent for the past two decades. Asymptomatic volunteers were men and women aged $30-70$ years, who had lived in these areas for at least the last 10 years and had never undergone an endoscopy procedure, nor been diagnosed with GCa, any other type of cancer, or any known gastric pathology. The second cohort included 56 patients from the *Hospital Universitario Nacional de Colombia* included in a larger study which was approved by the ethics committee (Act No. 007, Apr 29 2022). Patients were men and women who present symptoms suggestive of gastritis or other forms of gastric discomfort. However, none of them had a previous GCa diagnosis or any other type of lesion. A summary of these datasets is presented in Table 1.

Table 1. Demographic characteristics and OLGIM stage distribution for both data cohorts. The external test set corresponds exclusively to the second cohort.

Characteristics	**All cases** (N=205)	**Training** (N=73)	**Validation** (N=32)	**Internal Test** (N=44)	**External Test** (N=56)
Age (years)	53.68 ± 10.39	52.68 ± 10.68	56.62 ± 10.51	53.20 ± 9.61	64.73 ± 13.63
Female (%)	79 (53.02%)	44 (60.27%)	13 (40.62%)	22 (50.00%)	38 (67.86%)
OLGIM					
0	74 (36.10%)	16 (21.92%)	7 (21.88%)	10 (22.73%)	41 (73.21%)
1	74 (36.10%)	31 (42.47%)	14 (43.75%)	21 (47.73%)	8 (14.29%)
2	34 (16.59%)	14 (19.18%)	7 (21.88%)	8 (18.18%)	5 (8.93%)
3	14 (6.83%)	9 (12.33%)	2 (6.25%)	3 (6.82%)	0 (0.00%)
4	9 (4.39%)	3 (4.11%)	2 (6.25%)	2 (4.55%)	2 (3.57%)

Biopsy assessment

Five gastric biopsies were obtained according to the USS (Fig. 2) in both cohorts. Specifically, two biopsies were taken from the antrum (A_1, A_2), other two biopsies from the corpus (C_1, C_2), and a final biopsy from the incisura angularis (A_3) (antral-corpus transition zone). All biopsies were fixed in formalin and embedded in paraffin blocks and stained with H&E

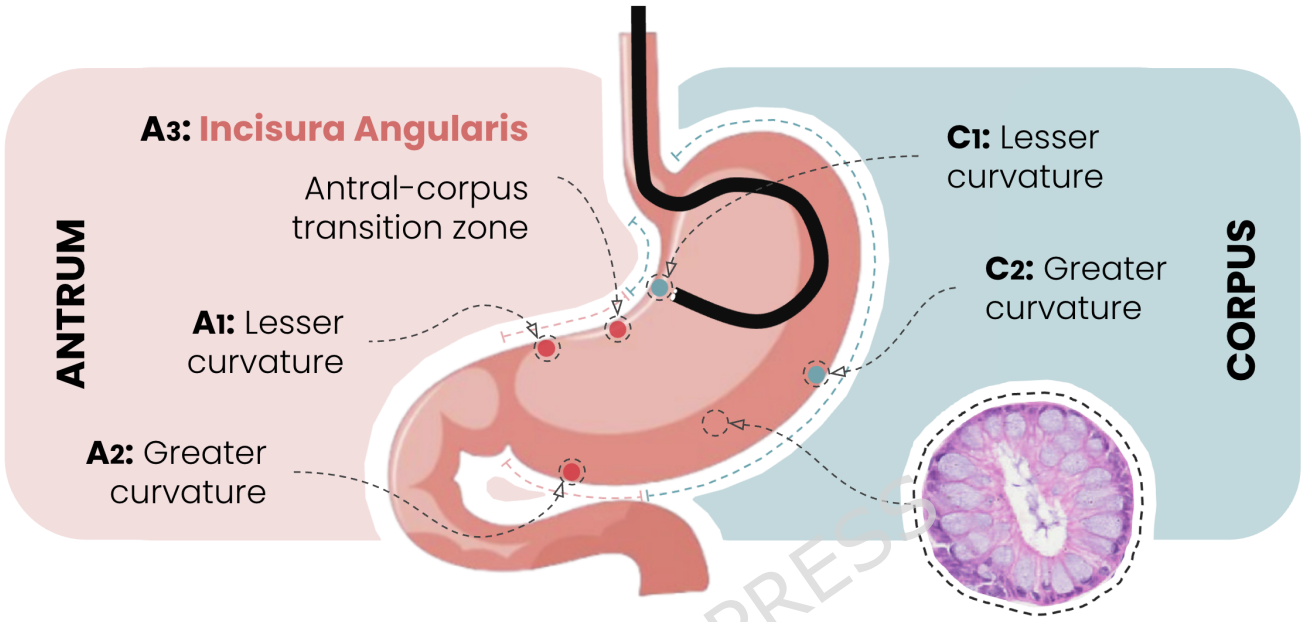

Figure 2. Biopsy places established for GCa detection according to the Updated Sydney System (USS). Antrum: Lesser curvature, Greater curvature and Antral-corpus transition zone (incisura angularis). Corpus: Lesser curvature and Greater curvature.

All cases were digitized using a whole-slide image (WSI) scanner (MoticEasyScan Pro), with a ×40 objective, corresponding to a spatial resolution of 0.255 μm per pixel (MPP). Each case comprised five biopsies (one per anatomical region of the stomach), resulting in a total of 205 WSI, i.e., 149 from the first cohort and 56 from the second. The first cohort was used to train and evaluate the deep learning models including 73 cases for training, 32 cases for validation and 44 cases for testing. The second cohort was used exclusively as an external testing set.

Each case was independently assessed by three experienced pathologists following the OLGIM system. This system assumes that variations regarding extent and topographical distribution of intestinal metaplasia reflect distinct clinic-pathological scenarios, each corresponding to a specific GCa risk. Pathologists performed blinded evaluations to estimate the proportion of intestinal metaplasia at five standardized biopsy sites: three from the mucosecreting area (two from the antrum and one from the incisura angularis) and two from the oxyntic mucosa (lesser and greater curvature of the corpus). After intestinal metaplasia was detected, a score was assigned as the estimated percentage of glands with goblet cells, ideally assessed in full-thickness (perpendicular) mucosa sections. Each biopsy was independently scored following a four-tier scale: absence of metaplasia = 0% (score 0), mild metaplasia = 1 – 30% (score 1), moderate metaplasia = 31 – 60% (score 2), and severe metaplasia = > 60% (score 3). The scores for the antral biopsies (A_1, A_2, A_3) and the corpus biopsies (C_1, C_2) were averaged separately to obtain a composite score for each region. Finally, risk was determined by combining mean scores of the antrum and corpus, as illustrated in Fig. 2.

Manual annotations of intestinal metaplasia

All cases were manually annotated by the most experienced pathologist, delineating those regions affected by intestinal metaplasia. Overall, these regions contain not only glands with intestinal metaplasia but also a non-negligible amount of stroma and surrounding tissues. The pathologist manually annotated intestinal metaplasia regions using a custom web software developed as part of the *Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño*, equipped with basic tools to annotate and visualize WSI. Annotations were saved according to the Web Annotation Data Model standard[41].

Additionally, all cases were evaluated by the three pathologists, who independently estimated the percentage of intestinal metaplasia in both antrum and corpus biopsies using the previously described four-tier scale. Based on these estimations, each pathologist assigned an OLGIM stage per case, derived from the previously mentioned heuristic combination. This staging reflects the presumed risk of gastric cancer progression associated with the distribution and severity of intestinal metaplasia.

Automatic detection of glandular regions

The initial step to detect and quantify intestinal metaplasia involves automatic segmentation of gastric glands. In this study, gland segmentation was performed applying a U-Net architecture with a ResNet18 backbone[42], originally trained to segment colorectal glands. This choice was motivated by the fact that colorectal and gastric glands exhibit remarkable morphological similarities[43], particularly by the presence of goblet cells associated with intestinal metaplasia, which by definition corresponds to structural changes resembling those observed in colorectal tissue[44].

The model was pre-trained using a dataset comprising 165 fields of view extracted from 16 H&E-stained whole slide images (WSI) from the GlaS database (Gland Segmentation in Colon Histology Images Challenge)[42]. Since gastric and colorectal glands are similar in appearance, this architecture processed the 205 WSI from both gastric cohorts. To adapt the model to the gastric domain, the most experienced pathologist manually annotated 2.434 glands from 15 randomly selected cases, including both metaplastic and non-metaplastic glands. This annotated dataset was used to fine-tune the U-Net architecture, achieving a Dice Score of 0.77 ± 0.22 with a test subset of 730 glands, that is, glands not included in the training phase.

Since the objective of this study is not only to segment glands but also to detect and quantify intestinal metaplasia, the segmented glands were used to define glandular regions. These regions were defined as bounding boxes surrounding groups of glands, within which a grid of fields of view is drawn.

Labeling automatically detected fields of view

Intestinal metaplasia regions, manually annotated by the most experienced pathologist, were superimposed onto the previously described grid of fields of view. Fields overlapping the annotated regions were labeled as intestinal metaplasia. Each of these fields of view corresponds to square patches of 256×256 pixels, extracted at $\times 40$ magnification. Regions outside the glandular region were excluded from the analysis.

The 149 cases from the first cohort were divided into training, validation, and internal testing sets, while the 56 cases from the second cohort were exclusively used as an external test set. In total, 476.351 fields of view were extracted, being 136.682 for training, 59.629 for validation, 89.945 for internal testing, and 190.095 for external testing.

Automatic classification of intestinal metaplasia

Deep learning models

Fields of view labeled as metaplasia or control classes were used to train and evaluate a set of state-of-the-art deep learning models. Specifically, four state-of-the-art convolutional neural networks pre-trained with the ImageNet dataset[45], ResNet50[46], DenseNet121[47] and ConvNeXtTiny[48], were adapted to classify microscopic gastric fields. ResNet50 was selected based on its remarkable performance in intestinal metaplasia classification, with predictions closely matching the assessments of experienced pathologists[49]. DenseNet121 was used as the basis to construct a specialized histopathology framework[50]. ConvNeXtTiny has demonstrated superior performance at identifying anatomical regions of the stomach in endoscopic images[51]. Furthermore, a foundational model, UNI2-h[52], pre-trained with TCGA[53], a large-scale histopathology image dataset, was used as a feature extractor for the classification of intestinal metaplasia.

Warm-up and fine-tuning

A multi-layer perceptron (MLP) was trained on top of each model to perform the classification of intestinal metaplasia. All models followed a two-stage training scheme consisting of a warm-up and fine-tuning phase. During the warm-up phase, the backbone of each model was frozen to preserve visual features learned from large datasets, such as ImageNet or TCGA. In this phase, only the MLP weights were updated. Subsequently, in the fine-tuning phase, selected layers of the pre-trained backbone were progressively unfrozen, enabling joint optimization of both the backbone and the classification head. Fine-tuning was performed with a lower learning rate to minimize overfitting and preserve the most relevant features of the original pretraining.

Hyper-parameter optimization

Each model was trained, validated and tested using the data partition presented in section *Biopsy assessment*, along with the corresponding fields of view described in section *Labeling automatically detected fields of view*.

The architecture of each model integrated a MLP consisting of a sequential block composed of a dropout layer, a fully-connected layer, and a batch normalization layer, followed by an output layer with a *softmax* activation function. Model training followed a two-phase scheme (warm-up and fine-tuning), each consisting of thirty epochs. A hyper-parameter optimization process was performed to determine the optimal learning rate and dropout values, based in three independent experimental runs. During the warm-up phase, only the MLP was trained, while the backbone of the pre-trained model remained frozen. In the subsequent fine-tuning phase, 80% of the pre-trained model layers remained frozen, while the remaining 20% (the deepest layers) were unfrozen and jointly optimized with the MLP to refine task-specific representations.

Data augmentation strategy

A significant class imbalance was noted for the dataset, with a ratio 9:1 for non-metaplasia vs metaplasia patches. To address this imbalance, an online data augmentation strategy was implemented during training. This strategy involved applying symmetrical rotations of 90° increments along the horizontal axis to artificially increase sample diversity and mitigate overfitting. Furthermore, class weights were adjusted before training, assigning greater importance to the minority class (*metaplasia*) in the loss function to compensate for the skewed class distribution.

Automatic quantification and scoring of intestinal metaplasia

Automatic quantification of intestinal metaplasia was performed independently by the four deep learning architectures at the level of the pre-established fields of view. Each field of view was set to a probability value of class membership (i.e., *metaplasia* or *control*), and model performance was assessed using standard metrics: Accuracy, Precision, Recall, F1-Score and Area Under the ROC Curve (AUC).

These regional proportions were then used to determine an intestinal metaplasia score according to the OLGIM system, by averaging the scores from the three antrum biopsies (A_1, A_2, A_3) and the two corpus biopsies (C_1, C_2). These scores reflect both the extent and topographic distribution of intestinal metaplasia, and were further used to assign the OLGIM stage by intersecting the antrum and corpus scores. Stages 0 - II were classified as low-risk, while stages III - IV were considered high-risk for GCa progression (Fig. 2).

Quantifying variability of manual and automatic estimations

Manual estimations were performed by three pathologists following OLGIM guidelines. Pathologists assigned scores approximating the extent and topographic distribution of intestinal metaplasia across the antrum, incisura, and corpus

Inter-observer variability

The 205 cases were independently evaluated by three pathologists following OLGIM guidelines. Pathologists assigned scores approximating the extent and topographic distribution of intestinal metaplasia across the antrum, incisura, and corpus, and assigned a score from 0 to 3. Based on these scores, an OLGIM stage was assigned: stages 0, I or II indicate the lowest risk, while stages III and IV indicate the highest risk. To assess inter-observer agreement, Fleiss' Kappa coefficient was calculated to measure overall consistency among the three pathologists, and Cohen's Kappa coefficient was used to assess pairwise agreement. Both metrics range from 0 (no agreement) to 1 (perfect agreement) with higher values indicating greater agreement.

Comparison between manual and automatic intestinal metaplasia percentages

Automatic scores calculated by the deep learning models were compared against the manual estimations provided by three experienced pathologists in two complementary approaches.

First, trends and variability of antrum and corpus intestinal metaplasia percentages are plotted to show differences between the three pathologists and four deep learning models. Second, a separate plot illustrates the distribution of the assigned OLGIM stages, allowing a direct comparison of staging variability between manual assessments and automated predictions.

Results

Automatic classification performance

Deep learning models were fine-tuned and evaluated in a binary classification task (presence vs. absence of intestinal metaplasia) using both cohorts. Specifically, models were trained, validated and tested using the first cohort of 149 cases, and externally tested with the second independent cohort of 56 cases. The classification results, summarized in Table 2, show mean and standard deviation for both internal and external test sets. These metrics were calculated across three experimental runs during the hyperparameter optimization phase and test sets were left aside from the

beginning to ensure unbiased evaluation. Under these conditions, ConvNeXtTiny achieved the highest performance when classifying fields of view, with F1-Score **0.80 ± 0.01**, and AUC **0.91 ± 0.01**, for the internal test set, and F1-Score **0.73 ± 0.01**, and AUC **0.83 ± 0.01** for the external test set. The UNI2-h foundational model, showed the poorest performance for intestinal metaplasia classification with F1-Score **0.72 ± 0.01** and AUC **0.84 ± 0.01**, for the internal test set, and F1-Score **0.62 ± 0.01** and AUC **0.75 ± 0.01** for the external test set.

Table 2. Performance for the field-of-view-level classification models with the internal (1st cohort) and external (2nd cohort) testing sets.

Architecture	Precision	Sensitivity	F1-Score	AUC
Internal test				
ConvNeXtTiny	0.79 ± 0.01	**0.80 ± 0.01**	**0.80 ± 0.01**	**0.91 ± 0.01**
ResNet50	0.82 ± 0.01	0.77 ± 0.01	0.79 ± 0.01	0.90 ± 0.01
DenseNet121	**0.83 ± 0.02**	0.74 ± 0.03	0.77 ± 0.02	0.89 ± 0.01
UNI2-h	0.75 ± 0.02	0.70 ± 0.02	0.72 ± 0.01	0.84 ± 0.01
External test				
ConvNeXtTiny	0.78 ± 0.01	**0.69 ± 0.01**	**0.73 ± 0.01**	**0.83 ± 0.01**
ResNet50	0.80 ± 0.01	0.64 ± 0.01	0.68 ± 0.01	0.79 ± 0.01
DenseNet121	0.81 ± 0.04	0.64 ± 0.02	0.68 ± 0.01	0.82 ± 0.01
UNI2-h	**0.81 ± 0.01**	0.51 ± 0.02	0.62 ± 0.01	0.75 ± 0.01

Automatic quantification and scoring of intestinal metaplasia

The fields of view predicted by the best-performing model were superimposed to the corresponding whole-slide images of both cohorts. Predictions were plotted in a color-code map: blue and red, the ones correctly classified, and orange and purple, the misclassified ones (Fig. 3). These prediction maps highlight in red areas where the model agrees with the expert-annotated regions. Within these annotated regions, areas in purple correspond to those in which the model disagrees with the expert. Interestingly, the model also identifies metaplastic areas (in orange) that have not previously been annotated. In general, a majority of misclassified fields of view were observed within large regions, which often contain a heterogeneous mix of metaplastic glands, control glands, stroma, and foveolar epithelium.

Quantifying variability of manual and automatic estimations

Results of the inter-observer agreement

Inter-observer agreement among pathologists and the best-performing deep learning model is presented in Figure 4. The Cohen's Kappa coefficient shows pairwise inter-observer agreement in both the antrum and corpus, for both internal and external test sets. Given that experts provided an accurate estimation of the intestinal metaplasia percentage, manual-automatic comparison was performed at a finer level by binning the entire range into incremental intervals of 10%, i.e., ten intervals of 10% instead of the usual 0%, 30%, 60% and more. This uniform partition was used to evaluate agreement between manual and automated estimations. Inter-observer agreement among pathologists was moderate, with Fleiss' Kappa = 0.31 for antrum, and 0.41 for corpus, therefore, pairwise Cohen's Kappa ranged 0.20 – 0.48. Agreement between the model and individual pathologists ranged 0.12 – 0.35. Variability of IM percentage estimates was evident, since pathologists tended to assign higher values than the automated model. These results underscore both the visual estimation variability and the relative consistency of the automated approach. In addition, Spearman's rank-order correlation was computed between the IM percentage estimates of the model and those of each pathologist, showing significant positive associations ($\rho = 0.732 \pm 0.011$ for OLGIM stage and $\rho = 0.893 \pm 0.022$ for IM percentage), indicating model prediction are associated with pathologists' estimation

Subplots in Figure 5 show the inter-observer agreement among pathologists, as well as between the pathologists and the best-performing deep learning model, regarding the OLGIM staging. Inter-observer agreement can be seen to increase when the categorical evaluation scale is simplified. In this work, expert agreement improved significantly when using a coarser grading scheme, the OLGIM staging system.

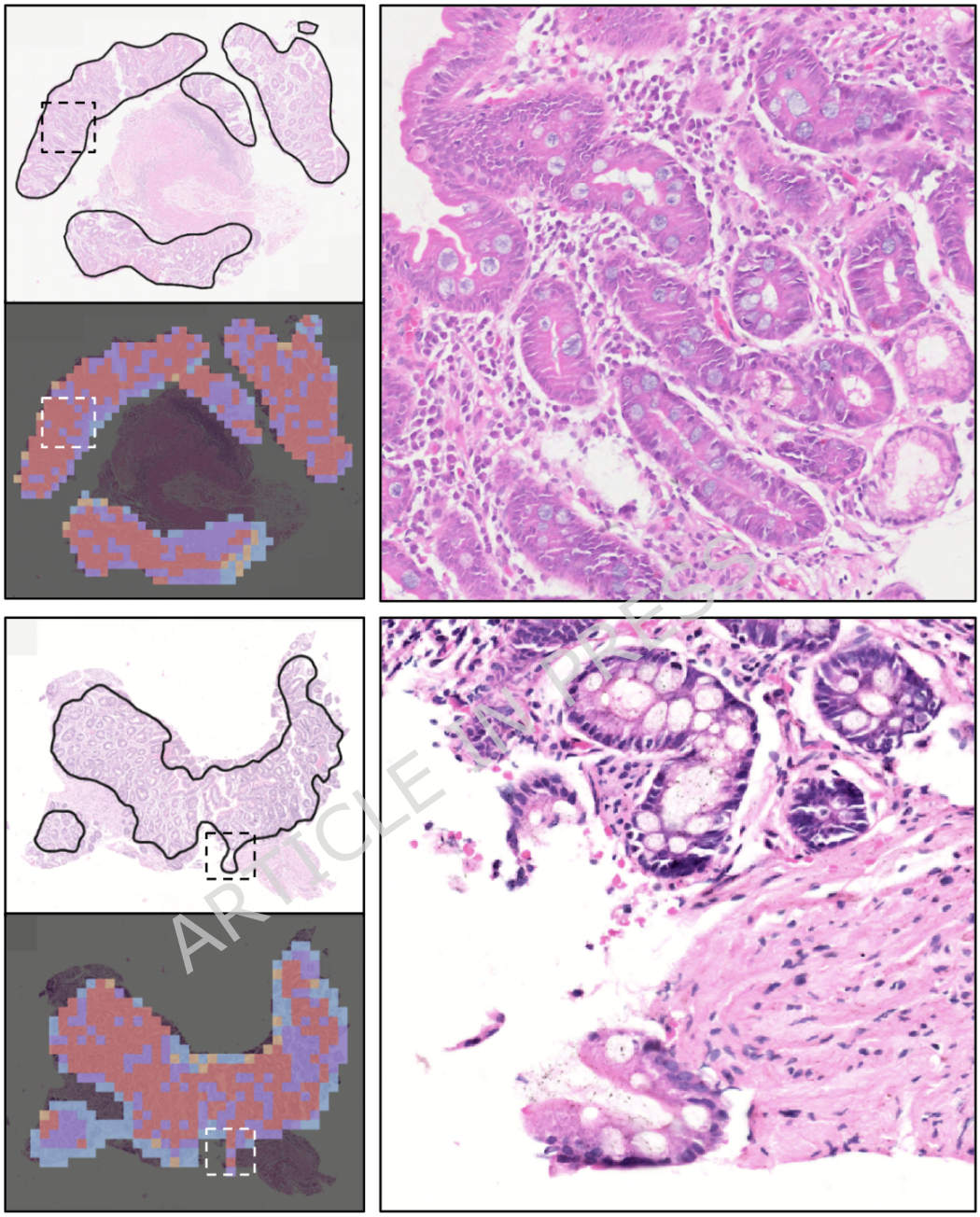

Figure 3. Representative examples comparing expert annotations and model predictions. For each case, the left subfigures display the whole-slide region with the most experienced pathologist's annotation (top) and the corresponding model prediction map (bottom). Marked boxes indicate areas of interest, which are shown at higher magnification on the right. The enlarged fields illustrate regions containing metaplastic glands, as well as portions of tissue annotated by the expert where not all structures correspond to intestinal metaplasia. Model predictions are color-coded as follows: correctly predicted metaplasia (red), correctly predicted control (blue), control regions incorrectly predicted as metaplasia (orange), and metaplastic regions incorrectly predicted as control (purple).

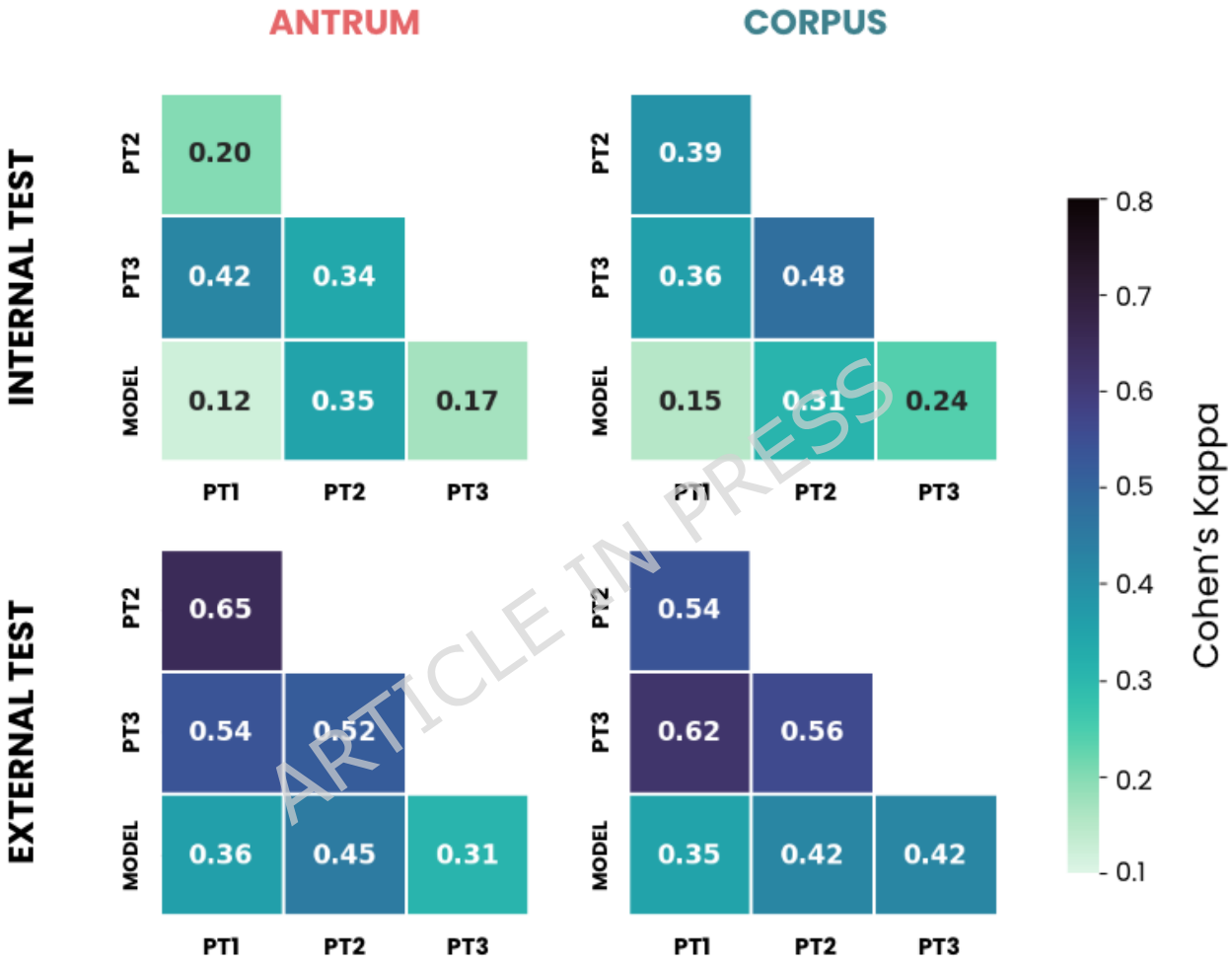

Figure 4. Inter-observer agreement in the assignment of intestinal metaplasia (IM) percentage intervals (10% bins across the full range) among pathologists and the best-performing deep learning model (ConvNeXtTiny) for the internal (first row) and external (second row) test sets. PT1, PT2 and PT3 correspond to the three pathologists, while MODEL refers to the automated model. Subplots in the first column show agreement for the antrum, and subplots in the second column correspond to the corpus.

/

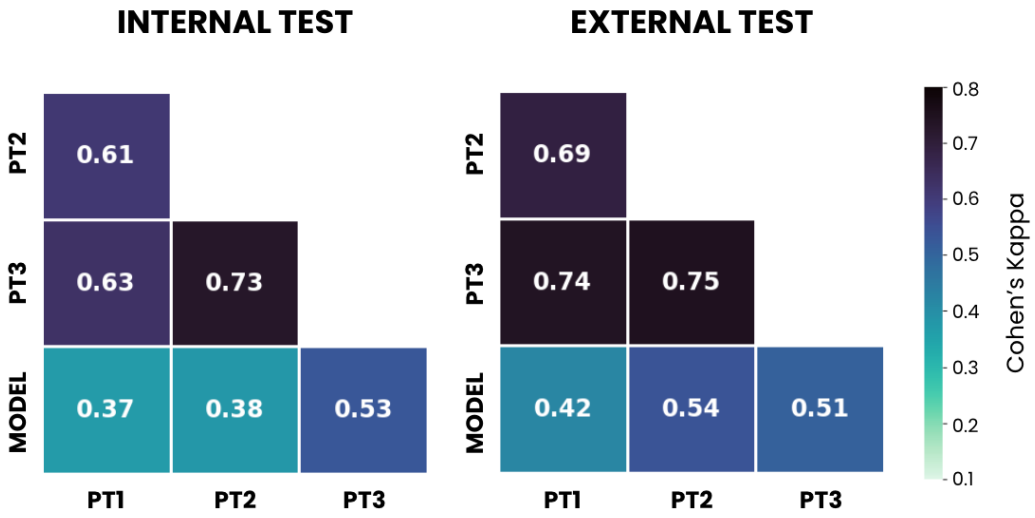

Figure 5. Inter-observer agreement in the assignment of OLGIM stages, compared with the automatic staging on the internal and external test datasets.

Variability of intestinal metaplasia percentages

Expert estimations of IM percentages showed the greatest variability in intermediate OLGIM stages, likely due to the larger number of possible antrum–corpus combinations that can yield the same stage. Both manual and automated assessments demonstrated a consistent increase in IM percentages with higher OLGIM stages. However, the automated model provided substantially lower variability across the four stages, while pathologists tended to assign higher percentages, reflecting a tendency to overestimate lesion extent. These differences are illustrated in Figure 6.

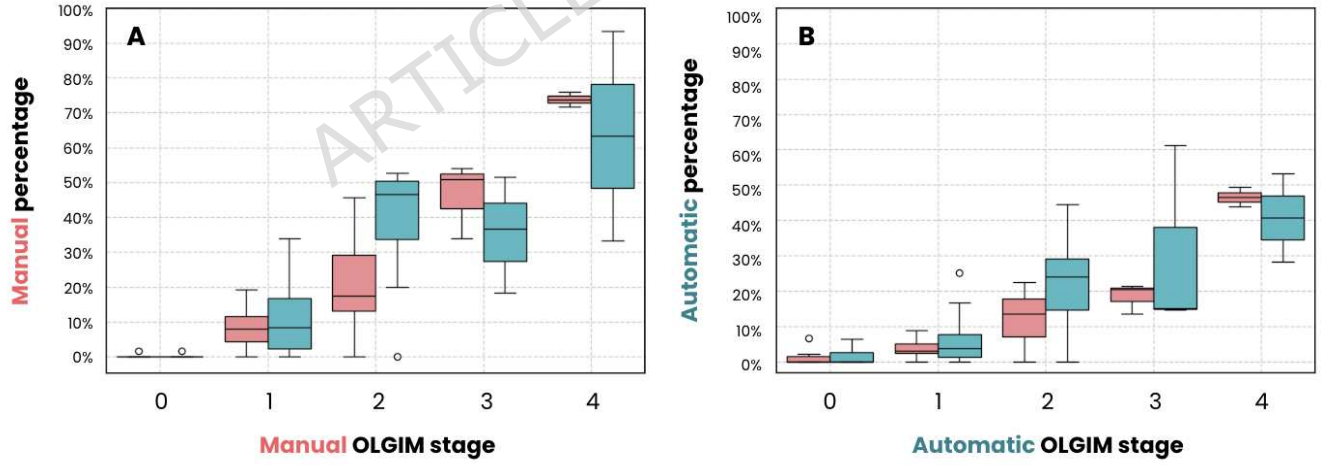

Figure 6. Distribution of manual (left) and automatic (right) intestinal metaplasia percentages in the antrum (red) and corpus (green), related to the consensus OLGIM stage assigned by pathologists.

Discussion

This paper has investigated the role of state-of-the-art deep learning models to detect and quantify intestinal metaplasia (IM) on digitized biopsy samples from patients and volunteers. Evaluation of these models demonstrated relatively accurate performance for computing stroma, glands, and the percentage of IM, with F1 = 0.80 and AUC = 0.91. Comparison of OLGIM scoring, pathologist assessments, and deep learning predictions highlighted the

inherent subjectivity of visual estimation, with only moderate inter-observer agreement (Fleiss' Kappa = 0.31 for the antrum and 0.41 for the corpus), consistent with previously reported values in the literature[54] (Fleiss' Kappa = 0.48). These findings confirm both the limitations of current risk systems, limiting the certainty of the medical decision and patient confidence, and the potential of AI tools to complement pathologist-based evaluation by serving as reproducible "second readers".

Several studies have investigated OLGIM limitations, and many pathologists worldwide hardly report OLGIM stages since they are aware of the conflict introduced by the typical disagreement[16], indeed, a moderate agreement is widely acknowledged[54,55]. An inevitable criticism of the OLGIM score is that it is inherently biased, that is, intermediate stages are obtained with a higher number of combinations and therefore these stages are more frequent, in other words "dices are charged" and confidence therefore is undermined. Additionally, pathologists tended to overestimate the extent of the lesion[18], typically assigning higher percentages than the model. In the present investigation, the analysis was restricted to cross-sectional biopsies, limiting thereby the impact of the present study. Moreover, annotations represented a bottleneck, since delineated regions often included not only metaplasia but also stroma and normal glands, introducing bias[56]. The multifocal nature of IM and its patch-like distribution further complicates assessment, as even standardized sampling with the five-biopsy protocol may fail to capture the full extent of lesions, affecting both manual and automated quantification. This work did not distinguish between complete and incomplete IM subtypes, which are morphologically and prognostically distinct, due to the lack of annotations at this level of detail. Finally, the participating pathologists were recently graduated, which may partly explain the relatively low inter-observer agreement observed.

It is important to recall that, in this work, OLGIM staging relied exclusively on IM since atrophy was difficult to evaluate, and even among experienced pathologists it has shown poor inter-observer reproducibility[19]. In contrast, IM has been recognized as the most reliable and prognostic component of OLGIM staging[57]. This methodological choice therefore prioritizes objectivity and reproducibility. Recent studies have further demonstrated the effectiveness of deep learning models at classifying and grading pathologies in H&E images[24,27,32]. Typically, pathologist knowledge, "the ground truth"[58], consists of delineated regions that train models to replicate visual annotations. However, variable biology along with subjective judgment inevitably introduce additional biases.

Despite the multifocal nature of IM, i.e. spreading along stomach regions, estimation of this condition is often reduced to a binary problem: presence or absence. Both IM extension and topographic distribution have been at the very base of risk estimation for progression. Correa et al. highlighted that IM comprises two main subtypes, complete and incomplete, associated with distinct morphological expressions of mucin enzymes. Complete IM resembles the small intestine, with enterocytes displaying a brush border, whereas incomplete IM is closer to colonic epithelium with irregular intra-cytoplasmic mucin droplets of variable size. In the present study, we did not differentiate between these subtypes because annotations at this level of detail were not available, and our analysis focused on overall IM quantification in line with OLGIM staging. We acknowledge this as a limitation, and future research with dedicated annotations and tailored AI methods may enable reliable distinction of IM subtypes, moving beyond binary characterization and towards a more continuous description of the phenomenon, which is a necessary step for reproducible quantification and personalized risk estimation.

In summary, deep learning-based quantification of IM demonstrates robust performance, reproducibility, and potential clinical relevance. While expert variability, annotation bias, and sampling constraints remain significant challenges, computational pathology offers a pathway for systematic and objective evaluation. With further validation in multi-center cohorts, follow-up studies, and the integration of IM subtyping, AI-based tools could transform gastric cancer prevention strategies, moving from subjective visual estimation to reproducible, data-driven, and patient-tailored risk assessment.

Conclusion

Deep learning models can reliably detect and quantify intestinal metaplasia with high performance, offering consistent estimates that may reduce the subjectivity of expert-dependent visual assessment. Further validation in larger, multi-center cohorts and the inclusion of complete and incomplete subtypes will be essential to capture the full morphological spectrum. In the long term, reproducible AI-based quantification of precancerous lesions has the potential to enhance gastric cancer risk stratification and guide more objective, personalized clinical decisions.

Author contributions

Conceptualization: FC, ACR, FAG, SEV and ER; Data curation: FC, MC, AS, JV and JQ; Methodology: FC, MC, ACR, FAG and ER; Formal analysis: FC, ACR, FAG, SEV and ER; Writing - original draft preparation: FC and ER; Writing - review and editing: FC, ACR, FAG, SEV and ER; Funding acquisition: ACR and ER; Resources:

ABU, MCB, YYC, ACR, FAG and ER; Supervision: FAG, SEV and ER; Project administration: ER. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partially supported by the project with code 110192092345 *Program for the Early Detection of Premalignant Lesions and Gastric Cancer in urban, rural and dispersed areas in the Department of Nariño* of call No. 920 of 2022 of MinCiencias. This work was partially supported by project 52895, titled *Proposal for the strategic plan for the establishment of the Center of Excellence (Inter-Sites) in Medicine and Artificial Intelligence (SemAI)*, from the National Call for Proposals Bank for the Consolidation of Centers of Excellence 2020-2021 at *Universidad Nacional de Colombia*. This work was partially supported by project BPIN 2019000100060 *Implementation of a Network for Research, Technological Development and Innovation in Digital Pathology (RedPat) supported by Industry 4.0 technologies* from FCTeI of SGR resources, which was approved by OCAD of FCTeI and MinCiencias.

Acknowledgements

Special thanks to the CIEDYN foundation and the project BPIN 20150000100064 Urkunina5000, from which whole slides of gastric tissue were recovered from asymptomatic volunteers and subsequently digitized to be used in the development of this work.

Data availability

The collection of internal gastric samples used in this study were obtained from the primary research project entitled *Investigación de la prevalencia de lesiones precursoras de malignidad gástrica y efecto de la erradicación de Helicobacter pylori colon prevención primaria de cáncer gástrico en el Departamento de Nariño*. This project was approved under Agreement No. 057 of 2017 by the Collegiate Body for Administration and Decision (OCAD Pacífico), and was funded by the Science, Technology, and Innovation Fund of the General System of Royalties/Government of Nariño. Internal and external data are publicly available at Harvard Dataverse[59].

Declarations

Conflict of interest

All authors declare no conflicts of interest.

Ethical considerations

This research study was conducted retrospectively using data provided by the CIEDYN foundation, a partner of the Urkunina5000 project, which contains information on ethical considerations in compliance with the Declaration of Helsinki. All patients signed informed consent forms. Additional ethics considerations was not required.

Informed consent

All asymptomatic volunteers selected for this study signed informed consent and all guarantees of anonymization were applied to their data.

References

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians.* **74(3)**, 229–263 (2022).
2. Ajani, J. *et al.* Gastric adenocarcinoma. *Nat. Rev. Dis. Prim.* **3**, 1–19 (2017).
3. Tan, P. & Yeoh, K. Genetics and molecular pathogenesis of gastric adenocarcinoma. *Gastroenterology.* **149**, 1153–1162 (2015).
4. Pucułek, M. *et al.* Helicobacter pylori associated factors in the development of gastric cancer with special reference to the early-onset subtype. *Oncotarget.* **9(57)**, 31146–31162 (2018).
5. Cai, Q. *et al.* Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study. *Gut.* **68(9)**, 1576–1587 (2019).
6. Misiewicz, J. The Sydney System: a new classification of gastritis. Introduction. *J. Gastroenterol. Hepatol.* **6(3)**, 207–208 (1991).
7. Dixon, M. *et al.* Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *Am. J. Surg. Pathol.* **20**, 1161–1181 (1996).
8. Rugge, M. *et al.* Gastritis staging in clinical practice: the olga staging system. *Gut.* **56(5)**, 631–636 (2007).

9. Crafa, P. *et al.* From Sidney to OLGA: an overview of atrophic gastritis. *Acta. Biomed.* **89(8-S)**, 93–99 (2018).

10. Correa, P. Gastric cancer: overview. *Gastroenterol. clinics North Am.* **42(2)**, 211–217 (2013).

11. Koulis, A. *et al.* Premalignant lesions and gastric cancer: Current understanding. *World J. Gastrointest. Oncol.* **11(9)**, 665–678 (2019).

12. Tjandra, D. *et al.* Gastric Intestinal Metaplasia: Challenges and the Opportunity for Precision Prevention. *Cancers.* **15**, 3913 (2023).

13. Marcos, P. *et al.* Endoscopic grading of gastric intestinal metaplasia on risk assessment for early gastric neoplasia: can we replace histology assessment also in the west? *Gut.* **69(10)**, 1762–1768 (2020).

14. Rugge, M. & Genta, R. Staging and grading of chronic gastritis. *Hum. Pathol.* **36(3)**, 228–233 (2005).

15. Capelle, L. *et al.* The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. *Gastrointest. Endosc.* **71(7)**, 1150–1158 (2010).

16. Fang, S. *et al.* Diagnosing and grading gastric atrophy and intestinal metaplasia using semi-supervised deep learning on pathological images: development and validation study. *Gastric Cancer.* **27**, 343–354 (2024).

17. Zhao, Q. *et al.* Deep learning model can improve the diagnosis rate of endoscopic chronic atrophic gastritis: a prospective cohort study. *BMC Gastroenterol.* **22(1)**, 133 (2022).

18. Yue, H. *et al.* The significance of OLGA and OLGIM staging systems in the risk assessment of gastric cancer: a systematic review and meta-analysis. *Gastric Cancer.* **21**, 579–587 (2018).

19. Isajevs, S. *et al.* Gastritis staging: interobserver agreement by applying OLGA and OLGIM systems. *Virchows. Arch.* **464**, 403–407 (2014).

20. Molaei, M. *et al.* Gastric atrophy: use of OLGA staging system in practice. *Gastroenterol. Hepatol. Bed. Bench.* **9(1)**, 25–29 (2016).

21. Rugge, M. *et al.* Gastritis OLGA-staging and gastric cancer risk: a twelve-year clinicopathological followup study. *Aliment. Pharmacol. Ther.* **31**, 1104–1111 (2010).

22. Mansour-Ghanaei, F. *et al.* OLGA- and OLGIM-Based Staging in the Patients with Gastritis and Endoscopy Indications. *Turk. J. Gastroenterol.* **33(2)**, 95–102 (2022).

23. Chen, C. *et al.* An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nat. Commun.* **12**, 1193 (2021).

24. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).

25. Kanavati, F. *et al.* Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci. Rep.* **10**, 9297 (2020).

26. Cruz-Roa, A. *et al.* High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PLoS ONE.* **13(5)**, e0196828 (2018).

27. Liu, M. *et al.* A Deep Learning Method for Breast Cancer Classification in the Pathology Images. *IEEE J. Biomed. Heal. Informatics.* **26(10)**, 5025–5032 (2022).

28. Wetstein, S. *et al.* Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Sci. Rep.* **12**, 15102 (2022).

29. Nir, G. *et al.* Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Med. Image. Anal.* **50**, 167–180 (2018).

30. Li, Y. *et al.* Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer. *IEEE Access.* **8**, 117714–117725 (2020).

31. Tolkach, Y. *et al.* High-accuracy prostate cancer pathology using deep learning. *Nat. Mach. Intell.* **2**, 411–418 (2020).

32. Wang, X. *et al.* Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. *Nat. Commun.* **12**, 1637 (2021).

33. Huang, B. *et al.* Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: A retrospective multicentre study. *EBioMedicine.* **73**, 103631 (2021).

34. Sharma, H. *et al.* Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput. Med. Imaging. Graph.* **61**, 2–13 (2017).

35. Veldhuizen, G. P. *et al.* Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study. *Gastric Cancer.* **26(5)**, 708–720 (2023).

36. White, J. *et al.* Identifying the pre-malignant stomach: from guidelines to practice. *Transl. Gastroenterol. Hepatol.* **7**, 8 (2022).

37. Pimentel-Nunes, P. *et al.* 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**, 365–388 (2019).

38. Hwang, Y. *et al.* Reversibility of atrophic gastritis and intestinal metaplasia after helicobacter pylori eradication — A prospective study for up to 10 years. *Aliment. Pharmacol. Ther.* **47**, 380–390 (2018).

39. Aumpan, N. *et al.* Predictors for regression and progression of intestinal metaplasia (IM): A large population-based study from low prevalence area of gastric cancer (IM-predictor trial). *PLoS One.* **16(8)**, e0255601 (2021).

40. Bedoya, A. *et al.* Proyecto Urkunina 5000. Investigación de la prevalencia de lesiones precursoras y del efecto de la erradicación de Helicobacter pylori como prevención primaria del cáncer gástrico en el departamento de Nariño. *Rev. Colomb. Cir.* **33(4)**, 345–352 (2019).

41. Sanderson, R. *et al.* Web Annotation Data Model: W3C Recommendation 23 February 2017. *In Proc. 5th Annu. ACM Web Sci. Conf.* 366–375 (2013).

42. Sirinukunwattana, K. *et al.* Gland segmentation in colon histology images: The glas challenge contest. *Med. image analysis.* **35**, 489–502 (2017).

43. Jass, J. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **50(1)**, 113–130 (2007).

44. Shah, S. C. *et al.* Histologic subtyping of gastric intestinal metaplasia: overview and considerations for clinical practice. *Gastroenterology* **158(3)**, 745–750 (2020).

45. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision pattern recognition.* 248–255 (2009).

46. He, K. *et al.* Deep residual learning for image recognition. *2016 IEEE Conf. on Comput. Vis. Pattern Recognit. (CVPR).* 770–778 (2016).

47. Huang, G. *et al.* Densely connected convolutional networks. *In Proc. IEEE conference on computer vision pattern recognition* 4700–4708 (2017).

48. Liu, Z. *et al.* A convnet for the 2020s. *In Proc. IEEE/CVF conference on computer vision pattern recognition* 11976–11986 (2021).

49. Caviedes, M. *et al.* An automatic classification of metaplasia in gastric histopathology images. *2023 19th Int. Symp. on Med. Inf. Process. Analysis (SIPAIM).* 1–4 (2023).

50. Riasatian, A. *et al.* Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Med. image analysis.* **70**, 102032 (2021).

51. Bravo, D. *et al.* Automatic Classification of Esophagogastroduodenoscopy Sub-Anatomical Regions. *2023 IEEE 20th Int. Symp. on Biomed. Imaging (ISBI).* 1–5 (2023).

52. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).

53. Tomczak, K. *et al.* The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp Oncol (Pozn).* **19(1A)**, A68–77 (2015).

54. Lerch, J. M. *et al.* Subtyping intestinal metaplasia in patients with chronic atrophic gastritis: an interobserver variability study. *Pathology.* **54(3)**, 262–268 (2022).

55. Salazar, B. *et al.* The OLGA-OLGIM staging and the interobserver agreement for gastritis and preneoplastic lesion screening: a cross-sectional study. *Virchows Arch.* **480**, 759–769 (2022).

56. Selvaraju, R. *et al.* Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. *Proc. IEEE Int. Conf. on Comput. Vis. (ICCV), 2017.* 618–626 (2017).

57. Rugge, M. *et al.* Operative link for gastritis assessment vs operative link on intestinal metaplasia assessment. *World journal gastroenterology* **17(41)**, 4596–4601 (2011).

58. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. pathology informatics.* **7(1)**, 29 (2016).

59. Cano, F. & others. Supplemented data for: Towards deep-learning based detection and quantification of intestinal metaplasia on digitized gastric biopsies: a multi-expert comparative study. *Harvard Dataverse* https://doi.org/10.7910/DVN/NVUCW8 (2025).