Binary Screening Gate NORMAL/ABNORMAL in Brazilian Upper Gastrointestinal Endoscopy: Safety Analysis Guided by Target Sensitivity

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 – UFG (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. *Automatic screening of Upper Gastrointestinal Endoscopy (UGE) images can reduce the specialized reading workload, but its adoption depends on quantifying how many pathologies would be missed when discarding images as normal. We evaluated a binary NORMAL/ABNORMAL gate on 1,984 UGE images (image-level analysis; patient and exam identifiers are unavailable) from two Brazilian centers, comparing five CNN architectures in 5-fold cross-validation. DenseNet-121 achieved the best performance (AUROC 0.887; AUPRC 0.903). At the 95% target sensitivity, the gate filters 24.2% of images (NPV 0.892), missing 28 of the 616 images with pathology (4.5%); relative miss-rates are higher for rare classes, and a global threshold does not guarantee the target in both centers.*

1. Introduction

Upper Gastrointestinal Endoscopy (UGE) is the standard exam for evaluating the mucosa of the esophagus, stomach, and duodenum, from gastritis screening to neoplasms [Fujishiro et al. 2026]. In medium- and high-complexity services, demand often exceeds the capacity for specialized interpretation, generating queues and delays in diagnosis. Computational reading support methods have been studied to increase process efficiency, without replacing the endoscopist's judgment.

The use of Convolutional Neural Networks (CNN) in UGE has grown over the last decade, with diagnostic performance comparable to specialists in specific tasks [Arribas et al. 2021, Lui et al. 2020], acknowledged even in position statements from international organizations [Fujishiro et al. 2026]. Nevertheless, gaps remain. Most studies come from Asian or European cohorts: in a prior systematic search conducted by the authors across seven bibliographic databases (1,621 works included), only two had a detectable Brazilian clinical context and none published an open UGE dataset. Furthermore,

the literature focuses on fine-grained classification of specific pathologies (neoplasm, ulcer, polyp) and devotes little attention to the prior step, namely the separation between normal and abnormal exams.

A binary NORMAL/ABNORMAL classifier can act as a pre-selection gate, forwarding to the specialist only exams with a high probability of abnormality and reducing reading workload. Works that propose screening rarely measure the clinical cost of the operation, i.e., which pathologies and in what proportion are missed when adopting the gate in production [Truong et al. 2024, Krenzer et al. 2026]. This safety, given by the rate of missed pathologies among exams discarded as normal, governs clinical acceptability and remains underanalyzed. Added to this is evaluation at a fixed operating point, at threshold 0.5 or at F1 maximization [Su et al. 2023, Ali et al. 2018]: in screening, the threshold must be derived from a minimum sensitivity defined by the clinician, rather than fixed in advance.

Given this, we investigate: is it possible to build a NORMAL/ABNORMAL gate from UGE images from a Brazilian service, with clinically acceptable sensitivity and measurable workload reduction, and what are the safety implications by pathology? Five CNN families (ResNet-50, EfficientNet-B0, ConvNeXt-Tiny, DenseNet-121, and MobileNetV3) were trained and evaluated in stratified 5-fold cross-validation on 1,984 images from two Brazilian centers. The best model was evaluated at three target sensitivities (90%, 95%, and 99%), with miss-rate by pathology, Negative Predictive Value (NPV), net benefit via Decision Curve Analysis (DCA), and heterogeneity by center. The contributions are: (1) systematic evaluation of a binary gate in UGE from a Brazilian context; (2) safety analysis guided by target sensitivity, with miss-rate by pathological category in absolute counts; and (3) identification of operational limitations, with emphasis on the gate's fragility in rare classes and the heterogeneity between centers under a global threshold. The remainder of the text covers related work (Section 2), methodology (Section 3), results (Section 4), and conclusions (Section 5).

2. Related Work

2.1. Automatic classification in UGE

Meta-analyses and systematic reviews show that Artificial Intelligence (AI) systems achieve performance comparable to specialists in the detection of neoplasms of the upper Gastrointestinal (GI) tract [Arribas et al. 2021, Lui et al. 2020]. A meta-analysis of randomized controlled trials confirmed gains in detection rate with AI assistance [Al Hayek et al. 2026], and a prospective study indicated a reduction in disparity between endoscopists of different experience levels [You et al. 2026]. On the architectural level, ImageNet pre-trained CNNs with clinical fine-tuning predominate: Yang et al. [Yang et al. 2022] covered multiple upper GI diseases with a single model (accuracy 78.7%, F1 0.78 on the Kyoto Score); Su et al. [Su et al. 2023] compared strategies for early gastric cancer; and Teramoto et al. [Teramoto et al. 2022] used a cascade model that first detects and then characterizes lesions. There is also a movement toward multimodal foundation models [He et al. 2025]. These studies optimize discrimination of specific findings at a fixed operating point, without addressing NORMAL/ABNORMAL pre-selection or the cost of discarding exams.

2.2. Binary screening, quality control, and operating point

Binary classification as pre-selection has received less attention. Choi et al. [Choi et al. 2022] verified the registration of mandatory anatomical regions during the exam, and Truong et al. [Truong et al. 2024] classified informative frames to discard frames without diagnostic content; both focus on acquisition quality, without measuring how many frames with pathological findings are eliminated. Zhang et al. [Zhang et al. 2024] reviewed quality control in UGE and the emerging role of AI, pointing to the lack of prospective validation. For diagnostic content, Ali et al. [Ali et al. 2018] detected gastric abnormalities with texture descriptors (accuracy 87%, AUC 0.91), without a clinical operating point or safety metrics. The choice of operating point is important in screening, but most studies report metrics at threshold 0.5 or at F1 maximization, without evaluating different sensitivity requirements. Evaluations in real-world settings [Chang et al. 2024, Osawa et al. 2026] report performance by finding and determinants of false positives, but do not adopt target sensitivity as a primary criterion nor quantify the per-class cost of discarded cases. Net benefit via DCA is rare in this literature, although it allows comparing system adoption with the strategies of referring all or none [Fujishiro et al. 2026].

2.3. Class imbalance, rare pathologies, and positioning

Class imbalance is structural in gastroscopy, with erosion and erythema being much more frequent than neoplasms and vascular ectasias. Krenzer et al. [Krenzer et al. 2026] addressed extreme imbalance with data augmentation and adapted loss functions (reporting a 40.4% F1 gain for the minority class), and Yin et al. [Yin et al. 2023] used few-shot learning to identify signet ring cell carcinoma. Neither adopts the safety perspective: how many rare cases are missed when reducing referrals? Table 1 contrasts the closest works along the dimensions that define this contribution. None covers, at the same time, binary screening, target-sensitivity operation, per-pathology safety, DCA, and Brazilian data.

Table 1. Closest works by contribution dimension. ✓ addresses; × does not address; partial partially addresses.

Work	Binary screening	Target sens. (operation)	Safety (miss-rate/pathology)	DCA	BR Data
[Yang et al. 2022]	×	×	×	×	×
[Teramoto et al. 2022]	×	×	×	×	×
[Choi et al. 2022]	partial	×	×	×	×
[Truong et al. 2024]	partial	×	×	×	×
[Ali et al. 2018]	✓	×	×	×	×
[Krenzer et al. 2026]	✓	×	partial	×	×
[Chang et al. 2024]	×	×	partial	×	×
[Osawa et al. 2026]	×	×	partial	×	×
This work	✓	✓	✓	✓	✓

Direct metric comparison is limited as cited works target different classification tasks (multiclass detection, cancer staging, quality control) on distinct cohorts, rather than binary NORMAL/ABNORMAL screening; Table 1 thus positions them along methodological dimensions. In Brazil, this gap is critical: local equipment, protocols, and population characteristics may differ from Asian and European cohorts, and data scarcity

prevents verifying if foreign systems preserve safety locally. This work formulates NORMAL/ABNORMAL screening as a target-sensitivity-oriented gate, quantifies per-pathology miss-rates in absolute counts, analyzes net benefit via DCA, and characterizes center heterogeneity.

3. Methodology

3.1. Data and ethical aspects

The images come from anonymized UGE databases of two Brazilian institutions: Hospital Estadual Alberto Rassi (HGG, Center 1) and Instituto do Aparelho Digestivo (IAD, Center 2) in Goiânia, Brazil. Approved by the Research Ethics Committee of Hospital Geral de Goiânia (Plataforma Brasil, CAAE 59898122.7.0000.0035; CNS 466/2012), the data contain no patient-identifiable information. The unit of analysis is the image: lacking patient/exam identifiers, images cannot be grouped by exam or partitioned by patient (revisited in Section 5). The final dataset has 1,984 images (717 from Center 1, 1,267 from Center 2). Each image received binary labels (NORMAL/ABNORMAL) and seven pathological categories (ERYTHEMA, POLYP, ULCER, EROSION, MICRONODULARITY, VASCULAR ECTASIA, NEOPLASIA) with 0/1 encoding, evaluated only for safety. Labeling was performed by a medical resident and validated by a specialist endoscopist over three months. Six records without labels were removed, and 4 images (0.2%) with both labels were treated as ABNORMAL=1. Pathological categories were not training targets. Supervised dataset access is available upon request to the authors.

Table 2 summarizes the dataset composition. Global ABNORMAL prevalence is 51.4% (1,019/1,984), varying from 35.1% in Center 1 to 60.5% in Center 2, motivating stratified analysis. EROSION is most frequent (321; 16.2%), followed by ERYTHEMA (199; 10.0%), ULCER (81; 4.1%), and POLYP (50; 2.5%); MICRONODULARITY (34; 1.7%), VASCULAR ECTASIA (4), and NEOPLASIA (4) are rare. In total, 616 images have active pathology labels; the remaining 409 ABNORMAL images carry acquisition-artifact annotations (light reflections or saliva/foam) without named pathology labels, representing non-specific visual disturbances or unlisted mucosal changes. This restricts safety granularity to those 616 images. VASCULAR ECTASIA and NEOPLASIA (4 cases each) are reported in absolute counts due to small sample size. Figure 1 illustrates dataset visual variability.

Table 2. Dataset composition by center and class.

Center	NORMAL	ABNORMAL	Total
Center 1	465	252	717
Center 2	500	767	1,267
Total	965	1,019	1,984

3.2. Formulation, cross-validation, and training

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the dataset, where x_i is a UGE image and $y_i \in \{0, 1\}$ indicates the ABNORMAL class. We seek a function $f_\theta : X \to [0, 1]$ that estimates the probability $\hat{y}_i = f_\theta(x_i)$ of the image being positive. The gate decision is

$$\hat{c}_i = \begin{cases} 1, & \text{if } f_\theta(x_i) \geq \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

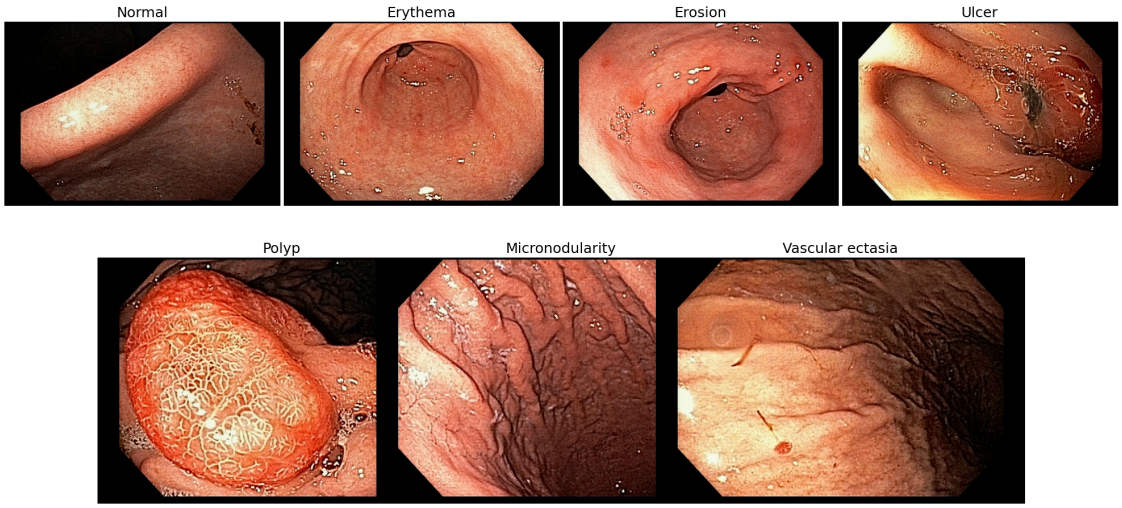


Figure 1. Dataset examples, one per category: Normal, Erythema, Erosion, Ulcer, Polyp, Micronodularity, and Vascular ectasia, the latter being the class with only 2 exclusive images. Each image was selected from those annotated with the indicated category. NEOPLASIA is not shown because all 4 of its cases present multiple simultaneous findings.

where τ is the decision threshold, defined from a target sensitivity rather than fixed at 0.5.

The data were split in stratified 5-fold cross-validation, preserving the NORMAL/ABNORMAL proportion. Each fold separates training, validation, and test sets (approximately 1,350, 238, and 397 images); each image appears exactly once in test, and the fold predictions are pooled to cover the entire dataset without overlap. One seed per fold was used (42 to 46). To reduce the risk of leakage between folds, pairwise visual similarity across all images was computed using pre-trained embeddings; no pair exceeded the cosine similarity threshold of 0.95, indicating the absence of near-duplicate frames, which represent the most common form of cross-fold contamination in endoscopic datasets without exam identifiers. Although semantically related but visually distinct frames from the same exam cannot be ruled out, the center-stratified bootstrap in Section 4 showed no underestimation of uncertainty from this source.

The threshold τ for each target sensitivity was defined on the validation partition of each fold and applied to its respective test set, in a nested scheme that avoids estimating and evaluating the operating point on the same partition. The test values of the five folds are then pooled. Section 4.4 compares this procedure with directly defining the threshold on the pooled predictions.

Five CNNs were evaluated via the `timm` library, initialized on ImageNet and adapted for a single output (logit). The learning rate (LR) differs between backbone (10^{-4}) and classification head (10^{-3}), both scaled by $\sqrt{\text{batch}/32}$; the optimizer was AdamW with weight decay 10^{-4}. The loss was binary cross-entropy with the positive class weight per fold (negative/positive ratio in training), compensating for the class imbalance between centers. Training ran for up to 40 epochs with early stopping at patience 8 by validation AUPRC, a ReduceLROnPlateau scheduler (half the LR after 4 epochs without improvement), and mixed precision when available. Data augmentation during training included resizing to 224×224, horizontal flipping, random resized crop (scale 0.85 to 1.0), and color jitter; at validation and test only resizing was applied, with Im-

ageNet normalization. The 224×224 resolution follows the ImageNet pre-training and facilitates comparison between architectures, but may attenuate fine patterns, such as vascular details, which is discussed when interpreting losses in subtle classes. The splits were fixed before training and shared between models.

3.3. Reproducibility

The code is available on GitHub[1].

3.4. Evaluation metrics

Metrics were computed on the pooled predictions, with 95% confidence intervals by bootstrap (2,000 resamples). **Discrimination:** AUROC and AUPRC; AUPRC guided model selection as it is less sensitive to prevalence imbalance. **Operating points:** Youden threshold (maximum of sensitivity + specificity − 1), specificity at target sensitivities of 90%, 95%, and 99%, and sensitivity at target specificities of 90% and 95%; at each point, sensitivity, specificity, Positive Predictive Value (PPV), NPV, F1, balanced accuracy, and MCC are reported. **Calibration:** Expected Calibration Error (ECE) with 15 bins, Maximum Calibration Error (MCE), and Brier score. **Safety:** filtered fraction, NPV, and number and rate of missed pathologies, i.e., exams with pathology discarded as NORMAL, by pathology and by center with global threshold. **Net benefit:** DCA against the treat-all and treat-none strategies.

4. Results

We present the discrimination of the five architectures and, for the best model, the calibration, operating points, and gate safety.

4.1. Discrimination and calibration

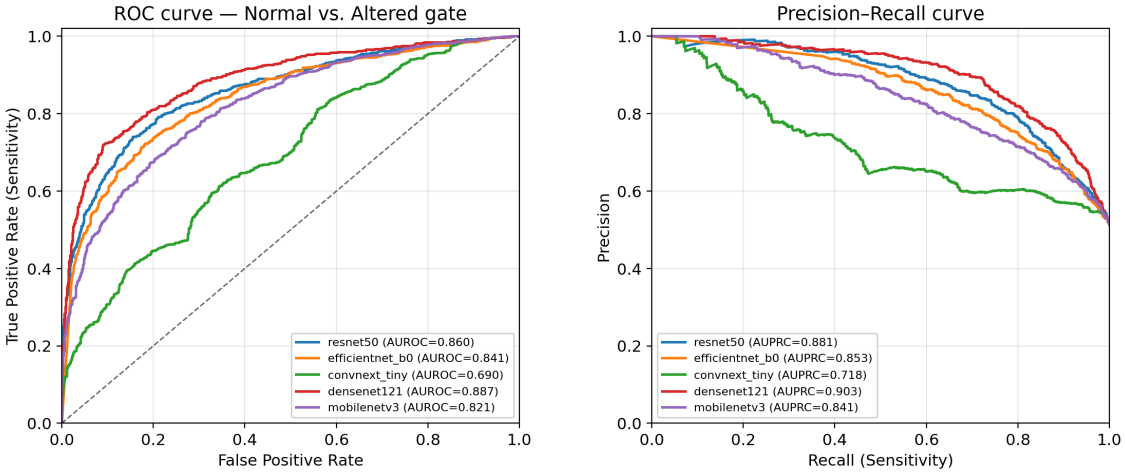
Table 3 summarizes the discrimination metrics. DenseNet-121 achieved the highest point estimates (AUROC 0.887; AUPRC 0.903) and was adopted as the reference. ResNet-50 ranked second (AUPRC 0.881), with slight Confidence Interval (CI) overlap: the DenseNet-121 advantage is consistent in point estimate but is not statistically separated from ResNet-50. ConvNeXt-Tiny performed considerably worse and showed high fold-to-fold variability (AUPRC 0.726 ± 0.108), a sign of convergence instability. Figure 2 shows the ROC and precision-recall curves.

Table 3. Discrimination of the five architectures (pooled estimate, 95% CI by bootstrap, n=1,984).

Model	AUROC	95% CI (ROC)	AUPRC	95% CI Precision-Recall (PR)
DenseNet-121	**0.887**	[0.873 – 0.901]	**0.903**	[0.887 – 0.917]
ResNet-50	0.860	[0.845 – 0.876]	0.881	[0.863 – 0.897]
EfficientNet-B0	0.841	[0.823 – 0.858]	0.853	[0.830 – 0.873]
MobileNetV3	0.821	[0.802 – 0.838]	0.841	[0.820 – 0.859]
ConvNeXt-Tiny	0.690	[0.666 – 0.712]	0.718	[0.689 – 0.746]

In calibration (Table 4), DenseNet-121 had the lowest Brier score (0.142) and competitive ECE (0.101). ConvNeXt-Tiny recorded the lowest ECE (0.068) but a high Brier score (0.225), suggesting concentration of probabilities in the central range rather than real discrimination. Figure 3 shows the reliability diagram of the reference model.

[1] https://github.com/Willgnner-Santos/safe-endoscopy-gate



(a) (b)

Figure 2. Discrimination of the five models (pooled predictions over the 5 folds): (a) ROC curves; (b) precision-recall curves.

Table 4. Calibration metrics by model.

Model	ECE	MCE	Brier
DenseNet-121	0.101	0.413	**0.142**
ResNet-50	0.112	0.318	0.161
ConvNeXt-Tiny	**0.068**	**0.150**	0.225
EfficientNet-B0	0.209	0.713	0.213
MobileNetV3	0.203	0.547	0.217

4.2. Operating points and gate safety

Table 5 details the operating points of DenseNet-121. At the Youden point, sensitivity and specificity are 72.3% and 88.9%. For screening, the primary target is sens≥95%: the obtained sensitivity is 94.9%, specificity drops to 44.4%, and NPV remains at 89.2%, meaning that 89.2% of discarded images are in fact normal. To quantify uncertainty at this point, 95% CIs were computed by bootstrap (2,000 resamples) on the pooled predictions: sensitivity [93.8%; 96.5%], specificity [43.5%; 49.7%], PPV [62.8%; 67.7%], NPV [87.5%; 92.6%], and filtered fraction [23.3%; 27.0%]. The trade-off of specificity for sensitivity is acceptable in screening, provided the per-pathology cost is documented.

Table 5. Clinical operating points of DenseNet-121 (pooled, n=1,984).

Operating point	Threshold	Sensitivity	Specificity	PPV	NPV	F1	MCC
Youden	0.693	0.723	0.889	0.873	0.753	0.791	0.619
Sens $\geq$ 90%	0.170	0.896	0.630	0.719	0.852	0.798	0.548
Sens $\geq$ 95%	**0.048**	**0.949**	0.444	0.643	**0.892**	0.767	0.458
Sens $\geq$ 99%	0.002	0.987	0.141	0.548	0.913	0.705	0.243
Spec $\geq$ 95%	0.887	0.605	0.947	0.924	0.694	0.731	0.584

Table 6 summarizes gate operation at the three targets, using the 616 images with pathology as risk cases. Sensitivity is measured over the 1,019 ABNORMAL images, while the safety miss-rate uses the 616 with named pathology, since the remaining ABNORMAL images did not have a specific category annotated. At sens≥95%, the gate filters 24.2% of images (480/1,984) and fails to refer 28 of the 616 with pathology (4.5%;

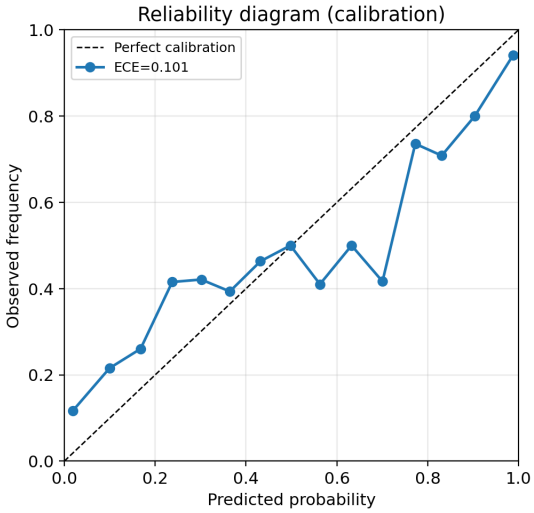


Figure 3. Reliability diagram of DenseNet-121. Each bar indicates the deviation between mean confidence and the fraction of positives in the interval.

95% CI by bootstrap [3.0%; 6.3%]). The 90% and 99% targets are less favorable: at 90%, 8.6% of pathologies are missed for only 11.8 additional percentage points of filtering, and at 99% the filtering drops to 7.5%. The corresponding safety-efficiency curve is shown in Figure 4(a).

Since the safety miss-rate considers only the 616 images with named pathology, we verified what happens with the 409 ABNORMAL images without a named pathology. At this operating point, 23 of them (5.6%) are filtered as NORMAL, a rate close to that of images with named pathology (4.3%); that is, the gate does not disproportionately discard unannotated abnormalities, and restricting the analysis to the 616 images does not underestimate the risk for the remaining cases. The robustness of the confidence interval was also assessed: a center-stratified bootstrap, which respects the structure of both sites, produces CIs practically identical to those of the simple bootstrap (sensitivity [93.8%; 96.4%]; miss-rate [3.0%; 6.3%]), indicating that uncertainty is not underestimated by aggregation at the image level.

Table 6. Gate safety and efficiency (DenseNet-121) at the three target sensitivity points.

Target sens.	Threshold	Obtained sens.	Spec.	NPV	Filtered	Lost pathologies
90%	0.170	89.6%	63.0%	0.852	36.0%	53/616 (8.6%)
95%	**0.048**	**94.9%**	44.4%	**0.892**	24.2%	**28/616 (4.5%)**
99%	0.002	98.7%	14.1%	0.913	7.5%	9/616 (1.5%)

In relative terms, the gate miss-rate is higher for rare and subtle classes (Table 7). EROSION, ULCER, and ERYTHEMA have miss-rates of up to 5.5%; POLYP reaches 10% and MICRONODULARITY 17.6%. VASCULAR ECTASIA and NEOPLASIA each miss 1 of 4 cases (25%), an unstable rate given the small number of examples. In absolute counts, however, most missed pathology instances belong to the more prevalent classes: ERYTHEMA (11) and EROSION (8) account for more lost instances than MICRONODULARITY (6), POLYP (5), and the two rarest categories (1 each).

Table 7. Miss-rate by pathology at the sens≥95% point (threshold=0.048, DenseNet-121).

Pathology	Total	Detected	Lost	Miss-rate
EROSION	321	313	8	2.5%
ERYTHEMA	199	188	11	5.5%
ULCER	81	79	2	2.5%
POLYP	50	45	5	10.0%
MICRONODULARITY	34	28	6	17.6%
VASCULAR ECTASIA	4	3	1	25.0%*
NEOPLASIA	4	3	1	25.0%*

* Unstable estimate due to extremely small number of cases (n=4).

The sum of lost pathology instances (34) exceeds the 28 lost images because one image may have more than one active pathology.

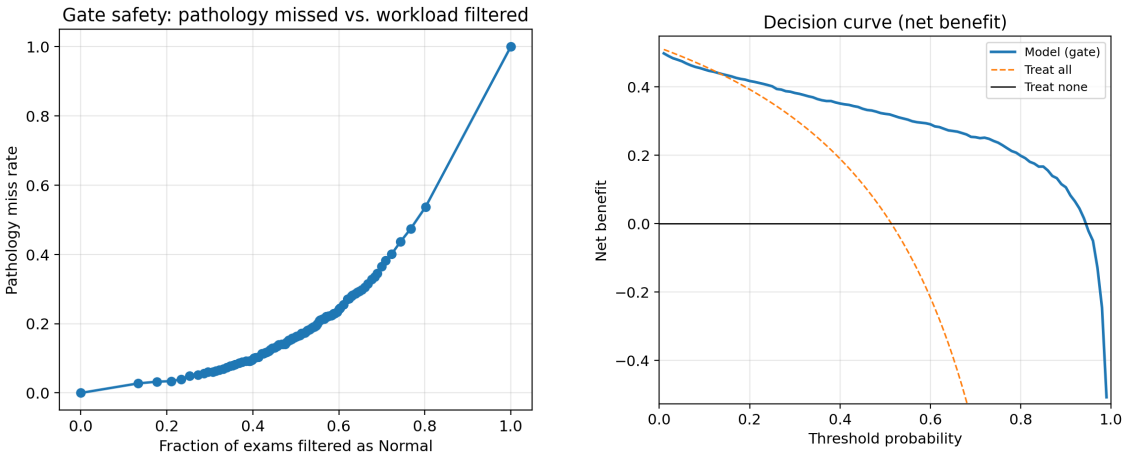
4.3. Center heterogeneity and net benefit

With the global threshold of 0.048, there is heterogeneity between centers. In Center 1 (ABNORMAL prevalence 35.1%), sensitivity is 87.7%, below the target, with 36.8% filtering and 17/136 pathologies missed (12.5%). In Center 2 (60.5%), it reaches 97.3%, above the target, with 17.0% filtering and 11/480 missed (2.3%). Calibration also differs by center (ECE 0.105 in Center 1 and 0.084 in Center 2). A single threshold does not guarantee the target in both centers, but two local corrections, both adjusted on validation data and applied to test, address the heterogeneity without retraining. The first is a per-center threshold: derived from local validation, it raises sensitivity to 94.0% in Center 1 (19.4% filtering) and 92.4% in Center 2 (30.7% filtering). The second is per-center temperature scaling, with the temperature adjusted on local validation; the procedure reduces ECE from 0.105 to 0.042 in Center 1 and from 0.084 to 0.028 in Center 2, and, combined with the local threshold, keeps sensitivity close to but still below the 95% target (94.0% and 92.7%). Both local corrections thus attenuate the heterogeneity and improve calibration, but did not consistently guarantee the 95% target in both centers. Per-center calibration with data from external institutions remains a future verification.

Figure 4 summarizes the gate operation: the safety-efficiency curve (a) relates the filtered fraction to the pathology miss-rate, and the decision curve (b) compares net benefit against the reference strategies. In DCA, net benefit is $\text{NB} = \text{TP}/n - \text{FP}/n \cdot p_t/(1-p_t)$, where p_t is the probability threshold that expresses the exchange ratio between harms (True Positive (TP) and False Positive (FP) counted at the image level, $n = 1{,}984$). In UGE screening, false negatives are more serious than false positives, placing clinical interest at low values of p_t. The gate outperforms both reference strategies between p_t values of 0.14 and 0.94.

4.4. Threshold selection robustness

Defining the threshold on the pooled predictions may introduce optimistic bias, since it estimates and evaluates the operating point on the same partition. To measure this effect, the analysis was repeated in the nested scheme: the threshold for each target sensitivity was obtained from the validation partition of each fold and applied to its test set, with subsequent pooling. Table 8 compares the two procedures at the primary point sens≥95%. The differences are small (sensitivity 95.3% vs. 94.9%; filtered fraction 23.2% vs. 24.2%;



(a) (b)

Figure 4. Gate operation (DenseNet-121): (a) safety-efficiency curve, with pathology miss-rate as a function of the filtered fraction and the operating points for target sensitivities of 90%, 95%, and 99%; (b) decision curve (DCA) against the treat-all and treat-none strategies.

miss-rate 4.9% vs. 4.5%) and the 95% intervals overlap, indicating that the threshold selection bias is negligible in this dataset and that the reported operating points are stable. At the fold level, sensitivity varied between 0.920 and 0.972 (Standard Deviation [SD] = 0.022), with corresponding thresholds in the interval [0.010; 0.152]; the operational sensitivity is more stable than the absolute threshold, since small variations in the tail of the score distribution alter τ more than the gate performance.

Table 8. Sens≥95% point under threshold defined on pooled predictions and under threshold defined on validation and applied to test (nested). 95% CI by bootstrap.

Procedure	Sens.	Spec.	NPV	Filtered	Lost pathol.
Pooled	0.949	0.444	0.892	24.2%	28/616 (4.5%)
Nested	0.953	0.428	0.896	23.2%	30/616 (4.9%)
95% CI (nested)	[0.939–0.965]	[0.397–0.461]	[0.868–0.922]	[21.4–25.1]%	[3.2–6.6]%

5. Conclusion

We investigated the feasibility of a binary NORMAL/ABNORMAL screening gate for UGE images from a Brazilian context, focusing on the per-pathology operational cost. Five CNNs were evaluated in stratified 5-fold cross-validation on 1,984 images from two centers, adopting target sensitivity as the configuration criterion. DenseNet-121 showed the best discrimination performance (AUROC 0.887, 95% CI [0.873–0.901]; AUPRC 0.903, 95% CI [0.887–0.917]). At sens≥95%, the gate filters 24.2% of images with an NPV of 0.892 and fails to identify 28 of the 616 images with pathology (4.5%); relative miss-rates were higher for rare and subtle categories (MICRONODULARITY, VASCULAR ECTASIA, NEOPLASIA), although most missed pathology instances in absolute counts occurred in more prevalent classes; for VASCULAR ECTASIA and NEOPLASIA (4 cases each), the percentages serve as qualitative alerts. DCA indicated positive net benefit between 0.14 and 0.94. Defining the threshold on validation and applying it to test (nested scheme) reproduced these numbers, indicating negligible threshold selection bias.

Limitations include the small volume for rare classes, which prevents reliable es-

timates for NEOPLASIA and VASCULAR ECTASIA, and the prevalence heterogeneity between centers, which causes a global threshold to fall short of the 95% target in Center 1 (87.7%). We showed that per-center threshold adjustment and temperature scaling, both calibrated on local validation, attenuate the heterogeneity and reduce calibration error, but did not consistently guarantee the 95% target in both centers; verification with external institutions is still required. The unit of analysis is the image, not the exam: the curated dataset carries no patient or exam identifier, which prevents aggregating multiple images per exam and partitioning by patient. Cosine similarity analysis did not find visual duplicates and the center-stratified bootstrap revealed no underestimation of uncertainty; nevertheless, leakage of frames from the same exam across folds cannot be excluded without an exam identifier. Miss-rates are per image: with exam-level aggregation (maximum or quantile across frames), safety estimates and utility bounds in DCA will likely differ. The study also did not evaluate external generalization by center exclusion, which requires retraining.

Future work includes collecting data with exam identifiers for exam-level aggregation and patient-level partitioning, leave-one-center-out cross-validation as a pipeline to measure external generalization without retraining on new data, the use of higher input resolutions or architectures designed for fine-grained processing such as Vision Transformers to reduce miss-rates in subtle classes like micronodularity, and formal safety strategies, such as selective prediction with statistical guarantees or conformal risk control, which could impose the sensitivity target with certified coverage even in rare classes. The results provide a quantitative reference point for screening gates in Brazilian UGE and support for designing them with safeguards appropriate to the characteristics of local data.

References

Al Hayek, M., Barberio, B., Al Hayek, O., Rugge, M., Di Pietro, M., Noviello, D., et al. (2026). Efficacy of artificial intelligence–assisted upper gastrointestinal endoscopy for neoplasm detection: a systematic review and meta-analysis of randomized controlled trials. *Gastrointestinal Endoscopy*.

Ali, H., Yasmin, M., Sharif, M., and Rehmani, M. H. (2018). Computer assisted gastric abnormalities detection using hybrid texture descriptors for chromoendoscopy images. *Computer Methods and Programs in Biomedicine*.

Arribas, J., Antonelli, G., Frazzoni, L., Fuccio, L., Ebigbo, A., van der Sommen, F., Ghatwary, N., et al. (2021). Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. *Gut*.

Chang, Y. H., Shin, C. M., Lee, H. D., Park, J., Jeon, J., Cho, S.-J., Kang, S. J., et al. (2024). Real-world application of artificial intelligence for detecting pathologic gastric atypia and neoplastic lesions. *Journal of Gastric Cancer*.

Choi, S. J., Khan, M. A., Choi, H. S., Choo, J., Lee, J. M., Kwon, S., Keum, B., and Chun, H. J. (2022). Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy. *Surgical Endoscopy*.

Fujishiro, M., Kakushima, N., Abe, S., Yip, H. C., Liu, X., et al. (2026). World Endoscopy Organization Position Statements for Artificial Intelligence in Endoscopic Diagnosis of Gastric Neoplasia. *Digestive Endoscopy*.

He, Y., Chen, Q., Liu, B., and Cao, Y. (2025). Foundational multi-task multimodal model for upper GI endoscopy. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Krenzer, A., Friedetzki, T., Dietsch, M., and Puppe, F. (2026). Classification of gastroscopy images under extreme class imbalance: A deep learning pipeline. In *Lecture Notes in Computer Science*.

Lui, T. K. L., Tsui, V. W. M., and Leung, W. K. (2020). Accuracy of artificial intelligence-assisted detection of upper GI lesions: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*.

Osawa, S., Yamada, T., Inui, W., Niwa, T., Takahashi, K., Egami, T., et al. (2026). Real-world evaluation of an AI-assisted diagnostic support system for early gastric cancer: Diagnostic performance, confidence stratification, and determinants of false-positive diagnosis. *Journal of Clinical Medicine*.

Su, X., Liu, Q., Gao, X., and Ma, L. (2023). Evaluation of deep learning methods for early gastric cancer detection using gastroscopic images. *Technology and Health Care*.

Teramoto, A., Shibata, T., Yamada, H., Hirooka, Y., Saito, K., and Fujita, H. (2022). Detection and characterization of gastric cancer using cascade deep learning model in endoscopic images. *Diagnostics*.

Truong, V. D., Nguyen, D. A., Manh, X. H., Nguyen, T. H., Pham, T. T., Nguyen, T. T., et al. (2024). Automatic classifying informative frames for upper gastrointestinal endoscopy diagnosis using deep learning method. In *Lecture Notes in Networks and Systems*.

Yang, H., Wu, Y., Yang, B., Wu, M., Zhou, J., Liu, Q., Lin, Y., Li, S., Li, X., Zhang, J., Wang, R., Xie, T., et al. (2022). Identification of upper GI diseases during screening gastroscopy using a deep convolutional neural network algorithm. *Gastrointestinal Endoscopy*.

Yin, M., Zhang, R., Lin, J., Zhu, S., Liu, L., Liu, X., Lu, J., Xu, C., et al. (2023). Identification of gastric signet ring cell carcinoma based on endoscopic images using few-shot learning. *Digestive and Liver Disease*.

You, H., Tao, X., Dong, Z., Li, Y., Huang, L., Li, J., Du, H., Deng, M., Wu, Z., et al. (2026). Artificial intelligence assistance narrows the experience gap in endoscopic reporting of gastric lesions: a prospective study. *Surgical Endoscopy*.

Zhang, L., Yao, L., Lu, Z., and Yu, H. (2024). Current status of quality control in screening esophagogastroduodenoscopy and the emerging role of artificial intelligence. *Digestive Endoscopy*.