

Article

Characterizing Multi-Domain Representations for Gastric Histopathological Image Classification

Zayra Reyes-Vera ^{1,*}, Jesús Yaljá Montiel-Pérez ^{1,*}, José Luis López-Bonilla ², Elsa Rubio-Espino ¹,
Juan Humberto Sossa-Azuela ¹, Víctor Hugo Ponce-Ponce ¹, Osvaldo Espinosa-Sosa ¹,
and Sergio Vidal-Beltrán ²

¹ Centro de Investigación en Computación, Instituto Politécnico Nacional, Ciudad de México 07738, Mexico; erubio@cic.ipn.mx (E.R.-E.); hsossa@cic.ipn.mx (J.H.S.-A.); vponce@cic.ipn.mx (V.H.P.-P.); espinosa@cic.ipn.mx (O.E.-S.)

² Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), Unidad Zacatenco, Instituto Politécnico Nacional, Ciudad de México 07738, Mexico; jlopezb@ipn.mx (J.L.L.-B.); svidalb@ipn.mx (S.V.-B.)

* Correspondence: zreyesv2025@cic.ipn.mx (Z.R.-V.); jyalja@cic.ipn.mx (J.Y.M.-P.)

Abstract

This study characterizes spatial, spectral, relational, and contextual representations for binary gastric histopathological image classification using FLIGENT-Net. The framework combines an ImageNet-pretrained convolutional backbone, patch-level fast Fourier transform (FFT) descriptors, graph convolution over a cosine-similarity graph, and a Transformer Encoder. Experiments use the 160-by-160 GasHisSDB subset, a persisted stratified 60%/20%/20% partition, and one training run per configuration with seed 42. In the nine completed configurations of the revised evaluation, FLIGENT-Net achieved 97.31% accuracy, 96.57% precision, 96.61% recall, 96.59% F1-score for the Abnormal class, and 0.9964 AUC-ROC. An additional native-grid FFT control yielded 97.43% accuracy and 96.74% F1 under an alternative spectral protocol that preserved the 160-by-160 sampling grid while changing the local FFT window sizes and spectral discretization. The conventional CNN baseline (ResNet18) obtained the highest accuracy and F1-score among these completed runs, at 98.51% and 98.13%, respectively. These results do not establish a predictive advantage for the complete multi-domain architecture; rather, they characterize how the added representation domains affect predictive behavior, graph topology, computational cost, and attribution response under a common protocol. The study contributes a reproducible integration of the four representation domains and a controlled empirical characterization of their combinations. Transformer attention visualizations describe information allocation among contextualized nodes, while a common input-gradient deletion protocol examines attribution behavior. Together, these analyses identify the behavior and limitations of the implemented fusion strategy and provide a basis for its further development. Performance estimates are conditional on the available patch-level split and should not be interpreted as patient-, slide-, or source-image-independent generalization; run-to-run variability is also not estimated because one fixed seed is used per configuration.



Academic Editors: Juan Ramón Torregrosa Sánchez and Jonathan Blackledge

Received: 3 September 2026

Revised: 19 September 2026

Accepted: 20 October 2026

Published: 23 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\) license](#).

Keywords: computational methods; computer-aided diagnosis; fast Fourier transform; frequency-domain image analysis; gastric cancer; graph neural networks; graph signal processing; histopathology image classification; multi-domain signal representation; vision transformers

MSC: 68T07; 68U10; 68T45

1. Introduction

Gastric cancer represents one of the main oncological pathologies worldwide. According to GLOBOCAN 2020 data [1–3], stomach cancer caused approximately 800,000 deaths (representing 7.7% of total cancer deaths), positioning itself as the fifth leading cause of cancer mortality in both men and women [4].

Beyond its clinical relevance, gastric histopathological image classification is fundamentally a computational signal processing problem. A digitized tissue image can be interpreted as a two-dimensional signal whose diagnostic information is not confined to a single representation domain: local pixel intensities describe spatial morphology, the Fourier spectrum encodes textural periodicity and roughness [5], similarity-based graph structures capture non-Euclidean relational dependencies through graph signal processing operators [6], and self-attention mechanisms model long-range contextual dependencies across the image [7]. Framed this way, the design of computational methods for gastric cancer detection becomes an instance of a broader mathematical problem: how to combine complementary spatial-, frequency-, and graph-domain signal representations into a single discriminative model. This perspective motivates the multi-domain characterization pursued in the present study, which extends beyond a purely diagnostic evaluation to examine the mathematical and computational behavior of each representation domain. Accordingly, the mathematical interest of the study lies not in introducing a new transform, graph operator, or attention mechanism but in analyzing the interaction of established operators acting on different representations of the same histological signal within one explicitly defined computational pipeline.

One of the causes of this mortality is late detection, since initial symptoms are usually nonspecific and are confused with common conditions such as colitis or gastritis. Generally, the patient seeks clinical review only when symptoms become persistent and acute, at which point an endoscopy is performed to inspect lesions in internal organs.

Upon finding irregularities, the intervention of a pathologist specialized in histopathology is required. Tissue biopsy specimens are processed into histopathological slides and subsequently digitized to produce whole slide images (WSIs) for microscopic examination and computer-aided analysis [8–10]. These digital slides enable detailed visualization of cellular morphology and tissue architecture, facilitating the identification of potential abnormalities. However, manual analysis of these samples is an exhaustive task; since they contain complex and subtle cellular structures, there is an inherent risk of human error or diagnostic disagreement, a phenomenon known as inter-observer variability [11,12]. This discrepancy in results can delay definitive diagnosis by requiring the opinion of multiple specialists. To mitigate these limitations, deep learning models and architectures have been developed for automated gastric cancer detection. These technologies emerge as vital support tools that allow experts to diagnose with greater certainty, optimizing early detection and reducing the burden of manual analysis. However, despite the effectiveness of convolutional neural networks (CNNs), these models present a fundamental limitation: their focus is mainly restricted to the analysis of local morphological features. In gastric histopathology, subtle textures and large-scale structural organization are critical indicators that are often lost in architectures that do not consider the frequency domain or global spatial relationships.

In response to the limitations of conventional approaches, FLIGENT-Net (Frequency-Localized Image Graph Encoding Network with Transformer) was developed based on the graph-Transformer framework of SCUBa-Net [13]. While SCUBa-Net provides the architectural foundation for combining graph-based representation learning and Transformer-based contextual modeling, FLIGENT-Net was redesigned for binary classification of gastric histopathological images and incorporates a frequency-domain representation into the fea-

ture learning process. The proposed architecture introduces the frequency-localized image (FLI)-extractor, which combines spatial patch representations with texture information obtained through the fast Fourier transform (FFT). The resulting representation is integrated into a graph through cosine-similarity-based connectivity and subsequently processed by graph convolutional layers before global feature aggregation. In this way, FLIGENT-Net extends the original graph-Transformer paradigm by incorporating complementary spatial and frequency-domain information into the representation used for classification.

2. Related Work

Automated classification of gastric malignancies from histopathological images has transitioned from manually designed visual descriptors toward convolutional, attention-based, ensemble, and hybrid representation-learning approaches. A recent scoping review of gastric-cancer pathology image analysis covering studies published from 2020 to 2025 found CNNs to remain the most common model family, while Vision Transformers, hybrid models, transfer learning, and ensemble strategies are increasingly represented; the same review also identified limited external validation, restricted dataset diversity, and insufficient interpretability as recurring limitations [14]. This section therefore reviews convolutional and feature-fusion pipelines, graph-based and Transformer-based contextual modeling, and frequency-domain representations, with emphasis on how these components relate to the four-domain formulation evaluated in FLIGENT-Net.

2.1. Convolutional Pipelines for Gastric Histopathology

Deep learning pipelines driven by convolutional neural networks (CNNs) have established robust performance baselines in digital pathology due to their ability to automatically extract hierarchical visual features [11,15–17]. In gastric tissue analysis, early efforts focused on adapting deep transfer learning models to distinguish malignant cells. For example, Iizuka et al. [18] developed convolutional models capable of classifying gastric and colorectal epithelial tumors, demonstrating that deep architectures could effectively isolate structural atypia. Similarly, Li et al. [19] conducted a comprehensive investigation of gastric cancer diagnosis using deep neural structures, validating the relevance of automated morphology extraction.

Following the release of the benchmark dataset GasHisSDB [20], multiple convolutional variants have been evaluated to refine classification accuracy. Hu et al. [21] compared classical classifiers with VGG16, Inception-V3, ResNet50, and ViT across the three GasHisSDB patch sizes, providing a useful reference for how model behavior changes with image size and training duration. Bratajaya et al. [22] leveraged CNN frameworks specifically for early stomach cancer detection. To maximize performance, ensemble approaches have also been introduced. In particular, Yong and Hum [23] studied transfer-learning variants, which were subsequently extended to a deep ensemble analysis on GasHisSDB [24]. Mudavadkar et al. [25] likewise evaluated ensembles of pretrained CNN and vision-model backbones across the supplied GasHisSDB resolutions.

A complementary direction is feature-level rather than decision-level fusion. Loddo et al. [26] compared handcrafted descriptors with deep features extracted from pretrained CNNs on GasHisSDB and evaluated several feature-fusion strategies with shallow classifiers. Their best reported accuracy was 95% with an SVM, and their cross-resolution experiments further emphasized that performance can change when the image-size distribution at test time differs from that used for model development. This work is particularly relevant to FLIGENT-Net because it illustrates the potential value of combining heterogeneous representations while also showing that the utility of a fused descriptor depends on the downstream classifier and evaluation protocol.

These studies establish strong spatial-feature baselines and demonstrate the value of transfer learning, feature fusion, and ensemble aggregation. However, their principal representations remain spatial or feature-aggregation based; they do not explicitly combine local Fourier descriptors, learned image-region relations, and global self-attention within one patch-level gastric histopathology architecture.

2.2. Graph Topological Abstractions and Global Attention Mechanisms

To represent spatial organization beyond regular convolutional grids, computational pathology increasingly models tissue regions, glands, nuclei, or image tiles as graph nodes connected by relational edges. Following early structural graph formulations in pathology [27], GNN methodology has been formalized more broadly for non-Euclidean representation learning [6,28,29]. In gastric histopathology, Xiang and Wu [30] used multiple-instance learning [31] together with graph representations of pathological image structure, illustrating how relational modeling can complement local appearance features.

The recent literature also shows that graph construction itself is an active research problem rather than a fixed design choice. Brussee et al. [32], in a dedicated review of GNNs in histopathology, identify hierarchical GNNs, adaptive graph-structure learning, multimodal GNNs, and higher-order GNNs as emerging directions. This perspective is relevant to the present work because FLIGENT-Net uses a deliberately simple, non-learned topology: pairwise cosine similarity between CNN descriptors followed by hard thresholding. The proposed graph should therefore be interpreted as one reproducible relational representation rather than as an optimized graph-learning mechanism.

Graph and Transformer processing have also been combined in recent WSI research. Shi et al. [33] proposed an integrative graph–Transformer framework for weakly supervised WSI classification in which a GCN models local neighboring relations and efficient global attention captures long-range dependencies. Their experiments were performed on TCGA-NSCLC, TCGA-RCC, and BRIGHT rather than on gastric patch classification, but the work is important conceptually because it demonstrates that graph–Transformer integration itself is not unique to FLIGENT-Net. The distinction in the present architecture is the explicit coupling of such relational and contextual processing with patch-level Fourier descriptors and CNN-derived spatial representations.

In parallel, Vision Transformers (ViTs) have become prominent in digital pathology because self-attention can model dependencies between spatially separated image regions [7]. Wang et al. [34] introduced TransPath, a self-supervised histopathology representation-learning framework that combines a convolutional network with a modified Transformer architecture to capture local and global information. Shao et al. [35] introduced TransMIL, a Transformer-based correlated multiple-instance learning framework for WSI classification. In gastric histopathology specifically, Chen et al. [36] proposed GasHis-Transformer, combining a convolutional local-information module with a position-encoded Transformer for global information modeling on the HE-GHI-DS gastric histopathology dataset. Armand et al. [37] subsequently evaluated a ViT-based approach for gastric-cancer detection and classification from weakly annotated histopathological images and extended the model to region-based prediction. These works establish the relevance of Transformer-based context modeling in gastric pathology, while none of the gastric approaches reviewed here explicitly couples Transformer context with a similarity-derived graph and a local FFT branch as done in the present framework.

2.3. Frequency and Spectral Domain Analysis for Histological Textures

Frequency-domain texture analysis provides a complementary representation of tissue architecture by emphasizing periodicity, roughness, and structural organization that may

not be captured directly by spatial features. Gibson and Gaydecki [5] demonstrated the utility of Fourier-domain texture measures for histological image analysis, while subsequent texture-analysis studies and reviews emphasized the complementary role of spatial and spectral descriptors in complex image-classification problems [38–40]. Despite this potential, explicit frequency-domain representations remain comparatively underexplored in recent gastric histopathology classification pipelines, which predominantly rely on spatial deep features.

2.4. The Multi-Domain Architectural Gap

The recent literature therefore contains several partial combinations of the representation mechanisms considered here. CNN and Transformer fusion has been explored for gastric histopathology [36]; graph–Transformer integration has been developed for WSI classification [33]; GasHisSDB studies have evaluated transfer learning, ensembles, and heterogeneous feature fusion [24–26]; and graph-based pathology analysis continues to evolve toward adaptive and hierarchical topology learning [32]. Accordingly, the contribution of FLIGENT-Net is not the isolated novelty of CNNs, GNNs, FFTs, or Transformers, nor the first use of any pairwise combination of these components.

Instead, FLIGENT-Net provides a controlled setting for studying the joint use of four explicit representation domains: CNN-derived spatial descriptors, patch-localized Fourier-magnitude descriptors, similarity-derived graph relations, and Transformer-based global context. Among the gastric and computational-pathology approaches reviewed here, we did not identify a method that simultaneously implements this particular four-part representation flow at image-patch level. This statement is deliberately scoped to the reviewed literature and should not be interpreted as an exhaustive claim over all computational-pathology publications.

The graph used in FLIGENT-Net is obtained by hard thresholding cosine similarities between learned CNN descriptors. The cosine function and threshold are not trainable parameters, and gradients are not propagated through the binary edge decision. The computational contribution therefore lies in the explicit mapping and integration of established operators within a reproducible architecture, together with an empirical characterization of their combinations rather than a claim that any individual operator is new.

Comparative Analysis of Architectures

To contextualize FLIGENT-Net’s contribution within the current landscape, Table 1 presents a structural comparison of representative gastric-histopathology and computational-pathology architectures discussed above. The matrix identifies the primary modeling mechanisms of each approach in terms of spatial extraction (CNN), explicit frequency analysis (FFT), topological modeling (GNN), global attention (ViT), and auxiliary strategies such as ensemble learning, transfer learning, and MIL; because the studies use different datasets and protocols, the table is not intended as a performance ranking.

Table 1 emphasizes that the relevant literature spans different architectural families and datasets. CNN-based and transfer-learning studies [18,19,22–26] primarily emphasize spatial features, pretrained representations, feature fusion, or decision-level ensembles. Gastric Transformer studies add explicit global-context modeling [36,37], while Xiang and Wu [30] and Shi et al. [33] demonstrate relational and graph–attention formulations in different pathology settings.

The table is intentionally structural rather than quantitative because these studies do not share a common dataset, split, label granularity, preprocessing pipeline, or initialization protocol. In particular, the presence of graph–Transformer and CNN–Transformer models in prior work means that FLIGENT-Net should not be framed as introducing those

pairwise combinations. Its specific experimental object is the four-domain flow in which patch-localized FFT descriptors are fused with CNN descriptors, graph relations are constructed from CNN similarity, and GCN-refined nodes are subsequently contextualized by a Transformer.

Table 1. Structural comparison of representative gastric-histopathology and computational-pathology approaches. The table summarizes primary modeling mechanisms and is not a performance ranking; the studies use different datasets and evaluation protocols. ✓ indicates that a component or strategy is present, whereas ✗ indicates that it is absent.

Study	CNN	FFT	GNN	ViT	Ensemble	MIL	Transfer
Iizuka et al. (2020) [18]	✓	✗	✗	✗	✗	✗	✓
Li et al. (2020) [19]	✓	✗	✗	✗	✗	✗	✓
Bratajaya et al. (2024) [22]	✓	✗	✗	✗	✗	✗	✗
Yong & Hum (2023) [23]	✓	✗	✗	✗	✗	✗	✓
Yong et al. (2023) [24]	✓	✗	✗	✗	✓	✗	✓
Xiang & Wu (2021) [30]	✓	✗	✓	✗	✗	✓	✗
Hu et al. (2022) [21]	✓	✗	✗	✗	✗	✗	✓
Mudavadkar et al. (2024) [25]	✓	✗	✗	✗	✓	✗	✓
Loddo et al. (2024) [26]	✓	✗	✗	✗	✗	✗	✓
Chen et al. (2022) [36], GasHis-Transformer	✓	✗	✗	✓	✗	✗	✗
Armand et al. (2024) [37]	✗	✗	✗	✓	✗	✗	✗
Shi et al. (2024) [33]	✗	✗	✓	✓	✗	✓	✗
FLIGENT-Net	✓	✓	✓	✓	✗	✗	✓

3. Materials and Methods

This section describes the materials, dataset, preprocessing procedures, and methodological framework used in this study. First, the GasHisSDB dataset and the experimental data partition are described. The preprocessing and image representation procedures are then presented, followed by the proposed multi-domain representation framework based on spatial, frequency-domain, graph-based, and global contextual information. The construction of the evaluated architectural configurations and their corresponding training and evaluation procedures are subsequently detailed. Finally, the metrics and experimental protocols used to assess classification performance, computational cost, interpretability, representation consistency, and faithfulness are presented.

3.1. Dataset GasHisSDB

The proposed FLIGENT-Net architecture was developed and evaluated using the publicly available GasHisSDB dataset [20]. GasHisSDB is a gastric histopathology image dataset composed of cropped histopathological images. The dataset paper describes 600 source images of 2048×2048 pixels; this count should not be interpreted as the number of patients or whole-slide images (WSIs). It contains two diagnostic categories: Normal, corresponding to healthy gastric tissue, and Abnormal, corresponding to gastric adenocarcinoma.

In this work, only the 160×160 pixel subset of GasHisSDB was employed. It has the largest patch dimensions among the supplied subsets; this should not be interpreted as a higher acquisition magnification or spatial sampling resolution.

The revised classification experiments use a single stratified partition with 60% training, 20% validation, and 20% test data, generated with seed 42 and saved as an explicit manifest. Every configuration reuses the same file lists. The loader checks file identities and content hashes; it does not silently replace unreadable images. Table 2 reports the available inventory and confirmed partition sizes. The nine saved manifests are identical

and their SHA-256 digests match the corresponding configuration records. Their index sets and image paths are disjoint across partitions and together cover all 33,284 records. All recorded image hashes are distinct; these hashes were audited from the manifests, without independently rehashing the source images. The ordered test paths and labels match all nine prediction reports.

Table 2. Dataset inventory and revised split verified from the nine saved manifests.

Subset	Normal	Abnormal	Total
Full subset	20,160	13,124	33,284
Training	12,096	7874	19,970
Validation	4032	2625	6657
Test	4032	2625	6657

The implemented split operates at image-patch level. It does not group samples by patient, source image, or slide, so patient- or slide-independent evaluation cannot be claimed. A content-hash check can identify byte-identical files, but cannot exclude related, overlapping, or recompressed patches originating from a common source. Dependence between such patches may lead to optimistic performance estimates and limits direct comparison with studies that use source-grouped evaluation.

The dataset publication describes random rotation of the cropped images, shuffling, and renaming [20]. These operations do not establish independence between patches from the same source. No patch-to-patient or patch-to-source-image mapping was identified in the dataset paper, the inspected repository documentation, or the experimental manifests. Although the dataset authors link an original-image repository, correspondence between those source images and the distributed 160×160 patches has not been established in the materials available for this study. Consequently, the results reported here must be interpreted as patch-level performance. Recovering or obtaining provenance information that permits source-grouped partitioning is therefore an explicit direction for future evaluation.

3.2. Data Preprocessing

A unified preprocessing pipeline was applied to the GasHisSDB dataset to ensure a consistent input representation during training and evaluation. The histopathological images were resized to 224×224 pixels to match the ResNet18 input configuration and the specified 7×7 region grid.

Subsequently, the RGB images were converted into PyTorch tensors and normalized using the ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]).

The same preprocessing pipeline was applied to the training, validation, and testing subsets. No additional data augmentation techniques were employed, ensuring that all experiments were conducted under identical preprocessing conditions.

In the primary nine-configuration comparison, the FFT branch receives this resized, channel-normalized RGB tensor. The grayscale descriptor therefore uses a weighted combination of normalized channels and represents spectral information from the preprocessed input rather than raw optical intensities. Grayscale provides a compact intensity-based descriptor, while the CNN retains three-channel input; this design does not establish that stain/color information is irrelevant. No stain normalization or additional augmentation is applied, keeping the specified pipeline fixed but limiting conclusions about robustness to staining and acquisition variation.

To examine whether resizing the 160×160 GasHisSDB input before spectral extraction affects the frequency branch, an additional FLIGENT-Net control was evaluated using a

native-grid spectral variant. The training configuration was unchanged from the reference FLIGENT-Net run (100 maximum epochs, batch size 16, learning rate 10^{-4} , positive-class weight 2.0, graph threshold $\tau = 0.70$, classification threshold 0.55, patience 12, seed 42, and ImageNet-pretrained CNN backbone), and the persisted split hash was identical to the reference run. For the spectral branch, the 160×160 input grid was partitioned with integer boundaries

$$b_k = \left\lfloor \frac{160k}{7} \right\rfloor, \quad k = 0, \dots, 7, \quad (1)$$

which yields $[0, 22, 45, 68, 91, 114, 137, 160]$. This produces a complete, non-overlapping 7×7 partition with regions of 22 or 23 pixels per side while preserving the same 49-node spatial indexing used by the main architecture. The resulting run is treated as a preprocessing control rather than as an additional architecture in the nine-configuration ablation table.

3.3. Experimental Setup

All experiments were conducted on a desktop workstation equipped with an Intel® Core™ i7-12700KF processor operating at 3.60 GHz, 32 GB of RAM, and an NVIDIA GeForce RTX 4070 SUPER graphics processing unit (GPU) with 12 GB of dedicated memory. The workstation operated under Windows 10 (64-bit), and the proposed framework was implemented in Python 3.11.6 using the PyTorch 2.7.1 deep learning library with CUDA 11.8 acceleration.

Table 3 summarizes the main configuration used throughout the experiments on GasHisSDB.

Table 3. Experimental configuration adopted for the GasHisSDB experiments.

Parameter	Value
Operating System	Windows 10 (64-bit)
CPU	Intel Core i7-12700KF @ 3.60 GHz
GPU	NVIDIA GeForce RTX 4070 SUPER (12 GB)
RAM	32 GB
Python	3.11.6
PyTorch	2.7.1
CUDA	11.8
Optimizer	Adam
Initial learning rate	1×10^{-4}
Batch size	16
Maximum epochs	100
Loss function	BCEWithLogitsLoss
Positive class weight (pos_weight)	2.0
Graph connectivity threshold (τ)	0.70
Classification threshold	0.55
Learning rate scheduler	ReduceLROnPlateau
Scheduler factor	0.5
Scheduler patience	5 epochs
Early stopping patience	12 epochs
Random seed	42

3.3.1. Computational-Cost Benchmarking

Inference cost was measured for the same nine revised configurations using the NVIDIA GeForce RTX 4070 SUPER described above. The benchmark used float32 inputs of size 224×224 , batch size 16, and the same persisted test split used for classification. All models were placed in evaluation mode and executed under `no_grad`. Twenty warm-up forward passes were performed before timing. Measurements were then collected over

three rounds of 60 synchronized forward passes, using three fixed test batches beginning at indices 192, 912, and 5232. The reported batch latency is synchronized wall-clock time for the complete original model forward pass and logit extraction with inputs already resident on the GPU; image decoding, preprocessing, and host-to-device transfer are excluded.

For each architecture, the analysis reports total parameters, mean and 95th-percentile batch latency, amortized latency per image, throughput, and peak PyTorch allocated GPU memory. The memory value is an inference-time allocation measurement rather than training memory or an `nvidia-smi` device-wide measurement. CuDNN deterministic execution was enabled and TF32 matrix multiplication was disabled during profiling. The benchmark used the same split SHA-256 identifier as the revised classification experiments. Thus, the measurements characterize relative inference cost under one controlled hardware/software configuration and should not be interpreted as hardware-independent deployment latency.

3.3.2. Graph-Topology Characterization Protocol

For the six graph-based configurations, graph topology was characterized on all 6657 test images using the binary adjacency matrix entering the first GCN block. Self-loops were excluded from these statistics so that they describe only connections between distinct image regions. Each graph contains $N = 49$ nodes. For image m , let E_m denote the number of undirected off-diagonal edges and $d_{m,i}$ the off-diagonal degree of node i . We report

$$\rho_m = \frac{2E_m}{N(N-1)}, \quad \bar{d}_m = \frac{1}{N} \sum_{i=1}^N d_{m,i} = \frac{2E_m}{N}, \quad (2)$$

where ρ_m is graph density and $\bar{d}_m$ is mean node degree. A node is counted as externally isolated when $d_{m,i} = 0$, and a graph is called edgeless when $E_m = 0$. The reported values are averages over the complete test split, together with the fraction of test images whose graph has no off-diagonal edges. These statistics characterize the topology induced by each trained checkpoint at the fixed cosine threshold $\tau = 0.70$; they are not parameter-matched interventions and do not by themselves establish that one topology is preferable.

3.4. Evaluation Metrics

The performance of the evaluated models was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC).

Precision, recall, and F1-score were computed with respect to the positive class (Abnormal). Precision quantifies the proportion of samples predicted as abnormal that were correctly classified, whereas recall measures the proportion of abnormal samples that were correctly identified. The F1-score corresponds to the harmonic mean of precision and recall for the Abnormal class:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (4)$$

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5)$$

Accuracy was computed considering both classes as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

AUC-ROC was computed from the predicted probabilities for the positive (Abnormal) class and summarizes the ability of the model to discriminate between Normal and Abnormal samples across different classification thresholds.

Unless otherwise stated, the F1-score reported in the comparative results corresponds to the positive (Abnormal) class and not to a macro- or weighted-averaged F1-score.

3.5. Training Strategy

The revised comparison contains nine unique configurations. CNN and ResNet18 denote the same baseline and are reported in one row. All models use a single output logit and the same binary loss. The patch-based controls are new, explicit implementations; they do not reuse the invalid historical component switches. Table 4 summarizes the representation and processing flow of each revised configuration.

Table 4. Definitions of the revised configurations. Transformer denotes the specified two-layer encoder, not a pretrained ViT-B/16.

Configuration	Representation and Processing
CNN (ResNet18)	ImageNet-pretrained ResNet18, global average pooling, one-logit head.
Patch-GCN	49 RGB patches of 32×32 pixels; linear projection of 3072 values to 512; cosine graph; three GCN blocks; pooling and classification.
Patch Transformer	The same patch projection; 512-to-768 projection, learned positions, two Transformer layers, and mean pooling.
Patch-GCN-Transformer	Projected RGB patches, GCN, learned positions, and Transformer, in that order.
CNN + GCN	49 ResNet18 spatial nodes, cosine graph, three GCN blocks, and pooling.
CNN + FFT + GCN	Spatial nodes define the graph; fused CNN/FFT nodes enter the GCN.
CNN + GCN + Transformer	Spatial nodes, GCN and position-aware Transformer; no FFT/fusion branch.
ResNet18 + FFT	Global CNN descriptor fused with the mean of 49 patch FFT descriptors; no graph or Transformer.
FLIGENT-Net	Fused CNN/FFT nodes, graph from CNN, GCN, and position-aware Transformer.

The CNN branches use ImageNet weights: torchvision ResNet18 weights for the two conventional ResNet baselines and timm ResNet18 weights for the graph-based CNN configurations. The patch projections and Transformer components are trained from random initialization. This difference in prior representation learning must be considered when comparing patch controls with CNN-based models. These controls are complementary architectural baselines, not parameter-matched interventions isolating only one factor.

3.5.1. Training Configuration

All revised runs use Adam with learning rate 10^{-4} , batch size 16, zero weight decay, and a maximum of 100 epochs. The loss is BCEWithLogitsLoss with positive-class weight 2.0. The graph threshold is $\tau = 0.70$ where applicable and the classification threshold is fixed at 0.55. ReduceLROnPlateau monitors validation loss, with factor 0.5 and patience five epochs. Early stopping uses patience 12; the checkpoint with the lowest validation loss is evaluated on test only after checkpoint selection finishes.

All revised experiments were conducted using the same fixed operating configuration: a positive-class weight of 2.0, a cosine-similarity threshold of $\tau = 0.70$, and a binary classification threshold of 0.55. These quantities serve different roles in the pipeline. The positive-class weight increases the contribution of errors associated with the Abnormal class during optimization; the cosine threshold converts pairwise CNN-feature similarities into off-diagonal graph edges; and the classification threshold converts the predicted Abnormal probability into a binary decision. With Abnormal treated as the positive class, the training partition contains 12,096 Normal and 7874 Abnormal images, giving $N_{\text{Normal}}/N_{\text{Abnormal}} = 1.536$. Therefore, $\text{pos_weight} = 2.0$ is not presented as the exact inverse-frequency class ratio, but as a fixed loss-weighting factor used consistently in the revised protocol.

These values were held constant across the corresponding revised configurations so that the architectural comparisons were performed under a common training and decision setting. In particular, no classification-threshold search is performed on the revised test predictions: the 0.55 decision threshold is applied unchanged when computing the reported test metrics. The three values are therefore treated as fixed operating parameters, not as globally optimal hyperparameters, and no sensitivity claim is made from the present experiments. The graph-topology analysis in Section 4.5 characterizes the practical consequence of the fixed $\tau = 0.70$ setting: because cosine-similarity distributions depend on the underlying representation space, the same numerical threshold does not impose comparable graph densities across architectures. A dedicated sensitivity study would require selecting candidate values using training/validation data and locking the operating point before independent test evaluation.

3.5.2. Reproducibility

A single stratified partition generated with seed 42 was persisted and reused without modification across the revised configurations. The Python, NumPy, and PyTorch random seeds were also set to 42, including CUDA generators when applicable. In the present study, the numerical value 42 is treated solely as a reproducibility convention for fixing the reported experimental setting; no statistical, model-selection, or representativeness significance is assigned to that particular seed. This protocol fixes the data partition and specified training/evaluation conditions; it does not by itself establish identical stochastic realizations across different architectures or bitwise determinism on every platform.

One training run was performed per completed configuration. Consequently, variability arising from initialization and other stochastic training processes was not estimated. Differences are interpreted descriptively within the evaluated protocol, not as evidence of statistically significant superiority, stability, or a general ranking across random initializations. The absence of evaluation across multiple seeds is therefore retained as an explicit limitation of the present study, and no mean or standard deviation across training seeds is reported.

3.6. FLIGENT-Net Architecture

FLIGENT-Net is a deep learning architecture designed for binary gastric histopathological image classification. The architecture integrates four complementary representation mechanisms: convolutional spatial feature extraction, frequency-domain analysis using the fast Fourier transform (FFT), graph-based relational modeling using a graph convolutional network (GCN), and global contextual modeling using a Transformer encoder. The proposed framework is inspired by the graph-Transformer interaction paradigm introduced in SCUBa-Net [13]; however, FLIGENT-Net was redesigned for binary classification on the GasHisSDB dataset and incorporates frequency-domain descriptors at the graph-node level.

The model name FLIGENT-Net uses FLIGENT as the acronym for Frequency-Localized Image Graph Encoding Network with Transformer; the suffix “-Net” is retained as a conventional architecture label. The term Localized refers specifically to the patch-wise FFT analysis performed over the regular 7×7 image grid. No mathematical linearization is performed in this stage: flattening each pooled 8×8 spectral matrix into a 64-dimensional descriptor is an ordered vectorization, followed by learnable projection and nonlinear feature fusion.

As illustrated in Figure 1, the input histopathological image is processed through two parallel representation branches. The CNN branch extracts spatial feature maps that are organized into a 7×7 grid, resulting in $N = 49$ spatial nodes with 512-dimensional descriptors. In parallel, the input image is partitioned into the same 7×7 spatial grid, and a two-dimensional FFT is independently computed for each 32×32 image patch. Each local spectrum is transformed into a 64-dimensional descriptor. The two branches are paired by their ordered 7×7 grid indices. CNN receptive fields can overlap and extend beyond a 32×32 FFT patch; grid alignment does not imply identical receptive-field support.

The FFT descriptors are subsequently projected from 64 to 512 dimensions and concatenated with the corresponding 512-dimensional CNN descriptors. A learnable fusion layer transforms the resulting 1024-dimensional representation into a 512-dimensional node embedding. The graph connectivity is independently determined from the CNN-derived spatial descriptors using pairwise cosine similarity and a predefined threshold τ . The fused node representations are then propagated through three GCN layers. Finally, the refined node representations are projected to a 768-dimensional space and processed as a sequence by a Transformer encoder to model global dependencies among image regions. Mean pooling of the resulting token representations produces a global representation that is forwarded to a binary classification layer.

The overall architecture is organized into three sequential modules:

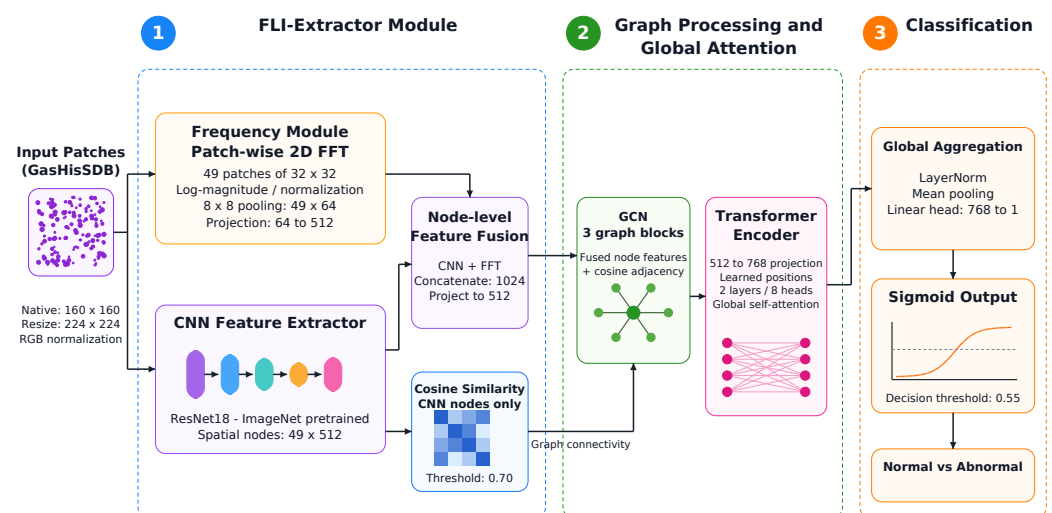


Figure 1. FLIGENT-Net architecture, organized into the three stages of the original design. Spatial and patch-wise spectral descriptors are fused into node attributes, while CNN descriptors determine graph connectivity. Three GCN blocks and a Transformer encoder precede global aggregation and binary classification. Colors are used only to distinguish processing modules and do not encode quantitative values.

- 1. Frequency-Localized Image (FLI)-Extractor Module:** Extracts complementary spatial and frequency-domain representations from the input tissue patch. The CNN produces a 7×7 spatial feature grid with 512-dimensional descriptors, while the

FFT branch produces one 64-dimensional spectral descriptor for each corresponding image region.

2. **Graph Processing and Global Attention:** The CNN-derived descriptors determine graph connectivity through cosine similarity. The spatial and spectral representations are fused to form the node attributes, which are subsequently refined through three GCN layers and processed by a Transformer encoder to model long-range dependencies among image regions.
3. **Classification Module:** Aggregates the Transformer representations using mean pooling and projects the resulting global representation through a fully connected layer to obtain the Abnormal-class probability.

3.6.1. FLI-Extractor Module

The Frequency-localized image (FLI)-extractor module obtains complementary spatial and frequency-domain representations from the input histopathological image. The spatial branch captures morphological information through a convolutional neural network, whereas the frequency-domain branch characterizes local spectral patterns using the fast Fourier transform (FFT). The two representations are subsequently aligned according to their spatial location and fused at the graph-node level.

Let the input histopathological image be denoted by

$$X \in \mathbb{R}^{H \times W \times C}, \quad (7)$$

where H and W denote the image height and width, respectively, and $C = 3$ represents the RGB channels. In the implemented configuration, the input image is resized to 224×224 pixels.

The FLI-Extractor can therefore be expressed as

$$F_{\text{FLI}}(X) = (F_{\text{CNN}}(X), F_{\text{FFT}}(X)), \quad (8)$$

where F_{CNN} denotes the spatial feature extraction branch and F_{FFT} denotes the frequency-domain feature extraction branch.

CNN Spatial Feature Extraction

The spatial branch uses an ImageNet-pretrained ResNet18 backbone as a feature extractor. The classification head is removed, and the resulting convolutional feature map is used to define the graph nodes. For an input image of size 224×224 , the final convolutional feature map has dimensions

$$F_{\text{CNN}}(X) \in \mathbb{R}^{512 \times 7 \times 7}. \quad (9)$$

The spatial dimensions are flattened while preserving their spatial ordering, producing

$$H = \{h_1, h_2, \dots, h_N\}, \quad H \in \mathbb{R}^{N \times d_s}, \quad (10)$$

where

$$N = 7 \times 7 = 49, \quad d_s = 512. \quad (11)$$

Thus, each image is represented by 49 spatial descriptors,

$$h_i \in \mathbb{R}^{512}, \quad i = 1, \dots, 49. \quad (12)$$

Each descriptor corresponds to one spatial position in the 7×7 CNN feature map and subsequently defines one node in the graph.

FFT Spectral Feature Extraction

To complement the spatial representation, the frequency-domain branch extracts local spectral descriptors using the two-dimensional fast Fourier transform (FFT). Unlike applying a single FFT to the entire image and obtaining one global descriptor, the proposed implementation preserves the same spatial organization used by the CNN branch.

The 224×224 input image is divided into a regular 7×7 grid, producing

$$N = 7 \times 7 = 49 \quad (13)$$

non-overlapping image patches. Since the input resolution is 224×224 , each patch has spatial dimensions

$$P_i \in \mathbb{R}^{32 \times 32}, \quad i = 1, \dots, 49. \quad (14)$$

Each RGB patch is converted to grayscale before the frequency-domain transformation. For a grayscale patch P_i , the two-dimensional discrete Fourier transform is defined as

$$F_i(u, v) = \sum_{x=0}^{31} \sum_{y=0}^{31} P_i(x, y) \exp \left[-i2\pi \left(\frac{ux}{32} + \frac{vy}{32} \right) \right]. \quad (15)$$

The magnitude spectrum is obtained as

$$M_i(u, v) = |\text{FFT2}(P_i)|. \quad (16)$$

The spectrum is shifted so that the low-frequency components are centered, followed by logarithmic compression:

$$\tilde{M}_i = \log(1 + \text{FFTShift}(M_i)). \quad (17)$$

The implemented normalization is

$$\overline{M}_i = \frac{\tilde{M}_i}{\max_{u,v} \tilde{M}_i(u, v) + 10^{-8}}. \quad (18)$$

Each normalized local spectrum is subsequently adaptively pooled to an 8×8 representation. The resulting matrix is flattened to obtain a 64-dimensional spectral descriptor:

$$s_i = \text{Flatten}(\text{AdaptivePool}_{8 \times 8}(\overline{M}_i)), \quad (19)$$

where

$$s_i \in \mathbb{R}^{64}. \quad (20)$$

Consequently, the complete FFT representation is

$$S = \{s_1, s_2, \dots, s_N\}, \quad S \in \mathbb{R}^{49 \times 64}. \quad (21)$$

For row r and column c , with $r, c \in \{0, \dots, 6\}$, define $i = 7r + c + 1$. The FFT patch is the preprocessed image region $X[32r : 32(r + 1), 32c : 32(c + 1), :]$, and the CNN descriptor is the vector at feature-grid position (r, c) . For pooled spectral matrix $B_i \in \mathbb{R}^{8 \times 8}$, the vector is defined by $(s_i)_{8a+b+1} = (B_i)_{a,b}$ for $a, b \in \{0, \dots, 7\}$. This operation is an ordered vectorization and grid-index association between the spectral and convolutional repre-

sentations; it should not be interpreted as a mathematical linearization of the underlying tissue patterns. Moreover, the receptive field of the corresponding CNN descriptor is not restricted to the spatial support of the associated FFT patch.

Node-Level Feature Fusion

Before graph processing, the CNN and FFT descriptors are transformed into a common feature space. The 64-dimensional FFT descriptors are projected to 512 dimensions using a learnable projection:

$$\hat{s}_i = \phi_{\text{FFT}}(s_i), \quad \hat{s}_i \in \mathbb{R}^{512}, \quad (22)$$

where ϕ_{FFT} denotes the FFT projection consisting of a fully connected transformation followed by normalization and a nonlinear activation.

The CNN descriptor and the projected FFT descriptor corresponding to the same spatial region are then concatenated:

$$q_i = [h_i; \hat{s}_i], \quad q_i \in \mathbb{R}^{1024}. \quad (23)$$

A learnable fusion transformation maps this concatenated representation back to a 512-dimensional node embedding:

$$x_i = \phi_{\text{fuse}}(q_i), \quad x_i \in \mathbb{R}^{512}. \quad (24)$$

The resulting node-feature matrix is therefore

$$X_G = \{x_1, x_2, \dots, x_{49}\} \in \mathbb{R}^{49 \times 512}. \quad (25)$$

The fused representation X_G , rather than the FFT descriptors alone, is used as the input node representation for graph processing.

The FFT itself has no learned parameters. The spectral projection contributes $64 \cdot 512 + 512 + 2 \cdot 512 = 34,304$ parameters, including affine LayerNorm. The fusion block contributes $1024 \cdot 512 + 512 + 2 \cdot 512 = 525,824$. Together they add 560,128 parameters (approximately 0.560 M); GELU and dropout add none. This accounts for the extra trainable capacity of the frequency branch and its fusion, rather than attributing parameters to the Fourier transform.

Graph Construction

The fused node representations are organized into a graph

$$G = (V, E), \quad (26)$$

where V denotes the set of 49 image-region nodes and E denotes the set of graph edges. Each node contains the fused spatial and frequency-domain representation corresponding to one spatial region of the input image.

The graph connectivity is determined using the CNN-derived spatial descriptors h_i . For two spatial descriptors h_i and h_j , their cosine similarity is defined as

$$\text{Sim}_{\cos}(h_i, h_j) = \frac{h_i^\top h_j}{\|h_i\|_2 \|h_j\|_2}. \quad (27)$$

An adaptive adjacency matrix is constructed by thresholding the pairwise cosine similarities:

$$A_{ij} = \begin{cases} 1, & \text{Sim}_{\cos}(h_i, h_j) > \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (28)$$

where τ denotes the cosine-similarity threshold. In the final experimental configuration,

$$\tau = 0.70. \quad (29)$$

Thus, the graph topology is dynamically determined from the learned CNN representations for each input image. The adjacency matrix is not an independently parameterized learnable matrix; instead, it is an adaptive graph obtained by thresholding similarities between learned spatial embeddings.

Self-loops are subsequently added to the graph:

$$\tilde{A}_{ij} = \max(A_{ij}, \delta_{ij}), \quad (30)$$

where δ_{ij} denotes the identity-matrix entry. This sets each diagonal entry to one without doubling an existing self-loop.

Graph Processing and Global Attention

The three graph blocks use the provided implementation rather than a generic GCN activation formula. Let $D_{ii} = \sum_j \tilde{A}_{ij}$ and $\hat{A} = (D + \epsilon I)^{-1/2} \tilde{A} (D + \epsilon I)^{-1/2}$ with $\epsilon = 10^{-10}$. The update is

$$U^{(l)} = (\hat{A}H^{(l)} + H^{(l)})W^{(l)} + \mathbf{1}(b^{(l)})^\top, \quad (31)$$

$$H^{(l+1)} = \text{BN}_l(\text{L2Norm}_{\text{row}}(U^{(l)})), \quad H^{(0)} = X_G. \quad (32)$$

Row normalization divides by the row norm with the numerical floor used by PyTorch. BatchNorm acts on feature channels across the valid nodes of the mini-batch during training and uses stored statistics during evaluation. The configured graph blocks have no extra ReLU or dropout. The $+H^{(l)}$ term supplies an additional self contribution beyond the diagonal already present in $\tilde{A}$; it is preserved here to match the implementation. The normalized adjacency is not the graph Laplacian. Hard edge thresholding is not differentiated, although the CNN descriptors are trained through the feature pathway.

After three GCN layers, the resulting representation is denoted by

$$H^{(3)} \in \mathbb{R}^{49 \times 512}. \quad (33)$$

The node representations are then projected into the Transformer embedding space:

$$T_0 = H^{(3)}W_T + \mathbf{1}b_T^\top + E_{\text{pos}}, \quad T_0 \in \mathbb{R}^{49 \times 768}, \quad (34)$$

where W_T and b_T are the projection parameters and $E_{\text{pos}} \in \mathbb{R}^{49 \times 768}$ is a learned positional embedding. The encoder uses two pre-norm layers, eight heads, a 3072-dimensional feed-forward block, GELU and dropout 0.1. A final LayerNorm is applied before mean pooling.

The resulting 49 vectors are interpreted as a sequence of region-level tokens and processed by a Transformer Encoder. The multi-head self-attention operation is defined as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (35)$$

where Q , K , and V denote the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the key vectors.

The Transformer encoder contains two encoder layers with eight attention heads. Its purpose is to model long-range dependencies among the 49 graph-refined image-region representations. Unlike the graph convolution operation, which propagates information according to the explicitly constructed graph topology, self-attention allows every token to interact with the other tokens in the sequence.

Global Representation and Classification

After Transformer processing, the resulting token representations are aggregated using mean pooling:

$$Z_{\text{global}} = \frac{1}{N} \sum_{i=1}^N T_i, \quad (36)$$

where $N = 49$ and $T_i \in \mathbb{R}^{768}$ denotes the Transformer output associated with the i -th image region. Therefore,

$$Z_{\text{global}} \in \mathbb{R}^{768}. \quad (37)$$

The global representation is subsequently passed to a fully connected classification layer:

$$z = W_c Z_{\text{global}} + b_c, \quad (38)$$

where W_c and b_c denote the learnable classification weights and bias, respectively, and z represents the classification logit.

The Abnormal-class probability is obtained using a sigmoid activation:

$$\hat{y} = \sigma(z) = \frac{1}{1 + \exp(-z)}, \quad (39)$$

where $\hat{y} \in [0, 1]$ represents the predicted probability that the input tissue belongs to the Abnormal class. During inference, the predicted class is determined by comparing $\hat{y}$ with the selected binary decision threshold.

Table 5 summarizes the forward-pass sequence of FLIGENT-Net.

Table 5. Forward-pass pseudocode for FLIGENT-Net; batch dimension omitted.

Step	Operation
1	Resize RGB input to 224×224 and apply the specified channel normalization.
2	Compute ResNet18 feature map; flatten its grid in row-major order to $H \in \mathbb{R}^{49 \times 512}$.
3	For $r, c = 0, \dots, 6$, extract the 32×32 patch, form the weighted grayscale tensor, FFT, shift, log-magnitude, and maximum normalization.
4	Average-pool each spectrum to 8×8 and vectorize to $s_{7r+c+1} \in \mathbb{R}^{64}$.
5	Project each spectral vector; concatenate with its CNN grid counterpart; fuse to $X_G \in \mathbb{R}^{49 \times 512}$.
6	Threshold pairwise CNN cosine similarities at 0.70 and set the diagonal to one.
7	Apply the three graph updates in Equation (31) with their normalizations.
8	Project to 768 dimensions, add learned positions, and apply the two-layer Transformer.
9	Apply LayerNorm, mean-pool tokens, obtain one logit, and use sigmoid; classify Abnormal at probability ≥ 0.55 .

3.7. Generative AI Assistance

ChatGPT (OpenAI, GPT-5.6) was used to assist with manuscript drafting and revision. ChatGPT was also used to prepare schematic illustrations of the architecture. The histopathological inputs were obtained from GasHisSDB; the reported numerical results and attribution and attention maps were produced by executing the computational workflows described in this study, rather than by generative image synthesis. The authors reviewed and edited the AI-assisted material and take responsibility for its accuracy and consistency with the reported methods and results. Product-level information, including the model version used, is provided in the Acknowledgments.

4. Experiments

The revised classification evaluation comprises nine completed configurations assessed on the common 6657-image test set using seed 42. The saved reports contain identical ordered test paths and labels: 4032 Normal and 2625 Abnormal images. Recalculation from the saved predictions and probabilities reproduces all five reported metrics, and the configuration records share the same split SHA-256 identifier and source-code hashes.

4.1. Ablation and Comparative Architectural Analysis

Table 6 summarizes the test performance of the nine revised configurations under the common split and seed-42 protocol.

Table 6. Revised test results with seed 42 and the common patch-level split. Precision, recall and F1 refer to Abnormal.

Architecture	Accuracy	Precision	Recall	F1	AUC
CNN (ResNet18)	98.51%	97.55%	98.70%	98.13%	0.9986
Patch-GCN	83.58%	74.68%	88.30%	80.92%	0.9177
Patch Transformer	87.44%	79.38%	92.08%	85.26%	0.9541
Patch-GCN-Transformer	84.27%	75.48%	89.03%	81.70%	0.9253
CNN + GCN	97.40%	96.09%	97.37%	96.73%	0.9964
CNN + FFT + GCN	97.13%	96.24%	96.50%	96.37%	0.9952
CNN + GCN + Transformer	97.54%	97.27%	96.46%	96.86%	0.9966
ResNet18 + FFT	98.30%	96.73%	99.05%	97.87%	0.9985
FLIGENT-Net	97.31%	96.57%	96.61%	96.59%	0.9964

Among the nine completed configurations, CNN (ResNet18) has the highest accuracy and Abnormal-class F1-score. FLIGENT-Net achieves 97.31% accuracy and 96.59% F1, compared with 98.51% and 98.13% for the CNN baseline. Its F1 is 1.53 percentage points lower using the unrounded values. These runs do not support a predictive advantage for the full architecture.

The Patch-GCN control obtained 83.58% accuracy, 74.68% precision, 88.30% recall, 80.92% F1, and 0.9177 AUC-ROC. This is the revised raw-patch projection plus GCN architecture, not the historical model with inactive component switches. Its lower performance relative to CNN-based configurations must be interpreted together with the different input representations and absence of ImageNet pretraining; it does not isolate the effect of graph convolution alone.

The Patch Transformer control obtained 87.44% accuracy, 79.38% precision, 92.08% recall, 85.26% F1, and 0.9541 AUC-ROC. Its F1 exceeded Patch-GCN by 4.33 percentage points in the observed runs. Both controls use randomly initialized patch projections; the Transformer here is the specified two-layer encoder, not an ImageNet-pretrained ViT-B/16. This comparison does not establish statistical significance or a general advantage across training seeds.

The Patch-GCN-Transformer control obtained 84.27% accuracy, 75.48% precision, 89.03% recall, 81.70% F1, and 0.9253 AUC-ROC. Its observed F1 was 0.78 percentage points higher than Patch-GCN and 3.56 points lower than the Patch Transformer. Thus, these single-run results do not show an advantage for combining graph processing and global attention over the standalone Patch Transformer.

ResNet18 + FFT achieved 98.30% accuracy, 96.73% precision, 99.05% recall, 97.87% F1, and 0.9985 AUC-ROC. Relative to ResNet18, recall increased by 0.34 percentage points, while precision decreased by 0.83 points, accuracy by 0.21 points, and F1 by 0.25 points (computed before rounding). In this run, spectral fusion shifted the precision–recall balance without improving overall accuracy or F1. ResNet18 + FFT had the highest Abnormal-class recall among the nine configurations, whereas ResNet18 retained the highest accuracy, precision, F1, and AUC.

Within the graph-based CNN configurations, adding the FFT branch and fusion changes F1 from 96.73% to 96.37%; adding the Transformer pathway to CNN + GCN changes it to 96.86%. FLIGENT-Net reaches 96.59%. These descriptive differences must not be interpreted as statistically significant effects, nor as proof of a general ranking across seeds. Head dimensions, pretraining sources, and the capacity added by fusion also limit a strict one-factor causal interpretation.

The spectral descriptor retains magnitude but discards phase, compresses the spectrum through logarithms and pooling, and normalizes by its maximum. These operations may discard useful distinctions or introduce redundancy with CNN features. Because resizing before FFT can also modify spectral content, the additional native-grid control described above was used to isolate this preprocessing choice under the same split and training settings. Its results are reported separately below. Learnable spectral weighting and wavelet-based representations remain possible future alternatives; neither is evaluated here.

4.2. Native-Grid FFT Control

The additional control evaluated FLIGENT-Net with spectral descriptors extracted from the native 160×160 input grid instead of from the resized 224×224 grid used in the primary protocol. All listed training hyperparameters and the persisted train/validation/test split were matched to the reference FLIGENT-Net run. The check-point was selected by minimum validation loss; this occurred at epoch 27, where the validation loss was 0.1022, the Abnormal-class F1-score was 0.9741, and AUC-ROC was 0.9974. Training stopped after epoch 39 under the configured patience of 12 epochs. Table 7 reports the controlled comparison with the primary resized-grid configuration.

Table 7. Controlled comparison of the spectral input grid for FLIGENT-Net. The primary model computes patch-wise FFT descriptors after resizing to 224×224 ; the additional control uses the native 160×160 grid with integer 7×7 boundaries. Precision, recall, and F1 refer to the Abnormal class.

Spectral Input	Accuracy	Precision	Recall	F1	AUC
Resized 224×224	97.31%	96.57%	96.61%	96.59%	0.9964
Native 160×160 grid	97.43%	96.73%	96.76%	96.74%	0.9966

The native-grid run yielded 3946 true negatives, 86 false positives, 85 false negatives, and 2540 true positives. Relative to the primary resized-FFT run, the observed changes were approximately +0.12 percentage points in accuracy, +0.16 in precision, +0.15 in recall, +0.15 in F1, and +0.0002 in AUC-ROC. The direction of change is therefore consistently positive across the reported metrics, but its magnitude is small. With one seed per condition,

these differences should be interpreted descriptively rather than as evidence of statistical significance.

This control indicates that the alternative native-grid spectral protocol can modestly change the measured performance. However, it is not a pure interpolation ablation: moving from the resized 224×224 grid to the native 160×160 grid also changes each local FFT support from 32×32 pixels to regions of 22 or 23 pixels per side and therefore changes the discrete frequency sampling and bin discretization of the local spectra. The observed increase from 97.31% to 97.43% accuracy should therefore be attributed to the spectral-protocol change as a whole, not specifically to the removal of interpolation. Other properties of the descriptor—including magnitude-only representation, logarithmic compression, maximum normalization, adaptive pooling, and the subsequent fusion mechanism—may also affect how much complementary information is provided to the CNN representation.

4.3. Training Dynamics and Extended Training

The saved validation-loss histories determine checkpoint selection. Training and validation curves are drawn from these same completed runs. A larger maximum epoch count alone does not establish a longer effective training duration when early stopping terminates both runs before the cap. The saved histories and test metadata agree on checkpoint selection by minimum validation loss. Training stopped 12 epochs after the selected epoch in all nine runs. Table 8 reports the observed training lengths; the maximum of 100 epochs was not reached.

Table 8. Verified training lengths and epochs selected by minimum validation loss.

Architecture	Selected Epoch	Total Epochs
CNN (ResNet18)	24	36
Patch-GCN	42	54
Patch Transformer	37	49
Patch-GCN-Transformer	58	70
CNN + GCN	20	32
CNN + FFT + GCN	8	20
CNN + GCN + Transformer	31	43
ResNet18 + FFT	19	31
FLIGENT-Net	21	33

Figure 2 shows the recorded training and validation losses without smoothing. In several CNN-based configurations, training loss continues to decrease while validation loss fluctuates, supporting checkpoint selection from validation rather than the final epoch. The patch-based graph controls retain higher training and validation losses. These observations describe the recorded runs; they do not establish stability across seeds or isolate the cause of the performance differences.

4.4. Computational Cost

Table 9 reports inference-time computational measurements for the nine revised configurations. These measurements correspond to the primary 224×224 protocol and do not include the additional native-grid FFT preprocessing control.

The measurements show that parameter count alone does not determine observed GPU latency. For example, the Patch Transformer contains 16.183 million parameters yet achieved the lowest mean batch latency in this benchmark, whereas the smaller Patch-GCN exhibited a higher mean latency and a substantially wider P95. This reflects implementation and hardware effects and is one reason to report measured timing rather than infer computational efficiency solely from model size.

The isolated FFT comparisons quantify the additional cost of the frequency pathway. Adding FFT and node-level fusion to CNN + GCN increases the parameter count from 11.969 to 12.529 million, exactly 0.560 million additional parameters, consistent with the projection-and-fusion accounting in Section 3.6. Mean batch latency increases from 5.887 to 6.131 ms (+4.14%), throughput decreases from 2717.7 to 2609.7 images/s (−3.97%), and peak allocated memory increases by 2.64 MiB. Similarly, adding the global FFT branch to ResNet18 increases parameters by 0.338 million, mean batch latency by 7.87%, and peak memory by 1.29 MiB while reducing throughput by 7.30%.

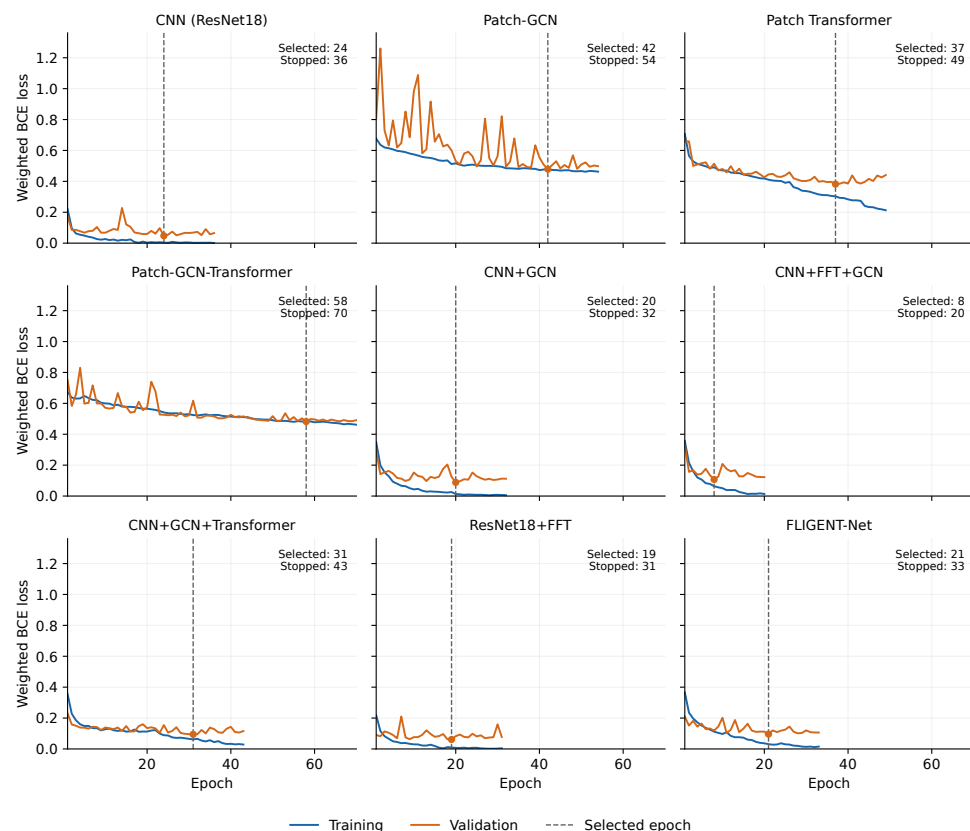


Figure 2. Training and validation weighted binary cross-entropy losses for the nine revised configurations. All panels share axis scales. Dashed lines mark the minimum-validation-loss epoch used for checkpoint selection, and orange dots mark the validation-loss value at that selected epoch; each curve ends at its observed stopping epoch. No test measurements are used to select these epochs.

Table 9. Computational cost of the nine revised configurations on an NVIDIA GeForce RTX 4070 SUPER using batch size 16. Mean and P95 refer to synchronized batch latency. The per-image value is the mean batch latency divided by 16. Peak memory is PyTorch allocated GPU memory during inference.

Architecture	Params. (M)	Mean Batch (ms)	P95 Batch (ms)	ms/Image	Images/s	Peak MiB
CNN (ResNet18)	11.177	3.927	5.026	0.245	4074.4	178.4
Patch-GCN	2.366	4.009	8.632	0.251	3990.5	64.7
Patch Transformer	16.183	3.472	4.400	0.217	4607.8	130.4
Patch-GCN-Transformer	16.974	5.529	6.505	0.346	2893.6	135.9
CNN + GCN	11.969	5.887	8.531	0.368	2717.7	181.4
CNN + FFT + GCN	12.529	6.131	7.452	0.383	2609.7	184.0
CNN + GCN + Transformer	26.577	7.920	8.997	0.495	2020.2	238.1
ResNet18 + FFT	11.515	4.236	5.126	0.265	3777.1	179.7
FLIGENT-Net	27.137	8.082	8.954	0.505	1979.7	240.3

For the closest full-architecture comparison, CNN + GCN + Transformer and FLIGENT-Net differ by the FFT projection/fusion pathway. FLIGENT-Net adds 0.560 million parameters (+2.11%), increases mean batch latency from 7.920 to 8.082 ms (+2.05%), reduces throughput from 2020.2 to 1979.7 images/s (−2.00%), and increases peak allocated memory by 2.14 MiB. Relative to ResNet18, the complete FLIGENT-Net has approximately 2.43 times as many parameters and 2.06 times the mean batch latency under this protocol, while the classification results in Table 6 do not show a corresponding predictive advantage. The computational results therefore reinforce the characterization finding that adding representation domains entails measurable resource cost and that additional architectural complexity should be evaluated jointly with predictive benefit.

4.5. Graph-Topology Results

Table 10 summarizes the graph statistics defined in Section 3.3.2. All graphs contain 49 nodes and can contain at most $\binom{49}{2} = 1176$ undirected off-diagonal edges.

Table 10. Graph topology on the full 6657-image test split at $\tau = 0.70$. Self-loops are excluded. Isolated nodes are nodes with no off-diagonal neighbors; edgeless graphs contain no off-diagonal edges.

Architecture	Mean Edges	Mean Degree	Mean Density	Mean Isolated Nodes	Edgeless Graphs (%)
Patch-GCN	1147.32	46.829	0.97561	0.025	0.00
Patch-GCN-Transformer	981.47	40.060	0.83458	0.096	0.00
CNN + GCN	5.98	0.244	0.00508	43.420	21.18
CNN + FFT + GCN	4.70	0.192	0.00399	43.215	18.22
CNN + GCN + Transformer	5.21	0.213	0.00443	42.896	16.75
FLIGENT-Net	6.49	0.265	0.00552	42.067	13.79

The fixed cosine threshold produces two markedly different topology regimes. The CNN-derived graph configurations are very sparse: their mean degrees are below 0.27 and, on average, approximately 42–43 of the 49 nodes have no external neighbor. FLIGENT-Net is the densest of these CNN-derived graphs, but its mean density remains only 0.00552; 13.79% of its test images contain no off-diagonal edge before self-loops are added. In such cases, the subsequent GCN still applies its learned transformations, normalization, self contribution, and self-loop pathway, but there is no message exchange between distinct regions through the adjacency matrix. More generally, the large fraction of isolated nodes in the CNN-derived configurations means that relational message passing is available only to a small subset of regions at this operating point. This provides a plausible architectural explanation for the limited incremental benefit observed from graph processing in the present experiments, although sparsity was not isolated as a causal factor and should not be treated as the sole explanation.

In contrast, the patch-projection controls are highly connected. Patch-GCN has mean density 0.97561 and mean degree 46.829, while Patch-GCN-Transformer has mean density 0.83458 and mean degree 40.060. Therefore, applying the same numerical threshold $\tau = 0.70$ does not enforce comparable graph sparsity across representation spaces. Cosine-similarity distributions depend on the learned or projected feature representation and on the trained checkpoint. These measurements should consequently be interpreted as a characterization of the induced topology rather than as evidence that dense or sparse graphs are intrinsically superior. They also show that the fixed global threshold is not representation-invariant and identify graph construction itself as a central design variable. Representation-aware alternatives such as k -nearest-neighbor connectivity, adaptive thresholds, minimum-degree constraints, or learned graph structure are therefore motivated by the observed topology rather than introduced only as generic future-work possibilities.

Importantly, these topology statistics characterize the relational stage of the GCN rather than the total cross-region interaction available in the full graph–Transformer models. After graph processing, the subsequent Transformer receives all 49 node representations and applies global self-attention, so every token can interact with every other token independently of whether an off-diagonal graph edge existed in the preceding GCN. Consequently, an edgeless or highly sparse graph restricts inter-region message passing in the GCN stage but does not eliminate global contextual interaction in FLIGENT-Net or in the other graph–Transformer configurations.

4.6. Interpretability Analysis

The revised comparison is restricted to ResNet18 and FLIGENT-Net, using the checkpoints evaluated in Table 6. A fixed sample of 100 Abnormal test images was randomly selected with seed 42 from the 2521 images correctly classified by both models at threshold 0.55. The recorded sample matches the selection fixed before this analysis. All reported interpretability results are conditional on this common-correct subset, not representative estimates for the entire test set or clinical population.

Both models use the same input-gradient definition,

$$a_{uv} = \sum_{c=1}^3 \left| \frac{\partial z}{\partial X_{uvc}} \right|, \quad (40)$$

where z is the Abnormal logit and X is the ImageNet-normalized 224×224 RGB input. Absolute gradients measure local sensitivity, not signed evidence supporting the class. The previous heterogeneous comparison of Grad-CAM with nodal activation magnitude is not retained as the primary comparison. This analysis does not evaluate the faithfulness of FLIGENT-Net’s separate nodal-magnitude output or its Transformer attention.

4.6.1. Quantitative Heatmap Evaluation

Pearson and Spearman correlations were computed between the flattened input-gradient maps of both models for each of the same 100 images. All 100 correlations were finite; no constant maps were reported. Mean Pearson correlation was 0.1092 (sample SD 0.1034, median 0.0989), and mean Spearman correlation was 0.1311 (sample SD 0.1262, median 0.1302). These values indicate limited agreement between the maps under the shared attribution definition. Low agreement alone establishes neither complementary diagnostic information nor superiority in interpretability. No pixel-level significance test treating spatially dependent pixels as independent observations is reported.

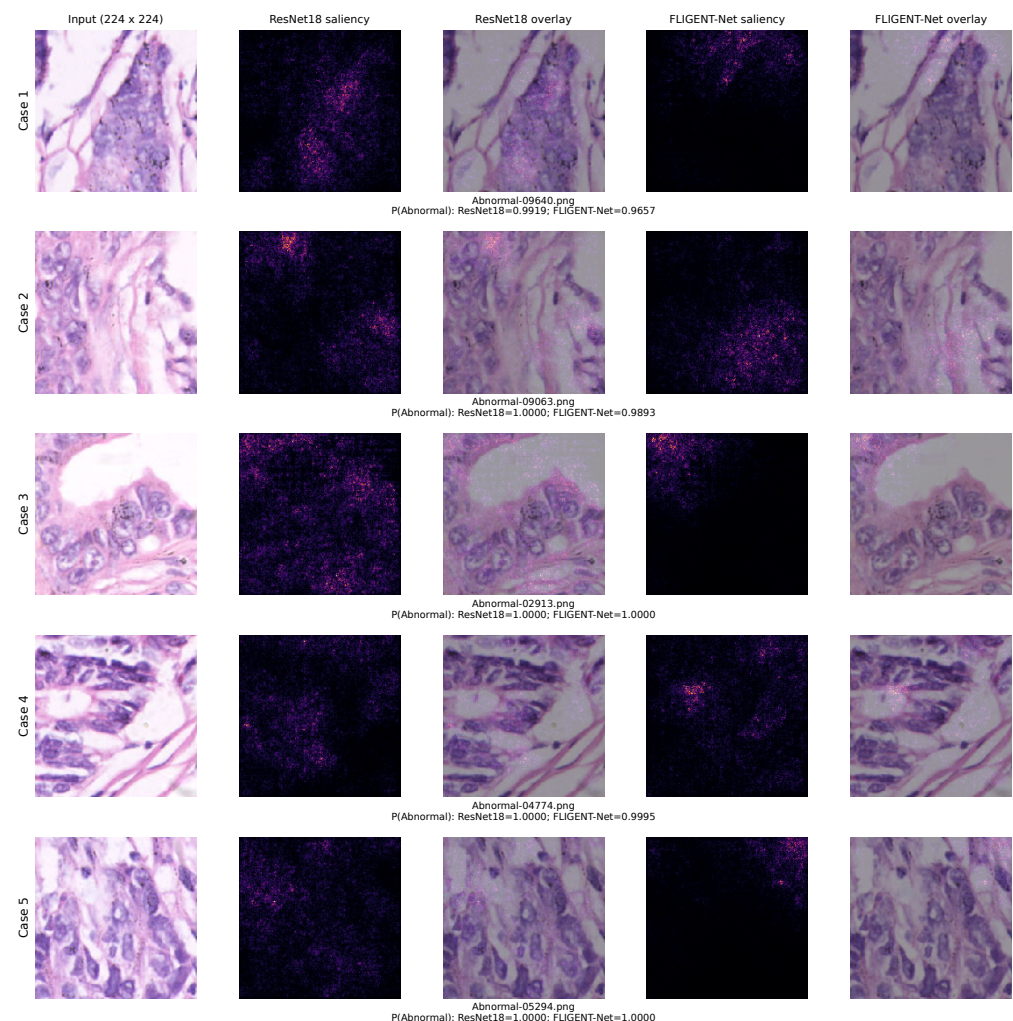
Figure 3 shows the first five entries of the fixed sample, without selection by visual appearance or deletion outcome. Each map is independently min–max scaled only for display and overlaid with opacity 0.4; equal colors across panels do not imply equal absolute gradient magnitudes. These are qualitative sensitivity maps, not tumor annotations or evidence of histological localization.

4.6.2. Transformer Attention Visualization

To characterize information aggregation in the global attention module, we extracted the self-attention weights of FLIGENT-Net on the same 100 selected images. Each of the two Transformer layers contains eight heads operating on 49 nodes. No classification token is used. For head h in the final layer, let $A_{ij}^{(h)}$ denote the attention weight from query node i to key node j . We summarize the mean attention received by each node as

$$r_j = \frac{1}{8 \cdot 49} \sum_{h=1}^8 \sum_{i=1}^{49} A_{ij}^{(h)}, \quad \sum_{j=1}^{49} r_j = 1. \quad (41)$$

The 49 values are arranged in the original 7×7 node grid. A uniform distribution corresponds to $r_j = 1/49$. This summary describes incoming attention in the final layer; it is not attention rollout or a class-specific attribution of the final prediction.



Same input-gradient attribution for both models. Independent min-max display scaling; colors do not compare absolute magnitudes. Five fixed Abnormal test examples; overlays are not tumor annotations or proof of localization.

Figure 3. Five preselected Abnormal test examples with the same input-gradient attribution for ResNet18 and FLIGENT-Net. Original inputs, independently scaled maps, and overlays are shown. Within each saliency map, brighter colors indicate larger normalized absolute-gradient magnitude; because each map is scaled independently, colors should not be compared across panels. These illustrations do not establish localization accuracy or explanation faithfulness. Histopathological inputs: MIA MIA Group, GasHisSDB [41], licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>); images were resized for analysis and presentation, and model-derived overlays were added.

Attention was extracted in evaluation mode from the original test batches with node positions preserved. The extraction procedure was verified by checking agreement with the scaled dot-product attention computation and concordance with the original model predictions. The visualizations use weights extracted from the original forward inputs; detailed numerical verification diagnostics are reported in Supplementary Note S1.

The Transformer attention maps are distinct from the model's auxiliary nodal-magnitude output. The latter is computed as $m_j = \|g_j\|_2$, where g_j is the embedding after the GCN and before the Transformer, and summarizes representation magnitude.

In contrast, r_j describes the attention received during Transformer processing. Neither quantity is the input-gradient map used in the Pixel Deletion evaluation below.

Figure 4 shows the first five cases of the fixed sample. The maps exhibit non-uniform attention distributions and different spatial patterns across the illustrated images, characterizing how the module weights contextualized node representations. These representations already incorporate graph processing and a preceding Transformer layer; highlighted grid positions therefore do not refer exclusively to the pixels beneath them. The maps complement the architectural description by revealing attention allocation, while classification performance is assessed separately through the controlled ablation. Tumor-localization accuracy and class-specific contributions cannot be inferred from these attention summaries alone.

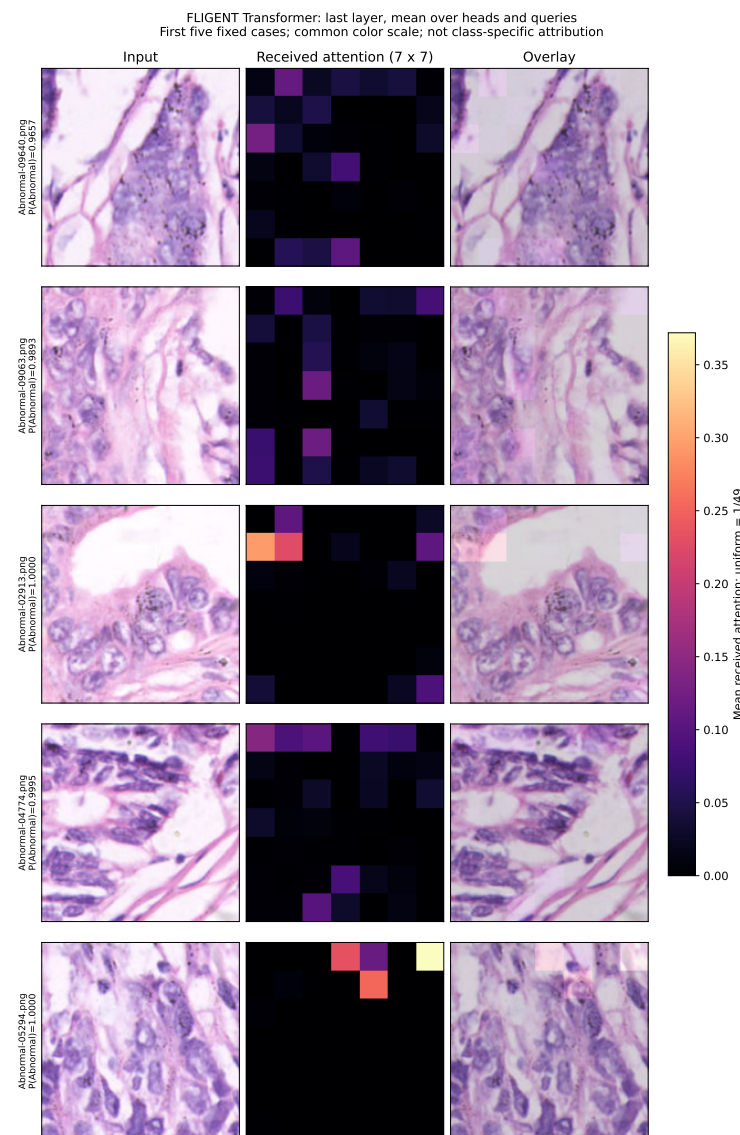


Figure 4. Mean received attention in the final Transformer layer of FLIGENT-Net for five preselected Abnormal images. Columns show the input, the 7×7 attention summary, and its overlay. A common color scale is used across cases without per-image min–max normalization. Higher values indicate greater received attention. Grid cells are enlarged by nearest-neighbor interpolation and overlaid with opacity 0.4. Histopathological inputs: MIA MIA Group, GasHisSDB [41], licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>); images were resized for analysis and presentation, and model-derived overlays were added.

4.7. Quantitative Evaluation of Interpretability Using Pixel Deletion

The initial gradient ranking was held fixed while deleting 0%, 5%, 10%, 20%, 30%, 40%, 50%, 75%, and 100% of the pixels. All three channels of a selected pixel were set to zero after normalization, equivalent to the ImageNet mean RGB color rather than black. Twenty random permutations per image provided a comparator, with the same permutations used for both models and nested masks across deletion fractions. These are perturbation repetitions, not independently trained model seeds. Models remained in evaluation mode. Each target retained its original test-batch position and unmodified companion images to control observed numerical differences caused by regrouping batches. The maximum discrepancy from saved original probabilities was zero for ResNet18 and 1.19×10^{-6} for FLIGENT-Net; gradient-enabled forward checks also passed the 10^{-4} tolerance.

For image i , let $P_i(0)$ be the original Abnormal probability. Curves were normalized per image as $q_i(r) = P_i(r)/P_i(0)$, integrated by the trapezoidal rule over $r \in [0, 1]$, and then averaged across images. Random AUCs were first averaged across the 20 permutations for each image. We define

$$\Delta\text{AUC}_i = \frac{1}{20} \sum_{k=1}^{20} \text{AUC}_{i,k}^{\text{random}} - \text{AUC}_i^{\text{gradient}}. \quad (42)$$

This quantity is based on deletion curves, not classification AUC-ROC. Positive values favor gradient-guided deletion for reducing Abnormal probability. Negative values and normalized curves exceeding one were retained. The common-correct selection ensures $P_i(0) \geq 0.55$, avoiding division by near-zero initial probabilities. Recalculation from the saved curves reproduces the individual and aggregate AUCs; the 0% and 100% deletion endpoints are consistent. Table 11 summarizes the mean normalized deletion AUCs and negative-gap counts.

Table 11. Pixel Deletion on the same 100 common-correct Abnormal test images. AUCs are means of per-image normalized deletion AUCs.

Model	Gradient AUC	Random AUC	ΔAUC	Negative Cases
ResNet18	0.1328	0.1503	+0.0175	33/100
FLIGENT-Net	0.3639	0.1386	−0.2253	91/100

Figure 5 shows the corresponding mean normalized deletion curves for gradient-guided and random deletion.

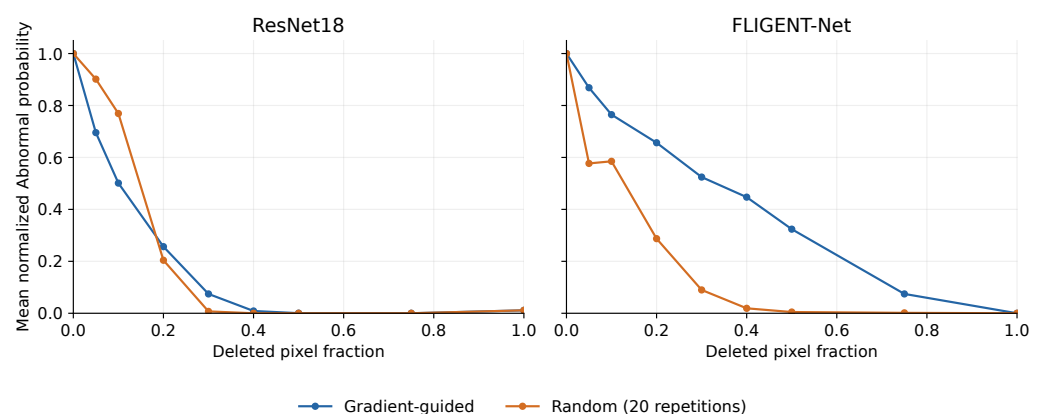


Figure 5. Mean per-image normalized deletion curves. Random curves average 20 shared permutations per image before averaging across images. No uncertainty across training seeds is estimated.

The Pixel Deletion analysis compared ResNet18 and FLIGENT-Net using the same input-gradient attribution, image subset, and perturbation protocol. ResNet18 obtained a mean ΔAUC of +0.0175, indicating a small average advantage of gradient-guided deletion over random deletion. FLIGENT-Net obtained -0.2253 , indicating that random deletion reduced its normalized Abnormal probability more effectively than deletion following the gradient ranking. Negative per-image gaps occurred in 33 and 91 of the 100 cases, respectively.

This difference highlights a limitation of the evaluated attribution–perturbation combination. Absolute input gradients measure local sensitivity around the original image, whereas cumulative deletion introduces finite changes. Furthermore, absolute values discard the direction of the gradient, so highly ranked pixels do not necessarily provide positive evidence for the Abnormal class. In FLIGENT-Net, each perturbed input also leads to recomputation of spectral descriptors and graph connectivity; the discrete edge-selection step is not differentiated by the gradient map. These properties may contribute to the discrepancy between the initial gradient ranking and the subsequent deletion response; however, their individual effects were not isolated experimentally.

The controlled comparison therefore identifies a boundary for interpreting the input-gradient maps: for FLIGENT-Net, their ranking did not outperform random deletion under the evaluated conditions. This finding complements the classification results by distinguishing predictive performance from the behavior of the selected attribution method. It does not evaluate the Transformer attention maps in Figure 4 or establish tumor-localization accuracy. The conclusions remain conditional on the 100 Abnormal images correctly classified by both models, the replacement scheme, and the evaluated checkpoints.

4.8. Comparison with Published GasHisSDB Results

Published GasHisSDB results are useful for external context, but they do not constitute a controlled ranking because the studies differ in split ratios, image-size subsets, data filtering, augmentation, pretrained weights, model-selection procedures, and reporting conventions. Table 12 therefore reports representative published results together with protocol notes rather than treating the numerical values as directly interchangeable.

Table 12. Representative published results relevant to GasHisSDB. Reported values are taken from the cited studies and should be interpreted together with their different experimental protocols. “Our split” denotes the persisted 60%/20%/20% patch-level partition used throughout the revised experiments.

Study	Data/Protocol	Representative Method	Accuracy (%)	F1 (%)
Hu et al. (2022) [21]	GasHisSDB, 160×160 ; study-specific train/validation/test protocol	ResNet50	96.09	95.10 (Abnormal)
Yong et al. (2023) [24]	GasHisSDB, 160×160 ; ensemble study	Ensemble-UA5	99.20	99.01
Mudavadkar et al. (2024) [25]	GasHisSDB, $80/120/160$; transfer-learning ensembles	Best reported ensemble across supplied resolutions	up to 99.72	–
Loddo et al. (2024) [26]	GasHisSDB; handcrafted/deep feature fusion without fine-tuning	Fused features + SVM	95.00	–
ResNet18 (this study)	GasHisSDB, 160×160 ; our split, seed 42	ImageNet-pretrained ResNet18	98.51	98.13
FLIGENT-Net (this study)	GasHisSDB, 160×160 ; our split, seed 42	CNN + FFT + GCN + Transformer	97.31	96.59

For the 160×160 subset, Hu et al. reported 96.09% accuracy for ResNet50, with 95.1% F1 for the Abnormal class [21]. Yong et al. reported 99.20% accuracy and 99.01% F1 for their best five-model unweighted ensemble on the same patch size [24]. Mudavadkar et

al. reported ensemble accuracies ranging up to 99.72% across the supplied GasHisSDB resolutions [25], whereas Loddo et al. obtained 95% accuracy with an SVM using fused handcrafted and deep features and additionally studied cross-resolution generalization [26]. These values show that the dataset supports substantially different performance levels under different modeling and evaluation choices.

A descriptive comparison of the reported accuracies further illustrates the effect of these heterogeneous protocols. Numerically, the ResNet18 baseline in the present study is 2.42 percentage points above the 96.09% ResNet50 result reported by Hu et al. and 0.69 points below the 99.20% Ensemble-UA5 result of Yong et al. FLIGENT-Net is 1.22 points above Hu et al. and 1.89 points below Yong et al. Relative to the 99.72% upper accuracy reported by Mudavadkar et al. across the supplied GasHisSDB resolutions, the corresponding differences are 1.21 points for ResNet18 and 2.41 points for FLIGENT-Net; relative to the 95.00% result of Loddo et al., they are +3.51 and +2.31 points, respectively. These arithmetic differences should not be interpreted as controlled performance gains or losses. Hu et al. evaluated a single CNN under a study-specific protocol, Yong et al. and Mudavadkar et al. used ensemble strategies, and Loddo et al. used fused handcrafted/deep features with an SVM. Consequently, the spread from approximately 95% to above 99% in Table 12 reflects not only model architecture but also differences in data partitioning, preprocessing, model selection, and evaluation design.

Under the revised common split used in the present study, ResNet18 achieved 98.51% accuracy and 98.13% F1, while FLIGENT-Net achieved 97.31% accuracy and 96.59% F1. These within-study results do not support a classification-performance advantage for the full multi-domain architecture over the CNN baseline. At the same time, direct claims of superiority or inferiority relative to the external studies in Table 12 would be unwarranted because the underlying protocols are not matched. The purpose of this comparison is therefore to position the observed results within the published GasHisSDB range and to motivate the controlled architectural characterization performed in this work.

5. Conclusions and Future Work

5.1. Conclusions

FLIGENT-Net provides a reproducible formulation for jointly representing gastric histopathology in four domains: CNN-derived spatial morphology, localized Fourier spectra, similarity-based graph relations, and Transformer-based global context. The principal contribution of the study is this explicit representation flow together with a controlled comparison of nine configurations under a common partition, training protocol, and evaluation scheme. FLIGENT-Net achieved 97.31% accuracy, 96.59% Abnormal-class F1, and 0.9964 AUC-ROC, whereas the simpler ResNet18 baseline achieved the highest accuracy and F1 in the comparison, at 98.51% and 98.13%, respectively. Thus, the complete multi-domain architecture is not presented as a predictive improvement over ResNet18; rather, it provides an experimental setting for characterizing what is added—and what is not gained—when spectral, relational, and contextual representations are integrated.

The complementary analyses make this characterization more explicit. The native-grid spectral control produced 97.43% accuracy, but it simultaneously changed the local FFT support from 32×32 to $22/23$ -pixel regions and altered the discrete spectral sampling, so the small difference cannot be attributed specifically to removal of interpolation. Computational profiling quantifies the resource trade-off introduced by the additional modules, while graph statistics show that a fixed cosine threshold produces representation-dependent connectivity regimes. Even when the CNN-derived graph is sparse, the subsequent Transformer still permits global interaction among all 49 node tokens, distinguishing limited GCN message passing from the global contextual stage. Finally, the common input-gradient

deletion protocol characterizes the behavior of the selected attribution method separately from predictive performance. Taken together, the reproducible formulation, controlled architectural comparisons, and component-level characterization provide a concrete basis for refining spectral fusion and graph construction in future work; the constraints on generalization and statistical inference are stated separately below.

5.2. Limitations

The evaluation uses one split and one run per configuration with seed 42; variability across runs is not estimated, and the numerical seed itself is not assigned statistical or model-selection significance. Accordingly, small observed differences among configurations should not be generalized beyond this fixed stochastic setting. The split is performed at image-patch level and therefore does not establish patient-, slide-, or source-image-independent generalization. Because no provenance mapping from the distributed patches to their originating source images was available in the materials examined, related patches from a common source cannot be excluded from different partitions, even though byte-identical files were not shared across them. This dependence could produce optimistic performance estimates and requires caution when comparing the reported results with studies based on patient- or source-grouped splits. Pretraining differs between CNN and patch-based controls, and the variants are not parameter matched. The positive-class weight, graph-connectivity threshold, and classification threshold were kept fixed throughout the revised experiments; their sensitivity was not evaluated under the present protocol. In particular, the same graph threshold generated sharply different densities across representation spaces, and no degree-matching, k -nearest-neighbor constraint, or learned graph-structure mechanism was evaluated. Consequently, the reported results characterize the architectures at this operating point rather than establishing optimal hyperparameter values or a topology-matched comparison. No stain-normalization comparison or augmentation study is reported. Native-grid versus resized FFT was evaluated in one additional seed-matched control, but the small observed difference was not replicated across multiple seeds or alternative spectral parameterizations. The reported latency and GPU-memory measurements are hardware- and implementation-specific inference measurements with GPU-resident inputs; they exclude image decoding, preprocessing, host-to-device transfer, and training-time memory. These limitations constrain causal claims about representation components, deployment cost, and external generalization.

5.3. Future Work

Future work should prioritize source-grouped evaluation. This requires recovering or obtaining a reliable mapping between the distributed GasHisSDB patches and their originating source images and, if patient or slide identifiers become available, grouping samples at the highest available provenance level so that related samples cannot cross the training, validation, and test partitions. The principal architectural comparisons should then be repeated under that grouped protocol. Evaluation across multiple training seeds and independent histopathology data sources is also needed to characterize variability and external generalization. A dedicated hyperparameter study may subsequently evaluate class weighting, graph connectivity, and the binary decision threshold using training/validation data only, with the selected operating point locked before independent test evaluation. The topology results additionally motivate representation-aware graph construction, including k -nearest-neighbor graphs, adaptive thresholds, minimum-degree constraints, and learnable graph-structure methods, so that relational connectivity is not determined solely by one fixed cosine threshold. In parallel, future studies should replicate the native-grid FFT comparison across multiple seeds and investigate improved spectral representations and

fusion strategies; learnable spectral filters and wavelet-based descriptors are candidate alternatives rather than demonstrated improvements. Model compression, more efficient fusion, and profiling on additional hardware are also relevant for reducing the computational overhead of the full multi-domain configuration. Interpretation methods should continue to be compared under a common attribution and perturbation protocol.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/math14213867/s1>: Supplementary Note S1: Verification of Transformer-Attention Extraction.

Author Contributions: Conceptualization, Z.R.-V. and J.Y.M.-P.; methodology, Z.R.-V. and J.Y.M.-P.; software, Z.R.-V.; validation, Z.R.-V., J.Y.M.-P., J.L.L.-B., E.R.-E., and J.H.S.-A.; formal analysis, Z.R.-V.; investigation, Z.R.-V.; resources, J.Y.M.-P., J.L.L.-B., E.R.-E., J.H.S.-A., V.H.P.-P., O.E.-S., and S.V.-B.; data curation, Z.R.-V.; writing—original draft preparation, Z.R.-V.; writing—review and editing, J.Y.M.-P., J.L.L.-B., E.R.-E., J.H.S.-A., V.H.P.-P., O.E.-S., and S.V.-B.; visualization, Z.R.-V.; supervision, J.Y.M.-P.; project administration, J.Y.M.-P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was a secondary computational analysis of the publicly available GasHisSDB histopathological image dataset. No participants were recruited, no new tissue samples were collected, and the authors did not access patient identities or identifiable clinical records. A separate ethics review was not sought because the study was limited to the reuse of these publicly distributed images and class labels. No formal institutional exemption is claimed. The original GasHisSDB publication reports that the research and preparation of the dataset were approved by the Ethical Committee at Northeastern University [20]; this approval pertains to the source dataset, not to a new approval for the present analysis.

Informed Consent Statement: No new informed consent was obtained for this secondary analysis of publicly distributed images; the present study involved no contact with participants.

Data Availability Statement: GasHisSDB is available from MIaMIA Group on Figshare at <https://doi.org/10.6084/m9.figshare.15066147> [41], under the Creative Commons Attribution 4.0 International license (<https://creativecommons.org/licenses/by/4.0/>; accessed on 19 October 2026). The dataset was downloaded using the script available at <https://github.com/engrkhubabahmad/gashissdb-dataset> (accessed on 19 October 2026), which retrieves the Figshare files. Histopathological images reproduced in this article are attributed to the dataset creators; resizing and model-derived overlays are identified in the relevant figure captions. Further inquiries about the analyses can be directed to the corresponding authors.

Acknowledgments: The authors acknowledge the institutional and academic support of the Instituto Politécnico Nacional (IPN), through the Secretaría de Investigación y Posgrado, and the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI). During the preparation of this study, the authors used ChatGPT and Codex (OpenAI, GPT-5.6) for assistance with writing, programming, interpretation of results, and schematic illustration preparation, as detailed in Section 3.7. The authors reviewed and edited the outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. <https://doi.org/10.3322/caac.21660>.
2. Bray, F.; Laversanne, M.; Sung, H.; Gortmaker, D.; Siegel, R.L. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. <https://doi.org/10.3322/caac.21834>.

3. Rawla, P.; Barsouk, A. Epidemiology of Gastric Cancer: Global Trends, Risk Factors and Prevention. *Prz. Gastroenterol.* **2019**, *14*, 26–38. <https://doi.org/10.5114/pg.2018.80001>.
4. Martínez-Galindo, M.; Zamarripa-Dorsey, F.; Carmona-Castañeda, A.; Angeles-Labra, A.; Peña-vera-Hernández, R.; Ugarte-Briones, C.; Blanco-Vela, C. Características Histopatológicas del Adenocarcinoma Gástrico en Pacientes Mexicanos. Experiencia de 10 Años en el Hospital Juárez de México. *Rev. Gastroenterol. México* **2015**, *80*, 21–26. <https://doi.org/10.1016/j.rgm.2014.11.002>.
5. Gibson, D.; Gaydecki, P. Definition and Application of a Fourier Domain Texture Measure: Applications to Histological Image Segmentation. *Comput. Biol. Med.* **1995**, *25*, 551–557. [https://doi.org/10.1016/0010-4825\(95\)00028-X](https://doi.org/10.1016/0010-4825(95)00028-X).
6. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. <https://doi.org/10.1109/TNNLS.2020.2978386>.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
8. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Krauss Silva, V.W.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat. Med.* **2019**, *25*, 1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>.
9. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* **2018**, *24*, 1559–1567. <https://doi.org/10.1038/s41591-018-0177-5>.
10. Echle, A.; Rindtorff, N.T.; Brinker, T.J.; Luedde, T.; Pearson, A.T.; Kather, J.N. Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. *Br. J. Cancer* **2021**, *124*, 686–696. <https://doi.org/10.1038/s41416-020-01122-x>.
11. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Moeskops, O.; van der Heijden, M.; Sánchez, C.I.; van Ginneken, B. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal.* **2017**, *42*, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
12. Khosravi, P.; Green, S.; Ho, D.J.; Choudhary, T.; Madabhushi, A.; Salgado, R. Artificial Intelligence in Gastrointestinal Pathology: A Review. *J. Pathol. Inform.* **2021**, *12*, 39. https://doi.org/10.4103/jpi.jpi_50_21.
13. Bui, D.C.; Song, B.; Kim, K.; Kwak, J.T. Spatially-Constrained and -Unconstrained Bi-Graph Interaction Network for Multi-Organ Pathology Image Classification. *IEEE Trans. Med Imaging* **2025**, *44*, 194–206. <https://doi.org/10.1109/TMI.2024.3436080>.
14. Xia, S.; Xia, Y.; Liu, T.; Luo, Y.; Pang, P.C.I. Application of Deep Learning Models in Gastric Cancer Pathology Image Analysis: A Systematic Scoping Review. *BMC Cancer* **2025**, *25*, 1257. <https://doi.org/10.1186/s12885-025-14662-3>.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
17. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. *Med Image Anal.* **2020**, *63*, 101693. <https://doi.org/10.1016/j.media.2020.101693>.
18. Iizuka, O.; Kanavati, F.; Kato, M.; Rambert, M.; Arihiro, K.; Tsuneki, M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci. Rep.* **2020**, *10*, 1504. <https://doi.org/10.1038/s41598-020-58467-9>.
19. Li, L.; Chen, M.; Zhou, Y.; Wang, J.; Wang, D. Research of Deep Learning on Gastric Cancer Diagnosis. In Proceedings of the 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), Fuzhou, China, 13–16 December 2020; pp. 1–3. <https://doi.org/10.1109/CSRSWTC50769.2020.9372583>.
20. Hu, W.; Li, C.; Li, X.; Rahaman, M.M.; Ma, J.; Zhang, Y.; Chen, H.; Liu, W.; Sun, C.; Yao, Y.; et al. GasHisSDB: A New Gastric Histopathology Image Dataset for Computer-Aided Diagnosis of Gastric Cancer. *Comput. Biol. Med.* **2022**, *142*, 105207. <https://doi.org/10.1016/j.compbiomed.2021.105207>.
21. Hu, W.; Chen, H.; Liu, W.; Li, X.; Sun, H.; Huang, X.; Grzegorzec, M.; Li, C. A Comparative Study of Gastric Histopathology Sub-Size Image Classification: From Linear Regression to Visual Transformer. *Front. Med.* **2022**, *9*, 1072109. <https://doi.org/10.3389/fmed.2022.1072109>.
22. Bratajaya, A.M.; Nabiilah, G.Z.; Nurimaba, D.P.; Moniaga, J.V. Leveraging Deep Learning for Early Detection of Stomach Cancer: A Convolutional Neural Network Approach. In Proceedings of the 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS), Surakarta, Indonesia, 29–30 November 2024. <https://doi.org/10.1109/ICORIS63540.2024.10903723>.
23. Yong, M.P.; Hum, Y.C. Histopathological Gastric Cancer Detection Using Transfer Learning. In Proceedings of the 2023 11th International Conference on Bioinformatics and Computational Biology (ICBCB), Hangzhou, China, 21–23 April 2023. <https://doi.org/10.1109/ICBCB57893.2023.10246524>.
24. Yong, M.P.; Hum, Y.C.; Lai, K.W.; Lee, Y.L.; Goh, C.H.; Yap, W.S.; Tee, Y.K. Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning. *Diagnostics* **2023**, *13*, 1793. <https://doi.org/10.3390/diagnostics13101793>.

25. Mudavadkar, G.R.; Deng, M.; Al-Heejawi, S.M.A.; Arora, I.H.; Breggia, A.; Ahmad, B.; Christman, R.; Ryan, S.T.; Amal, S. Gastric Cancer Detection with Ensemble Learning on Digital Pathology: Use Case of Gastric Cancer on GasHisSDB Dataset. *Diagnostics* **2024**, *14*, 1746. <https://doi.org/10.3390/diagnostics14161746>.
26. Loddo, A.; Usai, M.; Di Ruberto, C. Gastric Cancer Image Classification: A Comparative Analysis and Feature Fusion Strategies. *J. Imaging* **2024**, *10*, 195. <https://doi.org/10.3390/jimaging10080195>.
27. Yener, B. Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship. *Commun. ACM* **2016**, *60*, 74–84. <https://doi.org/10.1145/2934667>.
28. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
29. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
30. Xiang, X.; Wu, X. Multiple Instance Classification for Gastric Cancer Pathological Images Based on Implicit Spatial Topological Structure Representation. *Appl. Sci.* **2021**, *11*, 10368. <https://doi.org/10.3390/app112110368>.
31. Maron, O.; Lozano-Pérez, T. A Framework for Multiple-Instance Learning. *Adv. Neural Inf. Process. Syst.* **1998**, *10*, 570–576.
32. Brussee, S.; Buzzanca, G.; Schrader, A.M.R.; Kers, J. Graph Neural Networks in Histopathology: Emerging Trends and Future Directions. *Med Image Anal.* **2025**, *101*, 103444. <https://doi.org/10.1016/j.media.2024.103444>.
33. Shi, Z.; Zhang, J.; Kong, J.; Wang, F. Integrative Graph-Transformer Framework for Histopathology Whole Slide Image Representation and Classification. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2024*; Springer Nature Switzerland: Cham, Switzerland, 2024; Volume 15011, Lecture Notes in Computer Science, pp. 341–350. https://doi.org/10.1007/978-3-031-72120-5_32.
34. Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; Han, X. TransPath: Transformer-Based Self-Supervised Learning for Histopathological Image Classification. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*; Springer International Publishing: Basel, Switzerland, 2021; Volume 12908, Lecture Notes in Computer Science, pp. 186–195. https://doi.org/10.1007/978-3-030-87237-3_18.
35. Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; Zhang, Y. TransMIL: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; Volume 34, pp. 2136–2147.
36. Chen, H.; Li, C.; Wang, G.; Li, X.; Rahaman, M.M.; Sun, H.; Hu, W.; Li, Y.; Liu, W.; Sun, C.; et al. GasHis-Transformer: A Multi-Scale Visual Transformer Approach for Gastric Histopathological Image Detection. *Pattern Recognit.* **2022**, *130*, 108827. <https://doi.org/10.1016/j.patcog.2022.108827>.
37. Armand, T.P.T.; Bhattacharjee, S.; Kim, H.J.; Hussain, A.; Ali, S.; Choi, H.K.; Kim, H.C. Vision Transformer-Based Model for Gastric Cancer Detection and Classification Using Weakly Annotated Histopathological Images. In *Proceedings of the 2024 26th International Conference on Advanced Communications Technology (ICACT)*; IEEE: New York, NY, USA, 2024; pp. 413–418. <https://doi.org/10.23919/ICACT60172.2024.10471994>.
38. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. <https://doi.org/10.1109/TPAMI.2002.1017623>.
39. Tuceryan, M.; Jain, A.K. Texture Analysis. In *Handbook of Pattern Recognition and Computer Vision*, 2nd ed.; Chen, C., Pau, L., Landy, M., Eds.; World Scientific: Singapore, 1998; pp. 207–248.
40. Liu, L.; Chen, J.; Fieguth, P.; Zhao, G.; Chellappa, R.; Pietikäinen, M. From BoW to CNN: Two Decades of Texture Representation for Texture Classification. *Int. J. Comput. Vis.* **2019**, *127*, 74–109. <https://doi.org/10.1007/s11263-018-1125-z>.
41. MIaMIA Group. GasHisSDB, 2021. Dataset. Licensed Under CC BY 4.0. Available online: <https://figshare.com/articles/dataset/GasHisSDB/15066147> (accessed on 19 October 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.