Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LungColonCancerPred-NN: A Neural Network Approach for Lung and Colon Histopathology Prediction

outhman abbassi

outhmane.abbassi@um5r.ac.ma

Mohammed V University

Zeyneb Belganche

Mohammed V University

Wajih Rhalem

Mohammed V University

Nassim Kharmoum

Mohammed V University

Research Article

Keywords: Histopathology, Lung cancer, Colon cancer, LC25000, Graph neural networks, Vision Transformers, Data leakage, Deep learning

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11473288/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

LungColonCancerPred-NN: A Neural Network Approach for Lung and Colon Histopathology Prediction

Outhman Abbassi[1,2*] Zeyneb Belganche[3] Wajih Rhalem[2,3] Nassim Kharmoum[1,2]

[1]IPSS Research Team, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco

[2]Moroccan Society of Digital Health, Rabat, Morocco

[3]E2SN Research Team, ENSAM Rabat, Mohammed V University in Rabat, Rabat, Morocco

Abstract

Lung and colorectal cancers are the first and second causes of cancer death worldwide, and their diagnosis rests on the reading of hematoxylin and eosin (H&E) stained tissue. The LC25000 dataset is the most used public benchmark for automated classification of lung and colon histopathology, but its $25\,000$ images were produced by augmenting only $1\,250$ source images, so a random split places near copies of the same image in the training and test sets. We present LungColonCancerPred-NN, a compact triple-branch network (5.8M parameters) for five-class classification (colon adenocarcinoma, benign colon, lung adenocarcinoma, benign lung, and lung squamous cell carcinoma), and we evaluate it under a protocol designed to keep near copies apart. We combined LC25000 with a second public collection of the same classes ($40\,000$ files), removed $2\,076$ exact duplicates by MD5 hashing, grouped the remaining $37\,924$ images into $7\,795$ augmentation families using an EfficientNet-B0 embedding made invariant to the eight rotations and reflections of the square, and split the data $70/15/15$ by family. An audit of nearest-neighbor similarities confirmed the separation: the median similarity of a test image to its closest training image was 0.961, against 0.998 within the training set, and no test image exceeded the grouping threshold of 0.98. The model shares one EfficientNet-B0 feature map between a convolutional pooling branch, a graph branch that passes messages between similar feature positions, and a Transformer encoder branch, and a sigmoid gate weights the three embeddings before classification. On the held-out test set of $5\,497$ images, LungColonCancerPred-NN reached an accuracy of 99.25% (95% bootstrap CI 99.02-99.49), a macro-F_1 of 99.38% (99.19-99.57), Cohen's κ of 0.9905 and a macro-AUC of 0.9997. No malignant image was classified as benign, only two images were assigned to the wrong organ, and 33 of the 41 errors, squamous cell carcinomas were predicted as adenocarcinomas. This accuracy is higher than those reported by four recent (2025-2026) CNN, Vision Transformer, and CNN-Transformer methods on LC25000 (97.00-99.11%) for the three lung classes alone, while our model handles five classes under a split grouped by augmentation family.

Keywords: Histopathology · Lung cancer · Colon cancer · LC25000 · Graph neural networks · Vision Transformers · Data leakage · Deep learning

*Corresponding author: outhmane.abbassi@um5r.ac.ma

1 Introduction

Lung cancer was the most frequently diagnosed cancer in 2024, with about 2.6 million new cases and 1.9 million deaths, and colorectal cancer ranked third in incidence and second in mortality (Sung et al., 2026). In the United States alone, the two diseases account for a large share of the cancer deaths expected each year (Siegel et al., 2025). For both, the diagnosis and the histological type are set by a pathologist on stained tissue sections: in the lung, the distinction between adenocarcinoma and squamous cell carcinoma guides treatment, and in the colon the pathologist must separate adenocarcinoma from benign mucosa. This work is time-consuming, and it depends on the availability of trained pathologists, which motivates computer-aided classification of histopathology images.

Most published work on this problem uses the LC25000 dataset (Borkowski et al., 2019). It contains 25 000 colour images of 768×768 pixels in five classes of 5 000 images each. The images were generated from 750 lung and 500 colon source images (250 per class) by rotations and flips with the Augmentor package, so that each source image appears about twenty times. Recent studies on LC25000 report accuracies close to 100% with lightweight convolutional networks (Jim et al., 2025; AlShehri, 2025), dual-branch attention networks (Ochoa-Ornelas et al., 2025a), CNN-Transformer hybrids (Ramakrishnan et al., 2026; Sunitha and Vellingiri, 2026), graph-Transformer designs (Saremi and Ahmadkhan Kordbacheh, 2025), and fusions of handcrafted and deep features (Anusha and Reddy, 2025; Patharia et al., 2025). Other recent methods report lower values, between 97% and 99.1%, on the three lung classes (Ochoa-Ornelas et al., 2025b; Setiawan et al., 2025; Chaddad et al., 2025; Ramakrishnan et al., 2026). The structure of the dataset makes all these figures hard to read. When augmented copies of one source image are spread over the training and test sets, a model can score well by recognising images it has practically already seen. Diosdado et al. (2024) noted that the lack of traceability between LC25000 images and their sources can lead to data leakage, and Sunitha and Vellingiri (2026) discuss the same issue when interpreting accuracy on this benchmark. Leakage of this kind is a recognised cause of inflated results in medical imaging (Young et al., 2025; Jiménez-Sánchez et al., 2025).

Our previous version of this pipeline, trained on a random split, reached 99.4% accuracy. Before reporting results, we therefore rebuilt the evaluation. The two public collections we combine contain exact copies of each other, which we remove by hashing, and the augmented copies of a source image are not identical files, so we group them by similarity in an embedding that does not change under rotation or reflection. The split is then made by group, and we check it with a nearest-neighbor audit.

On the modelling side, histopathology images carry information at several scales. Convolutional networks capture the texture of nuclei and cytoplasm well. Relations between regions, such as the arrangement of glands or the way tumour cells form sheets, are better described by message passing between regions, as in graph neural networks (Mienye and Viriri, 2025; Ye et al., 2026), or by self-attention, as in Transformers (Vaswani et al., 2017). LungColonCancerPred-NN combines the three on a single EfficientNet-B0 feature map (Tan and Le, 2019), which keeps the model small.

The contributions of this work are the following.

- A leakage-controlled evaluation protocol for LC25000 and for collections derived from it. Exact copies are removed by MD5 hashing of the file bytes. Augmented copies, which differ in every byte, are grouped into families with an EfficientNet-B0 embedding averaged over the eight rotations and reflections of the square, linked by a cosine threshold chosen so that no family exceeds 80 images. The data are then split $70/15/15$ by family with stratification on the class, and the split is checked by comparing, for each test image, the similarity to its nearest training image with the similarities found inside the training set. On our data, this procedure removed 2 076 duplicates from 40 000 files and formed 7 795 families, and no test image had a training neighbor at or above the threshold.

- LungColonCancerPred-NN, a triple-branch network of 5.8M parameters built on one EfficientNet-B0

feature map. A convolutional branch pools local texture, a graph branch passes messages between regions with similar features, and a Transformer branch with positional embeddings relates regions by their place in the image; a sigmoid gate weights the three embeddings for each image before classification. The network is trained in two phases, first with the backbone frozen and then end to end, with label smoothing and stochastic weight averaging, and it is evaluated with three-view test-time augmentation.

- A detailed evaluation under this protocol on $5\,497$ test images from $1\,114$ families: 99.25% accuracy and 99.38% macroF_1 with bootstrap confidence intervals, Cohen's κ and macro-AUC, per-class precision, recall and specificity, and an analysis of the 41 errors at the level of the class, the organ, and the benign-malignant decision. We also compare these results with four methods published in 2025 and 2026, which report lower accuracies on the three lung classes alone.

The rest of the paper is organised as follows. Section 2 describes the two datasets, the deduplication and family construction, the grouped split and its audit, the preprocessing, the three branches and their fusion, and the training and evaluation settings. Section 3 first reports the leakage audit and the training curves, then the test performance, the per-class results and the error analysis, followed by qualitative predictions and the comparison with published work. Section 4 discusses the evaluation protocol, the role of each branch, the clinical reading of the errors, and the size, cost and calibration of the model, and Section 5 concludes.

2 Materials and methods

2.1 Datasets

We used two public collections from Kaggle. The first is LC25000 (Borkowski et al., 2019) (`andrewmvd/lung-and-colon-cancer-histopathological-images`), with $5\,000$ JPEG images in each of the five classes: colon adenocarcinoma (colon_aca), benign colon tissue (colon_n), lung adenocarcinoma (lung_aca), benign lung tissue (lung_n) and lung squamous cell carcinoma (lung_scc). The second is the lung and colon part of the "Cancer classification" collection (`maestroalert/cancer`), which provides $15\,000$ further images in the colon_aca, lung_aca and lung_scc classes. Folder names from both sources were mapped to the five labels, and all $40\,000$ files were readable.

2.2 Leakage-controlled data protocol

Exact de-duplication We computed the MD5 hash of the bytes of every file and kept the first occurrence of each hash. This removed $2\,076$ files: $1\,280$ duplicates inside LC25000 and 796 files of the second collection. The resulting pool contains $37\,924$ unique files ($23\,720$ from LC25000 and $14\,204$ from the second collection), with $9\,506$ colon_aca, $4\,772$ colon_n, $9\,454$ lung_aca, $4\,744$ lung_n and $9\,448$ lung_scc images.

Augmentation families Hashing only finds identical files. The copies that Augmentor produced from one source image differ by rotations and flips, and the same image may also have been re-encoded in the second collection. To find them, we represented each image by an embedding that does not change under these operations. Let ϕ be an ImageNet-pre-trained EfficientNet-B0 without its classifier and D_4 the group of the eight rotations and reflections of the square. For an image x (resized to 224×224 and normalised),

$$\mathbf{e}(x) = \frac{\bar{\phi}(x)}{\|\bar{\phi}(x)\|_2}, \qquad \bar{\phi}(x) = \frac{1}{8} \sum_{g \in D_4} \phi(g \cdot x). \tag{1}$$

Two images i and j are linked when j is among the 30 nearest neighbours of i, their cosine similarity $\mathbf{e}_i^\top \mathbf{e}_j$ is at least a threshold τ, and they share the same label. The connected components of this graph, found with a union-find structure, are the augmentation families. Long chains of links can merge unrelated images into one large component, so we tested $\tau \in \{0.90, 0.92, 0.94, 0.96, 0.98\}$ and kept the smallest threshold for which no family exceeds 80 images. Table 1 shows that low thresholds produce families of several thousand images; $\tau = 0.98$ is the first value that satisfies the rule, with 7 795 families of median size 2 and maximum size 62.

Table 1 Augmentation families were obtained for each similarity threshold τ on the 37 924 de-duplicated images. The selected threshold is the smallest one whose largest family has at most 80 images

τ	Families	Median size	Largest family	Selected
0.90	363	36	8 566	
0.92	719	34	7 066	
0.94	1 360	21	2 688	
0.96	2 864	6	186	
0.98	7 795	2	62	✓

Grouped split We split the families, not the images. A first stratified group split with seven folds set one fold apart as the test set, and a second split with six folds on the remainder set one fold apart for validation. This gave 26 976 training, 5 451 validation, and 5 497 test images ($71.1/14.4/14.5\%$), with 1 114 families in the test set and no family shared between any two sets. The test set contains 1 340 colon_aca, 668 colon_n, 1 404 lung_aca, 727 lung_n and 1 358 lung_scc images.

Leakage audit For every test image we computed the cosine similarity to its nearest training image in the embedding of Eq. (1), and for every training image the similarity to its nearest other training image. If near copies remained across the split, the two distributions would overlap. Section 3.1 reports the result.

2.3 Preprocessing and data augmentation

Images were converted to RGB, resized to 224×224 pixels, and normalized with ImageNet statistics. We did not apply contrast equalization or stain normalization so that the color information of the H&E stain is kept. Training augmentation consisted of horizontal and vertical flips ($p = 0.5$ each), rotations by multiples of 90° ($p = 0.5$), an affine transform (translation $\pm 6\%$, scale 0.90-1.10, rotation $\pm 15^\circ$; $p = 0.4$), one of brightness-contrast (± 0.15) or hue-saturation-value shifts (10, 20, 15; $p = 0.4$), Gaussian noise (standard deviation 0.01-0.02 of the range; $p = 0.15$) and coarse dropout of 2 to 6 patches of 6-20 pixels ($p = 0.25$). Tissue sections have no preferred orientation, which is why all rotations and flips were allowed.

2.4 Architecture

LungColonCancerPred-NN (Fig. 1) starts from an ImageNet-pre-trained EfficientNet-B0 (Tan and Le, 2019). Its last stage gives a feature map of $320 \times 7 \times 7$ for a 224-pixel input, which a 1×1 convolution and batch normalization project to $D = 256$ channels, $\mathbf{F} \in \mathbb{R}^{256 \times 7 \times 7}$. We write $\mathbf{x}_1, \ldots, \mathbf{x}_N$ for the $N = 49$ feature vectors of $\mathbf{F}$. The three branches below read the same $\mathbf{F}$.

Convolutional branch global average pooling gives $\mathbf{z}_{\text{cnn}} = \frac{1}{N} \sum_i \mathbf{x}_i$. This vector summarizes the local texture found anywhere in the image: nuclear size and density, chromatin, and the color of cytoplasm and stroma.

Graph branch The 49 positions form the nodes of a graph whose edges are set by feature similarity rather than by spatial distance. With s_{ij} the cosine similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, each node is linked to the $k+1=9$ nodes of highest similarity (itself included), and the edge weights are

$$a_{ij} = \frac{\max(s_{ij}, 0)}{\sum_{l \in \mathcal{N}(i)} \max(s_{il}, 0)}, \qquad j \in \mathcal{N}(i). \tag{2}$$

starting from $\mathbf{h}_i^{(0)} = \mathbf{x}_i$, each of $L = 3$ layers aggregates the neighbors and adds a residual connection,

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{(\ell)} + \mathrm{Dropout}\left(\mathrm{SiLU}\left(\mathbf{W}^{(\ell)} \sum_{j \in \mathcal{N}(i)} a_{ij}\, \mathbf{h}_j^{(\ell)}\right)\right), \tag{3}$$

and the read-out is $\mathbf{z}_{\text{gnn}} = \mathrm{LN}\big(\mathbf{W}_o \frac{1}{N}\sum_i \mathbf{h}_i^{(L)} + \frac{1}{N}\sum_i \mathbf{x}_i\big)$. Because the graph links positions that look alike, this branch lets, for instance, all regions with gland lumens or all regions of tumor nuclei exchange information, even when they are far apart in the image.

Transformer branch A learnable class token is prepended to the 49 vectors, a learnable positional embedding is added, and two pre-norm Transformer encoder layers with 4 heads, a feed-forward width of $4D$ and GELU activations apply multi-head self-attention (Vaswani et al., 2017),

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}. \tag{4}$$

The normalised output of the class token is $\mathbf{z}_{\text{tf}}$. Unlike the graph branch, this branch knows where each feature lies, and every position can attend to every other one, which suits the overall layout of the tissue.

Gated fusion and classifier A sigmoid gate assigns each branch a weight between 0 and 1 that depends on the image,

$$\mathbf{g} = \sigma\big(\mathbf{W}_g[\mathbf{z}_{\text{cnn}} \| \mathbf{z}_{\text{gnn}} \| \mathbf{z}_{\text{tf}}] + \mathbf{b}_g\big) \in (0,1)^3, \tag{5}$$

$$\mathbf{f} = \mathrm{Dropout}\big(\mathrm{SiLU}\big(\mathrm{LN}\big(\mathbf{W}_f[g_1\mathbf{z}_{\text{cnn}} \| g_2\mathbf{z}_{\text{gnn}} \| g_3\mathbf{z}_{\text{tf}}]\big)\big)\big), \tag{6}$$

with $\mathbf{f} \in \mathbb{R}^{256}$. The gates are independent, so the model can keep two or three branches at full weight rather than share a fixed total among them. A two-layer head ($256 \to 128 \to 5$, SiLU, dropout) gives the class logits. Dropout is 0.20 throughout, and the full model has 5.8M parameters.

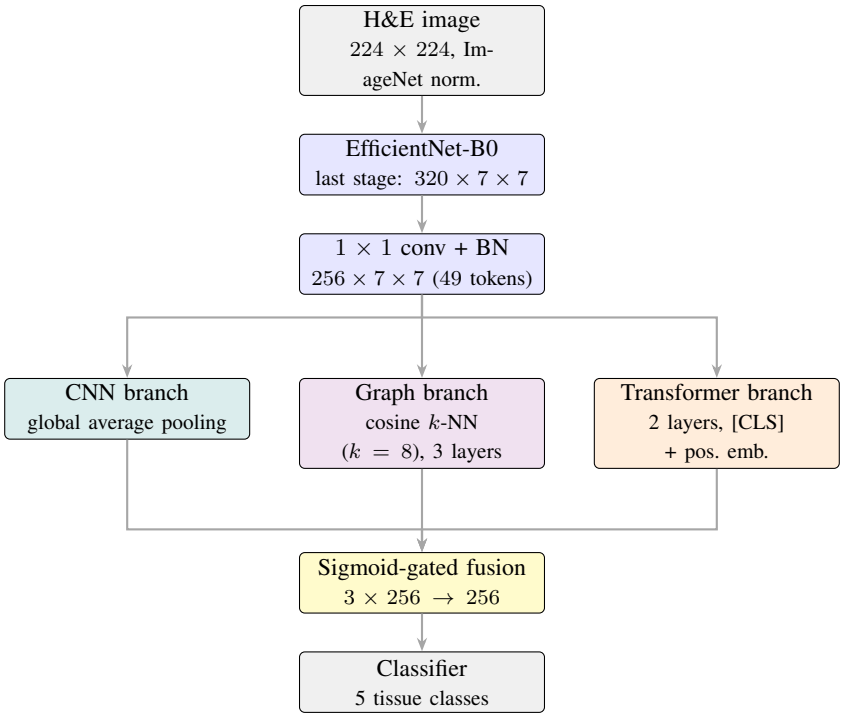


Fig. 1 Architecture of LungColonCancerPred-NN. A projected EfficientNet-B0 feature map feeds a convolutional pooling branch, a graph branch that passes messages between positions with similar features, and a Transformer branch; a sigmoid gate weights the three embeddings before classification into colon_aca, colon_n, lung_aca, lung_n and lung_scc

2.5 Training and evaluation

The loss was cross-entropy with label smoothing $\varepsilon = 0.08$. With $K = 5$ classes, the posterior that minimizes this loss for the true class is $1 - \varepsilon + \varepsilon/K = 0.936$, which is the confidence a well-fitted model should show. The classes are close to balanced after de-duplication (the benign classes have about half the images of the others), and we used no class weighting or resampling. Optimization used AdamW (Loshchilov and Hutter, 2019) with weight decay 0.02, batch size 32 and gradient clipping at norm 1.

Training had two phases. In phase 1 (4 epochs) the backbone weights were frozen and the other modules were trained at a learning rate of 8×10^{-4}. In phase 2 (22 epochs) the whole network was trained, with a learning rate of 2×10^{-5} for the backbone and 1.5×10^{-4} for the other modules. Both phases used a linear warm up of three epochs followed by cosine decay. From the 15th epoch of phase 2, stochastic weight averaging (SWA) (Izmailov et al., 2018) averaged the weights after each epoch with a constant learning rate of 1.5×10^{-5}, and the batch-normalization statistics of the averaged model were recomputed on the training set. After every epoch we computed the validation macro-F_1, kept the weights with the best value, and compared them with the SWA model at the end; early stopping with a patience of 8 epochs was available but was not triggered.

At test time we averaged the softmax outputs over the image and its horizontal and vertical flips,

$$\bar{\mathbf{p}}(x) = \tfrac{1}{3}\big[\mathrm{softmax}(f(x)) + \mathrm{softmax}(f(x^{\leftrightarrow})) + \mathrm{softmax}(f(x^{\updownarrow}))\big]. \tag{7}$$

We report accuracy, per-class precision, recall (sensitivity), specificity and F_1, macro-averaged F_1, Cohen's κ and the one-vs-rest macro-averaged area under the ROC curve (AUC). Confidence intervals for accuracy and macro-F_1 are 2.5th and 97.5th percentiles over $2\,000$ bootstrap resamples of the test set. Experiments

used PyTorch, `timm` and `albumentations` 2 with seed 42 on one NVIDIA Tesla T4 GPU. Table 2 summarises the settings.

Table 2 Main settings of LungColonCancerPred-NN and of the evaluation protocol

Component / setting	Value
Backbone	EfficientNet-B0 (ImageNet), last stage, 7×7 map
Projection / embedding D	1×1 conv + BN, $D = 256$
Graph branch	cosine k-NN in feature space, $k = 8$, 3 residual layers
Transformer branch	2 pre-norm layers, 4 heads, FFN $4D$, positional embedding
Fusion	sigmoid gate per branch, $3D \to 256$, LayerNorm, SiLU
Dropout / parameters	0.20 / 5.8M
Input	224×224 px, ImageNet normalisation, no stain normalisation
Optimiser	AdamW, weight decay 0.02, batch 32, clipping at 1.0
Phase 1 (frozen backbone)	4 epochs, LR 8×10^{-4}, warm-up + cosine
Phase 2 (fine-tuning)	22 epochs, LR 2×10^{-5} (backbone), 1.5×10^{-4} (rest)
Weight averaging	SWA from epoch 15 of phase 2, LR 1.5×10^{-5}
Loss	cross-entropy, label smoothing 0.08
Test-time augmentation	identity, horizontal flip, vertical flip
De-duplication	MD5 of file bytes (2 076 files removed)
Families	D_4-invariant EfficientNet-B0 embedding, top-30, $\tau = 0.98$
Split	stratified group split by family, $70/15/15$

3 Results

3.1 Leakage audit

Fig. 2 compares the two similarity distributions. Within the training set, almost every image has a near twin: the 5th, 25th, 50th, 75th and 95th percentiles of the similarity to the nearest other training image are 0.985, 0.995, 0.998, 0.999 and 1.000. This is what one expects from a dataset built by augmentation. For test images, the similarity to the nearest training image is clearly lower, with percentiles of 0.910, 0.944, 0.961, 0.971 and 0.978. None of the 5 497 test images has a training neighbor at or above $\tau = 0.98$, while most training images have one above 0.99. The two distributions barely overlap, so the test set measures the classification of images whose near copies were not seen during training.

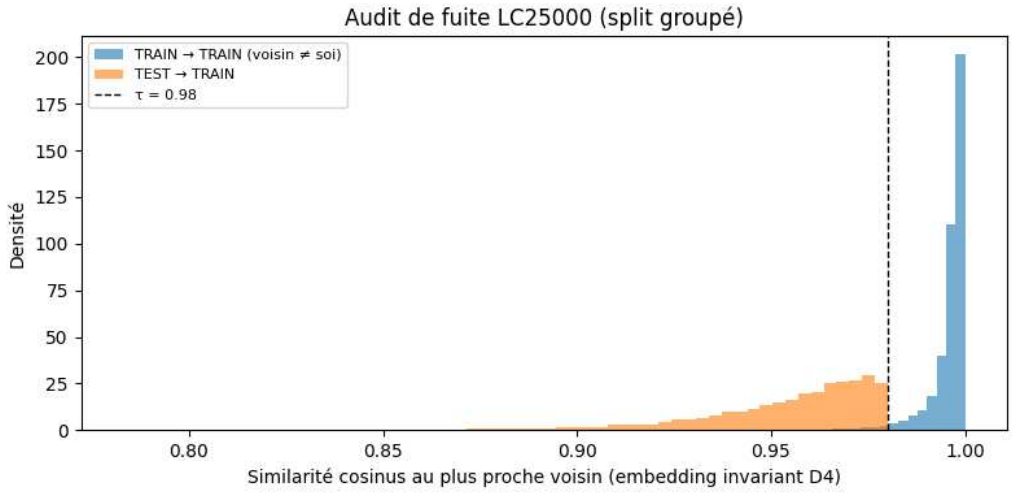


Fig. 2 Leakage audit of the grouped split. Cosine similarity, in the D_4-invariant embedding, of each training image to its nearest other training image (blue) and of each test image to its nearest training image (orange). The dashed line is the grouping threshold $\tau = 0.98$; no test image lies above it

3.2 Training behaviour

Fig. 3 shows the training curves. With the backbone frozen, the new modules reached a validation macro-F_1 of 0.953 after four epochs. Fine-tuning raised it steadily to 0.978 at the 19th epoch of phase 2, and the training loss decreased from 0.534 to 0.394. The validation loss followed the training loss closely and reached 0.40, near the floor imposed by label smoothing, with no sign of overfitting. The SWA model reached the same validation macro-F_1 (0.978) without exceeding the best checkpoint, which was kept as the final model (validation macro-F_1 0.9778).

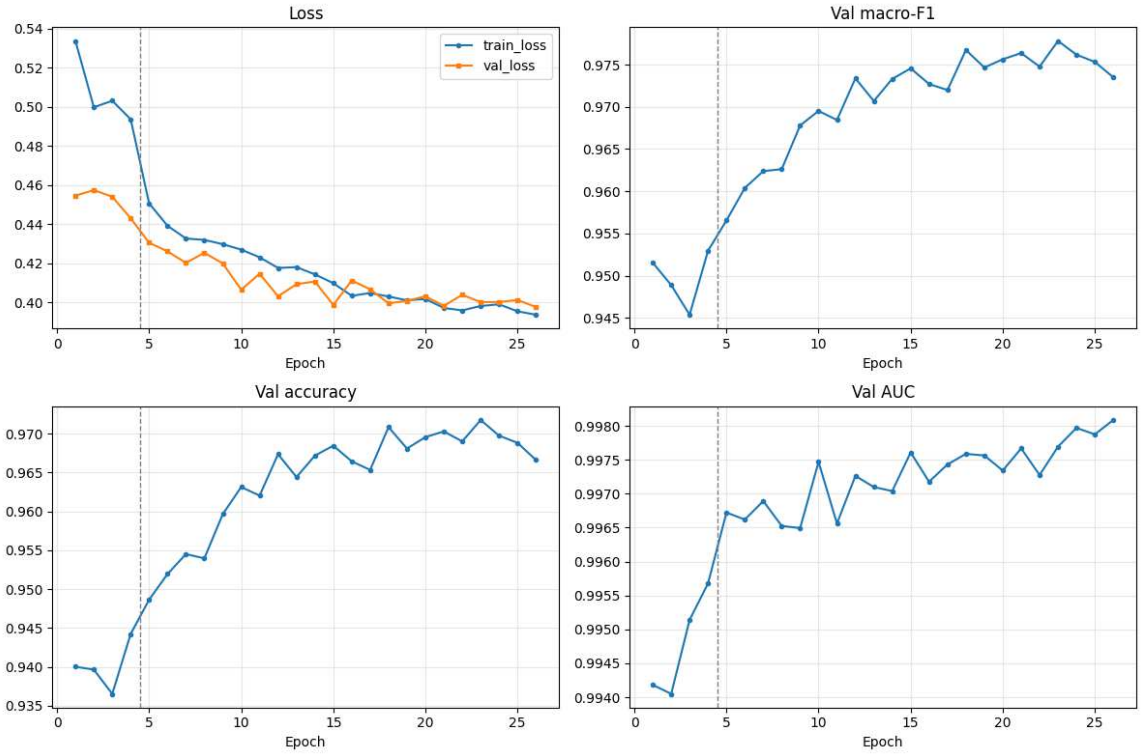


Fig. 3 Training curves of LungColonCancerPred-NN: training and validation loss, validation macro-F_1, accuracy and macro-AUC per epoch. The dashed line separates phase 1 (frozen backbone, epochs 1-4) from phase 2 (fine-tuning, with SWA from epoch 19 of the plot)

3.3 Test performance

Table 3 gives the results on the test set. The model classified 5 456 of the 5 497 images correctly: accuracy 99.25% (99.02-99.49), macro-F_1 99.38% (99.19-99.57), $\kappa = 0.9905$ and macro-AUC 0.9997. The test accuracy is higher than the validation accuracy without test-time augmentation (97.2% at the selected epoch); the two sets contain different families, and the three-view averaging of Eq. (7) is applied only at test time.

Table 3 Test performance of LungColonCancerPred-NN ($n = 5\,497$ images, 1 114 families, three-view TTA). Values in brackets are 95% bootstrap confidence intervals

	Accuracy (%)	Macro-F_1 (%)	Cohen's κ	Macro-AUC
LungColonCancerPred-NN	99.25 [99.02, 99.49]	99.38 [99.19, 99.57]	0.9905	0.9997

Table 4 gives the per-class results. Benign lung tissue was recognized without error. Benign colon tissue and colon adenocarcinoma have F_1 scores of 0.998. The two lung carcinomas are the hardest classes, with F_1 scores of 0.987 for adenocarcinoma and 0.986 for squamous cell carcinoma.

Table 4 Per-class test results of LungColonCancerPred-NN. Specificity is computed one class against the others from the confusion matrix

Class	Precision	Recall	Specificity	F_1	Support
Colon adenocarcinoma (colon_aca)	0.998	0.999	0.999	0.998	1 340
Benign colon (colon_n)	1.000	0.996	1.000	0.998	668
Lung adenocarcinoma (lung_aca)	0.977	0.998	0.992	0.987	1 404
Benign lung (lung_n)	1.000	1.000	1.000	1.000	727
Lung squamous cell carcinoma (lung_scc)	0.996	0.976	0.999	0.986	1 358
Macro average	0.994	0.994	0.998	0.994	5 497

3.4 Error analysis

The confusion matrix (Fig. 4) contains 41 errors, and they are not spread evenly. Thirty-three of them (80%) are squamous cell carcinomas predicted as adenocarcinomas; three adenocarcinomas went the other way. The other errors are three benign colon images predicted as colon adenocarcinoma and two colon adenocarcinomas predicted as lung squamous cell carcinoma.

Two coarser readings of the matrix matter in practice. At the level of the organ, only 2 of $5\,497$ images were assigned to the wrong organ (99.96% correct). At the level of malignancy, none of the $4\,102$ carcinomas was called benign, so the sensitivity for malignancy is 100%, and 3 of the $1\,395$ benign images were called malignant, a specificity of 99.78%. In a screening setting, this is the safer direction for errors to fall. The remaining confusion lies inside non-small-cell lung carcinoma, between two types that can share solid growth patterns when they are poorly differentiated and that pathologists often separate with immunohistochemistry.

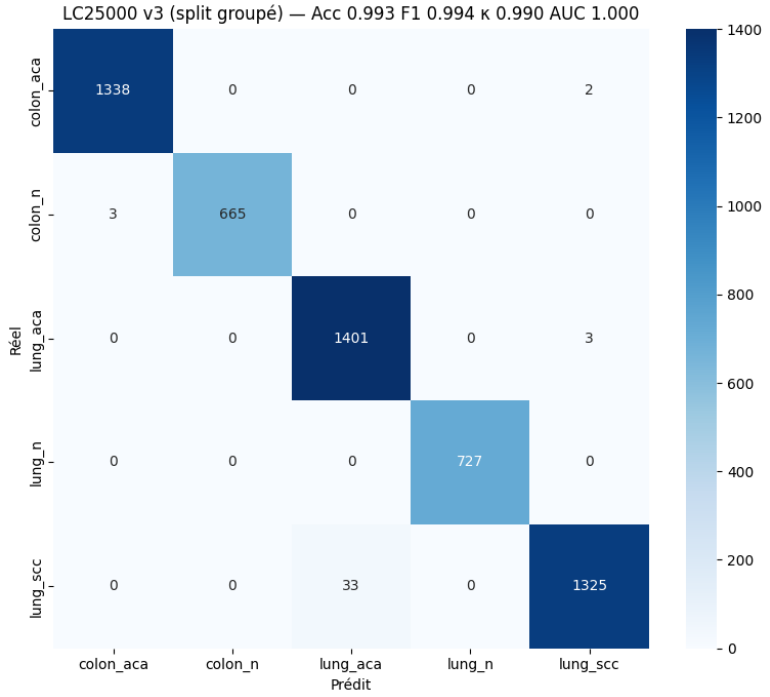


Fig. 4 Confusion matrix of LungColonCancerPred-NN on the test set ($n = 5\,497$, three-view TTA). Most errors are lung squamous cell carcinomas predicted as lung adenocarcinomas; no carcinoma is predicted as benign tissue

3.5 Qualitative predictions

Fig. 5 shows 20 test images drawn at random, with single-view predictions. All 20 are correct. Most confidences lie between 88% and 97%, around the label-smoothing optimum of 93.6%, so the model does not produce the saturated probabilities that often come with memorised training images. The lowest confidence (75%) is a lung adenocarcinoma with a solid arrangement of cells, the pattern behind most of the errors in Fig. 4. The sample also shows how much the stain varies, from pale pink colon mucosa to deep purple lung tumours. On a second random sample of 100 test images (Fig. 6), 99 predictions are correct.

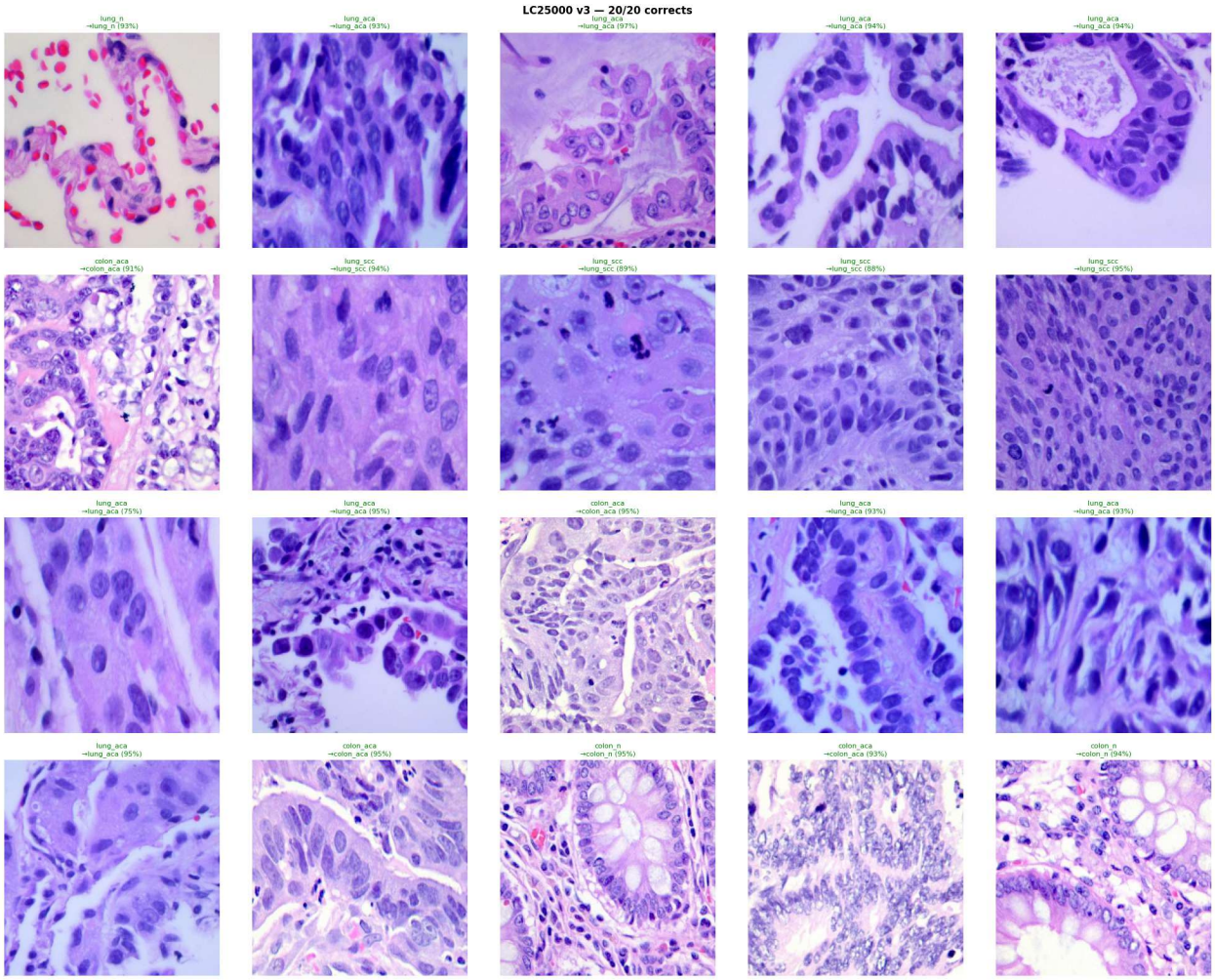


Fig. 5 Twenty randomly selected test images with ground truth (top) and single-view prediction with confidence (bottom). All twenty are correct

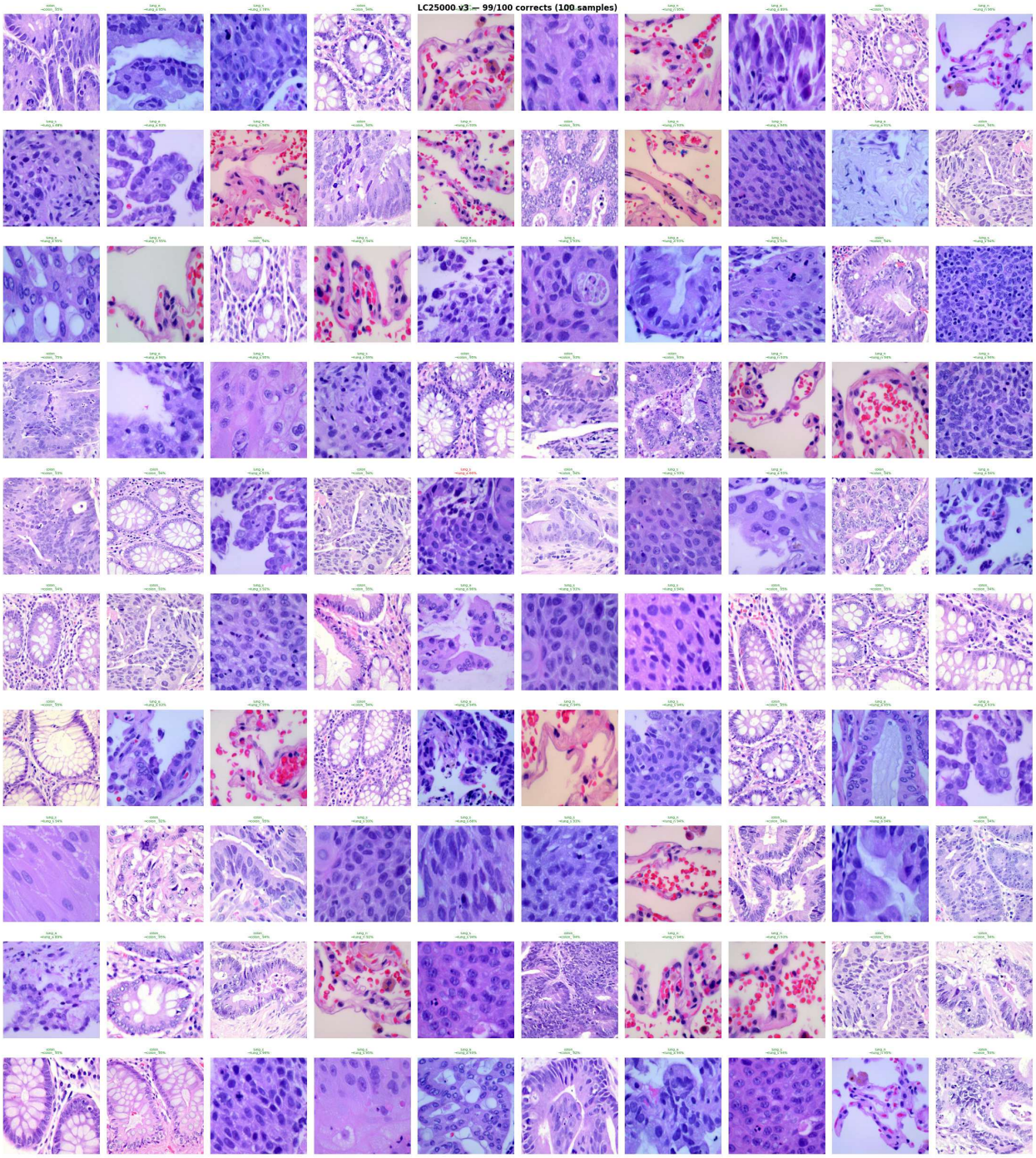


Fig. 6 One hundred randomly selected test images with single-view predictions; 99 of 100 are correct (green: correct, red: error)

3.6 Comparison with published approaches

Table 5 compares LungColonCancerPred-NN with four methods published in 2025 and 2026 and evaluated on LC25000. They cover the main families of models used on this benchmark: a fine-tuned MobileNet (Ochoa-Ornelas et al., 2025b), a Vision Transformer (Setiawan et al., 2025), a modified ResNet50 (Chaddad

et al., 2025) and a CNN-Transformer fusion network (Ramakrishnan et al., 2026).

Table 5 Comparison with methods published in 2025-2026 and evaluated on LC25000. Acc. and F_1 in %; "-" = not reported. Figures for other studies are those published, with the data and splits described in each paper; ours come from the family-grouped split of Section 2.2. ΔAcc is the difference between our accuracy and theirs, in percentage points

Study	Method	Data and task	Split	Acc.	F_1	ΔAcc
Setiawan et al. (2025)	Vision Transformer (ViT), CRISP-DM workflow	LC25000, lung, 3 classes (15 000 images)	random, 80/20	97.00	97.20	+2.25
Ochoa-Ornelas et al. (2025b)	MobileNet with transfer learning	LC25000 + NCI images, lung, 3 classes	–	97.65	–	+1.60
Chaddad et al. (2025)	ResNet50+, modified residual network	LC25000 subset (3 183 images), lung, 3 classes	random, 2 400/300/483	98.14	98.14	+1.11
Ramakrishnan et al. (2026)	CTFN, DenseNet-Transformer fusion with Grad-CAM and uncertainty estimation	LC25000, lung, 3 classes (15 000 images)	–	99.11	–	+0.14
LungColon-CancerPred-NN (ours)	EfficientNet-B0 + graph + Transformer, gated fusion, SWA, TTA	LC25000 + Cancer classification, MD5 de-duplicated, 5 classes (37 924 images)	grouped by family, 70/15/15	**99.25**	**99.38**	–

LungColonCancerPred-NN has the highest accuracy in the table, 0.14 to 2.25 percentage points above the other methods, and the highest F_1, 1.24 points above ResNet50+ and 2.18 points above the Vision Transformer. Three differences in the setting make this margin more meaningful than the raw numbers suggest.

The first is the task. All four methods classify the three lung classes, while our model separates five classes and must also tell the two organs apart. It does so with only two cross-organ errors in 5 497 images (Section 3.4).

The second is the size of the evaluation. ResNet50+ was tested on 483 images (Chaddad et al., 2025); our test set has 5 497 images from 1 114 augmentation families, and the bootstrap interval of our accuracy (99.02-99.49%) lies entirely above the accuracies of ResNet50+, MobileNet and the Vision Transformer.

The third is the split. A random split of LC25000 distributes about twenty augmented versions of each source image between training and test data (Borkowski et al., 2019; Diosdado et al., 2024), so part of the test set is close to images seen during training. Our test images have no near copy in the training set (Fig. 2). As far as their published descriptions show, the methods in Table 5 were tested on an easier task with a less strict split, and they still scored lower. Some image-level studies on LC25000 report accuracies up to 99.88% (AlShehri, 2025; Ochoa-Ornelas et al., 2025a; Jim et al., 2025); for the reason just given, these figures cannot be set against results obtained on a grouped split.

The architectures also differ in how they combine local and global information. The Vision Transformer works on image patches without a convolutional stem, MobileNet and ResNet50+ rely on convolution alone, and CTFN joins a DenseNet with Transformer layers. Our model adds a graph branch to the CNN-Transformer combination and lets a gate weight the three views for each image, with 5.8M parameters. Other recent studies on the same data combine handcrafted texture descriptors with an EfficientNet-B0 and Transformer fusion (Anusha and Reddy, 2025), apply super-resolution before DenseNet201 features and an SVM (Patharia et al., 2025), couple a SwinV2 Transformer with a ConvNeXtV2 network for the lung classes (Sunitha and Vellingiri, 2026), or join capsule, graph-attention and Transformer modules for the colon classes (Saremi and Ahmadkhan Kordbacheh, 2025). LC25000 is also used as a downstream benchmark for pathology foundation models, where the colon subset gave the most reproducible results (Afzaal et al., 2025).

4 Discussion

4.1 Evaluation on an augmented benchmark

LC25000 is convenient: it is balanced, freely available and large in appearance. Its effective size, however, is $1\,250$ source images (250 per class), from which the $25\,000$ files were generated by rotations and flips (Borkowski et al., 2019). Each source image therefore appears about twenty times. Combining LC25000 with other public collections can add further copies, because several collections on Kaggle are built from the same material. In our case the second collection contributed $15\,000$ files to three classes only (colon_aca, lung_aca and lung_scc), which is why these classes hold about $9\,500$ images each after merging, against about $4\,750$ for the two benign classes.

Exact copies were the first problem. MD5 hashing found 796 files of the second collection that were byte-identical to files already in the pool, and $1\,280$ exact duplicates inside LC25000 itself, probably cases where the augmentation produced the same output twice. Together they made up 5.2% of the $40\,000$ files. Hashing removes them at little cost, and we recommend it as a first step whenever public collections are merged, whatever the model.

Rotated and flipped copies need more than hashing, since two such copies differ in every byte. Averaging the embedding over the eight symmetries of the square (Eq. (1)) makes a copy rotated by a multiple of $90°$, or reflected, map to the same point as its source, up to the effects of JPEG compression and resizing. Files that were re-encoded in the second collection, and so escaped hashing, should also land next to their originals. The cosine threshold then decides which pairs are linked.

The choice of threshold is a trade-off between two failures. If τ is too low, links chain through visually similar tissue of the same class: at $\tau = 0.90$ the largest family holds $8\,566$ images, close to a quarter of the pool, and a split by such families would leave too few groups to stratify and would mix many unrelated images. If τ is too high, true copies stay in separate families and can end up on both sides of the split. Table 1 shows how the family structure changes between these extremes. From $\tau = 0.92$ to 0.98 the number of families grows from 719 to $7\,795$, and the largest family falls from $7\,066$ to 62 images. Our rule, the smallest threshold whose largest family has at most 80 images, selected $\tau = 0.98$, with $7\,795$ families of median size 2.

Because the rule fixes the families before any model is trained, the audit is the check that matters. The two distributions in Fig. 2 are almost disjoint: the 95th percentile of the test-to-train similarity (0.978) is lower than the 5th percentile of the train-to-train similarity (0.985). In other words, the closest training image of a typical test image is less similar to it than the closest neighbour of almost any training image is to that image. No test image reaches the threshold. The grouped split also kept the class balance: the test set holds 24.4% colon_aca, 12.2% colon_n, 25.5% lung_aca, 13.2% lung_n and 24.7% lung_scc images, against 25.1%, 12.6%, 24.9%, 12.5% and 24.9% in the de-duplicated pool.

The protocol also gives a way to read our own earlier result. With a random split, the same architecture reached 99.4% accuracy; with the grouped split it reaches 99.25%. The difference is small, which suggests that the model's accuracy does not rest mainly on recognising copies of training images. The audit costs one nearest-neighbor search over the embeddings and gives a direct, visual check of the split. We think a figure like Fig. 2, or at least the percentiles behind it, should accompany any result reported on this benchmark.

4.2 Role of the three branches

The three branches read the same 7×7 feature map, and each position of that map summarises a region of about 110×110 pixels of the original 768×768 image, enough to contain a few nuclei or part of a gland. The branches differ in how they combine these 49 regions.

Average pooling keeps the texture found anywhere in the image and discards its arrangement. This is

enough for the easiest distinctions. Benign lung tissue, with its thin alveolar walls, open air spaces and red blood cells, looks nothing like the dense carcinomas, and the model classified all 727 benign lung images correctly. Benign colonic mucosa, with regular crypts and goblet cells, is likewise set apart from adenocarcinoma, with three errors among 668 images.

The graph branch links each region to the eight regions most similar to it in feature space, wherever they lie, and passes information along these links in three rounds. A squamous cell carcinoma often shows scattered nests of keratinising cells separated by stroma; positions that fall on these nests are linked to one another and pooled into a shared representation, even when they are far apart. A spatial graph built on the grid would only connect adjacent regions, which is why we chose feature-space neighbors. The read-out adds the mean of the input features to the mean of the graph output, so the branch can fall back on plain pooling when the graph adds little.

The Transformer branch adds a learned position to each region and lets every region attend to all others through two layers of self-attention. It can thus relate a glandular lumen to the ring of epithelium around it, or a band of stroma to the tumour on either side, which are relations of position rather than of appearance. The class token collects these relations into one vector.

The gate of Eq. (5) computes a weight between 0 and 1 for each branch from the three embeddings of the image being classified. Since the weights are independent, the model is free to keep all three branches at full strength, or to turn one down for a given image, rather than splitting a fixed budget among them as a softmax would. The fused vector passes through a LayerNorm, so the scale of the gated embeddings does not drift during training.

The errors show where these views still meet their limit. Of the 41 errors, 36 fall between lung adenocarcinoma and squamous cell carcinoma, and they are not symmetric: 33 squamous cell carcinomas were called adenocarcinoma, and only 3 adenocarcinomas were called squamous cell carcinoma. When the model hesitates between the two, it leans toward adenocarcinoma. This is consistent with the lowest precision in Table 4 being that of lung_aca (0.977) and the lowest recall that of lung_scc (0.976). Poorly differentiated tumors of both types can grow as solid sheets of cells without glands or keratin, and at the field of view of LC25000 images the features that separate them may be absent; pathologists then often turn to immunohistochemistry. The other five errors involve the colon: three benign images called adenocarcinoma and two adenocarcinomas called lung squamous cell carcinoma.

4.3 Clinical reading of the results

A five-class accuracy treats all errors alike, but they do not carry the same consequences. Calling a carcinoma benign could delay treatment, and this never happened among the $4\,102$ carcinomas of the test set, so the sensitivity for malignancy is 100%. The reverse error, a benign image called malignant, happened three times among $1\,395$ benign images (specificity 99.78%); in practice it would lead to a second look by a pathologist rather than to harm. Organ assignment, which a pathologist always knows from the specimen, was wrong for only two images (99.96% correct), so the model's errors are almost all within an organ. The main remaining question, adenocarcinoma against squamous cell carcinoma in the lung, is the one where a pathologist would expect to need further tests.

4.4 Size, cost and calibration

With 5.8M parameters, the model is of the same order as the lightweight networks proposed for this task (Jim et al., 2025; Ochoa-Ornelas et al., 2025a). It trains on a single NVIDIA T4 GPU at about 4.5 minutes per epoch for $26\,976$ training images, a little over two hours for the 26 epochs of the two phases. The family construction adds one forward pass of EfficientNet-B0 over the pool for each of the eight symmetries, done

once before training. At inference, the three views of Eq. (7) triple the cost of a single forward pass, which remains small for a network of this size.

Label smoothing with $\varepsilon = 0.08$ sets the target confidence for the true class at 0.936 rather than 1.0, and weight averaging smooths the final weights. On the twenty images of Fig. 5, most confidences lie between 88% and 97%, and the lowest (75%) belongs to a solid lung adenocarcinoma, the pattern behind most errors. The probabilities therefore carry information about difficulty and can be used to rank cases, for example to send the least confident ones to a pathologist first. The test probabilities are saved by the pipeline, so calibration measures such as the expected calibration error can be computed directly from them.

5 Conclusion

LungColonCancerPred-NN combines convolutional, graph and Transformer views of one EfficientNet-B0 feature map through a sigmoid-gated fusion in a network of 5.8M parameters trained in two phases with stochastic weight averaging. We evaluated it on $37\,924$ lung and colon histopathology images from two public collections, after removing $2\,076$ exact duplicates by hashing and grouping the remaining images into $7\,795$ augmentation families with a rotation- and reflection-invariant embedding. The split was made by family, and a nearest-neighbor audit showed that no test image had a training neighbor at or above the grouping threshold, with test-to-training similarities well below those found inside the training set.

Under this protocol the model classified $5\,456$ of $5\,497$ test images correctly: 99.25% accuracy, 99.38% macro-F_1, $\kappa = 0.9905$ and a macro-AUC of 0.9997. No carcinoma was classified as benign, only two images were assigned to the wrong organ, and most 36 of the 41 errors fell between lung adenocarcinoma and squamous cell carcinoma. Its accuracy is above that of four methods published in 2025 and 2026 (97.00-99.11%), which were evaluated on the three lung classes only, while our model separates five classes.

The protocol itself (hashing, D_4invariant families, grouped split, and audit) does not depend on our architecture and can be applied to any model evaluated on LC25000 or on collections derived from it. The next step is to test the trained model, without retraining, on LungHist700 (Diosdado et al., 2024), which contains 691 lung images from 45 patients of a different hospital, acquired at $20\times$ and $40\times$.

Statements and Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Consent Not applicable; the study used previously published, de-identified public datasets.

Data availability LC25000 is available at `https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images` and the Cancer classification collection at `https://www.kaggle.com/datasets/maestroalert/cancer`. The lists of de-duplicated files, family assignments and split memberships will be released with the code

Code availability The implementation of LungColonCancerPred-NN and of the evaluation protocol will be available from the corresponding author upon reasonable request.

Author contributions Outhman Abbassi conceived and designed the study, implemented the models and the evaluation protocol, performed the experiments and analysis, and wrote the first draft of the manuscript.

Wajih Rhalem, Zeyneb Belganche and Nassim Kharmoum supervised the work and contributed to the methodology and the interpretation of the results. All authors read and approved the final manuscript.

References

Afzaal, U., Su, Z., Sajjad, U., Lu, H., Rezapour, M., Gurcan, M. N., & Niazi, M. K. K. (2025). *Hyperparameter optimization and reproducibility in deep learning model training* (arXiv:2510.15164). arXiv. https://doi.org/10.48550/arXiv.2510.15164

AlShehri, H. (2025). Dual-path convolutional neural network with squeeze-and-excitation attention for lung and colon histopathology classification. *Journal of Imaging, 11*(12), 448. https://doi.org/10.3390/jimaging11120448

Anusha, M., & Reddy, D. S. (2025). Fusion of classical and deep learning features with incremental learning for improved classification of lung and colon cancer. *Scientific Reports, 15*. https://doi.org/10.1038/s41598-025-24734-w

Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). *Lung and colon cancer histopathological image dataset (LC25000)* (arXiv:1912.12142). arXiv. https://doi.org/10.48550/arXiv.1912.12142

Chaddad, A., Peng, J., & Wu, Y. (2025). *Classification based deep learning models for lung cancer and disease using medical images* (arXiv:2507.01279). arXiv. https://doi.org/10.48550/arXiv.2507.01279

Diosdado, J., Gilabert, P., Seguí, S., & Borrego, H. (2024). LungHist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data, 11*, 1088. https://doi.org/10.1038/s41597-024-03944-3

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence* (pp. 876–885).

Jim, J. R., Rayed, M. E., Mridha, M. F., et al. (2025). XLLC-Net: A lightweight and explainable CNN for accurate lung cancer classification using histopathological images. *PLOS ONE, 20*(5), e0322488. https://doi.org/10.1371/journal.pone.0322488

Jiménez-Sánchez, A., Avlona, N. R., De Boer, S., et al. (2025). In the picture: Medical imaging datasets, artifacts, and their living review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 511–531). ACM. https://doi.org/10.1145/3715275.3732035

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mienye, I. D., & Viriri, S. (2025). Graph neural networks in medical imaging: Methods, applications and future directions. *Information, 16*(12), 1051. https://doi.org/10.3390/info16121051

Ochoa-Ornelas, R., Gudiño-Ochoa, A., Rosales-Aguayo, S. O., et al. (2025a). A lightweight cross-gated dual-branch attention network for colon and lung cancer diagnosis from histopathological images. *Medical Sciences, 13*(4), 286. https://doi.org/10.3390/medsci13040286

Ochoa-Ornelas, R., Gudiño-Ochoa, A., García-Rodríguez, J. A., & Uribe-Toscano, S. (2025b). Enhancing early lung cancer detection with MobileNet: A comprehensive transfer learning approach. *Franklin Open, 10*, 100222. https://doi.org/10.1016/j.fraope.2025.100222

Patharia, P., et al. (2025). Enhanced diagnosis of lung and colon cancer severity through deep feature analysis using DenseNet201 and SVM with histopathological images: A super-resolution approach. *Cancer Reports, 8*, e70439. https://doi.org/10.1002/cnr2.70439

Ramakrishnan, A. B., Vimal Cruz, M., & S, K. (2026). CTFN: A hybrid CNN-Transformer fusion network with explainability and uncertainty estimation for enhanced lung cancer classification. *Journal of Soft Computing and Data Mining, 7*(2), 330–348. https://doi.org/10.30880/jscdm.2026.07.02.018

Saremi, S., & Ahmadkhan Kordbacheh, A. (2025). *Multi-scale deep learning for colon histopathology: A hybrid graph-transformer approach* (arXiv:2509.02851). arXiv. https://doi.org/10.48550/arXiv.2509.

02851

Setiawan, W., Agustiono, W., & Pramudita, Y. D. (2025). Lung cancer classification using vision transformer: A CRISP-DM approach with histopathological imaging. *Communications in Mathematical Biology and Neuroscience, 2025*, Article 136. https://doi.org/10.28919/cmbn/9602

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians, 75*(1), 10–45. https://doi.org/10.3322/caac.21871

Sunitha, G., & Vellingiri, J. (2026). An explainable dual-branch transformer–ConvNeXt framework for robust lung histopathology classification. *Frontiers in Oncology, 16*, 1907746. https://doi.org/10.3389/fonc.2026.1907746

Sung, H., Bray, F., et al. (2026). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA: A Cancer Journal for Clinicians*, e70090. https://doi.org/10.3322/caac.70090

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).

Ye, J., Zeng, A., Pan, D., Chen, J., & Cheng, G. (2026). MedGNN: General medical image recognition network via GNN visual representations. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025* (LNCS 15975, pp. 332–342). Springer. https://doi.org/10.1007/978-3-032-05325-1_32

Young, V. M., Gates, S., Garcia, L. Y., & Salardini, A. (2025). Data leakage in deep learning for Alzheimer's disease diagnosis: A scoping review of methodological rigor and performance inflation. *Diagnostics, 15*(18), 2348. https://doi.org/10.3390/diagnostics15182348