

Article

Explainability and Feature Fusion for Colon CT Image Classification

Elham Tahsin Yasin ^{1, *} ¹ Computer Engineering Department, Architecture and Engineering Faculty, Sinop University, Sinop 57000, Türkiye

* Correspondence: iyasin@sinop.edu.tr

Abstract

Accurate classification of colon computed tomography (CT) images using deep learning has shown promising results. However, the contribution of handcrafted features, reliability of explainability methods, and potential influence of data leakage remain important concerns. This study evaluates six XGBoost-based classification strategies using 11,340 CT images from the ColonCancerCT-2025 dataset comprising 40 patients. The dataset was divided at the image level into training (7,937), validation (1,701), and testing (1,702) sets without verified patient separation. Deep features extracted using EfficientNet-B0 were compared with whole-image and Grad-CAM++ region-of-interest (ROI)-associated handcrafted features and their combinations. Four explainability methods were evaluated using deletion and insertion metrics. Additionally, duplicate detection, nearest-neighbor analysis, and body masking were performed. Experimental results showed that EfficientNet-B0 achieved 99.82% accuracy, while feature fusion provided no improvement. ROI-associated handcrafted features achieved 85.61% accuracy compared with 79.44% for whole-image features. LIME and Integrated Gradients produced the lowest, overlapping deletion AUCs. Duplicate analysis identified 134 test images (7.9%) with identical copies, while nearest-neighbor classification achieved 99.94% accuracy. Body masking reduced accuracy to 95.06%, whereas background-only classification achieved 0.74 ROC-AUC. These findings highlight the importance of patient-independent evaluation and leakage assessment before interpreting high classification performance as diagnostic reliability.

Keywords: computed tomography; colon cancer; explainable artificial intelligence; EfficientNet-B0; feature fusion; handcrafted features.

Citation: Yasin, E. T. (2026). Explainability and Feature Fusion for Colon CT Image Classification. *Impact in Computics*, 2, 7. <https://doi.org/10.65500/computics-2026-007>

Received: 2 September 2026 | Revised: 9 October 2026 | Accepted: 12 October 2026 | Published: 15 October 2026

Copyright: © 2026 by the authors. Licensee Impaxon, Malaysia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Computed Tomography (CT) provides anatomical and image-intensity information that can be used to aid in the computational analysis of colorectal disease. Deep learning and radiomics offer complementary approaches to represent this information, whereby convolutional neural networks learn representations directly from the image whereas handcrafted descriptors quantify predefined properties such as intensity distributions, texture and shape. The combination of these representations provides an opportunity to investigate

whether explicitly defined image characteristics provide information beyond that captured by a deep learning model. Explainable Artificial Intelligence (XAI) adds another perspective by identifying image regions that influence model predictions.

Radiomics has been investigated for clinically defined prediction tasks in colorectal cancer. Huang et al. developed a CT-based nomogram for preoperative prediction of lymph-node metastasis using a primary cohort of 326 patients and an independent cohort of 200 patients. Their approach combined a radiomics signature

with CT-reported nodal status and clinical information [1]. Such work demonstrates the potential of quantitative image descriptors within a defined clinical modelling framework and motivates further investigation of their relationship with learned image representations.

Among deep learning architectures, EfficientNet employs coordinated scaling of network depth, width and input resolution [2]. Its learned features can support classification through either the network's prediction head or a separate machine learning classifier. This flexibility enables comparisons between deep representations, handcrafted features and their combinations within a common classification framework. A central question is whether feature fusion improves predictive performance or introduces descriptors that largely overlap with information already captured by the network.

In addition to predictive performance, an important aspect for evaluating the performance of a classifier is understanding which image features affect its functioning. Various algorithms can be used to explain the classification process, each providing a unique perspective on the relationship between input data and classification results. Grad-CAM++ algorithm uses class-specific output probabilities from convolutional layers to generate class activation maps [3]. Integrated Gradients algorithm, on the other hand, attributes the prediction result to the input features by calculating the integral of the gradient from a baseline input to the calculated one [4]. GradientSHAP works similarly to Integrated Gradients, but the gradients are averaged over several randomly selected baseline inputs [5,6]. Finally, the LIME algorithm approximates the model's behavior near the input data with an interpretable linear model [7]. These four methods were selected because they represent distinct attribution paradigms: gradient-weighted class activation mapping (Grad-CAM++), axiomatic path-based attribution (Integrated Gradients), baseline-averaged gradient attribution (GradientSHAP) and local surrogate modelling (LIME). Comparing these methods can reveal differences in the regions and features emphasized by the same classifier. Deletion and insertion analyses assess the effect of perturbing highly attributed regions, while parameter-randomization tests examine the dependence of explanations on learned model parameters [8,9].

Explanation maps can also guide the selection of regions for quantitative feature extraction, connecting model interpretation with feature-based classification. This approach allows image descriptors associated with model-highlighted regions to be compared with descriptors

calculated over the whole image. Evaluating these representations alongside deep features can help clarify whether explanation-guided regional information provides additional predictive value and how its contribution relates to the behavior of the underlying classifier.

Accordingly, this study investigates an integrated framework combining EfficientNet-B0, multiple explanation methods and handcrafted image features for exploratory colon CT image classification using ColonCancerCT-2025 [10]. Throughout the study, the classification target is the class label assigned by the dataset to each image. Because individual images were not confirmed to contain visible tumor tissue and no lesion annotations were available, the task is image-level label classification rather than tumor detection or diagnosis. The objectives are to compare six feature strategies using XGBoost, evaluate Grad-CAM++, Integrated Gradients, GradientSHAP and LIME through perturbation-based analyses, and examine the image regions associated with model predictions. By considering classification performance, feature fusion and explanation behavior together, the study seeks to characterize the contribution of each component within a unified image-analysis workflow and to quantify the data-related threats to validity through duplicate, nearest-neighbor and body-masking analyses.

2. Related Work

Deep learning has been investigated for colorectal cancer detection and classification across CT and histopathological imaging. In CT, Ziegelmayer et al. [11] developed a three-dimensional convolutional neural network to distinguish colon carcinoma from acute diverticulitis and evaluated its use in supporting radiologist interpretation. Yao et al. [12] investigated colorectal cancer detection on contrast-enhanced CT without bowel preparation in a retrospective multicentre study. Their work extended evaluation beyond classification performance to the potential contribution of model assistance to radiologist detection. These studies provide clinically oriented examples of CT-based analysis, although their detection and differential-diagnosis tasks differ from classification of individual exported images.

Feature fusion has also been explored as a means of combining learned representations with explicitly defined image characteristics. In lung and colon histopathology, Attallah [13] combined features from lightweight convolutional networks with statistical and texture descriptors, followed by feature selection. Alsubai et al.

[14] integrated local binary pattern features with an Inception-ResNetV2-based approach and used SHAP for interpretation. In colorectal cancer imaging, Zhang et al. [15] developed a radiomics-boosted deep learning model using PET/CT for assessment of synchronous peritoneal metastasis. Beyond imaging, joint representation learning has been used to integrate heterogeneous data modalities; for example, scJVAE learns a shared latent space from multimodal single-cell data with a joint variational autoencoder [16]. These approaches motivate comparisons between individual and combined feature sets, while leaving the contribution of fusion to be established for each dataset and prediction task. To the author's knowledge, no prior work has compared deep, whole-image handcrafted and explanation-guided ROI handcrafted features within a single colon CT classification framework.

Explainability studies complement these developments by investigating the image regions associated with model predictions. Di Giammarco et al. [17] compare deep learning models for the task of colon histology classification using both predictive performance as well as visual explanations. As a result, they observe that high classification accuracy does not necessarily reflect informative explanation maps, which highlights the need to evaluate both model properties separately. This distinction is particularly relevant when attribution maps are subsequently used to guide regional feature extraction.

Together, these studies motivate an evaluation that connects feature representation, classification and model explanation. The present study brings these components together by comparing deep, whole-image handcrafted and Grad-CAM++ ROI-associated feature strategies within a common CT image-classification workflow. It also compares four explanation methods using deletion and insertion analyses, allowing the predictive contribution of feature fusion and the behavior of model explanations to be examined within the same experimental setting.

3. Materials and Methods

3.1 Study design and data source

This study conducted a secondary computational analysis of publicly available PNG CT images from the ColonCancerCT-2025: Abdominal CT Scans dataset on Kaggle [10]. The dataset description reports 40 patients, comprising 12 malignant and 28 non-malignant cases, and describes PNG images derived from DICOM scans. The metadata table distributed with the dataset lists 41 entries (12 malignant and 29 benign) and does not link patients to image files. The dataset licence was reported as CC BY 4.0.

The analyzed collection comprised 11,340 images assigned to cancer and non-cancer classes according to their enclosing classes. Only PNG images and these dataset-assigned labels were used for model training and evaluation. The labels represented the classification targets without confirming visible tumor tissue in each image; the terms cancer-labelled and non-cancer-labelled are therefore used to denote dataset labels, not confirmed tumor presence or absence. The distributed files did not include patient identifiers or a patient-to-image mapping, so patient independence across the training, validation and test partitions could not be verified and a patient-level split could not be constructed retrospectively. No expert lesion annotations or radiologist assessments of the explanation maps were available. The overall study workflow is presented in Figure 1.

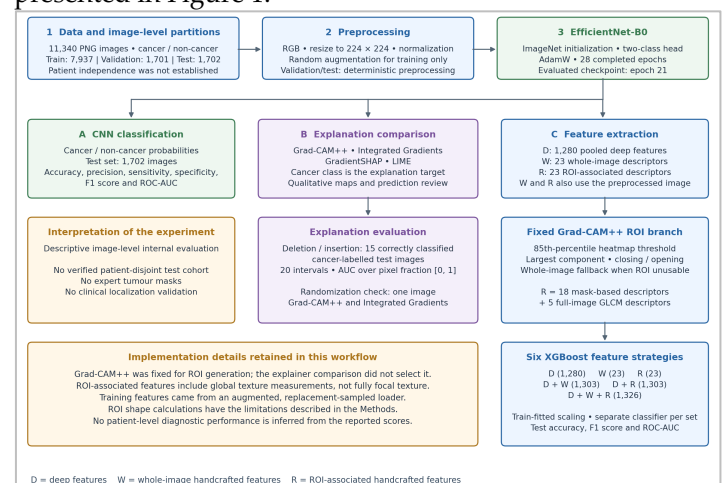


Figure 1. Overview of the classification, explanation and feature-fusion workflow. Grad-CAM++ was fixed for ROI extraction independently of the four-method explanation comparison. ROI-associated features comprised 18 mask-based descriptors and five whole-image GLCM texture descriptors.

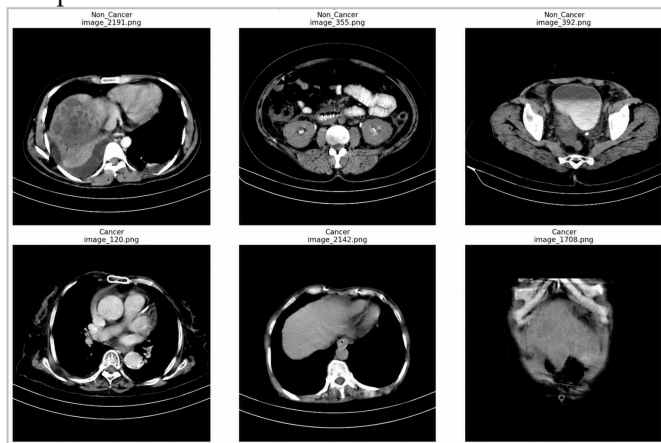
3.2. Image partitions and sample units

The 11,340 images were divided at the image level into training, validation and test sets containing 7,937, 1,701 and 1,702 images, respectively, corresponding approximately to a 70:15:15 split. The validation set was used for checkpoint selection and early stopping, while the test set was used for final classification evaluation. Table 1 summarizes the distribution of class-assigned labels across the three partitions. The partition was fixed by the folder structure used for training, and the file lists of each partition are provided with the code (see Code availability).

Table 1. Image counts by dataset-assigned class label and partition.

Partition	Cancer	Non-cancer	Total
Training	3,176	4,761	7,937
Validation	681	1,020	1,701
Test	681	1,021	1,702
Total	4,538	6,802	11,340

Each PNG image constituted one sample for classification evaluation. Random augmentation was applied during training after partitioning; the counts in Table 1 refer to the original image files. Performance was evaluated at the image level, with patient independence unverified as described in Section 3.1. Figure 2 illustrates examples from both dataset classes.

**Figure 2.** Sample CT images from the dataset.

3.3. Preprocessing and augmentation

Images were loaded as RGB arrays and resized to 224×224 pixels. Channel-wise normalization used the ImageNet means (0.485, 0.456 and 0.406) and standard deviations (0.229, 0.224 and 0.225). Validation and test preprocessing comprised resizing, normalization and tensor conversion, without random augmentation. Analysis was performed on two-dimensional, resized 8-bit PNG images without Hounsfield-unit calibration, voxel-spacing resampling or three-dimensional reconstruction. Pixel intensities therefore reflect the display transformation applied when the images were exported rather than calibrated tissue attenuation.

The training augmentation pipeline was configured to include horizontal flips, vertical flips and random rotations in multiples of 90° , each with a probability of 0.5. Additional geometric augmentation used ShiftScaleRotate with translation limits of $\pm 10\%$, scale variation of $\pm 15\%$ and rotation limits of $\pm 15^\circ$, applied with a probability of 0.5. An intensity-adjustment group selected one of contrast-limited adaptive histogram equalization (CLAHE), brightness/contrast adjustment or gamma adjustment with

an overall probability of 0.5. A second group selected Gaussian noise or Gaussian blur with an overall probability of 0.3. Coarse dropout was configured with a probability of 0.3. The augmentation settings are summarized in Supplementary Table S1. Vertical flips and 90° rotations produce orientations that do not occur in standard abdominal CT display and can therefore create anatomically implausible images. They were included as generic orientation regularization because the dataset contained heterogeneous image planes without orientation metadata. Their effect was not assessed by ablation, and they may have encouraged reliance on orientation-invariant image properties rather than on anatomical layout.

3.4. Network and training

An ImageNet-pretrained EfficientNet-B0 backbone was implemented using the timm library [18]. Its original classification layer was replaced with an identity layer to obtain a 1,280-dimensional pooled feature vector. A custom classification head comprised a dropout layer with a probability of 0.3, a fully connected layer with 256 units, a ReLU activation, a second dropout layer with a probability of 0.2 and a two-unit output layer. Softmax was used to convert the output logits into class probabilities. All backbone and classification-head parameters were included in optimization.

The network was trained using AdamW with an initial learning rate of 1×10^{-4} , weight decay of 1×10^{-4} and a batch size of 32. Training was configured for a maximum of 30 epochs, with cosine annealing over a 30-epoch period and a minimum learning rate of 1×10^{-6} . Class imbalance was addressed through inverse-frequency weighted sampling with replacement and class-weighted cross-entropy loss. The two mechanisms serve a similar purpose: the sampler alone approximately balanced the classes (one pass of the weighted training loader drew 4,049 cancer-labelled and 3,888 non-cancer-labelled samples), so the additional loss weighting shifted the effective emphasis further towards the cancer-labelled class. This combination was not ablated. Its expected consequence would be a bias towards sensitivity at the expense of specificity, although the test errors were nearly symmetric (one false positive and two false negatives). Early stopping monitored validation loss with a patience of seven epochs, while checkpoint selection was based on the highest validation accuracy. Training stopped after 28 epochs, and the checkpoint from epoch 21 was retained for evaluation. A random seed of 42 was used for the single training run. The same random seed was used for XGBoost (random_state =

42). LIME sampling and GradientSHAP baseline selection were not seeded separately; they drew from the globally seeded generators, so their exact samples depend on execution order. Table 2 summarizes the principal model and analysis settings.

Table 2. Principal model and analysis settings.

Component	Setting
Input	224 × 224 pixels; three RGB channels
Backbone	EfficientNet-B0 pretrained on ImageNet
Classification head	1,280 → 256 → 2 units; ReLU; dropout probabilities of 0.3 and 0.2
Optimizer	AdamW; initial learning rate 1×10^{-4} ; weight decay 1×10^{-4}
Learning-rate schedule	Cosine annealing; 30-epoch period; minimum learning rate 1×10^{-6}
Training	Batch size 32; maximum 30 epochs; 28 completed
Class balancing	Inverse-frequency weighted sampling with replacement and class-weighted cross-entropy
Early stopping	Validation loss; patience of seven epochs
Selected checkpoint	Epoch 21; validation accuracy 99.8824%
Grad-CAM++ target	Backbone conv_head layer; cancer class
ROI construction	85th-percentile threshold; largest connected component; 5 × 5 morphological closing followed by opening
Faithfulness evaluation	15 correctly classified cancer-labelled test images; 20 perturbation intervals
XGBoost	100 estimators; maximum depth 5; learning rate 0.1; random state 42

3.5. Hardware and software

Experiments were conducted on a Windows 11 laptop equipped with an Intel Core i7-12650H processor, an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of video memory and 64 GB of RAM. The implementation used Python 3.9.21 and Jupyter Notebook 7.3.2. Deep learning components were developed using PyTorch 1.12.1+cu116, torchvision 0.13.1+cu116 and timm 1.0.14, with CUDA 11.6 and cuDNN 8.3.2.

Image processing and augmentation used Albumentations 2.0.5, OpenCV 4.11.0.86 and Pillow 10.4.0. Numerical and feature analyses used NumPy 1.26.4, SciPy 1.13.1, pandas 2.2.3, scikit-learn 1.6.1 and scikit-image 0.24.0. Feature-based classification used XGBoost 2.1.4, while explanation methods were implemented using grad-cam 1.5.5, Captum 0.8.0 and LIME 0.2.0.1. Figures were generated using Matplotlib 3.9.4 and seaborn 0.13.2.

Feature extraction and the duplicate, nearest-neighbour and masking analyses (Sections 3.8, 3.9 and 3.11) applied the trained epoch-21 checkpoint on a CPU using Python 3.13.16, PyTorch 2.14.1, timm 1.0.14, grad-cam 1.5.5, XGBoost 2.1.4, scikit-learn 1.9.1, scikit-image 0.26.0 and

OpenCV 5.0.0. Test-set predictions in this environment were identical to those of the training environment (the same three errors; ROC-AUC 0.999997).

3.6. Explanation generation

Four methods were used to explain EfficientNet-B0 predictions: Grad-CAM++, Integrated Gradients, GradientSHAP and LIME [3–7]. Grad-CAM++ targeted the backbone's conv_head layer. Integrated Gradients used a zero-valued baseline in normalized input space, corresponding to the normalization means in the original RGB space. GradientSHAP was implemented using Captum (version 0.8.0) [6] with 30 background images obtained from the training loader. For Integrated Gradients and GradientSHAP, absolute attributions were summed across channels to produce spatial maps. Attribution maps were subsequently min–max normalized; the resulting maps represented attribution magnitude rather than signed contributions.

LIME operated on denormalized 8-bit RGB images, with normalization reapplied within the prediction wrapper. Positive local explanation weights were assigned to their corresponding superpixels to construct the heatmap. Qualitative visualizations used 500 perturbed samples, whereas quantitative faithfulness evaluation used 300. Integrated Gradients and GradientSHAP used 30 integration steps and 15 samples, respectively, during faithfulness evaluation. Qualitative comparisons used their respective default settings unless otherwise specified. Computational budgets therefore differed between methods: per image, LIME used 300 model evaluations, Integrated Gradients 30 gradient evaluations, GradientSHAP 15 and Grad-CAM++ a single forward and backward pass. Because LIME received the largest budget, this asymmetry could favour LIME in the perturbation-based comparison, and it is considered when interpreting Section 4.4.

The cancer class was the explanation target for all images, including those labelled or predicted as non-cancer. Confidence values displayed alongside the images referred to the predicted class.

3.7. Quantitative explanation evaluation

Explanation faithfulness was evaluated using deletion and insertion analyses [8]. The evaluation subset was formed from the first 15 correctly classified cancer-labelled images encountered in the test loader. This convenience subset was not randomly sampled and excluded non-cancer-labelled and misclassified images. In the explanation maps, pixel locations were ranked by

decreasing attribution value. For deletion, increasing fractions of the pixels were progressively replaced with the mean of the normalized input tensor. For insertion, the process started from a blurred version of the input image (51×51 Gaussian kernel with standard deviation 20), and the original pixels were gradually restored. Cancer-class softmax probabilities were recorded at 21 points spanning 20 equal intervals from zero to complete replacement or restoration.

Areas under the deletion and insertion curves were calculated using trapezoidal integration over the perturbed pixel fraction from 0 to 1. The saved summaries used division by 21 rather than 20; their means and standard deviations were therefore rescaled by 21/20 to obtain values on this fraction axis. Lower deletion AUC and higher insertion AUC indicated greater faithfulness under the specified perturbation protocol. Results were summarized as the mean, sample standard deviation and 95% confidence interval (CI) of the mean, based on the t distribution with 14 degrees of freedom, across the 15 evaluated images.

A supplementary parameter-randomization check was performed for Grad-CAM++ and Integrated Gradients using one cancer-labelled test image [9]. Explanation maps from the trained network were compared with those obtained after random initialization of convolutional and linear parameters. Similarity was quantified using Pearson correlation between the flattened maps. Because it involved one image and two methods, this check is reported as an illustrative sanity check rather than as an assessment of explanation robustness.

3.8. ROI construction and handcrafted descriptors

Grad-CAM++ was used as the fixed method for ROI construction, independently of the comparative explanation evaluation. Heatmap values at or above the 85th percentile were converted into a binary mask. The largest connected component was retained if its area was at least 100 pixels, followed by morphological closing and opening with a 5×5 kernel. A whole-image mask was used when a usable ROI could not be obtained. The final mask area depended on the heatmap distribution, connected-component selection and morphological processing. No anatomical mask or expert annotation was used to constrain ROI generation. The 85th-percentile threshold retains the most strongly attributed 15% of pixels, and the minimum component size and kernel were chosen to suppress isolated responses at 224×224 resolution. These parameters were set as fixed design choices, and no sensitivity analysis over alternative thresholds or

morphological settings was performed; the ROI-associated results are therefore specific to this configuration.

Denormalized images were converted to 8-bit grayscale before handcrafted feature extraction. Each feature set contained 23 descriptors: 15 intensity and statistical descriptors, three shape descriptors and five gray-level co-occurrence matrix (GLCM) descriptors. Whole-image features were obtained using a full-image mask. For the ROI-associated set, intensity descriptors were calculated from masked pixels and shape descriptors from the generated mask, while GLCM descriptors were calculated over the full grayscale image. Accordingly, this set combined 18 mask-based descriptors with five whole-image texture descriptors. Because the five GLCM descriptors are identical in the two handcrafted sets, the comparison between them is a partial ablation in which texture is held constant and only the intensity and shape descriptors differ; it does not evaluate ROI-restricted texture.

GLCMs were constructed using a one-pixel distance, four angles (0° , 45° , 90° and 135°), 256 intensity levels, symmetric construction and normalization. Contrast, dissimilarity, homogeneity, energy and correlation were averaged across the four angles. Grayscale values were used directly without bin-width discretization. Intensity entropy was calculated from a 64-bin histogram with an automatically determined range, and intensity energy was defined as the mean squared grayscale value. These were custom image descriptors; the complete feature inventory is provided in Supplementary Table S2.

Shape descriptors were calculated from a single connected region of the final mask. Area was defined as the pixel count of the largest 8-connected component, perimeter as the arc length of the external contour of the same component, and compactness as $4\pi \times \text{area} / \text{perimeter}^2$. This approach ensured that area and perimeter were derived from the same connected region, including cases where morphological opening divided the mask into multiple components. Such cases occurred in 77 of the 9,639 training and test images. Whole-image masks yielded constant shape descriptors.

3.9. Feature extraction and XGBoost comparisons

Deep and handcrafted features were extracted deterministically with EfficientNet-B0 in evaluation mode, using the validation and test preprocessing (resizing and normalization only) for all images. Each original training image contributed exactly one row (7,937 rows: 3,176 cancer-labelled and 4,761 non-cancer-labelled), and each test image contributed one row (1,702 rows), without

augmentation or replacement sampling. The source file name was stored with every row. Deep, whole-image and ROI-associated features were extracted together from each image, so all six strategies used the same aligned rows.

Six XGBoost strategies were evaluated: deep features alone; whole-image handcrafted features alone; ROI-associated handcrafted features alone; deep plus whole-image features; deep plus ROI-associated features; and all three feature sets combined. Their respective input dimensions were 1,280, 23, 23, 1,303, 1,303 and 1,326. Missing feature values were replaced with zero before standardization. For each strategy, a StandardScaler was fitted to the training matrix and then applied to the test matrix. Scaling is not required by tree-based models; it was retained only to keep one preprocessing pipeline across all strategies. Because tree splits depend on the ordering of values within each feature, a per-feature affine transformation does not change the partitions that XGBoost can learn, and scaling is not expected to affect the results.

All six strategies used the same XGBoost configuration: 100 estimators, a maximum tree depth of five, a learning rate of 0.1, a random state of 42 and log-loss as the evaluation metric [19]. Hyperparameters were fixed, and all six strategies were reported without validation-based selection of a preferred feature combination.

3.10. Outcomes and analysis scope

Cancer was designated as the positive class. CNN performance was evaluated using accuracy, precision, sensitivity, specificity, F1 score and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Feature-based classification strategies were compared using accuracy, F1 score and ROC-AUC. Explanation evaluation used deletion and insertion AUCs, supplemented by the parameter-randomization analysis.

Results were summarized at the image level for a single data split and training run. 95% CIs for accuracy, precision, sensitivity and specificity were calculated using the Wilson score method [20], and 95% CIs for ROC-AUC using the Hanley–McNeil standard error [21]. Paired classifier comparisons on the common test set used the exact McNemar test [22], computed from the observed discordant predictions of the two models. All intervals treat test images as independent. Because multiple images may originate from the same patient and patient identity was unknown, these intervals reflect test-sample variation only, are likely to be too narrow and do not capture variation due to training randomness or data partitioning.

CLAIM 2024 was used as a reporting reference for the deep learning components [23].

3.11. Duplicate, nearest-neighbor and masking analyses

Exact duplicates were identified by the MD5 hash of the decoded pixel values of every original 512×512 PNG file; file-level hashes gave identical groups. For similarity analysis, every test image was compared with all 9,638 training and validation images. Images were downsampled to 64×64 pixels, the most similar training or validation image was selected by Pearson correlation, and its structural similarity index (SSIM) [24] with the test image was computed at 224×224 pixels. A raw-pixel 1-nearest-neighbour classifier, which assigns each test image the label of its most correlated training or validation image without any learning, served as a reference for how much of the task could be solved by image lookup. CNN performance was recomputed on test subsets that excluded images with an identical copy in training or validation, and images whose nearest neighbor had an SSIM of at least 0.95, 0.90 or 0.80.

To test the contribution of non-anatomical regions, a body mask was generated automatically for each test image. Pixels brighter than air (grey value above 20 after 5×5 Gaussian smoothing) were selected, and a 7×7 elliptical opening detached the scanner table, cushions and burned-in annotations. Connected components of at least 300 pixels and at least 5% of the largest component were retained, and their convex hull, dilated by two pixels, defined the body region, which covered a median of 38% of the image (interquartile range 32–45%). The trained network was re-evaluated without retraining under two conditions: with all pixels outside the body set to zero (body only) and with all pixels inside the body set to zero (background only). Because masked images differ from the training images, a loss of accuracy with body-only input reflects both reliance on non-anatomical cues and distribution shift, whereas above-chance discrimination with background-only input directly shows that the background carries label information.

4. Results

4.1. Image and feature accounting

The original dataset classes contained 4,538 cancer-labelled and 6,802 non-cancer-labelled images. The deterministic exports contained one row per image (7,937 training and 1,702 test rows), each with its file name, 1,326 feature columns and the label. They contained no missing or infinite values, the five GLCM columns were identical in

the two handcrafted sets, and whole-image area, perimeter and compactness were constant.

Exact duplicates were frequent. Of the 11,340 images, 1,202 belonged to 595 groups of pixel-identical images (583 pairs and 12 triplets), leaving 10,733 distinct images. Of these groups, 586 were non-cancer-labelled and nine were cancer-labelled, and no group carried conflicting labels; 272 groups spanned more than one partition. In the test partition, 134 images (7.9%; 133 non-cancer-labelled and one cancer-labelled) had an identical copy in the training or validation partition, 111 of them in training. A further 12 test images duplicated another test image. Supplementary Table S3 gives the distribution of duplicate groups across partitions.

4.2. CNN training and image classification

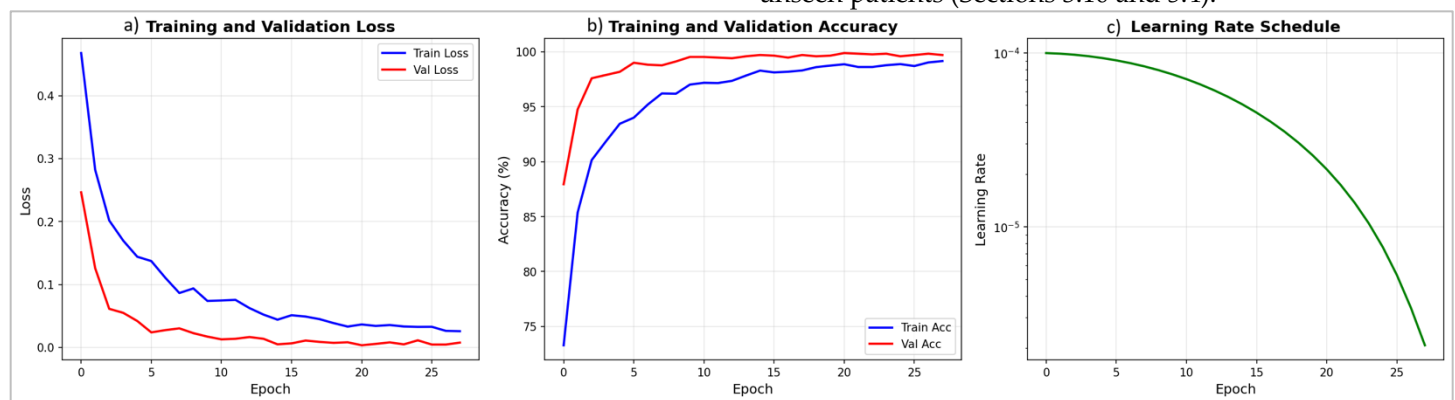


Figure 3. EfficientNet-B0 training history over 28 epochs: (a) training and validation loss, (b) training and validation accuracy, and (c) learning-rate schedule.

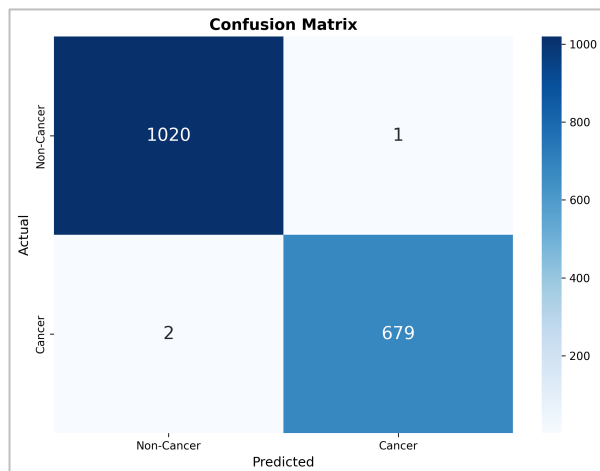


Figure 4. The confusion matrix corresponds to the epoch-21 checkpoint and 1,702 test images.

4.3. Feature representation comparisons

Table 3 compares the CNN classification head with the six XGBoost feature strategies trained on the deterministic features. ROI-associated handcrafted

Training stopped after 28 epochs, and the checkpoint from epoch 21 was selected based on its validation accuracy of 99.88%. Figure 3 presents the training and validation loss, accuracy trajectories and learning-rate schedule.

On the 1,702 test images, EfficientNet-B0 correctly classified 1,020 non-cancer-labelled images and 679 cancer-labelled images, yielding one false positive and two false negatives. Test accuracy was 99.82% (95% CI 99.48–99.94), precision 99.85% (99.17–99.97), sensitivity 99.71% (98.94–99.92), specificity 99.90% (99.45–99.98) and F1 score 99.78%. The ROC-AUC exceeded 0.9999 (95% CI 0.9999–1.0000). The corresponding confusion matrix is presented in Figure 4. These intervals treat images as independent and therefore understate the uncertainty that would apply to unseen patients (Sections 3.10 and 5.4).

features achieved an accuracy of 85.61% (95% CI 83.86–87.19), compared with 79.44% (77.45–81.29) for whole-image handcrafted features, a difference of 6.17 percentage points (105 images; exact McNemar $p < 10^{-8}$, 203 images correct only with ROI-associated features and 98 only with whole-image features). The corresponding ROC-AUC was 0.934 (0.921–0.948) versus 0.872 (0.854–0.891). Sensitivity was 76.21% versus 61.38%, and specificity 91.87% versus 91.48%. Because both sets used the same whole-image GLCM descriptors and whole-image shape descriptors were constant, this comparison holds texture fixed: the difference is attributable to the ROI-restricted intensity descriptors and/or the three ROI shape descriptors, whose separate contributions were not isolated.

Deep features alone achieved 99.82% accuracy (95% CI 99.48–99.94) and an F1 score of 99.78%, with two false negatives and one false positive. Adding whole-image features, ROI-associated features or both did not change the confusion-matrix counts, so fusion did not improve classification.

XGBoost using deep features misclassified the same three test images as the CNN classification head, with no discordant predictions. The deep representation therefore

supported the same near-ceiling agreement with dataset labels whether classified by the CNN head or by XGBoost.

Table 3. Test performance of the CNN classification head and six XGBoost feature strategies.

Model or feature set	Features	Accuracy % (95% CI)	F1 %	ROC-AUC (95% CI)
CNN classification head	—	99.82 (99.48–99.94)	99.78	>0.9999 (0.9999–1.0000)
Deep only	1280	99.82 (99.48–99.94)	99.78	>0.9999 (0.9998–1.0000)
Whole-image handcrafted	23	79.44 (77.45–81.29)	70.49	0.872 (0.854–0.891)
ROI-associated handcrafted	23	85.61 (83.86–87.19)	80.90	0.934 (0.921–0.948)
Deep + whole-image handcrafted	1303	99.82 (99.48–99.94)	99.78	>0.9999 (0.9998–1.0000)
Deep + ROI-associated handcrafted	1303	99.82 (99.48–99.94)	99.78	>0.9999 (0.9998–1.0000)
All features	1326	99.82 (99.48–99.94)	99.78	>0.9999 (0.9998–1.0000)

Note: Confidence intervals for accuracy were calculated using the Wilson method and those for ROC-AUC using the Hanley–McNeil approximation, treating images as independent. All XGBoost strategies were trained on one row per original training image, with features extracted without augmentation or resampling.

CIs for accuracy were calculated with the Wilson method and CIs for ROC-AUC with the Hanley–McNeil approximation, treating images as independent. All XGBoost strategies were trained on one row per original training image, extracted without augmentation or resampling. Both handcrafted feature strategies achieved

4.4. Explanation scores

Table 4 and Figure 5 summarize deletion and insertion AUCs across the 15 evaluated images. LIME achieved the lowest mean deletion AUC (0.466; 95% CI 0.351–0.581), followed by Integrated Gradients (0.634; 0.501–0.766), Grad-CAM++ (0.886; 0.828–0.944) and GradientSHAP (0.951; 0.934–0.967). The intervals for LIME and Integrated Gradients overlapped, whereas both were separated from those of Grad-CAM++ and GradientSHAP. Mean insertion AUCs were 0.997 for LIME and 0.996 for both GradientSHAP and Integrated Gradients, with overlapping intervals, and 0.968 for Grad-CAM++. The data therefore distinguish the two lower-deletion methods (LIME and Integrated Gradients) from the two higher-deletion methods, but do not establish a ranking of LIME above Integrated Gradients. Because LIME also received the largest perturbation budget and the subset contained only correctly classified cancer-labelled images, these scores may not generalize to non-cancer-labelled or misclassified images.

The insertion curves showed high initial cancer-class probabilities for the blurred baseline in several evaluated images. The high insertion AUCs should therefore be interpreted alongside the deletion results and baseline behavior.

4.5. Qualitative explanations and randomization

Figure 6 presents four examples illustrating differences among the explanation methods. These include a cancer-labelled image classified as non-cancer with a predicted-class confidence of 52.3%. Grad-CAM++ produced broad attribution regions, whereas Integrated Gradients and GradientSHAP showed more spatially distributed patterns. LIME highlighted discrete superpixel regions. In several examples, Grad-CAM++ and LIME assigned high attribution to areas outside the body, including the scanner table or support. This suggests potential sensitivity to contextual image features, although their causal contribution to predictions was not tested.

In the single-image parameter-randomization check, correlations between maps from the trained and randomly initialized models were 0.3063 for Grad-CAM++ and 0.1777 for Integrated Gradients. Both methods produced altered attribution patterns following randomization, indicating sensitivity to model parameters in the evaluated example.

Table 4. Deletion and insertion AUCs across 15 correctly classified cancer-labelled test images.

Method	Deletion AUC, mean $\pm$ SD [95% CI]	Insertion AUC, mean $\pm$ SD [95% CI]
Grad-CAM++	0.886 $\pm$ 0.105 [0.828, 0.944]	0.968 $\pm$ 0.083 [0.922, 1.000]
Integrated Gradients	0.634 $\pm$ 0.239 [0.501, 0.766]	0.996 $\pm$ 0.007 [0.992, 0.999]
GradientSHAP	0.951 $\pm$ 0.030 [0.934, 0.967]	0.996 $\pm$ 0.006 [0.993, 0.999]
LIME	0.466 $\pm$ 0.208 [0.351, 0.581]	0.997 $\pm$ 0.004 [0.995, 1.000]

SD: sample standard deviation across the evaluated images; CI: 95% confidence interval of the mean based on the t distribution ($n = 15$), truncated at 1. Lower deletion AUC and higher insertion AUC indicate greater faithfulness under the specified evaluation protocol.

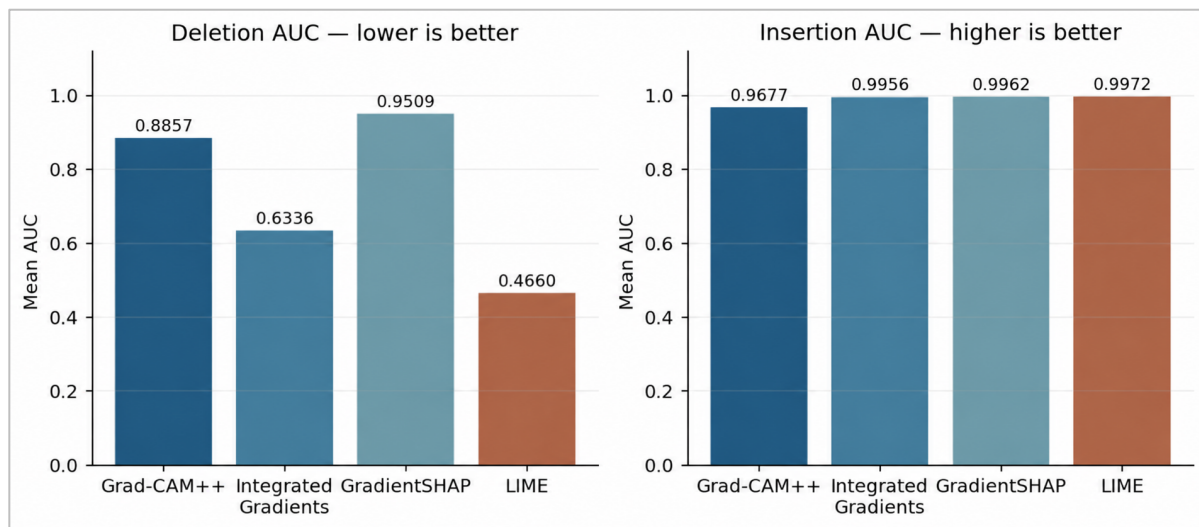


Figure 5. Mean deletion (left) and insertion (right) AUCs for four explanation methods across 15 correctly classified cancer-labelled test images.

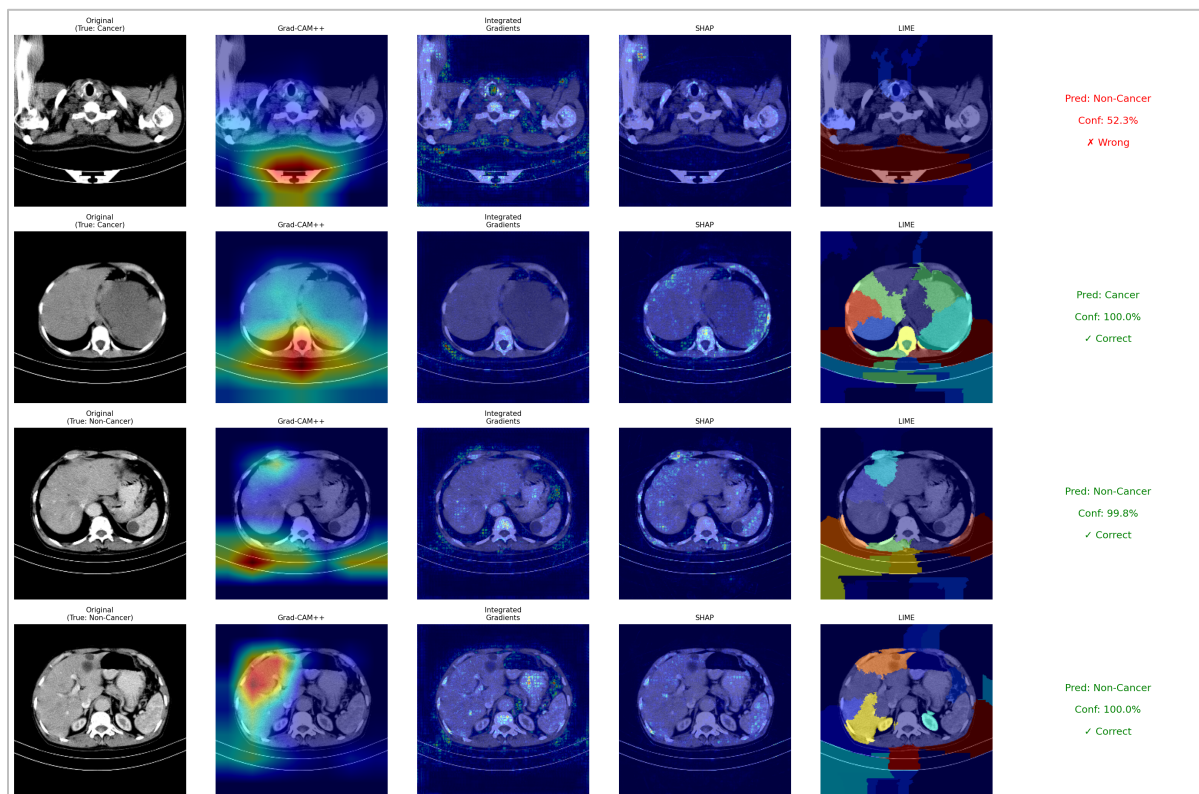


Figure 6. Qualitative comparison of four explanation methods. All maps target the cancer class, while confidence values refer to the predicted class. “True” denotes the dataset class label, and the “SHAP” column represents GradientSHAP. Integrated Gradients and GradientSHAP display absolute attribution magnitudes. Maps are normalized independently.

4.6. Duplicate-adjusted performance and background masking

Table 5 summarizes the analyses of Section 3.11. Excluding the 134 test images with an identical copy in training or validation left CNN accuracy at 99.81% (95% CI 99.44–99.93; 1,568 images), with the same three errors. The median SSIM between a test image and its nearest training or validation image was 0.80 (interquartile range 0.75–

0.86). Excluding all test images whose nearest neighbour had an SSIM of 0.80 or more left 836 images, on which accuracy was 99.64% (98.95–99.88). The raw-pixel 1-nearest-neighbour classifier, which involves no learning, reached 99.94% accuracy (99.67–99.99) on the full test set, with a single error, and 99.88% on the 836-image subset. Near-perfect agreement with the dataset labels was

therefore attainable by retrieving similar training images, so neither the CNN's accuracy nor its stability after duplicate removal can be taken as evidence of tumour-specific learning.

Masking the region outside the body reduced CNN accuracy from 99.82% to 95.06% (93.93–96.00) and the ROC-AUC to 0.986 (0.980–0.992); 83 of the 1,702 predictions changed (exact McNemar $p < 10^{-22}$). When only the background was retained, accuracy fell to 59.81%, close to the 60.0% proportion of non-cancer-labelled images, but the ROC-AUC remained 0.738 (0.713–0.763), with sensitivity 86.6% at specificity 41.9%. The background, which contains the scanner table, field-of-view boundaries and annotations, therefore carried label-related information, and part of the network's performance depended on regions outside the body. The same pattern was observed in every de-duplicated subset (Supplementary Table S4). Figure 7 shows examples, including a cancer-labelled image that was classified as non-cancer when only the body was shown ($P = 0.149$) but as cancer with $P = 1.000$ when only the scanner table and background remained.

5. Discussion

5.1. Deep representations and feature fusion

This study examined deep and handcrafted representations alongside multiple explanation methods within a common CT image-classification workflow. EfficientNet-B0 achieved high agreement with the supplied image labels, both through its classification head and through XGBoost applied to its deep features. The principal feature-fusion finding was that adding handcrafted descriptors to deep features did not improve classification.

These findings suggest that the handcrafted descriptors provided limited additional discriminative information under the evaluated configuration. Information overlap between the representations and the near-ceiling performance of the deep-feature baseline are possible explanations. Because the deep-only model made only three test errors, there was almost no headroom in which fusion could have produced a measurable gain. Neither possibility was directly tested, and the results do not imply that handcrafted features are unhelpful across other datasets or prediction tasks. Future work could test the overlap through correlation or canonical correlation analysis between handcrafted and deep features, or by training on handcrafted features combined with a reduced-

dimensional deep representation, so that 23 handcrafted descriptors are not numerically outweighed by 1,280 deep features. The results demonstrate the importance of comparing fusion strategies with a strong deep-feature baseline rather than assuming that a larger combined representation will improve classification.

The ROI-associated handcrafted set achieved higher accuracy and ROC-AUC than the whole-image handcrafted set, and the difference persisted after deterministic feature extraction and correction of the shape descriptors. Because both sets contained identical whole-image texture descriptors, this comparison functions as a partial ablation with texture held constant: the improvement is attributable to ROI-restricted intensity descriptors and/or ROI shape descriptors. It does not demonstrate a benefit from ROI-restricted texture analysis, and the separate contributions of intensity and shape would require a feature-group ablation.

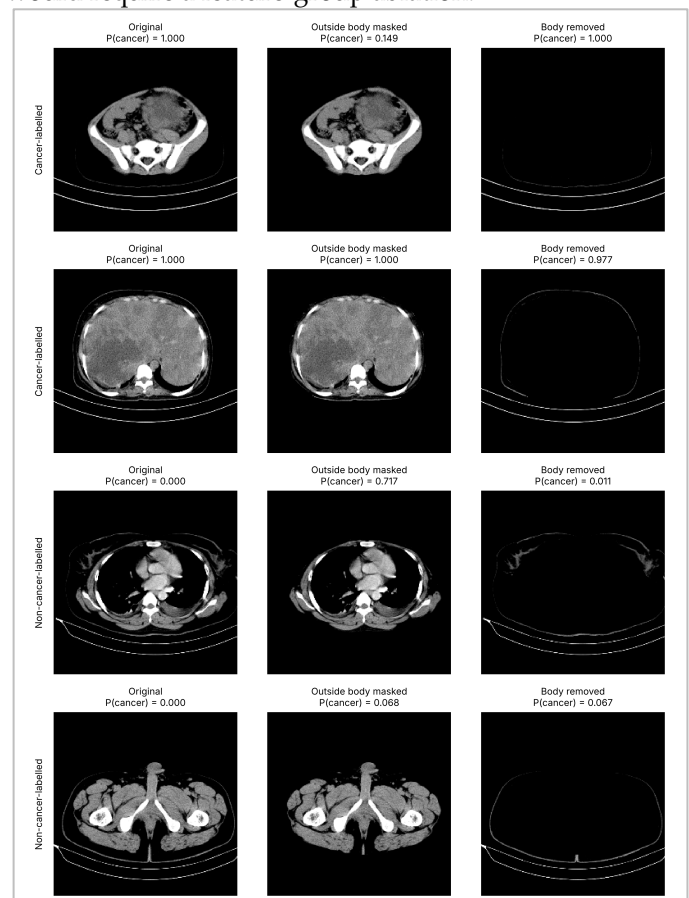


Figure 7. Examples of the body-masking experiment. Columns show the original test image, the image with all pixels outside the automatically generated body mask set to zero, and the image with the body removed. $P(\text{cancer})$ is the cancer-class probability of the trained EfficientNet-B0. A thin rim at the body boundary can remain in body-removed images.

Table 5. EfficientNet-B0 test performance after duplicate removal and under body masking, with a raw-pixel 1-nearest-neighbour reference.

Test set and condition	n	Accuracy % (95% CI)	Sens. / spec. %	ROC-AUC (95% CI)
Full test set	1,702	99.82 (99.48–99.94)	99.71 / 99.90	>0.9999 (0.9999–1.0000)
No identical copy in training or validation	1,568	99.81 (99.44–99.93)	99.71 / 99.89	>0.9999 (0.9999–1.0000)
Nearest-neighbor SSIM < 0.95	1,557	99.81 (99.44–99.93)	99.70 / 99.89	>0.9999 (0.9999–1.0000)
Nearest-neighbor SSIM < 0.80	836	99.64 (98.95–99.88)	99.40 / 99.80	>0.9999 (0.9997–1.0000)
Full test set, outside body masked	1,702	95.06 (93.93–96.00)	93.25 / 96.28	0.986 (0.980–0.992)
Full test set, body removed	1,702	59.81 (57.46–62.12)	86.64 / 41.92	0.738 (0.713–0.763)
Raw-pixel 1-NN, full test set	1,702	99.94 (99.67–99.99)	99.85 / 100.00	—
Raw-pixel 1-NN, SSIM < 0.80	836	99.88 (99.33–99.98)	99.70 / 100.00	—

Note: Rows 1–6 refer to the trained EfficientNet-B0 without retraining. SSIM: structural similarity between a test image and its most correlated training or validation image. Masking set the pixels outside (row 5) or inside (row 6) the automatically generated body mask to zero. The 1-NN reference assigns each test image the label of its most correlated training or validation image. CIs: Wilson (accuracy) and Hanley–McNeil (ROC-AUC), treating images as independent.

5.2. Interpretation of the explanation results

The explanation methods provided different views of the image regions associated with model predictions. Grad-CAM++ produced broad spatial responses, Integrated Gradients and GradientSHAP yielded more distributed attribution patterns, and LIME identified discrete superpixel regions. Their comparison therefore added information beyond the classification metrics by showing how the same prediction could be represented through different attribution approaches.

LIME obtained the most favourable mean deletion and insertion scores within the evaluated subset, but this ranking should be stated cautiously. Its deletion confidence interval overlapped that of Integrated Gradients, insertion scores were nearly indistinguishable among LIME, Integrated Gradients and GradientSHAP, and the subset comprised 15 non-randomly selected, correctly classified cancer-labelled images. LIME also received the largest computational budget. In addition, the blurred insertion baseline already produced high cancer-class probabilities in several cases, reducing the ability of insertion AUC to distinguish explanation quality. Differences in attribution sign, spatial resolution, baseline selection and computational settings also affect comparisons between methods [3–9]. The results consequently support considering deletion, insertion and visual inspection together when assessing model explanations, rather than selecting a single preferred method. Grad-CAM++ remained the fixed ROI generator; its use was independent of the comparative explanation scores.

Several visualizations highlighted the scanner table, support structures or other regions outside the body. Such patterns identify a potential source of contextual dependence that would not be apparent from accuracy

alone. The masking experiment (Section 4.6) showed that these regions influenced the predictions: masking them changed 83 test predictions, and the background alone discriminated the dataset labels (ROC-AUC 0.74). The explanation maps therefore flagged a contextual dependence that accuracy alone did not reveal, which supports their use as an audit tool. The parameter-randomization check also produced changes in the explanation maps, indicating sensitivity to model parameters in the evaluated example. This complements visual inspection, although testing on more images and all four methods would be needed to assess consistency [9].

5.3. Relationship with previous work

Previous studies have investigated radiomics, deep learning and feature fusion for a range of colorectal imaging tasks. Huang et al., for example, combined a CT radiomics signature with clinical information and CT-reported nodal status to predict lymph-node metastasis [1]. The present study instead evaluated classification of exported PNG images using dataset-assigned labels. These differences in prediction target, input representation and evaluation unit prevent direct numerical comparison, but both approaches examine the contribution of quantitative image representations to prediction.

The present results add a specific observation to the feature-fusion literature. When a fine-tuned CNN representation already separated the dataset labels almost perfectly (three test errors), appending 23 handcrafted intensity, shape and texture descriptors left the test predictions unchanged, with identical confusion matrices for all three fusion strategies. Fusion has been reported as beneficial in other settings [13,15]; the present result indicates that such benefits depend on the strength of the deep baseline and the headroom available and should be

evaluated against a strong deep-only comparator. The multi-method explanation analysis further illustrates how predictive performance and attribution behavior can be considered together within the same workflow.

The handcrafted measurements used here characterize resized, 8-bit image appearance rather than calibrated tissue properties. They should therefore be interpreted as image descriptors rather than established clinical biomarkers. Clearly specifying the data source, label definition, feature construction and evaluation unit is essential to placing these findings in context. CLAIM 2024 informed reporting of the deep learning components [23].

5.4. Limitations and future directions

The principal limitation is image-level partitioning without verified patient separation. The 11,340 images originate from 40 patients, so the number of images does not represent an equivalent number of independent clinical cases. The analyses in Section 4.6 show that this concern is substantive. Overall, 7.9% of test images had an identical copy in training or validation, and a raw-pixel nearest-neighbor lookup classified 99.94% of test images correctly. Most test images therefore have closely similar, same-label counterparts in training, as expected when adjacent slices of the same examinations are distributed across partitions. Removing exact and near-duplicate images did not reduce accuracy, but this does not establish generalization, because the remaining test images still had same-label neighbors that sufficed for near-perfect lookup. The near-ceiling results (three errors among 1,702 test images for

both the CNN head and XGBoost on deep features) should therefore be read as agreement with dataset labels within this split, not as estimates of performance on unseen patients.

Label definition is a second, related threat. Labels were assigned by dataset class, and no lesion annotations were available. If the cancer-labelled classes contain whole examinations, many cancer-labelled images would not show tumor tissue, and near-perfect classification of such images could not be explained by tumor appearance; it would instead imply reliance on patient-, examination- or acquisition-level characteristics such as scanner, protocol, field of view, patient positioning or background. The class composition makes such differences plausible: if each patient contributed images to one class only, malignant cases contributed on average about 378 images (4,538 from 12 patients) and non-malignant cases about 243 (6,802 from 28 patients), which may reflect differences in examination coverage or slice spacing between the two groups. The explanation maps showing attention to the scanner table and regions outside the body are consistent with this possibility. The masking experiment confirmed a partial dependence on non-anatomical regions (Section 4.6). Shortcut cues therefore contribute to the reported performance; because the body region alone still supported 95.06% accuracy, the extent of tumor-specific discrimination cannot be determined without lesion annotations. Table 6 summarizes the validity concerns, the evidence available in this study and their likely effect on the reported results.

Table 6. Validity concerns, evidence available in this study and likely effect on the reported results.

Validity concern	Evidence available in this study	Likely effect on reported results
Patient overlap across partitions	Image-level split of 11,340 images from 40 patients; no patient identifiers; raw-pixel 1-NN accuracy 99.94%	Test metrics measure agreement with labels of similar training images, not generalization; image-level CIs too narrow
Duplicate or identical images	1,202 images in 595 pixel-identical groups; 134 test images (7.9%) with an identical copy in training or validation	Removal did not change CNN accuracy (99.81%); does not resolve the near-neighbor overlap
Label definition	Class-assigned labels; no lesion confirmation or annotation	Model may learn patient-, examination- or acquisition-level differences rather than tumor appearance
Contextual or shortcut cues	Attribution to scanner table and background; masking outside the body: accuracy 99.82% → 95.06%; background alone: ROC-AUC 0.74	Partial reliance on non-anatomical cues confirmed
Single split, run and seed	One partition and one training run	Training- and partition-related variability not captured
XGBoost training features	One row per original image, extracted without augmentation or resampling; file names retained	No sampling-related bias in the XGBoost comparisons
ROI feature design	GLCM computed on the whole image; shape descriptors from a single connected component	Contribution of ROI-restricted texture not evaluated
Explanation evaluation scope	15 correctly classified cancer-labelled images; unequal budgets; randomization on one image	Method ranking provisional; generalization to other images untested

Several implementation choices further define the scope of the findings. Evaluation used a single split, training run and random seed, so variability due to training and partitioning is not reflected in the confidence intervals. The ROI-associated representation combined regional descriptors with whole-image texture. The body mask was generated automatically and was not verified manually; it removed the table and surrounding background but can retain a thin rim at the body boundary and, in coronal and sagittal images, some background within the convex hull. Quantitative explanation evaluation was restricted to 15 correctly classified cancer-labelled images with unequal computational budgets, and randomization was assessed on one image for two methods. Robustness of predictions and explanations to image perturbations was not established.

Future work should therefore follow a defined validation protocol. First, data independence should be established through a patient-level split, or repeated patient-level cross-validation with multiple seeds, on a dataset with patient identifiers, together with an independent external test set. Second, duplicate images should be removed before partitioning, and partitions should be formed so that no image has a near-identical counterpart in another partition. Third, models should be trained on body-masked images so that non-anatomical cues are unavailable, and tumor-specific discrimination should be assessed on lesion-confirmed slices. Fourth, ROI-specific texture and an intensity, shape and texture ablation should be evaluated. Fifth, explanation evaluation should use a larger random sample from both classes, including misclassified images, with matched computational budgets, parameter randomization for all four methods on multiple images, and Gaussian noise, blur and contrast perturbations to assess explanation stability. Sensitivity analyses of the ROI threshold and of the class-balancing strategy would further clarify the effect of these design choices.

6. Conclusion

This study evaluated EfficientNet-B0 representations, handcrafted feature fusion and four explanation methods for exploratory colon CT image classification. Deep features achieved the highest accuracy, while adding handcrafted descriptors did not change the test predictions. ROI-associated handcrafted features performed better than whole-image handcrafted features, a difference attributable to regional intensity and shape descriptors rather than texture. Explanation analysis

revealed differences in attribution patterns and highlighted contextual regions that were not apparent from classification metrics alone. Specifically, high accuracy (99.82%) coexisted with attention to non-anatomical regions, underscoring the need for explanation audits even when performance is near ceiling. Duplicate and nearest-neighbor analyses showed that near-perfect accuracy was attainable by image lookup across partitions, and masking showed partial reliance on regions outside the body. The reported performance therefore reflects agreement with dataset labels within one image-level split; patient-independent testing, training without non-anatomical cues and anatomically validated ROI analysis remain necessary before clinical application.

Funding: This work did not receive any external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Author contributions: E.T.Y.: Conceptualization, methodology, software, formal analysis, investigation, visualization, writing – original draft, and writing – review and editing.

Data Availability Statement: The source dataset is publicly available through Kaggle [10] at <https://www.kaggle.com/datasets/orvile/coloncancerct-2025-abdominal-ct-scans/data> under a CC BY 4.0 licence, as reported by the dataset provider.

Code Availability: The code for model training, explanation, feature extraction and the XGBoost comparisons, the additional analyses (duplicate detection, nearest-neighbour similarity, body masking and deterministic feature extraction) and the partition file lists are publicly available at <https://github.com/elhamTY/Colon-Study>.

Acknowledgments: None.

Conflicts of Interest: The authors declare no conflicts of interest

AI Use Declaration: Generative AI tools and Grammarly was used solely for language editing and proofreading in the complete manuscript. The authors reviewed and approved all content and take full responsibility for the manuscript.

References

1. Huang, Y.Q.; Liang, C.H.; He, L.; Tian, J.; Liang, C.S.; Chen, X.; Ma, Z.L.; Liu, Z.Y. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis

- in Colorectal Cancer. *Journal of Clinical Oncology* **2016**, *34*, 2157–2164. <https://doi.org/10.1200/JCO.2015.65.9128>
2. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; *Proceedings of Machine Learning Research* **2019**, *97*, 6105–6114. Available online: <https://proceedings.mlr.press/v97/tan19a.html>
3. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. <https://doi.org/10.1109/WACV.2018.00097>
4. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; *Proceedings of Machine Learning Research* **2017**, *70*, 3319–3328. Available online: <https://proceedings.mlr.press/v70/sundararajan17a.html>
5. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, **2017**; pp. 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
6. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; Reblitz-Richardson, O. Captum: A Unified and Generic Model Interpretability Library for PyTorch. *arXiv* **2020**, arXiv:2009.07896. <https://doi.org/10.48550/arXiv.2009.07896> (Captum version 0.8.0 was used in this study.)
7. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
8. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-Box Models. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018. <https://doi.org/10.48550/arXiv.1806.07421>
9. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31*; Curran Associates: Red Hook, NY, USA, **2018**; pp. 9505–9515. <https://doi.org/10.48550/arXiv.1810.03292>
10. Orville. ColonCancerCT-2025: Abdominal CT Scans [Dataset]. Kaggle. Available online: <https://www.kaggle.com/datasets/orville/coloncancerct-2025-abdominal-ct-scans/data>
11. Ziegelmayr, S.; Reischl, S.; Havrda, H.; Gawlitza, J.; Graf, M.; Lenhart, N.; Nehls, N.; Lemke, T.; Wilhelm, D.; Lohöfer, F.; et al. Development and Validation of a Deep Learning Algorithm to Differentiate Colon Carcinoma From Acute Diverticulitis in Computed Tomography Images. *JAMA Network Open* **2023**, *6*, e2253370. <https://doi.org/10.1001/jamanetworkopen.2022.53370>
12. Yao, L.; Li, S.; Tao, Q.; Mao, Y.; Dong, J.; Lu, C.; Han, C.; Qiu, B.; Huang, Y.; Huang, X.; et al. Deep Learning for Colorectal Cancer Detection in Contrast-Enhanced CT without Bowel Preparation: A Retrospective, Multicentre Study. *EBioMedicine* **2024**, *104*, 105183. <https://doi.org/10.1016/j.ebiom.2024.105183>
13. Attallah, O. Multi-Domain Feature Incorporation of Lightweight Convolutional Neural Networks and Handcrafted Features for Lung and Colon Cancer Diagnosis. *Technologies* **2025**, *13*, 173. <https://doi.org/10.3390/technologies13050173>
14. Alsubai, S. Transfer Learning Based Approach for Lung and Colon Cancer Detection Using Local Binary Pattern Features and Explainable Artificial Intelligence (AI) Techniques. *PeerJ Computer Science* **2024**, *10*, e1996. <https://doi.org/10.7717/peerj-cs.1996>
15. Zhang, D.; Zheng, B.; Xu, L.; Wu, Y.; Shen, C.; Bao, S.; Tan, Z.; Sun, C. A Radiomics-Boosted Deep-Learning for Risk Assessment of Synchronous Peritoneal Metastasis in Colorectal Cancer. *Insights into Imaging* **2024**, *15*, 150. <https://doi.org/10.1186/s13244-024-01733-5>
16. Wani, S.A.; Khan, S.A.; Quadri, S.M.K. scJVAE: A Novel Method for Integrative Analysis of Multimodal Single-Cell Data. *Computers in Biology and Medicine* **2023**, *158*, 106865. <https://doi.org/10.1016/j.compbiomed.2023.106865>
17. Di Giammarco, M.; Martinelli, F.; Santone, A.; Cesarelli, M.; Mercaldo, F. Colon Cancer Diagnosis by Means of Explainable Deep Learning. *Scientific Reports* **2024**, *14*. <https://doi.org/10.1038/s41598-024-63659-8>
18. Wightman, R. PyTorch Image Models. GitHub Repository, **2019**. <https://doi.org/10.5281/zenodo.4414861> (timm version 1.0.14 was used in this study.)
19. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
20. Wilson, E.B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* **1927**, *22*, 209–212. <https://doi.org/10.1080/01621459.1927.10502953>
21. Hanley, J.A.; McNeil, B.J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
22. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. <https://doi.org/10.1007/BF02295996>
23. Tejani, A.S.; Klontzas, M.E.; Gatti, A.A.; Mongan, J.T.; Moy, L.; Park, S.H.; Kahn, C.E., Jr.; CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence* **2024**, *6*, e240300. <https://doi.org/10.1148/ryai.240300>
24. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* **2004**, *13*, 600–612. <https://doi.org/10.1109/TIP.2003.819861>

Disclaimer: All views, interpretations, and data presented in Impaxon publications are the sole responsibility of the respective authors. These do not necessarily reflect the opinions of Impaxon or its editorial team. Impaxon and its editors assume no liability for any harm or loss arising from the use of information, procedures, or materials discussed in the published content.

Publisher’s Note: Impaxon remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Material

Table S1. Image augmentation and preprocessing pipelines (Albumentations 2.0.5).

Step	Transform	Effective parameters	Probability
Training pipeline (applied in this order)			
1	Resize	224 × 224 pixels	1.0
2	HorizontalFlip	—	0.5
3	VerticalFlip	—	0.5
4	RandomRotate90	Rotation by $k \times 90^\circ$, $k \in \{0, 1, 2, 3\}$	0.5
5	ShiftScaleRotate	Shift $\pm 10\%$ (x and y); scale $\pm 15\%$; rotation $\pm 15^\circ$; bilinear interpolation; constant (zero) border fill	0.5
6	OneOf (intensity)	One of: CLAHE (clip limit 1–2, 8×8 tiles); RandomBrightnessContrast (brightness ± 0.2 , contrast ± 0.2); RandomGamma (gamma 80–120)	0.5 (group)
7	OneOf (noise/blur)	One of: GaussNoise (per-channel, zero mean, SD 0.20–0.44 of the maximum intensity) ^a ; GaussianBlur (kernel 3–7, sigma 0.5–3.0)	0.3 (group)
8	CoarseDropout	1–2 rectangular holes, each 10–20% of image height and width, filled with zero	0.3
9	Normalize	ImageNet mean (0.485, 0.456, 0.406) and SD (0.229, 0.224, 0.225)	1.0
10	ToTensorV2	Conversion to a PyTorch tensor	1.0
Validation and test pipeline			
1	Resize	224 × 224 pixels	1.0
2	Normalize	As in training	1.0
3	ToTensorV2	As in training	1.0

Probabilities for OneOf groups apply to the group as a whole; within a group, one transform is selected uniformly at random. Parameters are those in effect in Albumentations 2.0.5; unlisted settings used library defaults.

^a The code specified GaussNoise with var_limit = (10, 50) and CoarseDropout with max_holes = 8, max_height = 16 and max_width = 16. These legacy argument names are not recognized by Albumentations 2.0.5 and were ignored, so the library defaults shown above were applied during training.

Table S2. Handcrafted image descriptors (23 per feature set).

No.	Descriptor	Definition	Region
Intensity and first-order statistics (15) - calculated from the pixels inside the mask			
1	Mean	Arithmetic mean of grey values	Mask
2	Standard deviation	Population standard deviation of grey values	Mask
3	Minimum	Lowest grey value	Mask
4	Maximum	Highest grey value	Mask
5	Median	50th percentile of grey values	Mask
6	Skewness	Sample skewness (Fisher–Pearson coefficient, SciPy default)	Mask
7	Kurtosis	Excess kurtosis (Fisher definition, SciPy default)	Mask
8	Entropy	Shannon entropy (natural logarithm) of a 64-bin histogram spanning the observed grey-value range	Mask
9	Energy	Mean squared grey value, $\sum x^2/N$	Mask
10	Range	Maximum – minimum	Mask
11	10th percentile	10th percentile of grey values	Mask
12	25th percentile	25th percentile of grey values	Mask
13	75th percentile	75th percentile of grey values	Mask
14	90th percentile	90th percentile of grey values	Mask
15	Interquartile range	75th percentile – 25th percentile	Mask

No.	Descriptor	Definition	Region
Shape (3) - calculated from the binary mask			
16	Area	Number of pixels in the largest 8-connected component of the mask	Mask
17	Perimeter	Arc length of the external contour of the same component (OpenCV findContours, RETR_EXTERNAL, CHAIN_APPROX_SIMPLE) ^b	Mask
18	Compactness	$4\pi \times \text{area} / \text{perimeter}^2$, from the same component	Mask
Texture: grey-level co-occurrence matrix (5) - always calculated from the full image			
19	GLCM contrast	$\sum P(i,j)(i-j)^2$	Full image
20	GLCM dissimilarity	$\sum P(i,j) i-j $	Full image
21	GLCM homogeneity	$\sum P(i,j) / [1 + (i-j)^2]$	Full image
22	GLCM energy	$\sqrt{\sum P(i,j)^2}$, i.e. the square root of the angular second moment	Full image
23	GLCM correlation	$\sum P(i,j)(i - \mu_i)(j - \mu_j) / (\sigma_i \sigma_j)$	Full image

All descriptors were calculated on 8-bit grayscale images (224×224 pixels) obtained by denormalizing the network input and converting it with OpenCV (RGB to grayscale). Grey values were used without bin-width discretization. *Mask*: for the whole-image set, a mask covering all 224×224 pixels; for the ROI-associated set, the Grad-CAM++ ROI mask (85th-percentile threshold, largest connected component of at least 100 pixels, 5×5 morphological closing then opening; the whole-image mask when no usable ROI was obtained). GLCMs used a distance of one pixel, angles of 0° , 45° , 90° and 135° , 256 grey levels, symmetric construction and normalization; each property was averaged over the four angles (scikit-image graycoprops). Feature columns are prefixed *whole_* or *xai_* in the exported matrices.

^b For the whole-image mask, the shape descriptors are constant: area = 50,176 pixels, perimeter = 892 pixels and compactness ≈ 0.792 . For ROI masks, area, perimeter and compactness all refer to the largest connected component of the final mask (manuscript Section 3.8).

Table S3. Groups of pixel-identical images by the partitions in which their members occur.

Partitions of group members	Groups	Images	Non-cancer-labelled groups	Cancer-labelled groups
Training + Training	291	582	291	0
Training + Validation	133	266	133	0
Training + Test	108	216	108	0
Validation + Test	23	46	23	0
Validation + Validation	17	34	17	0
Test + Test	11	22	11	0
Training + Training + Validation	6	18	2	4
Training + Training + Training	4	12	0	4
Training + Test + Test	1	3	1	0
Training + Training + Test	1	3	0	1
Total	595	1,202	586	9

Duplicates were identified by the MD5 hash of the decoded pixel values of each original 512×512 PNG file; file-level MD5 hashes gave the same groups. No group contained images with different labels. In total, 134 test images had an identical copy in the training or validation partition.

Table S4. EfficientNet-B0 performance on de-duplicated test subsets under the original, body-only and background-only conditions, with the raw-pixel 1-nearest-neighbour reference.

Test subset	Images (cancer)	Original acc. %	Body only: acc. % / ROC-AUC	Background only: acc. % / ROC-AUC	1-NN acc. %
Full test set	1,702 (681)	99.82	95.06 / 0.986	59.81 / 0.738	99.94
No identical copy in training/validation	1,568 (680)	99.81	94.90 / 0.986	63.27 / 0.761	99.94
NN SSIM < 0.95	1,557 (677)	99.81	94.86 / 0.985	63.13 / 0.760	99.94
NN SSIM < 0.90	1,453 (632)	99.79	94.98 / 0.987	63.18 / 0.762	99.93
NN SSIM < 0.80	836 (336)	99.64	95.33 / 0.988	59.69 / 0.757	99.88

NN SSIM: structural similarity between a test image and its most correlated training or validation image (64×64 Pearson correlation search; SSIM at 224×224). Body only: pixels outside the automatically generated body mask set to zero. Background only: pixels inside the mask set to zero. The trained network was not retrained. 1-NN: label of the most correlated training or validation image.