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.

Benchmarking CNN, Transformer, Hybrid, and Foundation Models for Multi-Label Bronchoscopic Finding Classification on the BERD Dataset

İlkay Sibel Kervancı

skervanci@gantep.edu.tr

Gaziantep University

Research Article

Keywords: Bronchoscopy, Deep learning, Multi-label classification, Hybrid CNN-Transformer, Foundation models, Class imbalance, Grad-CAM

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11268987/v1

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

Additional Declarations: No competing interests reported.

Benchmarking CNN, Transformer, Hybrid, and Foundation Models for Multi-Label Bronchoscopic Finding Classification on the BERD Dataset

İlkay Sibel Kervancı[1*]

[1*]Department of Computer Engineering, Gaziantep University, University Bulv, Gaziantep, 27310, Gaziantep, Türkiye.

Corresponding author(s). E-mail(s): skervanci@gantep.edu.tr;

Abstract

Background: Automated classification of bronchoscopic findings is hindered by severe class imbalance, visual similarity between findings, and scarce expert-annotated data. This study benchmarked convolutional, Transformer, hybrid, and foundation models for multi-label bronchoscopic finding classification.

Methods: ResNet-50, EfficientNet-B3, ViT-B/16, Swin-T, and Hybrid-EffSwin (average-pooling fusion of EfficientNet-B3 and Swin-T) were trained on 23 finding categories of the Bronchoscopy Examination Report Dataset (BERD) using identical hyperparameters and patient-level data splits. ConvNeXt-T and two foundation models (DINOv2 ViT-B/14 and BiomedCLIP ViT-B/16) served as additional baselines. Ablations examined fusion strategy, loss function, and decision thresholds. Models were compared using patient-level bootstrap confidence intervals (CIs) and Holm-corrected Wilcoxon tests on patient-level summaries.

Results: On the independent test set (316 images, 180 patients), Hybrid-EffSwin achieved the highest macro F1-score (0.376), Matthews correlation coefficient (0.339), and area under the ROC curve (0.803). Across three training seeds, Hybrid-EffSwin ranked first in every run (macro F1 = 0.363 $\pm$ 0.013); its advantage over ResNet-50 and ViT-B/16 was significant in all runs, over EfficientNet-B3 in one of three, and over Swin-T in none (ΔF1 = +0.020 to +0.025). Among additional baselines, DINOv2 performed comparably to Hybrid-EffSwin (Hybrid-EffSwin minus DINOv2: ΔF1 = +0.013; 95% CI −0.025 to 0.052) and achieved the best exact-match accuracy and calibration,

whereas ConvNeXt-T was significantly inferior; a late-fusion ensemble of separately trained EfficientNet-B3 and Swin-T performed close to Hybrid-EffSwin. Average-pooling fusion performed at least as well as residual-projection and concatenation fusion, and validation-selected thresholds changed macro F1 only modestly. **Conclusions:** Hybrid CNN-Transformer, hierarchical Transformer, and self-supervised foundation models formed the strongest tier for multi-label bronchoscopic classification and could not be distinguished on the current test set. Multi-institutional validation is warranted.

Keywords: Bronchoscopy, Deep learning, Multi-label classification, Hybrid CNN-Transformer, Foundation models, Class imbalance, Grad-CAM

Background

Bronchoscopy is a cornerstone diagnostic and therapeutic procedure in pulmonary medicine, enabling direct visualization of the tracheobronchial tree for the assessment of airway pathologies including neoplasms, infections, inflammatory conditions, and structural abnormalities [1]. The accurate and timely interpretation of bronchoscopic findings is critical for clinical decision-making; however, this task demands substantial expertise and remains challenging even for experienced clinicians. The increasing volume of bronchoscopic procedures performed globally, combined with a shortage of specialized pulmonologists in many healthcare settings, has created a pressing need for automated support systems capable of assisting clinicians during bronchoscopic examination [2]. Computer-aided diagnosis (CAD) systems powered by deep learning have demonstrated remarkable promise across a wide spectrum of medical imaging tasks [3], and their application to bronchoscopy represents a rapidly emerging area of investigation.

Early deep learning studies applied to bronchoscopy focused primarily on anatomical navigation and binary pathology detection [4]. More recent works have extended these approaches to multi-class airway disease classification using convolutional neural network (CNN)-based architectures, including EfficientNet-based models for multinodular airway disease differentiation [5] and multi-task DenseNet frameworks for

simultaneous tuberculosis and lung cancer classification [6]. Sun et al. proposed MARN, a multiscale attention residual network trained on the BI2K dataset comprising 2,900 bronchoscopic images from 615 patients, which achieved an accuracy of 0.978, an F1-score of 0.977, and an area under the curve (AUC) of 0.998 for binary lesion detection [7]. Despite these advances, existing bronchoscopy AI systems remain limited to narrow, institution-specific classification problems with restricted pathological coverage, and, to our knowledge, none has addressed fine-grained multi-label classification across a comprehensive set of bronchoscopic findings.

The introduction of the Vision Transformer (ViT) demonstrated that a pure Transformer applied directly to sequences of image patches can match or surpass convolutional networks on large-scale image classification tasks [8], shifting the deep learning landscape toward attention-based architectures. ViTs and their hierarchical variants, such as the Swin Transformer, have substantially advanced the state of the art in medical image classification [9]. Unlike CNNs, which excel at capturing local texture features through convolutional operations, Transformers model long-range global dependencies through self-attention mechanisms. Hybrid architectures that combine the complementary strengths of CNNs and Transformers have demonstrated superior performance across multiple medical imaging domains, including brain tumor classification [10] and breast ultrasound analysis [11]. Related dual-branch architectures that combine Swin Transformer representations with complementary feature streams have also been reported for retinal vessel segmentation [12] and for gastric intestinal metaplasia classification in endoscopic images [13]. However, such hybrid approaches have not yet been systematically evaluated for bronchoscopic finding classification.

A critical bottleneck in bronchoscopy AI research has been the absence of large-scale, publicly available, fine-grained annotated datasets. The recently introduced Bronchoscopy Examination Report Dataset (BERD) addresses this gap by providing 6330 expert-annotated bronchoscopic images derived from 3692 clinical examination

reports, with multi-label pathological annotations spanning 24 distinct bronchoscopic findings [14]. Although the original BERD publication demonstrated the dataset's utility for automated report generation using multimodal large language models, its potential for pure image classification remains unexplored. Pure image classification represents a more computationally accessible and clinically deployable approach, requiring neither textual annotations nor large language model inference at deployment time. Artificial intelligence has increasingly been applied across airway management, from YOLO-based anatomical segmentation of the upper airway [15] using the Upper Airway Anatomical Landmark (UAAL) dataset [16] to preoperative prediction of difficult laryngoscopy from facial photographs with hybrid AI models [17]. The present study extends this line of work to the lower airway, introducing a fine-grained multi-label classification framework based on BERD.

To our knowledge, this study presents the first systematic evaluation of deep learning architectures for multi-label bronchoscopic finding classification using BERD. Five architectures spanning three paradigms are benchmarked: CNN-based models (ResNet-50, EfficientNet-B3), Transformer-based models (ViT-B/16, Swin-T), and a hybrid CNN-Transformer model (Hybrid-EffSwin) that fuses EfficientNet-B3 and Swin-T features through element-wise average pooling of projected feature representations, a design selected through an ablation study comparing it with residual-projection and concatenation fusion. Model performance is evaluated using F1-score, area under the receiver operating characteristic curve (AUC-ROC), Matthews correlation coefficient (MCC), and area under the precision–recall curve (AUPRC), with uncertainty and statistical significance assessed through patient-level bootstrap confidence intervals and Holm-corrected patient-level Wilcoxon signed-rank tests. Gradient-weighted class activation mapping (Grad-CAM) is employed to provide visual interpretability of model decisions. The principal contributions of this work

are as follows: (1) an independent application of BERD to automated bronchoscopic finding classification, extending beyond the dataset's original report-generation purpose; (2) a systematic comparison of CNN, Transformer, and hybrid CNN-Transformer architectures for multi-label bronchoscopy classification, complemented by contemporary CNN and foundation-model baselines, providing reference results for future bronchoscopy AI systems; and (3) a rigorous statistical evaluation framework based on patient-level bootstrap confidence intervals, multiplicity-corrected non-parametric pairwise tests, replication across training seeds, and validation-based threshold selection, with design choices verified on the validation subset.

Methods

Dataset

This study utilized the Bronchoscopy Examination Report Dataset (BERD), a publicly available, expert-annotated bronchoscopic image dataset introduced by Luo et al. [14]. BERD comprises 6330 representative bronchoscopic images derived from 3692 clinical examination reports, annotated by experienced bronchoscopists. Finding labels were assigned through a semi-automated workflow in which a locally deployed large language model proposed labels from the clinical captions based on a clinician-defined category list, after which all labels were reviewed and refined by the bronchoscopists and verified by a senior expert [14]. Each image is accompanied by a free-text caption, an anatomical location label, one or more pathological finding labels provided as a comma-separated string, and a patient identifier; images from the same patient share the same identifier [14]. The official BERD release is partitioned into a training set of 6014 images from 3512 patients and a test set of 316 images from 180 patients, with no patient overlap between the two sets.

The severe class imbalance characteristic of the BERD dataset is illustrated in Fig. 1. Exploratory data analysis revealed 24 distinct pathological labels across the

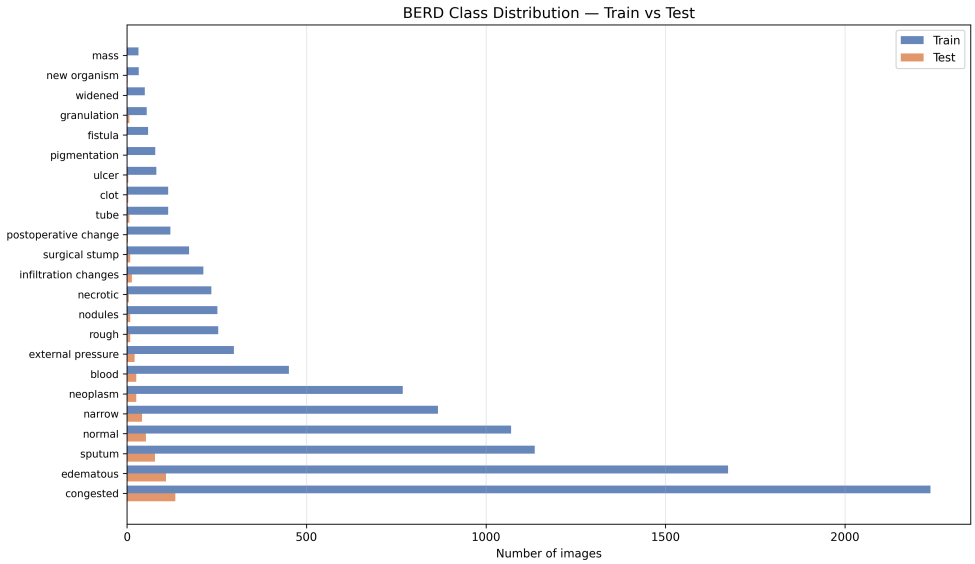


Fig. 1 Class distribution across the 23 BERD finding categories in the training pool (6012 images; training and validation subsets combined) and the test set (316 images).

combined training and test sets. One class, diverticulum ($n = 2$ in the training set), was excluded from the classification task. With only two training samples, reliable model optimization and meaningful evaluation of this class are not feasible; such extreme minority classes are known to contribute noise rather than signal during training [18], and cannot be assessed with statistical validity on the test set. Removing this label left two training images without any remaining finding label; these were excluded, reducing the training pool to 6012 images from 3511 patients. The final label space thus comprised 23 classes. The class distribution was markedly imbalanced: the most frequent class (congested) appeared in 37.2% of training images, while the rarest retained class (mass) appeared in only 0.5%. The mean number of labels per image was 1.72 (training) and 1.79 (test), with 46.0% of training images carrying two or more simultaneous labels, confirming the multi-label nature of the classification task.

Data preprocessing and augmentation

All images were resized to 224×224 pixels to comply with the input requirements of ImageNet-pretrained backbone architectures. Pixel intensities were normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]).

Training images underwent the following augmentation pipeline, designed to reflect the variability inherent in bronchoscopic image acquisition across different endoscopic systems and lighting conditions: (1) random resize and crop (resize to 256×256, random crop to 224×224); (2) random horizontal flip ($p = 0.5$) and vertical flip ($p = 0.2$); (3) random rotation ($\pm 15°$); (4) color jitter (brightness ± 0.25, contrast ± 0.25, saturation ± 0.10, hue ± 0.05); (5) random autocontrast ($p = 0.4$), a global contrast adjustment mimicking variation in endoscope exposure; and (6) Gaussian blur (kernel size 3, $\sigma \in [0.1, 1.5]$; $p = 0.2$) to simulate motion-induced blurring. No augmentation was applied to test images beyond resizing and normalization.

To further address class imbalance during training, two complementary strategies were employed. First, a weighted random sampler assigned higher sampling probabilities to images containing rare-class labels, ensuring adequate exposure to underrepresented findings. Second, a weighted binary cross-entropy (BCE) loss function was used, with per-class positive weights computed on the training subset as the ratio of negative to positive samples. To prevent gradient instability caused by extreme imbalance, positive weights were capped at an empirically chosen value of 50.0. Batch-level MixUp augmentation ($\alpha = 0.4$, application probability = 0.3) was applied during training; MixUp interpolates pairs of randomly selected training images and their corresponding label vectors, improving decision boundary smoothness and generalization in multi-label settings.

Model architectures

Building on prior segmentation-based work on endoscopic airway imaging [15], the present study shifts toward a classification paradigm enabled by the availability of labeled pathological annotations in BERD. Five deep learning architectures representing three distinct paradigms were evaluated.

CNN-based baselines

ResNet-50 [19] is a 50-layer residual network with skip connections that mitigate vanishing gradients and serves as the standard CNN baseline in medical image classification benchmarks (24.6 million parameters including the classification head). The final global average pooling output (2048-dimensional) was passed through a dropout layer ($p = 0.4$), a fully connected layer ($2048 \to 512$) with GELU activation, and a second dropout layer ($p = 0.2$) before the 23-dimensional output layer. EfficientNet-B3 [20] employs compound scaling of depth, width, and resolution; with 11.5 million parameters, it achieves competitive performance at substantially lower computational cost than ResNet-50. Its 1536-dimensional feature vector was processed through an identical classification head.

Vision Transformer-based models

ViT-B/16 [8] applies a pure Transformer to non-overlapping 16×16 image patches. The 768-dimensional [CLS] token output was passed through layer normalization, dropout ($p = 0.4$), a linear layer ($768 \to 512$) with GELU activation, and a second dropout layer ($p = 0.2$) before the output layer. The Swin Transformer (Swin-T) [21] introduces hierarchical feature maps and shifted-window self-attention, preserving both local and global contextual information; its 768-dimensional output was processed through an analogous classification head.

Hybrid CNN-Transformer architecture

The hybrid model (Hybrid-EffSwin) integrates EfficientNet-B3 as the CNN branch and Swin-T as the Transformer branch, combining local texture extraction with global contextual reasoning. The 1536-dimensional EfficientNet-B3 feature vector F_{Eff} and the 768-dimensional Swin-T feature vector F_{Swin} were independently projected to a shared 512-dimensional space by a linear layer followed by layer normalization (LN) and GELU activation:

$$f_{\text{cnn}} = \text{GELU}\big(\text{LN}(W_1 F_{\text{Eff}} + b_1)\big), \quad W_1 \in \mathbb{R}^{512\times 1536}, \tag{1}$$

$$f_{\text{tr}} = \text{GELU}\big(\text{LN}(W_2 F_{\text{Swin}} + b_2)\big), \quad W_2 \in \mathbb{R}^{512\times 768}. \tag{2}$$

The two 512-dimensional feature vectors were fused by element-wise average pooling:

$$f_{\text{fused}} = \frac{f_{\text{cnn}} + f_{\text{tr}}}{2}. \tag{3}$$

The fused representation was passed through dropout ($p = 0.15$) and a linear output layer ($512 \to 23$). Unlike the single-backbone models, whose heads contain a 512-unit hidden layer, Hybrid-EffSwin maps the fused representation directly to the output layer, because its two projection layers already act as a hidden layer.

This fusion step, which adds no parameters beyond the two projection layers, was selected following a dedicated ablation study (see *Ablation study design*) comparing average pooling with two higher-capacity alternatives, residual-projection and concatenation fusion, under identical training hyperparameters. Average pooling achieved the highest validation macro F1 among the three variants, consistent with the principle that additional learnable fusion parameters may not be reliably optimized under the limited training data available in this study (2809 training patients).

All backbone weights were initialized from ImageNet-pretrained models (ResNet-50: IMAGENET-1K_V2; EfficientNet-B3, ViT-B/16, and Swin-T: IMAGENET-1K_V1). Initialization from ImageNet-pretrained weights follows the prevailing transfer-learning practice for medical image classification with limited annotated data [22]. For ViT-B/16, the SWAG-pretrained variant (IMAGENET-1K_SWAG_E2E_V1) was initially considered but requires a 384×384 input; the standard IMAGENET-1K_V1 weights were therefore used to maintain a consistent 224×224 input resolution across all five architectures.

Training strategy

To ensure a fair architectural comparison, identical hyperparameters (learning rate, optimizer, scheduler, and batch size) were applied to all five models rather than performing per-model hyperparameter search, isolating the effect of architectural choice from optimization tuning. All models were trained for a maximum of 70 epochs using the AdamW optimizer [23] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The learning rate was reduced on plateau (factor = 0.5, patience = 5 epochs, monitored metric: validation macro F1-score), and early stopping was applied with a patience of 15 epochs.

The BERD training set was partitioned at the patient level into a training subset (80%; 4826 images, 2809 patients) and a validation subset (20%; 1186 images, 702 patients). The validation subset was used exclusively for early stopping and model selection, and patient-level partitioning prevented images from the same patient from appearing in both subsets. The held-out test set (316 images, 180 patients) was reserved for final evaluation and was not used for any training or model selection decision.

Gradient clipping (maximum norm = 1.0) was applied to prevent gradient explosion, and automatic mixed-precision (AMP) training was employed to reduce memory

usage and accelerate training. A batch size of 32 was used throughout. All primary experiments were run with a fixed random seed of 42, chosen before replicate runs were planned, so that this run constitutes the pre-specified primary analysis rather than a selected one; because cuDNN autotuning was enabled, bitwise reproducibility across GPU runs is not guaranteed.

To validate the fixed hyperparameter selection, a supplementary search was conducted using the Tree-structured Parzen Estimator (TPE) sampler with Hyperband pruning as implemented in Optuna [24, 25]. For each architecture, 15 trials of up to 30 epochs were evaluated on the validation subset using architecture-specific search spaces, with the first trial seeded with the fixed configuration (learning rate = 1×10^{-4}, maximum positive weight = 50.0). The test set was not used at any stage of the search; results are reported in the Discussion. To quantify variability due to training stochasticity, the fixed configuration of each of the five primary architectures was additionally trained with two further training seeds (7 and 123), yielding three runs per architecture. The patient-level data split was held fixed (seed 42) across runs, so that only the stochastic elements of training varied (initialization of the classification heads, sampler order, dropout masks, and MixUp sampling).

Unless stated otherwise, all results refer to the seed-42 run; seed-replicate results (Table 4) are summarized as mean ± standard deviation (SD). All experiments were conducted on a single NVIDIA GeForce RTX 4080 GPU (17.2 GB VRAM) under Windows 11 using PyTorch 2.5.1. Training time per model ranged from approximately 37 to 70 minutes.

Ablation study design

Two complementary ablation studies were conducted to isolate the contribution of individual design choices. First, to assess the effect of the loss function, the four single-backbone architectures (ResNet-50, EfficientNet-B3, ViT-B/16, and Swin-T)

were trained under both weighted BCE and Asymmetric Loss (ASL; $\gamma_{\text{neg}} = 4.0$, $\gamma_{\text{pos}} = 0.0$, clip $= 0.05$) [26], with all other hyperparameters held constant. In the ASL runs, the capped per-class positive weights were retained, so that only the loss formulation changed. Hybrid-EffSwin was trained only under BCE, consistent with its selection via the fusion strategy ablation described below.

Second, to validate the average-pooling fusion strategy selected for Hybrid-EffSwin, three fusion configurations were evaluated under identical training hyperparameters: average pooling (Eq. (3)); cross-attention, in which the projected CNN feature vector f_{cnn} serves as the query and the projected Transformer feature vector f_{tr} provides the key and value,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V, \tag{4}$$

where $Q = f_{\text{cnn}}$, $K = V = f_{\text{tr}}$, and $d_k = 512/8 = 64$ (8 attention heads), followed by a residual connection and layer normalization. Because both inputs are single pooled vectors (sequence length 1), the attention weights are identically 1 and the query has no effect; this variant therefore reduces to a learned projection of the Transformer feature added residually to the CNN feature and is referred to as residual-projection fusion below. The third configuration was concatenation, in which the two feature vectors are concatenated and projected through a linear layer. Fusion strategy and loss function were selected on the validation subset; test-set results of both ablations are reported for completeness.

Additional baselines

To place the primary comparison in the context of current models, three additional baselines were evaluated as a secondary analysis: ConvNeXt-T [27], a modernized CNN pretrained on ImageNet-1k; DINOv2 ViT-B/14 [28], a general-purpose foundation

model pretrained by self-supervision on 142 million natural images; and BiomedCLIP ViT-B/16 [29], the image encoder of a biomedical vision–language foundation model pretrained on 15 million image–text pairs from PubMed Central. Each backbone was followed by a classification head analogous to those of the primary models and trained with the same loss, augmentation, sampling, and early-stopping settings. ConvNeXt-T was trained with the fixed learning rate of 1×10^{-4}. For the two foundation models, the pretrained backbone was fine-tuned with a tenfold lower learning rate (1×10^{-5}) while the classification head retained 1×10^{-4}, following standard fine-tuning practice for large pretrained encoders; a preliminary DINOv2 run with a uniform learning rate of 1×10^{-4} failed to train on the validation subset (best validation macro F1 = 0.212 after 70 epochs, compared with 0.410 after the adjustment), indicating that the pretrained representation had been degraded. Because these baselines were added after the primary comparison had been specified, each was trained with a single seed and compared with Hybrid-EffSwin and Swin-T by patient-level paired bootstrap only, outside the Holm-corrected family of primary comparisons. To separate the effect of joint training from that of combining two backbones, a late-fusion ensemble was additionally formed by averaging the predicted probabilities of the separately trained EfficientNet-B3 and Swin-T models of the same seed.

Evaluation metrics

Model performance was assessed on the held-out test set ($n = 316$) at a fixed classification threshold of $\theta = 0.5$. Six complementary metrics were selected to capture distinct aspects of multi-label classification quality, as no single metric is sufficient when class distributions are markedly imbalanced.

Accuracy was computed as exact subset match, the strict criterion under which a prediction is correct only if all 23 labels are simultaneously assigned correctly. The F1-score was reported using macro-averaging, which weights each class equally regardless

of its prevalence and is therefore more sensitive to performance on rare pathological categories than micro-averaging, which is dominated by frequent classes such as congested and edematous.

Because two retained classes (mass and widened) had no positive instances in the test set, per-class F1-score, MCC, AUPRC, and AUC-ROC are undefined for these classes. All macro-averaged metrics were therefore computed over the 21 classes with at least one positive test instance, whereas exact-match accuracy and Hamming loss were computed over all 23 labels. Macro-averaged values over all 23 classes, in which the two absent classes contribute structural zeros, are reported in Additional file 1 (Table S1). An MD5 check identified three test images that were pixel-identical to training images despite belonging to different patients. Excluding these images reduced macro F1 by less than 0.001 and exact-match accuracy by less than 0.01 for all five architectures, without altering the model ranking (Additional file 1, Table S3); they were therefore retained in the primary analysis to preserve the official BERD split.

AUC-ROC, macro-averaged across the 21 evaluable classes, was used as a threshold-independent measure of discriminative ability. Because AUC-ROC can remain misleadingly high under severe class imbalance, it was complemented by AUPRC, which is more sensitive to performance on minority classes and is considered more informative than AUC-ROC in imbalanced settings [30]. Macro-averaged AUPRC is equivalent to the mean average precision (mAP) commonly reported in multi-label image classification benchmarks and is denoted AUPRC (mAP) throughout this manuscript.

MCC, macro-averaged across the 21 evaluable classes, was included as a complementary measure of classification reliability; macro F1 over the 21 evaluable classes was pre-specified as the primary endpoint. Unlike the F1-score, MCC incorporates

all four entries of the confusion matrix and is not inflated by majority-class performance, making it particularly informative for long-tailed label distributions [31]. It is computed for each class as

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{5}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives for the binary one-vs-rest classification of each label. A macro MCC of 0 indicates performance no better than chance, and a value of 1 indicates perfect classification.

Finally, Hamming loss, defined as the proportion of individual label predictions that are incorrect across all samples and classes, was reported as a label-wise error measure. Per-class decision thresholds were additionally selected on the patient-level validation subset by searching $\theta \in [0.05, 0.95]$ in steps of 0.01 and choosing, for each class, the threshold that maximized validation F1; the selected thresholds were then applied unchanged to the test set. Per-class F1-score, AUC-ROC, and MCC were also reported to identify architecture-specific strengths across individual pathological categories.

Grad-CAM visual interpretability

To provide qualitative insight into the spatial regions driving each model's classification decisions, Grad-CAM [32] was applied to the four architectures that yield spatial feature maps (ResNet-50, EfficientNet-B3, Swin-T, and Hybrid-EffSwin). Grad-CAM produces a coarse localization heatmap by using the gradients of a target class score, flowing into the final feature-map layer, to weight the spatial activation maps of that

layer:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{6}$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\Big(\sum_k \alpha_k^c A^k \Big), \tag{7}$$

where A^k denotes the activation of the k-th feature map of the target layer, y^c is the pre-sigmoid score for class c, Z is the number of spatial locations in the feature map, and α_k^c is the global-average-pooled gradient (neuron importance weight) for channel k with respect to class c. This implementation follows Selvaraju et al. [32] and has previously been applied across multiple CNN architectures for medical image interpretability [33].

For ResNet-50 and EfficientNet-B3, Grad-CAM was applied to the output of the final convolutional block preceding global average pooling. For Swin-T, the final-stage feature map, returned in channel-last format by the torchvision implementation, was permuted to a channel-first layout prior to CAM computation. For Hybrid-EffSwin, Grad-CAM was computed exclusively on the final convolutional block of the EfficientNet-B3 branch: because the fusion step averages two already-pooled feature vectors, no spatial map exists downstream of either backbone's internal pooling, making the pre-pooling output of the CNN branch the only architecturally valid hook point. This CNN-branch view therefore shows where the convolutional pathway attends when the fused model makes its prediction and does not capture the window-based attention patterns of the Swin-T branch, which would require a separate attention-rollout analysis. Although Grad-CAM has been adapted to Vision Transformer token representations for chest radiograph analysis [34], ViT-B/16 was excluded from this analysis, as its patch-embedding and [CLS]-token design does not

natively produce a spatial feature map compatible with gradient-based CAM localization. The example shown was not curated: the first test image containing the congested label, in dataset order, was visualized.

Statistical analysis

Because the 316 test images originate from 180 patients and images from the same patient are not independent, uncertainty was quantified using a patient-level (cluster) bootstrap with 2000 resamples, in which patients rather than images were resampled with replacement. Percentile 95% confidence intervals (CIs) were computed for the macro F1, MCC, and AUPRC of each model and for paired differences (ΔF1, ΔMCC) between models evaluated on the same resamples; bootstrap p-values were obtained as twice the smaller tail proportion of the difference distribution. Because image-level tests assume independence, pairwise differences were additionally assessed at the patient level: per-patient mean example-based F1 (computed per image as $2|y \cap \hat{y}|/(|y| + |\hat{y}|)$) and per-patient exact-match rate were compared with the two-sided Wilcoxon signed-rank test; image-level Wilcoxon and McNemar tests are reported in Additional file 1 (Table S4). Within each bootstrap resample, macro metrics were averaged over the classes with at least one positive instance in that resample. P-values from the ten pairwise model comparisons were adjusted using the Holm method; bootstrap CIs are reported without multiplicity adjustment as estimates of effect size. For the seed replicates, paired bootstrap comparisons between Hybrid-EffSwin and each baseline were repeated within each seed. Statistical significance was set at $\alpha = 0.05$.

Use of large language models

Claude (Anthropic) and Gemini (Google) were used to assist with code development and English-language editing. All generated code was reviewed, tested, and validated by the author, and all text was reviewed and edited by the author, who takes full

Table 1 Primary performance metrics on the BERD test set ($n = 316$ images, 180 patients), fixed hyperparameters, $\theta = 0.5$. F1, MCC, AUPRC, and AUC-ROC are macro-averaged over the 21 classes with at least one positive test instance; accuracy and Hamming loss are computed over all 23 labels. 95% CIs from patient-level bootstrap (2000 resamples). Bold = best per column; ↓ = lower is better.

Model	Acc	F1 macro	MCC	AUPRC (mAP)	AUC-ROC	Hamming ↓	95% CI (F1)
ResNet-50	0.199	0.299	0.251	0.343	0.741	0.091	[0.253, 0.363]
EfficientNet-B3	0.177	0.324	0.280	0.340	0.799	0.096	[0.278, 0.387]
ViT-B/16	0.155	0.312	0.261	0.344	0.787	0.104	[0.274, 0.370]
Swin-T	0.247	0.351	0.312	0.362	0.715	0.078	[0.298, 0.428]
Hybrid-EffSwin	**0.275**	**0.376**	**0.339**	**0.385**	**0.803**	**0.074**	[0.316, 0.453]

responsibility for the content of this publication. No large language model was used to generate data or results.

Results

Overall model performance

Table 1 presents the performance of all five architectures on the held-out test set. Hybrid-EffSwin achieved the highest macro F1-score (0.376), MCC (0.339), exact-match accuracy (0.275), AUPRC (0.385), and AUC-ROC (0.803), while attaining the lowest Hamming loss (0.074). Swin-T ranked second (F1 = 0.351, MCC = 0.312), followed by EfficientNet-B3 (0.324), ViT-B/16 (0.312), and ResNet-50 (0.299). The patient-level 95% CIs of Hybrid-EffSwin and Swin-T overlapped substantially. The comparatively low macro AUC-ROC of Swin-T (0.715) despite its high micro-averaged AUC (0.906) was driven by classes with fewer than five positive test instances, for which per-class AUC is unstable; over the 15 classes with at least five positives, Swin-T reached a macro AUC-ROC of 0.822 and Hybrid-EffSwin 0.842. Supplementary metrics, including macro-averaged values over all 23 classes, are reported in Additional file 1 (Table S1).

Figure 2 summarizes these results graphically, with Hybrid-EffSwin achieving the best value on all six metrics. The micro-averaged ROC curves (Fig. 3) show

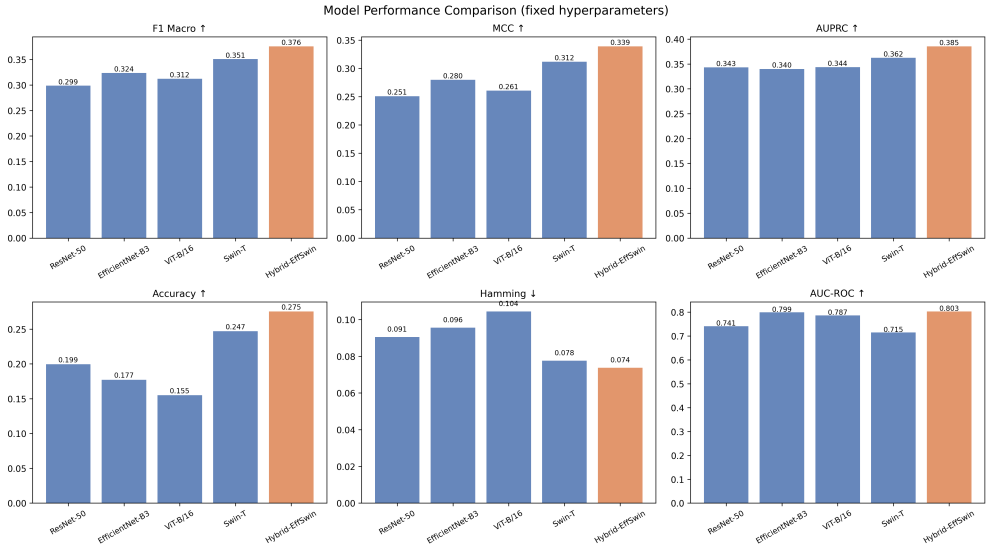


Fig. 2 Performance comparison of all five architectures across six primary metrics on the BERD test set (fixed hyperparameters).

Hybrid-EffSwin (AUC = 0.911) and Swin-T (AUC = 0.906) achieving the highest discriminative performance, and the micro-averaged precision–recall curves (Fig. 4) show the same ranking; macro-averaged AUPRC values are reported in Table 1.

Per-class performance

Per-class F1-scores for all five architectures are presented in Additional file 1 (Table S2). High-performing classes (mean F1 > 0.50) included blood (mean = 0.690, range 0.60–0.78), congested (0.592, range 0.57–0.62), edematous (0.592, range 0.56–0.63), tube (0.584, range 0.42–0.71), normal (0.580, range 0.54–0.60), sputum (0.564, range 0.53–0.62), and neoplasm (0.526, range 0.45–0.62). These categories correspond to the most prevalent findings in the training set and exhibit visually distinctive endoscopic characteristics.

The per-class F1 distribution is visualized in Fig. 5. Four evaluable classes, fistula ($n = 1$), pigmentation ($n = 1$), new organism ($n = 2$), and postoperative change ($n = 3$), had F1 = 0.000 across all architectures, and mass and widened had no positive test

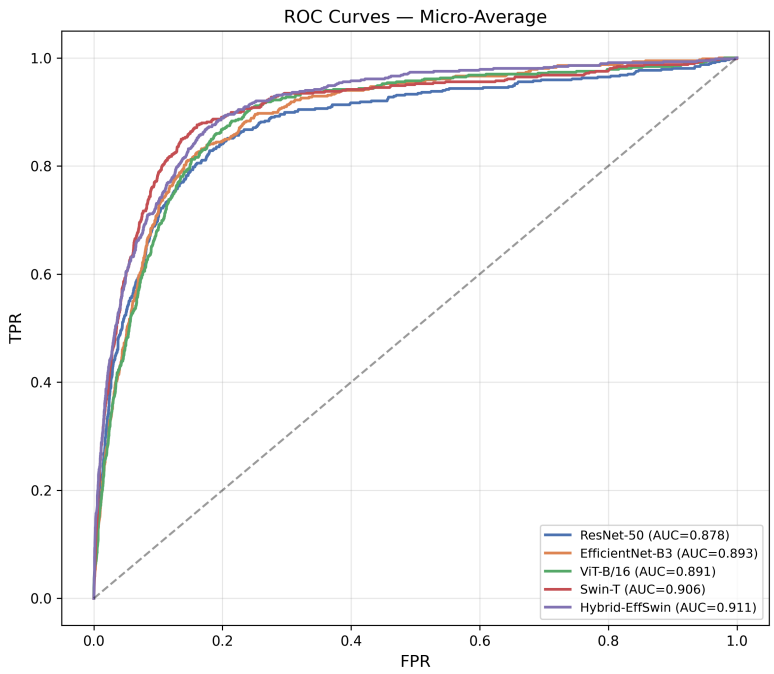


Fig. 3 Micro-averaged ROC curves for all five architectures on the BERD test set.

instances, reflecting the severe rarity of these categories. The per-class MCC pattern was highly consistent with F1 ($r = 0.96$ across all model–class pairs; Additional file 1, Table S2). These correspond to the rarest categories in both the training and test sets, confirming that class imbalance remains the primary performance bottleneck.

Notable architecture-specific findings emerged from this per-class analysis. Of the 17 classes with non-zero F1, Hybrid-EffSwin achieved the highest F1-score for six (blood 0.78, narrow 0.45, neoplasm 0.62, surgical stump 0.67, tube 0.71, and ulcer 0.33) and tied with Swin-T for first on two (granulation 0.22 and normal 0.60), whereas Swin-T led on five (congested 0.62, edematous 0.63, nodules 0.48, rough 0.30, and sputum 0.62). EfficientNet-B3 and ViT-B/16 each led on two categories (infiltration changes and necrotic, and clot and external pressure, respectively), whereas ResNet-50 did not achieve the highest F1 for any individual category, although it remained competitive across most classes.

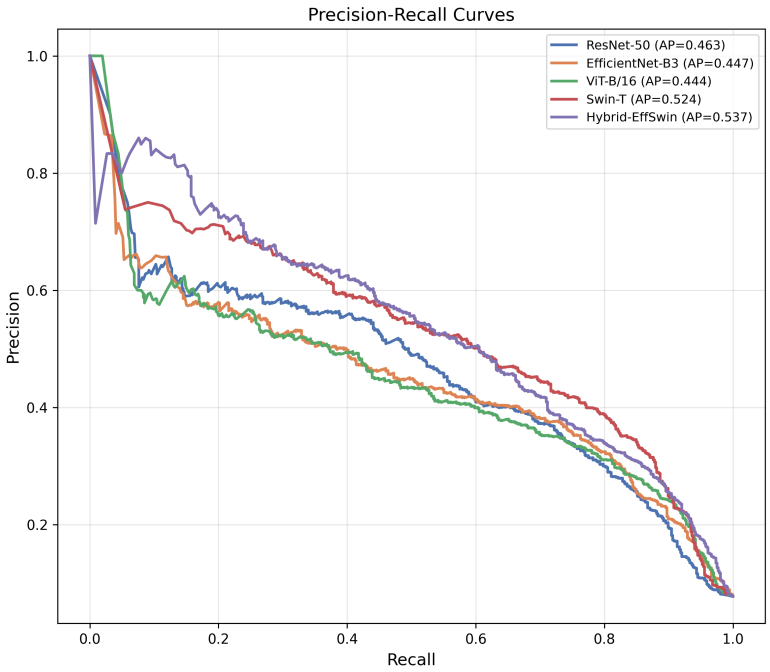


Fig. 4 Micro-averaged precision–recall curves for all five architectures on the BERD test set.

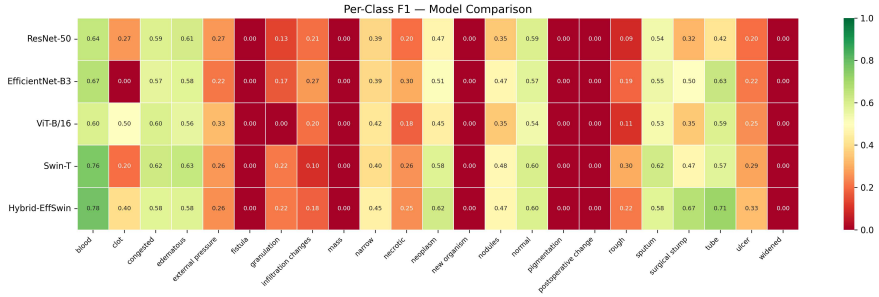


Fig. 5 Per-class F1-score heatmap across all five architectures and 23 BERD finding categories; mass and widened had no positive test instances and are shown for completeness only.

Class-level confusion matrices for the six most prevalent categories (Fig. 6) reveal a distinct error asymmetry between the normal class and the five pathological categories. For normal, recall was high (90.6%; 48/53) with a low false-negative rate (9.4%), but precision was comparatively low (44.4%), reflecting 60 false-positive predictions; the model therefore frequently over-assigned the normal label when no strong

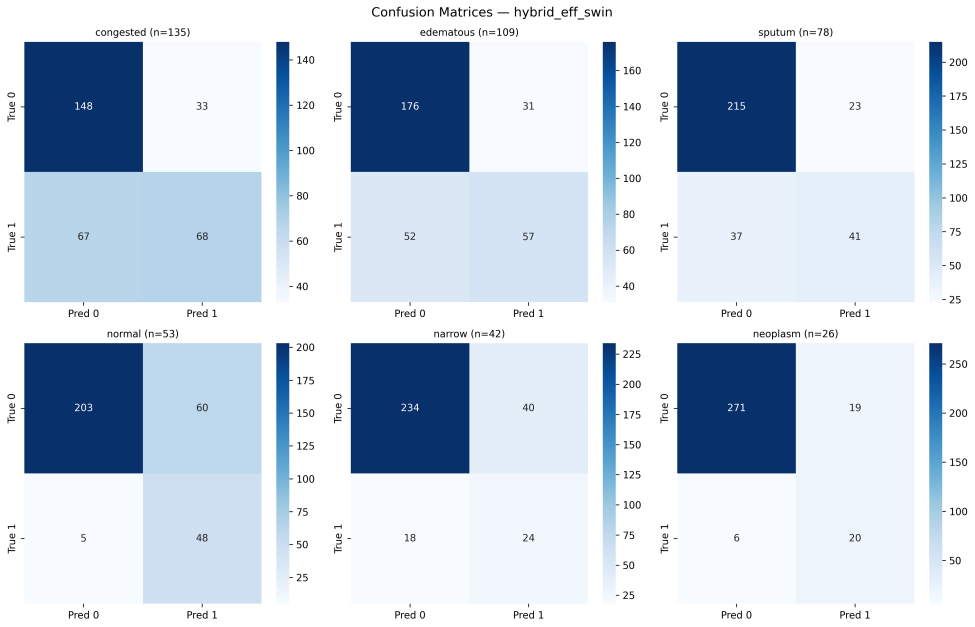


Fig. 6 Confusion matrices for Hybrid-EffSwin on the six most frequent BERD finding categories (test set).

pathological feature was detected. In contrast, the five pathological categories showed the opposite pattern: recall ranged from 50.4% (congested) to 76.9% (neoplasm), with congested, edematous, sputum, and narrow each missing 43–50% of true positive cases at the default threshold ($\theta = 0.5$). This asymmetry, a conservative bias toward "no finding" alongside substantial under-detection of active pathology, is clinically consequential, as it suggests that a meaningful proportion of true pathological findings would be missed under default deployment settings, motivating the threshold calibration analysis described below.

Statistical significance

Patient-level paired bootstrap analysis (Table 3) showed that Hybrid-EffSwin achieved higher macro F1 than ResNet-50 (ΔF1 = +0.077, 95% CI 0.039–0.120), ViT-B/16 (+0.063, 0.017–0.112), and EfficientNet-B3 (+0.052, 0.009–0.094), whereas its difference from Swin-T was not significant (+0.025, −0.011 to 0.061; $p = 0.182$). Swin-T

Table 2 Patient-level pairwise Wilcoxon signed-rank tests (180 patients) on per-patient mean example-based F1 and per-patient exact-match rate. *P*-values adjusted with the Holm method across 10 comparisons; $^{*}p$ (Holm) < 0.05. Image-level tests are given in Additional file 1 (Table S4).

		Example-based F1			Exact-match rate		
Model 1	Model 2	W	p	p (Holm)	W	p	p (Holm)
ResNet-50	EfficientNet-B3	3195.5	0.255	1.000	388.5	0.425	0.463
ResNet-50	ViT-B/16	3860.5	0.971	1.000	208.0	0.029	0.142
ResNet-50	Swin-T	2180.0	<0.001	0.005*	224.5	0.033	0.142
ResNet-50	Hybrid-EffSwin	2335.0	0.001	0.010*	390.0	0.004	0.025*
EfficientNet-B3	ViT-B/16	4022.5	0.331	1.000	305.5	0.232	0.463
EfficientNet-B3	Swin-T	3154.0	0.039	0.196	232.5	0.009	0.056
EfficientNet-B3	Hybrid-EffSwin	2380.0	0.016	0.096	226.5	<0.001	0.001*
ViT-B/16	Swin-T	2539.5	0.001	0.010*	150.5	<0.001	0.001*
ViT-B/16	Hybrid-EffSwin	3020.0	0.003	0.019*	180.0	<0.001	<0.001*
Swin-T	Hybrid-EffSwin	2977.5	0.588	1.000	468.5	0.149	0.446

Table 3 Paired differences in macro F1 and MCC (first model minus second model) with patient-level bootstrap 95% CIs (2000 resamples), $\theta = 0.5$. Bootstrap p-values are unadjusted.

Comparison	ΔF1 [95% CI]	p	ΔMCC [95% CI]	p
Hybrid-EffSwin − ResNet-50	+0.077 [0.039, 0.120]	<0.001	+0.088 [0.046, 0.134]	<0.001
Hybrid-EffSwin − EfficientNet-B3	+0.052 [0.009, 0.094]	0.019	+0.059 [0.013, 0.104]	0.013
Hybrid-EffSwin − ViT-B/16	+0.063 [0.017, 0.112]	0.009	+0.078 [0.031, 0.131]	0.003
Hybrid-EffSwin − Swin-T	+0.025 [−0.011, 0.061]	0.182	+0.027 [−0.012, 0.067]	0.192
Swin-T − ResNet-50	+0.052 [0.020, 0.091]	0.001	+0.061 [0.025, 0.105]	<0.001
Swin-T − EfficientNet-B3	+0.027 [−0.006, 0.066]	0.118	+0.032 [−0.003, 0.072]	0.080
Swin-T − ViT-B/16	+0.038 [−0.010, 0.091]	0.098	+0.051 [0.001, 0.107]	0.046
EfficientNet-B3 − ResNet-50	+0.025 [−0.004, 0.060]	0.107	+0.029 [−0.004, 0.068]	0.088
ViT-B/16 − ResNet-50	+0.014 [−0.026, 0.057]	0.486	+0.010 [−0.034, 0.055]	0.679
EfficientNet-B3 − ViT-B/16	+0.011 [−0.034, 0.059]	0.585	+0.019 [−0.028, 0.073]	0.416

outperformed ResNet-50 (+0.052, 0.020–0.091), but its differences from EfficientNet-B3 and ViT-B/16 had CIs including zero. The three single-paradigm baselines did not differ significantly from one another, and ΔMCC estimates showed the same pattern.

The patient-level tests (Table 2) were consistent with the bootstrap analysis. After Holm correction, per-patient example-based F1 differed between Hybrid-EffSwin and ResNet-50 ($p = 0.010$) and ViT-B/16 ($p = 0.019$), and between Swin-T and ResNet-50 ($p = 0.005$) and ViT-B/16 ($p = 0.010$); per-patient exact-match rate additionally separated Hybrid-EffSwin from EfficientNet-B3 ($p = 0.001$). Hybrid-EffSwin and Swin-T did not differ on either measure (p (Holm) = 1.000 and 0.446), indicating that they form a top tier whose members cannot be distinguished on the current test set.

Table 4 Robustness to training seed. Mean ± SD over three training runs (seeds 42, 7, and 123) with a fixed patient-level split; macro metrics over the 21 evaluable classes, $\theta = 0.5$. Rank = mean rank in macro F1 across seeds. The last three columns give ΔF1 (Hybrid-EffSwin minus model) within seed 42 (S42), seed 7 (S7), and seed 123 (S123); *95% patient-level bootstrap CI excludes zero. Bold = best per column.

					ΔF1 (Hybrid − model)		
Model	F1 macro	MCC	AUPRC	Rank	S42	S7	S123
ResNet-50	0.307±0.011	0.265±0.014	0.341±0.015	4.3	0.077*	0.041*	0.047*
EfficientNet-B3	0.328±0.011	0.284±0.010	0.347±0.007	2.7	0.052*	0.020	0.030
ViT-B/16	0.309±0.005	0.259±0.006	0.332±0.011	4.7	0.063*	0.050*	0.048*
Swin-T	0.340±0.011	0.297±0.014	0.354±0.010	2.3	0.025	0.020	0.022
Hybrid-EffSwin	**0.363±0.013**	**0.322±0.016**	**0.371±0.014**	**1.0**	–	–	–

Robustness to training seed

Across three training runs, Hybrid-EffSwin ranked first in macro F1 in every run (mean F1 = 0.363 ± 0.013; mean rank 1.0), followed by Swin-T (0.340 ± 0.011; mean rank 2.3) and EfficientNet-B3 (0.328 ± 0.011; mean rank 2.7), with ViT-B/16 (0.309 ± 0.005) and ResNet-50 (0.307 ± 0.011) ranking last (Table 4). Within each seed, the advantage of Hybrid-EffSwin over ResNet-50 and ViT-B/16 was statistically supported in all three runs, whereas its advantage over EfficientNet-B3 was supported in one of three runs. The margin over Swin-T was consistently positive (ΔF1 = +0.020 to +0.025) but its 95% CI included zero in every run, confirming that the two architectures form a top tier that cannot be separated on the current test set.

Threshold calibration

Applying per-class thresholds selected on the validation subset changed test-set macro F1 by a mean of +0.016 across the five architectures (range −0.005 to +0.037; Table 5, Fig. 7). The largest gains were observed for the weaker baselines (ResNet-50: +0.037; EfficientNet-B3 and ViT-B/16: +0.020), whereas the gain was small for Swin-T (+0.010) and absent for Hybrid-EffSwin (−0.005). The model ranking was unchanged, with Hybrid-EffSwin (F1 = 0.371) and Swin-T (F1 = 0.361) remaining

Table 5 Per-class threshold calibration. θ_{val} denotes per-class thresholds selected on the validation subset and applied unchanged to the test set; ΔF1 = F1(θ_{val}) − F1(0.5). Macro F1 over the 21 evaluable classes. Bold = best per column.

Model	F1 ($\theta = 0.5$)	F1 (θ_{val})	ΔF1
ResNet-50	0.299	0.335	+0.037
EfficientNet-B3	0.324	0.344	+0.020
ViT-B/16	0.312	0.333	+0.020
Swin-T	0.351	0.361	+0.010
Hybrid-EffSwin	**0.376**	**0.371**	−0.005

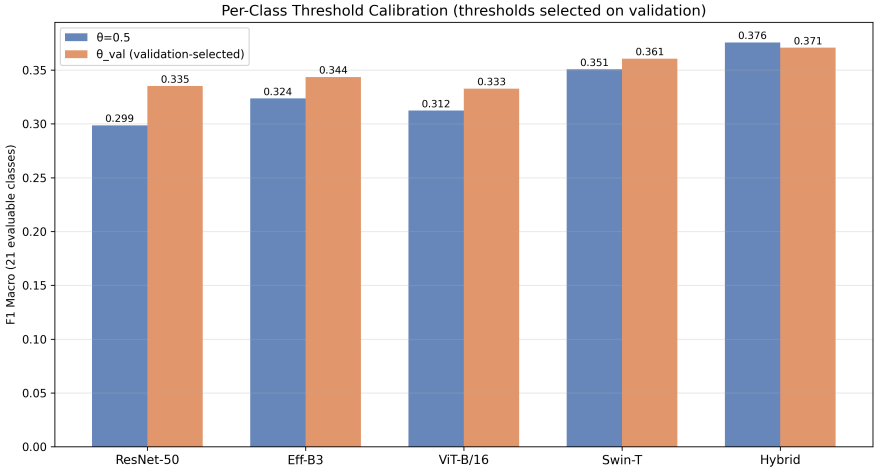


Fig. 7 Test-set macro F1-score at the default threshold ($\theta = 0.5$) and at per-class thresholds selected on the validation subset (θ_{val}) for all five architectures.

the two best-performing architectures, indicating that validation-based threshold calibration yields modest, model-dependent benefits and does not alter the comparative conclusions drawn at $\theta = 0.5$.

Table 6 Fusion strategy ablation for Hybrid-EffSwin (test set, $n = 316$). Δ values are average pooling minus the alternative, with patient-level bootstrap 95% CIs. Bold = best per column.

Variant	F1 macro	MCC	ΔF1 [95% CI]	ΔMCC [95% CI]
Average pooling (proposed)	**0.376**	**0.339**	–	–
Residual-projection	0.336	0.287	+0.039 [−0.006, 0.088]	+0.052 [0.002, 0.102]
Concatenation	0.330	0.287	+0.045 [−0.001, 0.093]	+0.052 [0.003, 0.103]

Ablation studies

Fusion strategy ablation

To validate the average-pooling fusion strategy selected for Hybrid-EffSwin, three fusion configurations were evaluated under identical training hyperparameters: average pooling, residual-projection fusion, and concatenation with linear projection (see *Ablation study design*). As shown in Table 6, average pooling achieved the highest macro F1 and MCC among the three strategies. It was also best on the validation subset on which the choice was based (macro F1 = 0.407 vs. 0.401 and 0.385). The MCC advantage over residual-projection (ΔMCC = +0.052, 95% CI 0.002–0.102) and concatenation (+0.052, 0.003–0.103) was statistically supported, whereas the corresponding F1 differences (+0.039 and +0.045) had 95% CIs marginally including zero. As illustrated in Fig. 8, average pooling also yielded higher exact-match accuracy (0.275 vs. 0.196 and 0.193, respectively) and lower Hamming loss (0.074 vs. 0.087 and 0.092, respectively), indicating that its advantage extended to both set-level and label-wise error measures.

Loss function ablation

The four single-backbone architectures were trained under both class-weighted BCE and ASL, with all other hyperparameters held constant (see *Ablation study design*). As shown in Table 7, BCE yielded higher macro F1 than ASL for three of the four architectures at $\theta = 0.5$, although the difference was statistically supported only for

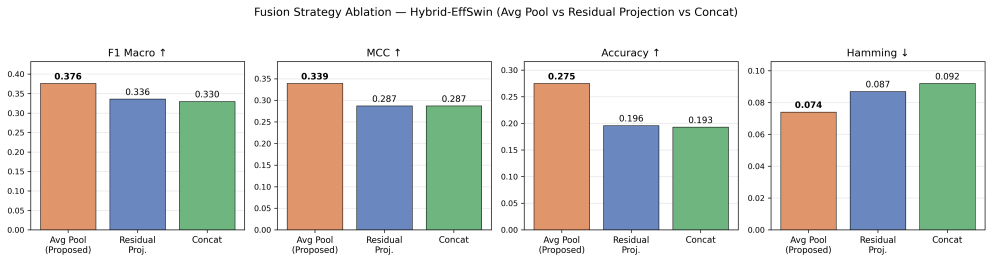


Fig. 8 Fusion strategy ablation for Hybrid-EffSwin: average pooling vs. residual projection vs. concatenation across four metrics (macro F1, MCC, accuracy, and Hamming loss).

Table 7 BCE vs. Asymmetric Loss ablation (test set, $n = 316$). ΔF1 = F1(BCE) − F1(ASL) with patient-level bootstrap 95% CI.

Model	F1 (BCE)	F1 (ASL)	ΔF1 [95% CI]	AUC-ROC (BCE)	AUC-ROC (ASL)
ResNet-50	0.299	0.300	−0.001 [−0.038, 0.038]	0.741	0.780
EfficientNet-B3	0.324	0.294	+0.029 [−0.012, 0.074]	0.799	0.764
ViT-B/16	0.312	0.274	+0.039 [0.006, 0.070]	0.787	0.807
Swin-T	0.351	0.330	+0.020 [−0.033, 0.083]	0.715	0.797

ViT-B/16 (ΔF1 = +0.039, 95% CI 0.006–0.070); for ResNet-50 the two losses were equivalent (ΔF1 = −0.001). On the validation subset, BCE also yielded higher macro F1 than ASL for all four architectures. In contrast, ASL yielded higher threshold-independent AUC-ROC for three of the four architectures (e.g., Swin-T: 0.797 vs. 0.715), suggesting that its effect lies in ranking rather than in calibration at the default threshold. As shown in Fig. 9, BCE produced lower (better) Brier scores for all four architectures (ResNet-50: 0.074 vs. 0.107; EfficientNet-B3: 0.078 vs. 0.097; ViT-B/16: 0.084 vs. 0.107; Swin-T: 0.067 vs. 0.108), indicating that ASL's asymmetric down-weighting of easy negatives came at the cost of probability calibration in this dataset. Hybrid-EffSwin was not evaluated under ASL, as its fusion strategy was already established via the fusion ablation; extending the loss-function comparison to the hybrid architecture is left for future work.

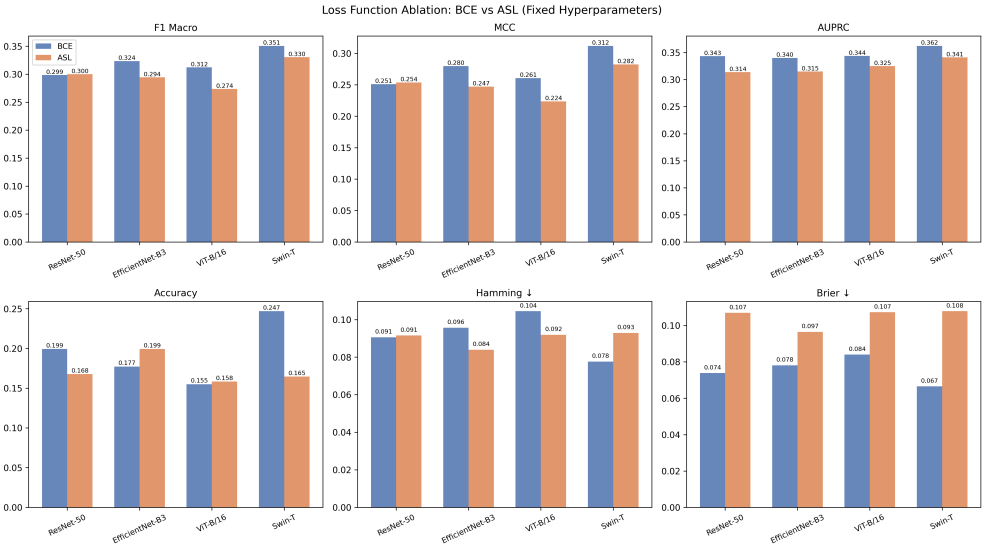


Fig. 9 Loss function ablation (BCE vs. Asymmetric Loss) across six metrics (macro F1, MCC, AUPRC, accuracy, Hamming loss, and Brier score) for four architectures under fixed hyperparameters.

Comparison with contemporary and foundation models

Hybrid-EffSwin retained the highest macro F1-score and MCC among all eight evaluated models (Table 8). DINOv2 ViT-B/14 performed comparably (macro F1 = 0.362; ΔF1 = +0.013, 95% CI −0.025 to 0.052) and achieved the highest exact-match accuracy (0.307), the lowest Hamming loss (0.069), and the lowest Brier score (0.062) of all models. BiomedCLIP ViT-B/16 (macro F1 = 0.340) did not differ significantly from Hybrid-EffSwin (ΔF1 = +0.036, 95% CI −0.010 to 0.102) or Swin-T, whereas ConvNeXt-T (macro F1 = 0.324) was significantly inferior to Hybrid-EffSwin (ΔF1 = +0.052, 95% CI 0.004–0.103) and performed at the level of EfficientNet-B3. A late-fusion ensemble of separately trained EfficientNet-B3 and Swin-T (macro F1 = 0.359, MCC = 0.319) performed close to Hybrid-EffSwin: Hybrid-EffSwin was numerically better in all three seeds (ΔF1 = +0.017, +0.007, and +0.020), but none of these differences was significant (seed 42: 95% CI −0.019 to 0.052).

Table 8 Comparison with contemporary and foundation-model baselines (secondary analysis; test set, $n = 316$, $\theta = 0.5$, single training run). Macro metrics over the 21 evaluable classes. ΔF1 = Hybrid-EffSwin minus model with patient-level bootstrap 95% CI. Bold = best per column; $\downarrow$ = lower is better. IN-1k = ImageNet-1k; LVD-142M = self-supervised pretraining on 142M natural images; PMC-15M = 15M biomedical image–text pairs.

Model	Pretraining	F1 macro	MCC	Exact	Brier ↓	ΔF1 [95% CI]
Hybrid-EffSwin	IN-1k	**0.376**	**0.339**	0.275	0.063	–
DINOv2 ViT-B/14	LVD-142M	0.362	0.318	**0.307**	**0.062**	0.013 [−0.025, 0.052]
Swin-T	IN-1k	0.351	0.312	0.247	0.067	0.025 [−0.011, 0.061]
BiomedCLIP ViT-B/16	PMC-15M	0.340	0.292	0.180	0.074	0.036 [−0.010, 0.102]
ConvNeXt-T	IN-1k	0.324	0.275	0.187	0.070	0.052 [0.004, 0.103]

Training dynamics and convergence

Training was run for up to 70 epochs with early stopping on validation macro F1 (patience = 15 epochs). Early stopping was triggered for ViT-B/16 (epoch 36; best validation epoch 21), Hybrid-EffSwin (epoch 47; best validation epoch 32), and Swin-T (epoch 64; best validation epoch 49), whereas ResNet-50 and EfficientNet-B3 reached the 70-epoch limit (Additional file 1, Figure S1). All architectures showed a rapid decrease in training loss during the first 10 epochs. ViT-B/16 reached a comparatively high validation macro F1 early in training but fluctuated considerably and did not improve after epoch 21, consistent with the tendency of pure Vision Transformers to overfit limited training data in the absence of convolutional inductive biases [8]; this aligns with reported differences in the representations learned by CNNs and Vision Transformers [35, 36]. Swin-T attained the highest validation macro F1 among the single-backbone models, benefiting from the locality bias of shifted-window self-attention [21], whereas the convolutional backbones (ResNet-50 and EfficientNet-B3) improved more gradually and plateaued at lower values [19, 20]. Hybrid-EffSwin showed the fastest rise in validation macro F1 during the first 10 epochs and generally the lowest training loss, reaching a peak validation performance comparable to

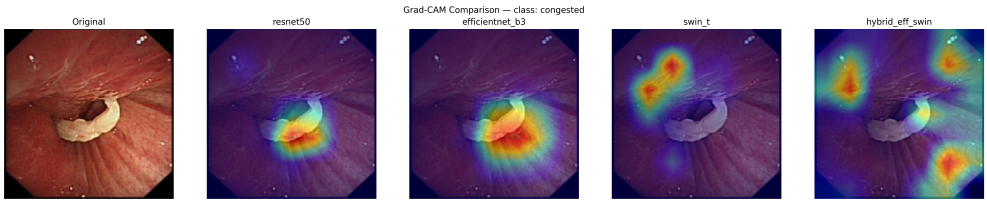


Fig. 10 Representative Grad-CAM visualizations for the congested class across the four architectures with spatial feature maps. From left to right: original endoscopic image, ResNet-50, EfficientNet-B3, Swin-T, and Hybrid-EffSwin. Warmer colors indicate regions of higher class-discriminative activation. ViT-B/16 is not shown, as its architecture does not natively produce a spatial feature map compatible with gradient-based CAM localization.

that of Swin-T; this is consistent with the test-set results, in which the two architectures could not be statistically distinguished (Table 3). MixUp interpolation has been reported to stabilize optimization in long-tailed multi-label settings [37].

Grad-CAM results

Representative Grad-CAM visualizations for the congested class (Fig. 10) show that ResNet-50 and EfficientNet-B3 converged on similar regions along the airway lumen boundary, whereas Swin-T attended to a distinct mucosal surface region on the opposite side of the frame. Hybrid-EffSwin exhibited two areas of high activation: one spatially consistent with the attention of ResNet-50 and EfficientNet-B3, and a second, comparably strong focus near the image periphery that coincided with a specular highlight in the original endoscopic frame rather than an evident pathological feature.

These visualizations provide qualitative, illustrative evidence only and do not constitute a quantitative localization validation: no expert-annotated ground-truth regions were available for formal overlap analysis (e.g., intersection-over-union or pointing-game accuracy), the observations are based on a limited set of representative examples rather than a systematic review of the full test set, and the known sensitivity of gradient-based saliency methods to model parameterization should be considered when interpreting these results. Comparable limitations of Grad-CAM,

including coarse spatial resolution, architecture dependence, and the risk of over-reliance on heatmaps without clinical validation, have been noted in earlier medical imaging studies [38]. The occasional attention to non-pathological image features, such as specular reflections, underscores the importance of expert visual validation before any clinical deployment of interpretability outputs.

Discussion

This study evaluated deep learning architectures for multi-label bronchoscopic finding classification on the publicly available BERD dataset. To address the dual challenge of local mucosal texture recognition and global airway structural reasoning, Hybrid-EffSwin combined EfficientNet-B3 and Swin-T via element-wise average pooling. Among the five primary models, Hybrid-EffSwin achieved the highest performance on all primary metrics (F1 = 0.376, MCC = 0.339, accuracy = 0.275, AUPRC = 0.385, AUC-ROC = 0.803). Among single-paradigm baselines, Swin-T ranked second (F1 = 0.351, MCC = 0.312) and was not statistically distinguishable from Hybrid-EffSwin (bootstrap ΔF1 = +0.025, 95% CI −0.011 to 0.061; Table 3), whereas Hybrid-EffSwin outperformed ResNet-50, EfficientNet-B3, and ViT-B/16. This pattern suggests that hierarchical Transformer architectures already capture much of the representational benefit motivating the hybrid design, with the addition of a CNN branch offering a small numerical gain that could not be confirmed statistically on the current test set. This pattern was stable across training seeds: Hybrid-EffSwin ranked first in all three runs and its margin over Swin-T was consistently positive but never statistically significant, whereas its advantage over EfficientNet-B3 proved less robust than in the primary run (Table 4). A similar hybrid CNN-Transformer approach was recently applied to multi-label retinal disease classification, achieving a mean average precision of 92.86% on the ODIR-5K dataset [39], supporting the broader complementarity

of CNN local features and Transformer global context in multi-label medical imaging. A recent systematic review identified only 20 eligible bronchoscopy AI studies from 1,196 candidates [2], indicating that existing evidence remains largely confined to binary or single-label classification settings. Although direct comparison is limited by differences in modality, label space, and data scale, the mean average precision values obtained in the present study (AUPRC (mAP) = 0.340–0.385 across architectures; Table 1) are comparable to those reported by the nine top-performing teams in the CXR-LT challenge on 26 thorax disease classes (mAP = 0.279–0.372) [40].

Per-class thresholds selected on the validation subset produced only modest and model-dependent changes in test-set macro F1 (Table 5; mean ΔF1 = +0.016, range -0.005 to $+0.037$). The benefit was concentrated in the weaker baselines and was negligible for the two strongest architectures, suggesting that the default threshold was already well suited to these two architectures. Because thresholds were selected without access to the test set, these estimates are free of the optimistic bias that arises when thresholds are tuned on the evaluation data.

The additional baselines refine this picture. A self-supervised foundation model (DINOv2) reached the same performance tier as Hybrid-EffSwin and Swin-T and provided the best-calibrated probabilities and the highest exact-match accuracy, suggesting that large-scale pretraining can compensate for the absence of an explicit CNN branch. This benefit was, however, contingent on a reduced backbone learning rate, as uniform fine-tuning at the rate used for the primary models substantially degraded performance. Domain-specific biomedical pretraining (BiomedCLIP) did not outperform general-purpose pretraining, plausibly because bronchoscopic images are scarce in biomedical image–text corpora, and a modernized CNN (ConvNeXt-T) did not close the gap between convolutional baselines and the top tier. Hybrid-EffSwin thus remains competitive with considerably larger foundation models at less than half their

parameter count (39.4M vs. 87.0M and 86.5M). The late-fusion ensemble shows, however, that most of the benefit of combining a CNN and a hierarchical Transformer can be obtained without joint training; the small but consistent, non-significant advantage of Hybrid-EffSwin suggests at most a modest additional gain from learning a shared representation. The fusion strategy ablation (Table 6) indicated that element-wise average pooling performed at least as well as residual-projection fusion (ΔF1 = +0.039, ΔMCC = +0.052) and concatenation with linear projection (ΔF1 = +0.045, ΔMCC = +0.052), with statistically supported MCC gains. This is consistent with the principle of parsimony under limited training data (2809 training patients): residual-projection and concatenation fusion introduce additional learnable parameters that may require more data to optimize reliably, whereas average pooling, which adds no fusion parameters, imposes a stronger inductive bias suited to data-constrained clinical imaging. The loss function ablation (Table 7) showed that class-weighted BCE yielded higher F1 than ASL at $\theta = 0.5$ for ViT-B/16 (ΔF1 = +0.039), EfficientNet-B3 (+0.029), and Swin-T (+0.020), with only the ViT-B/16 difference statistically supported, and equivalent performance for ResNet-50 (ΔF1 = -0.001). For the architectures favoring BCE, the simultaneous application of ASL's negative down-weighting ($\gamma_{\text{neg}} = 4.0$) and positive-weight capping (maximum weight = 50.0) may create gradient conflicts, as both mechanisms independently suppress easy-negative gradients; this interaction warrants closer examination in future multi-label bronchoscopy studies.

Hyperparameters identified by TPE with Hyperband pruning modestly outperformed the fixed configuration for three of five architectures, namely ResNet-50 (ΔF1 = +0.037), EfficientNet-B3 (+0.031), and, marginally, ViT-B/16 (+0.003), whereas the fixed hyperparameters performed slightly better for Swin-T (ΔF1 = +0.009) and Hybrid-EffSwin (+0.012). The consistently small magnitude of these differences, including for the two-backbone Hybrid-EffSwin architecture, suggests that the fixed configuration was already a reasonably well-chosen starting point; nonetheless, the

modest advantage of the searched configurations for the three single-backbone architectures indicates that per-architecture hyperparameter search remains worthwhile when computational budget allows [25].

The Grad-CAM analysis (Fig. 10) provided qualitative support for the plausibility of model attention, with ResNet-50 and EfficientNet-B3 attending to the airway lumen boundary in the example shown. However, the occasional attention of Hybrid-EffSwin to a specular highlight rather than a pathological feature illustrates a broader limitation of gradient-based saliency methods in endoscopic imaging, where illumination artifacts can resemble discriminative texture patterns; this observation reinforces the need for expert visual validation prior to any clinical deployment of interpretability outputs.

Several limitations warrant acknowledgment. The held-out test set (316 images, 180 patients) is adequate for overall benchmarking but limits per-class evaluation of rare categories: four evaluable classes had F1 = 0.000 across all architectures and two classes had no positive test instances (fistula and pigmentation: $n = 1$; new organism: $n = 2$; postoperative change: $n = 3$; mass and widened: $n = 0$). Seed replicates were limited to three training runs per architecture and varied only training stochasticity, not the data split; with three runs, between-seed variability can be described but not formally tested. Cross-validation over a pooled patient set, token-level cross-attention, and input resolutions above 224×224 were not evaluated, and no comparison with bronchoscopists was performed; at the observed performance level, none of the models is suitable for unsupervised clinical use. The additional baselines were evaluated with a single training run and, for the foundation models, a reduced backbone learning rate; more extensive tuning could change their relative standing. In addition, the official BERD split contains 34 pixel-identical image pairs between our training and validation subsets and three between the training and test sets despite distinct patient identifiers, which may reflect the reuse of images across examination reports; their

effect on test performance was negligible (Additional file 1, Table S3). Because BERD labels were derived from report captions rather than from independent image-level annotation, some residual label noise cannot be excluded, particularly for rare and visually ambiguous categories. The BERD dataset was collected at a single institution, which may limit generalizability to bronchoscopic images acquired with different equipment or clinical conventions. Grad-CAM visualizations provide qualitative, illustrative evidence only, as no expert-annotated ground-truth regions were available for formal overlap analysis. Finally, consistent with a recent comprehensive review of AI in bronchoscopy and endobronchial ultrasound [41], the absence of external multi-institutional validation represents a pervasive limitation across the field.

Conclusions

This study evaluated deep learning architectures for multi-label bronchoscopic finding classification on the BERD dataset. Five architectures spanning CNN, Transformer, and hybrid paradigms were benchmarked across 23 pathological categories under severe class imbalance. Hybrid-EffSwin, combining EfficientNet-B3 and Swin-T via element-wise average pooling, achieved the highest performance among the five primary architectures on all primary metrics (F1 = 0.376, MCC = 0.339, AUC-ROC = 0.803), ranked first in all three training runs, and consistently outperformed ResNet-50 and ViT-B/16, although its advantages over EfficientNet-B3 and Swin-T were not robustly significant. A self-supervised foundation model (DINOv2) performed comparably to Hybrid-EffSwin and achieved the best calibration, whereas biomedical pretraining (BiomedCLIP) and a modernized CNN (ConvNeXt-T) offered no advantage. Weighted BCE yielded higher F1 than ASL at the default threshold in three of the four architectures evaluated under both losses, whereas ASL improved threshold-independent discrimination. Parameter-efficient average pooling performed at least as well as residual-projection and concatenation fusion under the data constraints of

the BERD benchmark, and validation-based threshold calibration yielded only modest gains without changing the model ranking. A late-fusion ensemble of the two constituent backbones performed similarly to Hybrid-EffSwin. These results provide reference performance levels for future automated bronchoscopy assistance systems.

Supplementary information. Additional file 1: Table S1. Supplementary performance metrics. Table S2. Per-class F1-score and MCC. Table S3. Sensitivity analysis excluding pixel-identical test images. Table S4. Image-level Wilcoxon and McNemar tests. Figure S1. Training dynamics.

Abbreviations.

AI	Artificial intelligence
AMP	Automatic mixed precision
ASL	Asymmetric loss
AUC-ROC	Area under the receiver operating characteristic curve
AUPRC	Area under the precision–recall curve
BCE	Binary cross-entropy
BERD	Bronchoscopy Examination Report Dataset
CAD	Computer-aided diagnosis
CI	Confidence interval
CNN	Convolutional neural network
GELU	Gaussian error linear unit
Grad-CAM	Gradient-weighted class activation mapping
LN	Layer normalization
mAP	Mean average precision
MCC	Matthews correlation coefficient
SD	Standard deviation
TPE	Tree-structured Parzen Estimator
UAAL	Upper Airway Anatomical Landmark
ViT	Vision Transformer

Declarations

Ethics approval and consent to participate. Not applicable. This study exclusively analysed the publicly available, de-identified Bronchoscopy Examination Report Dataset (BERD). The original data collection was approved by the Clinical Research and Laboratory Animal Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University (Ethical Review No. [2024]517), which permitted data collection, annotation, subsequent research, and publication [14]. No new human data were collected, and no interaction with patients took place.

Consent for publication. Not applicable.

Availability of data and materials. The dataset analysed during the current study is available in the Science Data Bank repository, https://doi.org/10.57760/sciencedb.28018 [14]. The code, including the training pipeline and the evaluation scripts used to produce all reported results, is available at https://github.com/SibelKervanci/BERD-multilabel-bronchoscopy.

Competing interests. The author declares no competing interests.

Funding. This research was funded by the Scientific Research Projects Coordination Unit of Gaziantep University (grant number BAP BNG.25.01). The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Authors' contributions. ISK conceived and designed the study, developed the software, performed the experiments and statistical analyses, interpreted the results, prepared the figures, and wrote and revised the manuscript. The author read and approved the final manuscript.

Acknowledgements. Not applicable.

Clinical trial number. Not applicable.

References

[1] Shen YC, Chen CH, Tu CY. Advances in Diagnostic Bronchoscopy. Diagnostics. 2021 Oct;11(11):1984. https://doi.org/10.3390/diagnostics11111984.

[2] Cold KM, Vamadevan A, Laursen CB, Bjerrum F, Singh S, Konge L. Artificial intelligence in bronchoscopy: a systematic review. European Respiratory Review. 2025 Apr;34(176):240274. https://doi.org/10.1183/16000617.0274-2024.

[3] Lepcha DC, Goyal B, Dogra A, Alkhayyat A, Sahu PK, Ali A, et al. Deep Learning in Medical Image Analysis: A Comprehensive Review of Algorithms, Trends, Applications, and Challenges. Computer Modeling in Engineering & Sciences. 2025;145(2):1487–1573. https://doi.org/10.32604/cmes.2025.070964.

[4] Keuth R, Heinrich M, Eichenlaub M, Himstedt M. Airway label prediction in video bronchoscopy: capturing temporal dependencies utilizing anatomical knowledge. International Journal of Computer Assisted Radiology and Surgery. 2024 Jan;19(4):713–721. https://doi.org/10.1007/s11548-023-03050-6.

[5] Chen C, Tang F, Herth FJF, Zuo Y, Ren J, Zhang S, et al. Building and validating an artificial intelligence model to identify tracheobronchopathia osteochondroplastica by using bronchoscopic images. Therapeutic Advances in Respiratory Disease. 2024 Jan;18:17534666241253694. https://doi.org/10.1177/17534666241253694.

[6] Setayeshi R, Vahidi J, Kozegar E, Tan T. An end-to-end multi-task deep learning framework for bronchoscopy image classification. Multimedia Systems. 2024 Dec;30(6):361. https://doi.org/10.1007/s00530-024-01579-3.

[7] Sun W, Yan P, Li M, Li X, Jiang Y, Luo H, et al. An accurate prediction for respiratory diseases using deep learning on bronchoscopy diagnosis images.

Journal of Advanced Research. 2025 Oct;76:423–438. https://doi.org/10.1016/j.jare.2024.11.023.

[8] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR); 2021. Available from: https://openreview.net/forum?id=YicbFdNTTy.

[9] Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Medical Image Analysis. 2023 Apr;85:102762. https://doi.org/10.1016/j.media.2023.102762.

[10] Jayaraman G, Meganathan S, Shah SSM, Anuradha M, Sundararajan RK, Rajakumar R. A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification. Scientific Reports. 2025 Nov;15(1):45769. https://doi.org/10.1038/s41598-025-28636-9.

[11] Zhang Q, Li R, Tang P, Zhang J, Yang G, Lin Z, et al. A Hybrid CNN-Transformer Deep Learning Model for Differentiating Benign and Malignant Breast Tumors Using Multi-View Ultrasound Images. Technology in Cancer Research & Treatment. 2026 Apr;25:15330338261447344. https://doi.org/10.1177/15330338261447344.

[12] Bakkouri I, Bakkouri S. UGS-M3F: unified gated swin transformer with multi-feature fully fusion for retinal blood vessel segmentation. BMC Medical Imaging. 2025 Mar;25(1):77. https://doi.org/10.1186/s12880-025-01616-1.

[13] Rokhsati H, Fasihfar Z, Shahriari HR, Rezaee K. Dual-stream token fusion with Swin Transformer and lesion-aware tokens for gastric metaplasia classification in

IoMT-assisted deployment. BMC Medical Imaging. 2026 Jun;26(1):405. https://doi.org/10.1186/s12880-026-02477-y.

[14] Luo X, Huang X, Liang X, Wang J, Gu J, Yi D, et al. Towards Automated Reporting: A Bronchoscopy Report Dataset for Enhancing Multimodality Large Language Models. Scientific Data. 2026 Feb;13(1):339. https://doi.org/10.1038/s41597-026-06692-8.

[15] Atmaca M, Kervancı IS, Olgun N. A Reproducible COCO-Polygon Quality-Control Pipeline Improves Segmentation Stability in Endoscopic Airway Imaging. Diagnostics. 2026 Apr;16(8):1160. https://doi.org/10.3390/diagnostics16081160.

[16] Hao R, Zhang Y, Tang Z, Zhou Y, Seenivasan L, Chan CPL, et al. Upper Airway Anatomical Landmark Dataset for Automated Bronchoscopy and Intubation. Scientific Data. 2025 Dec;12(1):1907. https://doi.org/10.1038/s41597-025-06169-0.

[17] Bayhan BU, Kervancı IS, Kişi TG, Balık BC, Atmaca M, Altinsoy KE. Predicting difficult laryngoscopy via three-dimensional graphical modeling and hybrid artificial intelligence using standard facial photographs in elective surgery patients. BMC Anesthesiology. 2026 Aug;In press. https://doi.org/10.1186/s12871-026-04160-9.

[18] Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks. 2018 Oct;106:249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

[19] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 770–778. Available from: http://ieeexplore.ieee.org/document/7780459/.

[20] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). vol. 97 of Proceedings of Machine Learning Research; 2019. p. 6105–6114.

[21] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. p. 9992–10002. Available from: https://ieeexplore.ieee.org/document/9710580/.

[22] Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Medical Imaging. 2022 Dec;22(1):69. https://doi.org/10.1186/s12880-022-00793-7.

[23] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (ICLR); 2019. Available from: https://openreview.net/forum?id=Bkg6RiCqY7.

[24] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA: ACM; 2019. p. 2623–2631. Available from: https://dl.acm.org/doi/10.1145/3292500.3330701.

[25] Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research. 2018;18(185):1–52.

[26] Ridnik T, Ben-Baruch E, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric Loss for Multi-Label Classification. In: 2021 IEEE/CVF International

Conference on Computer Vision (ICCV); 2021. p. 82–91.

[27] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 11976–11986.

[28] Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, et al. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research. 2024;Available from: https://openreview.net/forum?id=a68SUt6zFt.

[29] Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. NEJM AI. 2025;2(1):AIoa2400640. https://doi.org/10.1056/AIoa2400640.

[30] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015 Mar;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.

[31] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Dec;21(1):6. https://doi.org/10.1186/s12864-019-6413-7.

[32] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020 Feb;128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7.

[33] Apostolopoulos ID, Papathanasiou ND, Papandrianos NI, Papageorgiou EI, Panayiotakis GS. Deep Learning Assessment for Mining Important Medical Image Features of Various Modalities. Diagnostics. 2022 Sep;12(10):2333. https:

//doi.org/10.3390/diagnostics12102333.

[34] Vanitha K, Mahesh TR, Kumar VV, Guluwadi S. Enhanced tuberculosis detection using Vision Transformers and explainable AI with a Grad-CAM approach on chest X-rays. BMC Medical Imaging. 2025 Mar;25(1):96. https://doi.org/10.1186/s12880-025-01630-3.

[35] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do Vision Transformers See Like Convolutional Neural Networks? In: Advances in Neural Information Processing Systems. vol. 34; 2021. p. 12116–12128.

[36] Guo J, Han K, Wu H, Tang Y, Chen X, Wang Y, et al. CMT: Convolutional Neural Networks Meet Vision Transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022. p. 12165–12175. Available from: https://ieeexplore.ieee.org/document/9879701/.

[37] Han S, Choi E, Lim C, Shim H, Lee J. Long-tail Mixup for Extreme Multi-label Classification. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta GA USA: ACM; 2022. p. 3998–4002. Available from: https://dl.acm.org/doi/10.1145/3511808.3557632.

[38] Kumaran S Y, Jeya JJ, Mahesh TR, Khan SB, Alzahrani S, Alojail M. Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam. BMC Medical Imaging. 2024 Jul;24(1):176. https://doi.org/10.1186/s12880-024-01345-x.

[39] Zhao J, Zhu J, He J, Cao G, Dai C. Multi-label classification of retinal diseases based on fundus images using Resnet and Transformer. Medical & Biological Engineering & Computing. 2024 Nov;62(11):3459–3469. https://doi.org/10.1007/

s11517-024-03144-6.

[40] Holste G, Zhou Y, Wang S, Jaiswal A, Lin M, Zhuge S, et al. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. Medical Image Analysis. 2024 Oct;97:103224. https://doi.org/10.1016/j.media.2024.103224.

[41] Winiarski S, Radziszewski M, Wiśniewski M, Cisek J, Wąsowski D, Plewczyński D, et al. Integrating Artificial Intelligence in Bronchoscopy and Endobronchial Ultrasound (EBUS) for Lung Cancer Diagnosis and Staging: A Comprehensive Review. Cancers. 2025 Aug;17(17):2835. https://doi.org/10.3390/cancers17172835.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.docx
- TableS2.docx
- TableS4.docx
- TableS3.docx
- FigureS1.png
- SupplementALL.docx