IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 February 2026, accepted 23 February 2026, date of publication 4 March 2026, date of current version 10 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3670363

RESEARCH ARTICLE

Clinically Oriented CNN–Transformer Architectures for Reliable Bronchoscopic Recognition of Lung Lesions and Anatomical Structures

ROLYPH ERWAN NTOUTOUME NGUEMA[1], MOHAMAD FOROUZANFAR[2], (Senior Member, IEEE), AND ALI TRAORE[1]

[1]Department of Biomedical Engineering, Université Polytechnique de l'Ouest Africain (UPOA), Dakar 50465, Senegal
[2]Department of Systems Engineering, École de Technologie Supérieure (ETS), Montréal QC H3C 1K3, Canada

Corresponding author: Rolyph Erwan Ntoutoume Nguema (rolypherwan4@gmail.com)

ABSTRACT Bronchoscopy is central to diagnosing central lung cancers but remains limited by reliance on operator expertise and variability in visual interpretation. In this work, we adapt and evaluate CNN–Transformer hybrid models for the classification and segmentation of bronchoscopic images, with a particular emphasis on clinically realistic patient-level evaluation. These models combine convolutional blocks, which capture fine-grained local features, with Transformer components that encode long-range dependencies and global context, yielding feature representations well suited to the complexity of bronchoscopic images. The primary objective of this study is to adopt CNN-Transformer hybrid architectures for bronchoscopic lesion and landmark recognition, while evaluating their performance under clinically relevant data partitioning conditions. We evaluate our methods on BM-BronchoLC, a publicly available dataset of 2,921 annotated bronchoscopic images, and present two complementary frameworks: MedViT, a convolution-enhanced vision transformer for multi-label classification, and FCB-SwinV2, a dual-branch design coupling a convolutional encoder with a SwinV2 Transformer U-Net decoder for semantic segmentation. To directly address the study objective, we compare the performance of both models under random image-level splitting and rigorous patient-level partitioning, which prevents leakage of visual patterns between training and testing sets and provides a more clinically realistic evaluation. MedViT achieves 94.7% accuracy (AUC 0.95) for anatomical landmarks under random splitting and preserves comparative performance with 93% (AUC 0.91) under patient-level separation. For lung lesions, results remain competitive at 82.3% (AUC 0.79) and 80% (AUC 0.69), respectively. FCB-SwinV2 yields Dice scores of 0.42 for landmarks and 0.33 for lesions with random splitting, which decline to 0.38 and 0.32 under patient-level evaluation. These results show that while the models maintain overall solid performance, they also exhibit a consistent drop under patient-level validation, underscoring the risk of overestimation when relying solely on random splitting. This controlled comparison between the two evaluation protocols demonstrates that despite the expected decrease in performance when removing data leakage, the proposed architectures remain competitive and generalize effectively to unseen patients. These findings indicate that our adapted CNN–Transformer architectures provide useful baselines for BM-BronchoLC and show encouraging signs of generalization to unseen patients, while also illustrating the performance differences between random and patient-level evaluation. They also reinforce that proper patient-level evaluation is central to the reliability of AI systems, and should be systematically adopted to avoid inflated performance estimates. All code and models are released to support reproducibility and foster future research.

INDEX TERMS Bronchoscopy, classification, CNN-transformer, data leakage, deep learning, segmentation.

I. INTRODUCTION

With a 5-year survival rate of only 15-20%, mainly due to late diagnosis and with half of the cases being discovered

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

at an advanced stage [1], lung cancer is one of the leading causes of death worldwide, with approximately 1.8 million deaths and 2.2 million new cases detected in 2020 [2].

Early recognition of lung cancer lesions is therefore key to improving patient prognoses. Among the multiple diagnostic methods, bronchoscopy stands out by allowing direct visualization of the respiratory tracts and enabling histopathological examinations for confirmation [3]. For central tumors, this medical imaging technique has a sensitivity of up to 88%, but its effectiveness is reduced for peripheral lesions, with sensitivity ranging between 50% and 78% depending on the specific methods used [4].

However, the visual interpretation of bronchoscopic images largely relies on the clinician's experience, leading to significant inter-observer variability [5]. This constitutes a major limitation of this diagnostic tool, as it increases the risk of false negatives and missed lesions, particularly for subtle abnormalities [6]. Thus, integrating Artificial Intelligence-based solutions appears to be a relevant approach to optimizing the interpretation and accuracy of bronchoscopy diagnoses.

AI models have demonstrated high potential for the automated recognition of anatomical structures and cancerous lesions. Specifically, hybrid CNN-Transformer approaches combine the ability of CNNs to extract local features with the power of Transformers to capture global features, allowing for advanced analysis of medical images. Their growing adoption in medical imaging demonstrates their relevance in lung disease classification [7] (98.97% accuracy), brain and gastrointestinal tumor detection [8], as well as organ segmentation in CT scans [9] and polyp segmentation in endoscopy [10]. This suggests that hybrid CNN–Transformer approaches can complement or improve upon traditional architectures in certain imaging tasks, particularly when both local detail and global context are required.

In parallel, recent studies have also explored endoscopic super-resolution techniques to enhance the visual quality of medical images, either by increasing the resolution of stereoscopic endoscopic views [11] or by improving the segmentation of surgical instruments through deep-learning-based approaches [12].

In bronchoscopy, many recent studies have demonstrated the contribution of machine learning and deep learning models to the analysis of bronchoscopic images and clinical decision support: Deng et al. [13] designed a model inspired by ResNet achieving 95% accuracy in distinguishing malignant from benign lesions, while Feng et al. [14] extracted HSV texture features to classify lung cancer subtypes with 86% accuracy. A multi-task model based on DenseNet, capable of differentiating cancer from tuberculosis with 90.6% accuracy, was introduced by the work of Setayeshi et al. in 2024 [15]. For navigation, Banach et al. [16] used a cycle-GAN to estimate depth and align bronchoscopic images with 3D scans, and Chang et al. proposed ESFPNet [17], a real-time segmentation model for autofluorescence lesions achieving a Dice index of 0.76. Yoo et al. trained EfficientNet-B1 [18] to identify bronchial structures, outperforming physicians in accuracy, and Cold et al. demonstrated [17] that AI could guide operators in real time, increasing the number of explored segments (+6 on average). For performance evaluation, Cold et al. developed AIBA, an automated system correlated with human assessments, and Cold et al. [20] proved that AI assistance optimizes bronchoscopy quality, making procedures more comprehensive and standardized.

Nevertheless, one methodological limitation may affect many of these advances: most published works do not explicitly specify their data partitioning strategy, which suggests that random image-level splitting could have been applied. Such practice would implicitly allow data from the same patient to appear in both training and testing sets, raising the risk of overestimated performances due to data leakage. Although often overlooked, this methodological issue questions the true generalization ability of reported models and underscores the importance of adopting stricter patient-level evaluation protocols. In addition, beyond this methodological concern, the field still suffers from a lack of large, publicly available, and richly annotated datasets dedicated to bronchoscopy. This scarcity of benchmark data has limited the reproducibility and objective comparison of AI models, and it has constrained their clinical translation.

In this regard, the BM-BronchoLC dataset, published in 2024 [21], represents a major breakthrough in this field. It is one of the first publicly available, richly annotated datasets specifically designed for bronchoscopic image analysis, enabling the training and evaluation of segmentation and classification models for cancerous lesions and anatomical landmarks. This dataset provides a valuable reference for testing AI architectures and evaluating the performance of lung lesion and anatomical landmark recognition algorithms under consistent and reproducible conditions.

To address the inherent complexities of bronchoscopic image interpretation, this study investigates the application of hybrid CNN–Transformer models for the automated classification and segmentation of lung lesions and anatomical structures. Two established architectures are adapted for this purpose: MedViT, a vision transformer designed for multi-label classification, and FCB-SwinV2, a hybrid segmentation network combining convolutional decoders with Transformer-based encoders. Both models are customized to account for the specific visual and structural challenges of bronchoscopic imagery.

The central contribution of this study is to evaluate CNN–Transformer hybrid architectures for bronchoscopic lesion and landmark recognition under clinically relevant data partitioning conditions. By explicitly comparing random and strict patient-level evaluation protocols, we aim to quantify the impact of data leakage and assess model generalization in a realistic clinical setting, while demonstrating that the adapted models remain competitive and clinically relevant.

Performance is rigorously assessed using a range of metrics—including mean accuracy, AUC, Dice Score, Recall, and confusion matrix analysis—to provide a detailed

understanding of model behavior, reliability, and potential clinical utility. This evaluation framework is designed to support a realistic interpretation of model performance in bronchoscopic practice.

II. METHODS

A. DATASET

BM-BronchoLC stands out as the first public bronchoscopic dataset providing detailed information on the precise localization and identification of anatomical landmarks and airway lesions [21]. It consists of 2,921 images from 208 patients, acquired at Bach Mai Hospital in Vietnam, one of the largest medical centers in the country. Captured using an Olympus bronchoscopy system, these images adhere to strict quality criteria, including a minimum resolution of 480 × 480 pixels and standard white light illumination, ensuring optimal visualization of bronchial structures. Only images free from excessive blur, underexposure, or artifacts were retained.

The annotation, performed by two experienced bronchoscopists (≥5 years of experience) and validated by a senior expert (>10 years of experience), ensures high inter-observer consistency. The dataset covers 11 anatomical landmarks, ranging from the trachea to the segmental bronchi, as detailed in **Table 2**. For lesion classification, 7 types of bronchial abnormalities were annotated—including tumors, mucosal infiltrations, erythema, anthracosis, ulcerations, hypervascularization, and stenoses—as described in **Table 1**, following the classifications of reference bronchoscopic atlases [22].

BM-BronchoLC guarantees a faithful representation of bronchial structures and pulmonary lesions by relying on strict image selection criteria and rigorous annotation. The images are captured in standard white light with uniform resolution, ensuring consistent and optimal quality for AI-assisted analysis. This approach standardizes model training while minimizing biases related to acquisition variations.

TABLE 1. List of lung lesions annotated in BM-BronchoLC.

Class	Description
Mucosal erythema	Abnormal redness indicating inflammation
Anthracosis	Carbon deposits due to inhalation of pollutants
Stenosis	Pathological narrowing of the bronchial pathways
Mucosal edema of carina	Swelling of the mucosal tissue at the carina
Mucosal infiltration	Thickening of the bronchial mucosa
Vascular growth	Abnormal vascularization of bronchial walls
Tumor	Abnormal mass suspected to be malignant

B. IMAGE PREPROCESSING

For classification, images were resized to 224 × 224 pixels, followed by normalization in accordance with ImageNet distributions, thereby facilitating the use of the model's pretrained weights. Conversely, the segmentation task required rescaling to 384 × 384 pixels, accompanied by normalization between 0 and 1, promoting gradient stabilization during training. The masks were converted into binary maps, simplifying the distinction between structures of interest and the background.

Data augmentation was applied exclusively to the training sets but with distinct strategies: in classification, horizontal and vertical flips, random rotations (±20°), and color variations were introduced to enhance sample diversity, whereas in segmentation, more advanced transformations, including transpositions, affine distortions (rotation, shear, scaling), and contrast and brightness adjustments, were implemented to reflect the variability of clinical conditions.

To prevent excessive class imbalance in classification, underrepresented labels (fewer than 20 occurrences) were filtered out, ensuring a more balanced learning process in accordance with the reference BM-BronchoLC benchmark protocol [21], which applies the same threshold to avoid unstable training on extremely rare labels. Finally, a systematic verification of image-mask correspondence was implemented in segmentation, ensuring annotation integrity and preventing any inconsistency between inputs and their ground truths.

TABLE 2. List of annotated anatomical landmarks in BMBronchoLC.

Class	Description
Vocal cords	Fold structures in the larynx that produce sound
Main carina	Ridge at the bifurcation of the trachea separating the openings of the main bronchi
Intermediate bronchus	Bronchial passage between the right main bronchus and lobar bronchi
Right superior lobar bronchus	Branch of the right main bronchus leading to the superior lobe
Right inferior lobar bronchus	Branch of the right main bronchus leading to the inferior lobe
Right middle lobar bronchus	Branch of the right main bronchus leading to the middle lobe
Left inferior lobar bronchus	Branch of the left main bronchus leading to the inferior lobe
Left superior lobar bronchus	Branch of the left main bronchus leading to the superior lobe
Right main bronchus	Main right branch of the bronchial tree
Left main bronchus	Main left branch of the bronchial tree
Trachea	Main airway connecting the larynx to the bronchi

C. CUSTOMIZED MODELS

The main contribution of this work is the clinical adaptation of these classification and segmentation models for bronchoscopic imaging, through modified output heads, tailored preprocessing and annotation, and training strategies suited to the constraints of BM-BronchoLC. These changes were necessary for clinical relevance, but the internal architectures of the models were preserved to retain their original strengths.

➢ Classification Model

MedViT is a classification model based on Vision Transformers, introduced by Manzari [24] and designed for medical image analysis. It adopts a hybrid CNN-Transformer architecture, combining a convolutional patch embedding with a Transformer backbone to leverage both local and global image features [24]. Unlike traditional ViTs that directly segment images into sequences of patches, MedViT incorporates a convolutional preprocessing step to structure spatial information before encoding by the Transformer [24]. Its architecture is based on a hierarchy of Efficient Convolutional Blocks (ECB) and Local Transformer Blocks (LTB), integrating optimized multi-head attention and depthwise convolutions to enhance robustness and computational efficiency. Classification is performed through a linear head applied to the [CLS] token, providing a global representation of the image. The general architecture of the MedViT model is illustrated in Figure 1, showing its different stages, from the initial processing of bronchoscopic images to the final class prediction.

The adaptation of MedViT to the BM-BronchoLC dataset required several structural adjustments. To effectively handle the 11 anatomical landmarks and 7 types of bronchial lesions, the classification head was replaced with an architecture adapted to the new dataset classes. The use of pre-trained weights on ImageNet accelerated the model's convergence, optimizing its learning despite the limited amount of available data. To mitigate the impact of class imbalance, dynamic weighting was applied to the loss function, preventing the model from favoring majority classes. An early stopping mechanism was also integrated, stopping training when no significant improvement in the F1-score was observed, thus preventing overfitting to the training data. For a more detailed performance evaluation, multiple metrics were used, including mean accuracy, AUC-ROC, precision, recall, and F1-score, ensuring a more suitable analysis for medical challenges.

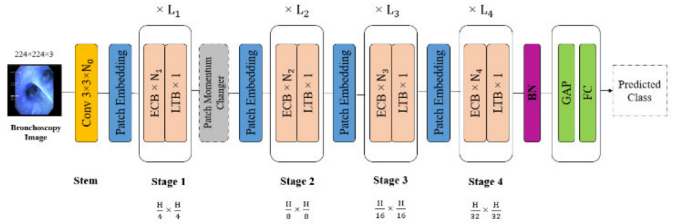

FIGURE 1. Overall architecture of the classification model.

Figure 1 presents the architecture of the proposed MedViT-based classification model, highlighting the complementary roles of its convolutional and Transformer components. The initial convolutional stem extracts fine-grained bronchoscopic features—mucosal texture, illumination gradients, vascular patterns, and local lesion boundaries—providing strong localized representations essential for detecting subtle abnormalities. A sequence of multi-scale Transformer blocks then performs global contextual reasoning by modeling long-range dependencies across the bronchial tree. Through multi-head self-attention, the model captures airway topology, bifurcation geometry, and the spatial organization of bronchial segments, enabling disambiguation of visually similar structures or lesions occurring in different anatomical contexts. This hybrid design effectively combines local texture extraction with global structural modeling. Unlike purely convolutional networks, which remain limited by their receptive fields [25], [26], and pure Vision Transformers, which require large datasets to avoid underfitting [26], [27], MedViT leverages both convolutional inductive biases and attention-based global context. This allows robust representation learning even with the moderate size of BM-BronchoLC. Finally, the multi-label classification head operates on the fused multi-scale representation to simultaneously identify multiple co-occurring anatomical landmarks or lesions, a capability well suited to the complex visual conditions of bronchoscopy.

➢ Segmentation Model

Developed by FITZGERALD [28], FCB-SwinV2 Transformer is based on a hybrid architecture combining Swin Transformer V2 [29] for global feature extraction and an optimized CNN decoder for segmented reconstruction. The SwinV2 U-Net branch, used as an encoder, applies a hierarchical attention mechanism with sliding windows, effectively capturing spatial dependencies while limiting computational complexity [28]. This approach ensures better modeling of bronchial structures by accounting for morphological variations and low contrasts typical of bronchoscopic images. Figure 2 illustrates the general architecture of FCB-SwinV2, highlighting the combination of the Transformer encoder and convolutional blocks for feature fusion and refinement before generating the segmentation mask. Additionally, the Feature Coupling Branch (FCB) merges information extracted by the SwinV2 encoder with the refined spatial details from the decoder [28]. Unlike purely convolutional architectures, this structure leverages the strengths of both paradigms: the Transformer captures global relationships, while the FCB refines local segmentation for better delineation of complex bronchial structures. This fusion is achieved through feature concatenation followed by convolutional refinement, ensuring precise contour reconstruction.

Several adjustments were also made to adapt it to the bronchoscopic images of BM-BronchoLC. The model output was configured to produce binary masks, enabling a clear distinction between annotated regions and the background. Additionally, the annotation processing pipeline was optimized to ensure a rigorous correspondence between each image and its associated mask, leveraging the dataset's CSV files.

To enhance the model's robustness against contrast variations and poorly differentiated bronchial textures, specific transformations were integrated, improving the network's ability to adapt to different clinical cases. To complement performance evaluation, multiple metrics were used to analyze

segmentation quality, including Dice Score, Intersection over Union (IoU), Surface Dice, and recall, ensuring a more rigorous assessment of predictions and bronchial structure contours.

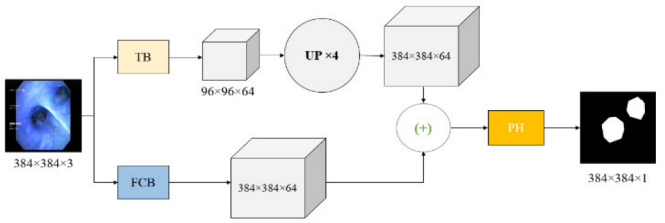

FIGURE 2. Overall architecture of the segmentation model.

D. MODEL TRAINING

Training was conducted on the Graham and Cedar clusters of the Digital Research Alliance of Canada, typically using 1 GPU per task, 128 GB of RAM, and 32 CPU cores per task. The training duration ranged from 2 to 14 hours, depending on the model and task.

For all experiments, the dataset was systematically divided into three distinct subsets: training, validation, and test. The training set was used to fit the model parameters, while the validation set played a key role in hyperparameter tuning, model selection, and early stopping decisions. Finally, the test set was strictly reserved for final performance evaluation, ensuring unbiased assessment. This protocol ensures that hyperparameters were optimized using only the validation data, preventing information leakage from the test set.

The models were implemented with PyTorch/Python and optimized using AdamW, a dynamic scheduler, and an early stopping strategy to improve convergence. In classification, the BCEWithLogitsLoss function was used to handle the multi-label nature of the annotations. In segmentation, a combination of Dice Loss and BCE Loss was applied to refine segmentation quality. The complete details of the hyperparameters are presented in **Table 3**.

Table 3 summarizes the hyperparameters used for both the classification and segmentation models, and each choice is aligned with the characteristics of BM-BronchoLC and the specific objectives of the tasks. For MedViT, the input size of 224 × 224 follows the ViT-Base configuration while preserving sufficient spatial detail for bronchoscopic textures. Learning rates (1e-4 for anatomical landmarks and 3e-5 for lung lesions) were determined empirically to stabilize training under heterogeneous class distributions. AdamW was adopted for all models to improve convergence with Transformer-based components, while the ReduceLROnPlateau scheduler prevented early stagnation, with patience values adapted to task complexity. For FCB-SwinV2 segmentation, a lower learning rate (1e-6) was required due to the deeper attention hierarchy and the sensitivity of Dice-based losses. Batch sizes were constrained by GPU memory, and early stopping was guided by validation Dice to limit overfitting. Finally, dropout was applied only to MedViT, as multi-label classification benefitted from additional regularization, whereas segmentation performance degraded when dropout was introduced.

TABLE 3. Hyperparameters of classification and segmentation models.

Hyperparameters	Anatomical Landmarks Classification	Lung Lesions Classification	Anatomical Landmarks Segmentation	Lung Lesions Segmentation
Model	MedViT (ViT-Base)	MedViT (ViT-Base)	FCB-SwinV2 Transformer	FCB-SwinV2 Transformer
Input size	224x224	224x224	384x384	384x384
Learning rate	1e-4	3e-5	1e-6	1e-6
Optimizer	AdamW	AdamW	AdamW	AdamW
Scheduler	ReduceLROnPlateau (factor=0.5, patience=5)	ReduceLROnPlateau (factor=0.5, patience=5)	ReduceLROnPlateau (factor=0.6, patience=10)	ReduceLROnPlateau (factor=0.6, patience=10)
Loss function	BCEWithLogitsLoss	BCEWithLogitsLoss	Dice Loss + BCE Loss	Dice Loss + BCE Loss
Dropout	30%	40%	None	None
Batch size	8	8	2 (train) / 1 (val)	2 (train) / 1 (val)
Number of epochs	200	100	150	100

E. EVALUATION METRICS

➢ **Classification Metrics**

Mean Accuracy (**MA**): The average of the accuracies calculated for each individual class.

$$MA = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{1}$$

where N is the total number of classes and A_i is the accuracy for class i.

This metric is used in multi-label classification problems to measure the overall performance of the model across all classes. Rather than relying on a simple proportion of correct predictions, it first calculates the accuracy for each class individually by comparing the number of correct predictions to the total occurrences of that class. In this formulation, A_i represents the accuracy computed specifically for class i, obtained by dividing the number of correctly predicted samples of that class by its total number of instances in the dataset. This class-wise computation is particularly important in bronchoscopy, where the distribution of anatomical structures and lesion categories is highly imbalanced. By averaging across all classes, the mean accuracy prevents dominant anatomical classes from overshadowing visually rare or underrepresented lesion types and provides a more balanced estimate of model performance. Unlike global accuracy, which aggregates all predictions without accounting for class imbalance, MA ensures that each class contributes equally to the evaluation, making it especially suitable for heterogeneous medical datasets such as BM-BronchoLC.

AUC-ROC measures the model's ability to distinguish between positive and negative classes by analyzing the true positive rate versus the false positive rate across various

classification thresholds. A value close to 1 indicates excellent discrimination.

Other metrics used to evaluate classification models include **precision**, **recall**, and **F1-score**. These metrics are computed using the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

TP (True Positives) refers to the number of examples correctly identified as belonging to the positive class, while **FP** (False Positives) corresponds to instances incorrectly classified as positive. Finally, **FN** (False Negatives) represents examples that are actually positive but were misclassified as negative.

➢ **Segmentation Metrics**

The performance of segmentation models is evaluated using several metrics, including Dice Score, Intersection over Union (IoU), Surface Dice, and Recall. These metrics quantify the quality of segmentation by comparing the model's predictions to the ground truth annotations. The formulas used are as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

TP (True Positives) corresponds to pixels correctly segmented as belonging to the target class, while **FP** (False Positives) refers to those incorrectly classified as belonging to this class. Conversely, **FN** (False Negatives) represents pixels that actually belong to the target class but were not detected by the model.

Surface Dice is a variant of the Dice Score that considers contour accuracy. All these metrics help evaluate the quality of the produced segmentations and identify errors, thereby facilitating the optimization of the model's performance. In addition to Dice and IoU, we report the 2D Surface Dice metric, which provides a boundary-aware assessment of segmentation quality within the planar resolution constraints of bronchoscopic images.

III. EXPERIMENTS AND RESULTS

A. EXPERIMENT 1: RANDOM DATA SPLITTING

In this first approach, taken directly from the reference article of the BM-BronchoLC dataset [19], images were randomly distributed following a classic split: 80% for training, 10% for validation, and 10% for testing. This method maintains a good balance between classes in each set but does not take into account the patient origin of the images. In other words, the same person may have images distributed across multiple subsets, introducing an implicit overlap between training and testing.

Figure 3 summarizes the detailed results of our evaluation. Overall, the models demonstrate strong performance within this framework. For classification tasks, anatomical landmarks achieve an AUC-ROC of 0.95 and a mean accuracy of 0.95, while bronchial lesions reach an AUC of 0.79. In segmentation, anatomical landmarks obtain a Dice score of 0.42 with a recall of 0.57, compared to scores of 0.33 and 0.47 for cancerous lesions.

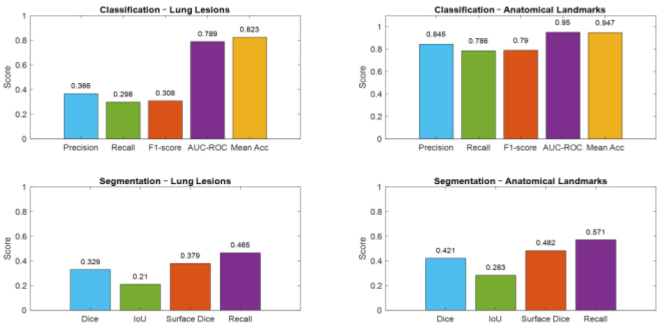

FIGURE 3. Classification results (top) and segmentation results (bottom) obtained with random data splitting. Left: cancerous lesions; Right: anatomical landmarks.

This difference in performance between anatomical structures and lesions can be partly explained by the intrinsic composition of the BM-BronchoLC dataset: anatomical landmarks account for 2,132 annotated images, whereas lesions represent only 789 samples. This nearly 3:1 ratio provides a markedly richer and more diverse training distribution for anatomical classes, enabling the model to learn more stable and discriminative representations. Conversely, lesion classes are not only smaller in number but also more visually heterogeneous—varying in shape, texture, illumination, and pathological appearance—which naturally increases their classification difficulty under a random split.

A similar performance gap between anatomical landmarks and lesions was also reported in the original BM-BronchoLC benchmark [19], confirming that this tendency is an intrinsic property of the dataset rather than a behavior specific to our model.

B. EXPERIMENT 2: STRICT PATIENT-BASED SPLITTING

Recognizing the importance of subject-level separation to prevent data leakage, the second experiment employs a patient-based split strategy. This ensures that no images from the same individual appear across different subsets. While the overall distribution remains the same as in the previous approach (80% training, 10% validation, and 10% testing), the allocation is performed at the patient level to eliminate any overlap between training, validation, and testing sets.

In total, the BM-BronchoLC dataset includes 208 patients (106 lung-cancer cases and 102 non-cancer cases). Using a fixed random seed (2024), the strict patient-wise split yields

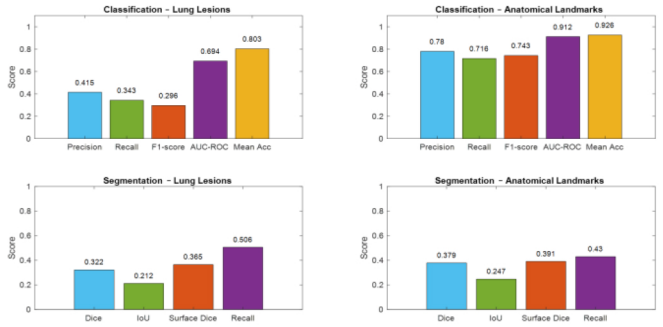

FIGURE 4. Performance results obtained with strict patient-based splitting: scores for classification (top) and segmentation (bottom). Left: cancerous lesions; Right: anatomical landmarks.

165 patients for training, 21 for validation, and 22 for testing. Because the split is performed at the patient identifier level, all images belonging to a given subject—including near-duplicate frames or images acquired within the same bronchoscopic session—remain in a single subset, fully preventing leakage across splits. This information, together with the complete splitting script released in the public repository, ensures exact reproducibility of the partitioning protocol.

The performances obtained by the models are presented in Figure 4. In classification, MedViT achieves a mean accuracy of 0.80 for bronchial lesions and 0.93 for anatomical landmarks. The associated AUC-ROC scores are 0.69 and 0.91, respectively. For segmentation, FCB-SwinV2 achieves a Dice score of 0.32 on lesions and 0.38 on landmarks, with recalls of 0.51 and 0.43. All complementary metrics are detailed in Figure 4.

The decrease in performance observed under the patient-based split is directly linked to the removal of data leakage present in the random strategy. Because images from the same patient often share anatomical geometry, illumination patterns, and endoscopic trajectories, their presence across both training and testing sets artificially inflates model performance. By enforcing a strict subject-level separation, the patient-based split requires the model to generalize to entirely unseen anatomical and pathological presentations. This results in more realistic, yet naturally lower, performance estimates—particularly for lesions, whose appearance varies substantially from one patient to another.

C. IMPACT OF DATA LEAKAGE

The use of a random splitting strategy without considering patient origin exposes the model to similar images across training and testing sets. This introduces data leakage bias, leading to a biased performance estimate. This practice, still common in some studies, is particularly problematic because several recent works [7], [8], [14], [15], [18], [21] do not clearly specify whether the split was performed by patient, video, or sequence, nor do they detail their cross-validation method. This lack of methodological clarity significantly increases the risk of bias and undermines the evaluation of the model's true generalization capability in real clinical scenarios.

Our experiments reveal significant differences between the two strategies. In classification, the AUC-ROC decreases by 12.1% for lung lesions and by 4.0% for anatomical landmarks when moving from random splitting to strict patient-based separation. In segmentation, the impact is even more pronounced on anatomical landmarks, with a drop of 24.7% in recall and 18.9% in Surface Dice. These differences are summarized in Figure 5, which illustrates the score variations (Δ) between the two approaches for all evaluated metrics.

Interestingly, some metrics such as the Recall for lesion classification show a slight improvement (15.1%) in the absence of leakage, suggesting that some models may learn to generalize better when the evaluation is rigorous. This highlights the complexity of learning dynamics and the importance of strict validation.

These results clearly demonstrate that evaluation on data shared between training and testing leads to an overestimation of performance, offering an illusion of robustness. Only strict patient-based separation provides a reliable measure of the model's ability to generalize in real clinical contexts, where prediction accuracy is essential.

Task	Target	Metric	Random Split	Patient-Level Split	Δ	Δ%
Classification	Lung Lesions	Mean Accuracy	**0.823**	0.803	0.020	2.43%
		AUC-ROC	**0.789**	0.694	0.095	12.05%
		F1-score	**0.308**	0.296	0.012	3.88%
		Recall	0.298	**0.343**	-0.045	-15.07%
		Precision	0.366	**0.415**	-0.049	-13.39%
Classification	Anatomical Landmarks	Mean Accuracy	**0.947**	0.926	0.021	2.22%
		AUC-ROC	**0.95**	0.912	0.038	4.00%
		F1-score	**0.79**	0.743	0.047	5.94%
		Recall	**0.786**	0.716	0.070	8.90%
		Precision	**0.845**	0.78	0.065	7.70%
Segmentation	Lung Lesions	Dice	**0.329**	0.322	0.007	2.13%
		IoU	0.21	**0.212**	-0.002	-0.95%
		Surface Dice	**0.379**	0.365	0.014	3.69%
		Recall	0.465	**0.506**	-0.041	-8.82%
Segmentation	Anatomical Landmarks	Dice	**0.421**	0.379	0.042	10.00%
		IoU	**0.283**	0.247	0.036	12.73%
		Surface Dice	**0.482**	0.391	0.091	18.89%
		Recall	**0.571**	0.43	0.141	24.70%

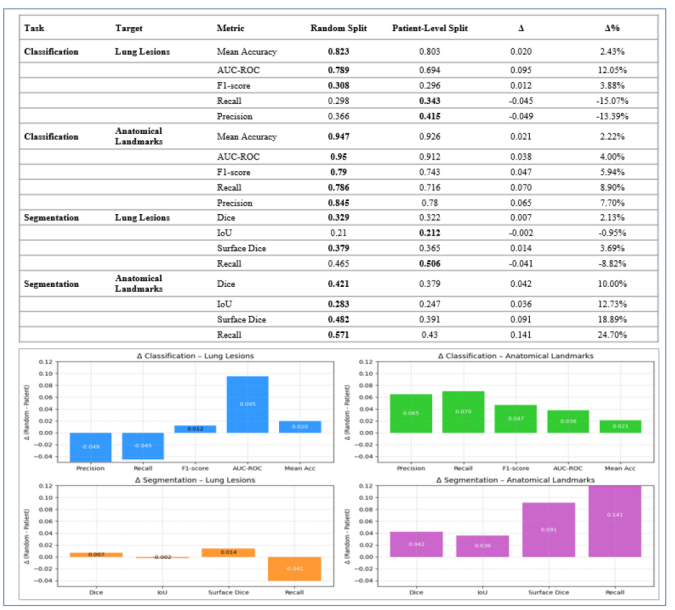

FIGURE 5. Comparison of model performance under random and patient-level split. Positive values indicate overestimation due to data leakage, while negative values reflect better generalization in the absence of leakage.

IV. DISCUSSION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, largely due to late diagnoses and reliance on visual interpretation during procedures such as bronchoscopy [1], [2]. Although bronchoscopy is a front-line diagnostic tool, it suffers from significant inter-operator variability and challenges in detecting subtle or poorly visible abnormalities [5]. In this context, the development of robust artificial intelligence models capable of automating

the recognition of pulmonary lesions and anatomical landmarks represents a highly relevant opportunity to support clinicians, reduce diagnostic delays, and improve the reliability of procedures. With this in mind, we proposed two hybrid architectures combining the strengths of convolutional networks and Transformers to address two key tasks in bronchoscopy: multi-label classification using MedViT and binary segmentation using FCB-SwinV2. These models were specifically adapted to the visual complexity of bronchoscopic images from the BM-BronchoLC dataset. In addition, we emphasized a realistic clinical evaluation by comparing conventional random data splitting with strict patient-based separation, which more accurately reflects real-world deployment conditions. This methodological contribution is just as essential: our results not only benchmark CNN–Transformer architectures for bronchoscopy but also underline the risks of overestimated performance when evaluation relies solely on random splitting.

The proposed hybrid CNN–Transformer architectures present several advantages for bronchoscopic image analysis. Convolutional layers effectively capture local details such as mucosal textures, vascular patterns, and lesion boundaries, while Transformer components model long-range dependencies and the global spatial organization of the bronchial tree. This dual representation is particularly suited to bronchoscopic imaging, where fine-grained features coexist with complex anatomical structures. Compared to purely convolutional architectures, which remain constrained by their limited receptive fields [25], [26], the inclusion of attention mechanisms provides explicit global contextual reasoning, improving inter-patient stability and the discrimination of visually similar abnormalities. Conversely, pure Vision Transformers generally require large-scale datasets to avoid underfitting or over-smoothing [26], [27], whereas the incorporation of convolutional blocks introduces beneficial inductive biases that enable effective feature extraction even with a modest dataset such as BM-BronchoLC.

These strengths, however, come with certain limitations. The addition of Transformer blocks increases computational cost and memory usage, which may restrict applicability in real-time clinical environments. Moreover, attention-based components tend to be more sensitive to hyperparameter settings and require careful regularization; insufficient tuning can lead to optimization instability or representation over-smoothing—a critical consideration for clinical adoption, where transparency and interpretability remain essential.

The performance differences between anatomical landmarks and pulmonary lesions, observed both in our experiments and in the original BM-BronchoLC benchmark [21], can be largely explained by intrinsic properties of the dataset. Anatomical structures are annotated in 2,132 images, whereas lesions appear in only 789 samples, providing a much richer and more diverse training space for anatomical classes. This imbalance naturally favors better performance on structural classes. Additionally, lesion categories exhibit much higher visual heterogeneity—variations in shape, texture, illumination, and pathological expression—making them inherently more difficult to classify.

Beyond class imbalance, the segmentation of pulmonary lesions remains intrinsically challenging due to the visual characteristics of bronchoscopic imagery. Lesions often exhibit weak contrast with surrounding mucosa, diffuse or irregular boundaries, and substantial inter-patient variability in shape, size, and appearance. Unlike anatomical landmarks, which are structurally consistent and spatially constrained, lesion contours are frequently ambiguous and gradually transition into healthy tissue, making precise boundary delineation particularly difficult.

This behavior is reflected in the segmentation metrics, where lesion Dice scores remain limited (0.329 under random split and 0.322 under patient-level split) and IoU values stay low (0.210 and 0.212, respectively), compared to higher values obtained for anatomical landmarks (Dice up to 0.421 and IoU up to 0.283 under random splitting). Moreover, Surface Dice values for lesion segmentation remain modest (0.379 and 0.365), indicating limited boundary localization accuracy. The relatively small variations observed between random and patient-level splits for lesion segmentation further suggest that performance is primarily constrained by intrinsic lesion characteristics and inter-patient variability rather than by data leakage or model instability. These results highlight that fine-grained lesion boundary localization constitutes a fundamental challenge in bronchoscopy, even when using hybrid CNN–Transformer architectures and fully supervised annotations.

Furthermore, the discrepancy between random and patient-based splits is closely linked to the presence of data leakage in random splitting. Because images from the same patient may appear in both training and testing sets, the model benefits from patient-specific visual cues such as airway geometry or illumination patterns, artificially inflating performance. In strict patient-based splitting, these cues are eliminated, forcing the model to generalize to entirely unseen anatomical and pathological presentations, which yields more realistic but naturally lower performance estimates—particularly for lesions. These observations reinforce the essential role of patient-level protocols in evaluating real-world clinical generalizability.

The results indicate that the proposed approach shows promising potential. MedViT achieved a mean accuracy of 94.74 % for anatomical landmarks and 82.30 % for pulmonary lesions, suggesting its robustness despite the multi-class nature of the task and the frequent visual overlaps between abnormalities. FCB-SwinV2, although more modest in segmentation performance (Dice scores of 42.06 % for anatomical structures and 32.93 % for lesions), provides a coherent baseline for targeted morphological detection tasks. Moreover, these performances remained generally stable—although slightly reduced—under patient-level partitioning, indicating that the evaluation strategy notably influences the reported results and that clinically realistic validation protocols are crucial for establishing confidence

in the performance of prediction models in AI-assisted bronchoscopy.

Compared to approaches from the BM-BronchoLC benchmark [21], our methods showed competitive results. MedViT delivered solid performance, achieving comparable and in some cases slightly superior results to CNN-based models (UNet++ [30], [31]) and multitask Transformer models (ESFPNet [17]) in classifying anatomical structures, while maintaining accuracy within the same range as leading approaches for lesion classification. In contrast, FCB-SwinV2 lagged behind in pure segmentation performance, likely due to its lack of integration with a classification branch and its single-task specialization. These results are summarized in **Table 4**, which presents a direct comparison of our models' performance with those of the benchmark approaches across both tasks. To ensure fairness, we strictly followed the BM-BronchoLC partitioning code, ensuring that all models were evaluated on the same images with the same random splits. However, published models such as UNet++ and ESFPNet were trained with different preprocessing, losses, and augmentations, which may affect the results. Therefore, while the comparisons are valid, they should be interpreted with caution. This discrepancy in training protocols could be considered as a direction for future work, where retraining the models under a unified protocol would allow for a more exhaustive and consistent benchmarking process.

A closer examination of **Table 4** further clarifies these trends. The superior classification performance of MedViT can be attributed to its hybrid design, which effectively combines convolutional inductive biases with long-range Transformer reasoning [24]—an advantage over both CNN-based models like UNet++ [30], [31] and multitask Transformer architectures such as ESFPNet [17]. Conversely, the lower segmentation performance of FCB-SwinV2 can be explained by the absence of shared representation learning typically available in multitask frameworks. Models such as ESFPNet [17] benefit from complementary supervision between classification and segmentation, which is particularly advantageous for detecting small, low-contrast, or visually ambiguous lesion regions. These observations highlight that while hybrid architectures appear especially well suited for classification, reaching state-of-the-art segmentation performance may require multitask or strongly regularized designs capable of leveraging cross-task consistency.

Our findings also align with recent studies using other datasets. Deng et al. [13] reported a 95.1 % accuracy in a binary classification task using a ResNet, while Setayeshi et al. [15] achieved 90.6 % with a multitask DenseNet. Although these results are strong, they relate to simpler tasks (binary or tri-class classification), and their protocols do not necessarily enforce strict patient separation. This limits the conclusions regarding real-world deployment, as inter-patient variability is not properly accounted for. For segmentation, Chang et al. [17] reported a Dice score of 0.756 on autofluorescence images—a modality with inherently higher contrast than white light imaging. Yoo et al. [18], using EfficientNet-B1 to localize bronchial structures in videos, reached 86 % accuracy compared to 94.74 % for MedViT, which situates our results within a similar performance range while emphasizing the importance of patient-level validation for fair and meaningful evaluation.

TABLE 4. Comparison of the performance of our models (MedViT and FCB-SwinV2) with other CNN and Transformer approaches (UNet++ and ESFPNet) [21] for bronchoscopic classification and segmentation.

Model (Task)	Anatomical landmarks (Dice / MA %)	Lung lesions (Dice / MA %)	Type of task
ESFPNet (Segmentation Only)	75,42	55,36	Segmentation
ESFPNet (Multi-task)	75,85	56,25	Segmentation
UNet++ (Segmentation Only)	71,28	46,51	Segmentation
UNet++ (Multi-task)	72,42	45,33	Segmentation
FCB-SwinV2 Transformer	**42,06**	**32,93**	**Segmentation**
ESFPNet (Classification Only)	93,98	85,91	Classification
ESFPNet (Multi-task)	91,14	86,11	Classification
UNet++ (Classification Only)	88,58	82,29	Classification
UNet++ (Multi-task)	88,77	82,49	Classification
MedViT	**94,74**	**82,30**	**Classification**

Despite the promise of our approach, several limitations must be considered. First, the study was based on data from a single clinical center, which limits the diversity of acquisition conditions (equipment, practices, patient profiles) and raises concerns about generalizability to other settings. Second, only white-light images were used, without including complementary modalities such as autofluorescence or narrow-band imaging, which are commonly employed to enhance lesion visibility. This limitation reduces the richness of the information available to the models, particularly in visually complex cases.

Another limitation lies in the architecture design: classification and segmentation were treated as independent tasks, handled by separate models without shared representation learning. However, in a clinical context, a multitask approach may be more appropriate—allowing the representations learned during classification to guide segmentation. This type of architecture could improve system consistency and potentially enhance segmentation performance on challenging structures. Although multitask models have been explored in prior work [21], current results suggest there is still room for improvement, especially in refining interactions between classification and segmentation branches.

In addition, while Dice and Surface Dice metrics provide meaningful overlap- and boundary-based evaluations, a more exhaustive geometric assessment—using advanced boundary-aware measures such as HD95 or distance-based metrics expressed in physical units—would allow a finer

characterization of segmentation accuracy, particularly for elongated and topologically complex structures like the bronchial tree. Such metrics would offer complementary insight into boundary precision and anatomical plausibility and therefore represent a relevant direction for future work.

Similarly, although the CNN–Transformer architectures employed in this study are well-established in the literature and were adopted without structural modification, the absence of a systematic ablation analysis constitutes a limitation. In particular, future investigations could benefit from controlled one-factor ablation studies under fixed random seeds, examining the impact of key design and training choices such as class weighting strategies, ImageNet pretraining, input resolution, feature coupling mechanisms (e.g., FCB), the contribution of Transformer blocks, and alternative loss formulations (e.g., BCE+Dice versus Focal-Tversky). Such analyses would enable a more granular understanding of how these components influence performance in bronchoscopic imaging, beyond the scope of the present methodological focus.

The analysis of the confusion matrices, shown in Figure 6, further confirms the challenges faced by the models, particularly when dealing with visually similar classes. Recurrent misclassifications occurred between mucosal infiltrations, stenoses, and vascular growths—often difficult to distinguish due to variations in viewing angle, image quality, or patient anatomy. For anatomical landmarks, most errors involved symmetrical or adjacent structures such as the left and right lobar bronchi, especially under strict patient-based evaluation. These findings highlight the difficulty of inter-patient generalization and point to the need for more robust learning mechanisms. Once again, they stress the necessity of evaluation frameworks that reflect real deployment, since random splitting would fail to capture these inter-patient sources of error.

These observations are consistent with the more detailed patterns revealed by the confusion matrices. For lung lesions, the model's errors are concentrated among categories that share overlapping textures or illumination characteristics—such as mucosal infiltration, stenosis, and vascular growths—making them particularly prone to ambiguity. For anatomical landmarks, confusion is strongest between symmetric or spatially adjacent bronchi (e.g., left versus right lobar bronchi), where structural similarity significantly increases classification difficulty. These effects become even more pronounced under patient-based splitting, where the model can no longer rely on patient-specific visual cues and must instead generalize across unseen anatomical variations. This reinforces the intrinsic difficulty of bronchoscopy and highlights the importance of designing architectures capable of capturing subtle discriminative cues beyond local texture alone.

As illustrated in Figures 7 and 8, these examples show classification errors and model predictions in the tasks of cancerous lesion and anatomical landmark classification. The figures present images where errors occurred, particularly in classes where the model made the most frequent confusions.

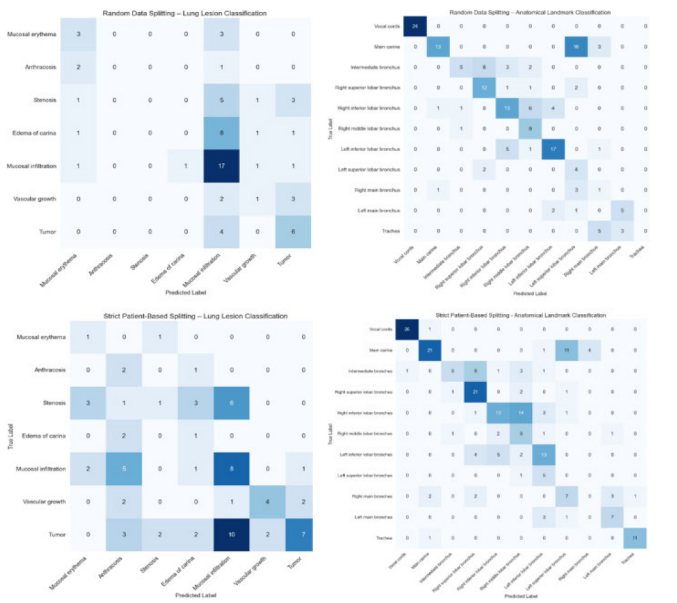

FIGURE 6. Confusion matrices obtained with MedViT for the classification of lung lesions (left) and anatomical landmarks (right), under two splitting strategies: random (top) and patient-based (bottom).

These errors were mostly observed between visually similar classes, such as Stenosis and Mucosal infiltration, as well as Right inferior lobar bronchus and Right middle lobar bronchus. These errors highlight the model's difficulty in handling multiclasse images, where multiple labels must be assigned simultaneously. The images demonstrate that, while the classes are present, the model fails to recognize all the classes in situations where objects overlap visually. Errors are particularly noticeable in cases where two classes coexist in an image, but the model only predicts one. For example, the model predicted Mucosal infiltration but omitted Stenosis (Figure 7), despite their co-occurrence in the image. This suggests that the model struggles to identify relationships between multiple classes that appear in the same area of interest. This issue is even more pronounced when the boundaries between classes are not distinct enough.

These errors are not simply due to poor prediction, but also to the model's inability to differentiate classes in complex contexts where multiple regions of interest belong to similar classes. This highlights the multiclasse nature of complex images as a significant challenge for current models, which are not always able to manage this overlap of labels without losing precision.

To shed light on the impact of partitioning on prediction stability, we present in Tables 5 and 6 the per-class performance with 95% confidence intervals (calculated using the Wilson score method). Visually distinct and relatively frequent structures, such as the Vocal cords (Precision/Recall: 1.00 → 0.96), Right Superior Lobar Bronchus (F1: 0.65 → 0.68), and Trachea (Precision/Recall: 0.92/0.92 at Patient-level), show notable stability between the two protocols. In contrast, some rare classes or those

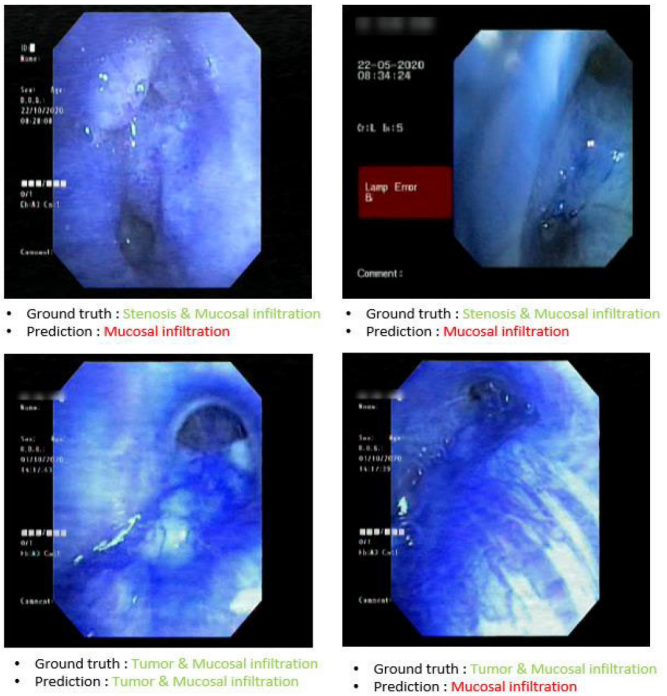

FIGURE 7. Error in anatomical landmark classification: confusion between main carina, left superior lobar bronchus, and right inferior lobar bronchus.

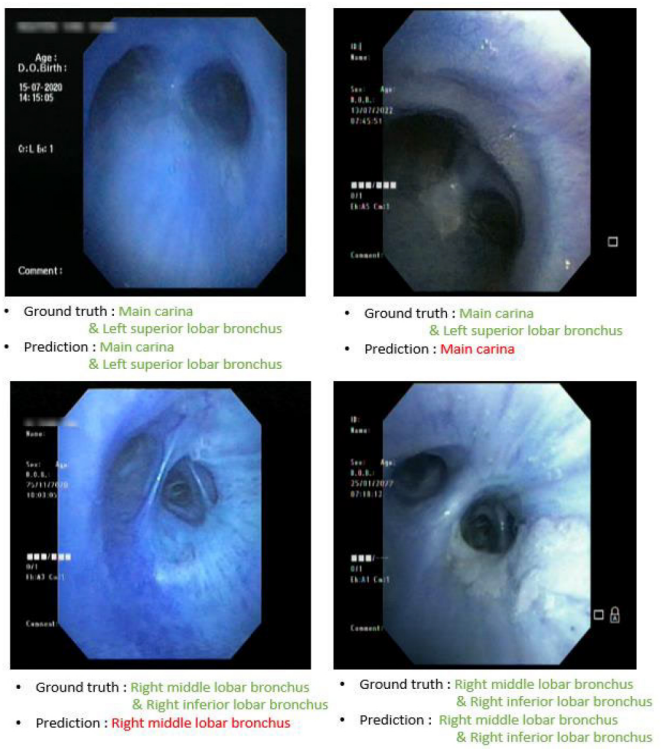

FIGURE 8. Error in cancerous lesion classification: confusion between stenosis, mucosal infiltration, and tumor.

morphologically similar to other segments exhibit more pronounced variations, such as Right Middle Lobar Bronchus (F1: 0.64 → 0.40), Main Carina (Recall: 0.41 → 0.57), or Left Inferior Lobar Bronchus (Recall: 0.67 → 0.00), reflecting their low support and high inter-patient variability. Overall, patient-level partitioning primarily amplifies fluctuations in underrepresented classes, but these trends should be interpreted with caution due to the higher uncertainty associated with these categories.

Aligned with our objective of developing models capable of automatically recognizing cancerous lesions and anatomical landmarks in bronchoscopic images, several research directions can be considered. First, extending the validation to multi-center datasets would allow a more comprehensive assessment of model robustness across diverse clinical environments. Second, incorporating additional bronchoscopic imaging modalities, such as autofluorescence or narrow-band imaging, could further support the analysis of subtle anomalies that may be less apparent in white-light imaging. Developing multitask architectures that jointly handle classification and segmentation could strengthen the coherence of the model and leverage cross-task learning.

Furthermore, since bronchoscopy primarily produces video sequences, evaluating the models on temporal data is necessary to assess the stability of predictions and their real-time applicability. Additionally, incorporating continual learning mechanisms would allow models to adapt progressively to new clinical cases without full retraining. Finally, integrating non-visual clinical data, such as patient history, symptoms, or estimated anatomical location, could refine predictions—particularly in ambiguous cases. Together, these perspectives aim to enhance the precision, robustness, and clinical relevance of the proposed models in real-world bronchoscopic applications. Beyond technical improvements, future research must also standardize evaluation protocols around strict patient-level validation, so as to ensure that performance estimates reflect true generalization capacity and to avoid misleadingly optimistic results.

Additionally, approaches such as enriched attention mechanisms [32], adaptive contrastive learning [33], as well as generative and semantic prompt-based modeling for weakly supervised semantic segmentation [34], [35], could also be explored to improve the discriminative ability and generalization capacity of the models, particularly in contexts characterized by limited or highly heterogeneous data.

In parallel, the clinical deployment of such models requires careful consideration of robustness and operational validity. Bronchoscopic imaging is subject to frequent artefacts—including glare, motion blur, illumination shifts, and the presence of blood or smoke—as well as variability between devices. While the BM-BronchoLC benchmark already exhibits realistic visual heterogeneity, the absence of controlled stress-testing or multi-center validation still limits a full assessment of resilience. Exploring robustness under these perturbations, as well as cross-site generalization, represents an important extension for future work. Likewise, incorporating clinically oriented operating points—such as fixed-FPR sensitivity, decision curves, or explicit clinical cost analyses of false positives and false negatives—would help

TABLE 5. Per-class performance with 959% wilson confidence intervals for anatomical landmark classification under random and patient-level splitting.

Class	Precision (Random)	Precision (Patient)	Recall (Random)	Recall (Patient)	F1-Score (Random)	F1-Score (Patient)	Support (Random)	Support (Patient)
Vocal cords	1.0000 [0.859–1.000]	0.9630 [0.832–0.999]	1.0000 [0.859–1.000]	0.9630 [0.832–0.999]	1.0000	0.9630	24	27
Main carina	0.8130 [0.667–0.958]	0.8080 [0.673–0.942]	0.4060 [0.243–0.570]	0.5680 [0.411–0.724]	0.5410	0.6670	32	37
Intermediate bronchus	0.4550 [0.233–0.696]	0.5000 [0.273–0.727]	0.3130 [0.109–0.516]	0.2730 [0.091–0.455]	0.3704	0.3530	16	22
Right superior lobar bronchus	0.5710 [0.357–0.786]	0.5680 [0.394–0.741]	0.7500 [0.562–0.938]	0.8400 [0.704–0.976]	0.6490	0.6770	16	25
Right inferior lobar bronchus	0.5420 [0.358–0.725]	0.6190 [0.442–0.796]	0.5200 [0.322–0.718]	0.4060 [0.234–0.578]	0.5310	0.4900	25	32
Right middle lobar bronchus	0.5000 [0.233–0.767]	0.3100 [0.138–0.534]	0.9000 [0.714–1.000]	0.5630 [0.313–0.813]	0.6429	0.4000	10	16
Left superior lobar bronchus	0.7391 [0.566–0.912]	0.4643 [0.270–0.658]	0.7083 [0.532–0.885]	0.5420 [0.333–0.750]	0.7234	0.5000	24	24
Left inferior lobar bronchus	0.1600 [0.033–0.387]	0.0000 [0.000–0.000]	0.6667 [0.333–1.000]	0.0000 [0.000–0.000]	0.2581	0.0000	6	6
Right main bronchus	0.1000 [0.002–0.445]	0.0000 [0.000–0.000]	0.2000 [0.000–0.717]	0.0000 [0.000–0.000]	0.1333	0.0000	5	15
Left main bronchus	0.6250 [0.323–0.927]	0.6364 [0.364–0.909]	0.6250 [0.323–0.927]	0.6364 [0.364–0.909]	0.6250	0.6364	8	11
Trachea	0.0000 [0.000–0.000]	0.9170 [0.615–0.998]	0.0000 [0.000–0.000]	0.9170 [0.615–0.998]	0.0000	0.9170	8	12

TABLE 6. Per-class performance with 95% wilson confidence intervals for lung lesion classification under random and patient-level data splitting.

Class	Precision (Random)	Precision (Patient)	Recall (Random)	Recall (Patient)	F1-Score (Random)	F1-Score (Patient)	Support (Random)	Support (Patient)
Mucosal erythema	0.3750 [0.118–0.719]	0.1670 [0.009–0.641]	0.5000 [0.157–0.843]	0.5000 [0.063–0.937]	0.4290	0.2500	6	2
Anthracosis	0.0000 [0.000–0.000]	0.1333 [0.012–0.527]	0.0000 [0.000–0.000]	0.6670 [0.194–1.000]	0.0000	0.2222	3	3
Stenosis	0.0000 [0.000–0.000]	0.2500 [0.092–0.515]	0.0000 [0.000–0.000]	0.0670 [0.002–0.304]	0.0000	0.1053	10	15
Edema of carina	0.0000 [0.000–0.000]	0.1250 [0.011–0.494]	0.0000 [0.000–0.000]	0.3330 [0.043–0.846]	0.0000	0.1818	11	3
Mucosal infiltration	0.4250 [0.276–0.523]	0.3200 [0.190–0.484]	0.8095 [0.600–0.923]	0.4710 [0.256–0.697]	0.5577	0.3810	21	17
Vascular growth	0.3333 [0.063–0.738]	0.6667 [0.354–0.982]	0.1667 [0.004–0.585]	0.4444 [0.181–0.739]	0.2222	0.5333	6	9
Tumor	0.4286 [0.207–0.677]	0.7000 [0.500–0.899]	0.6000 [0.323–0.877]	0.2692 [0.124–0.463]	0.5000	0.3889	10	26

contextualize model performance within triage or assistive-review workflows. By grounding these evaluations in collaboration with clinicians, such analyses would clarify the operational value of AI systems in bronchoscopy and constitute a natural continuation of the present study.

Regarding clinical integration, the evaluated CNN–Transformer architectures are based on established designs commonly used in medical image analysis, suggesting that their computational complexity is compatible with near–real-time inference. While inference latency was not explicitly measured in this study, the models operate at the frame level and could be envisioned as assistive tools providing visual cues for lesion or anatomical landmark recognition during or immediately after bronchoscopy. A detailed analysis of real-time performance, hardware optimization, and workflow-level integration remains an important direction for future work.

In summary, while our adapted hybrid CNN–Transformer architectures demonstrate promising and competitive performance, the central contribution of this work lies in the clinically grounded adaptation of these models to bronchoscopic lesion and anatomical landmark recognition, together

with a systematic evaluation under clinically relevant data partitioning strategies.

Establishing such an evaluation framework enables a clearer interpretation of performance differences between random and patient-level splits, represents an important step toward obtaining more reliable estimates of model generalizability, and provides a solid basis for future methodological extensions. Although further validation and refinement are required before clinical deployment, this work contributes to ongoing efforts toward developing AI-assisted bronchoscopy systems by emphasizing evaluation practices that better reflect real-world clinical conditions, whose performance remains robust when evaluated under patient-level data separation.

DATA AND CODE AVAILABILITY

- The BM-BronchoLC dataset analyzed in this study is publicly available and can be freely accessed at: https://doi.org/10.6084/m9.figshare.24243670.v3.
- All scripts used for data preprocessing, model training, evaluation, and figure generation are openly available on GitHub: [GitHub Repository]

REFERENCES

[1] K. Chaitanya Thandra, A. Barsouk, K. Saginala, J. Sukumar Aluru, and A. Barsouk, "Epidemiology of lung cancer," *Współczesna Onkologia*, vol. 25, no. 1, pp. 45–52, 2021, doi: 10.5114/wo.2021.103829.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] A. N. Wilkinson and S. Lam, "ABC du dépistage du cancer du poumon: Information clé pour les médecins de première ligne," *Can. Family Physician*, vol. 67, no. 11, pp. 823–829, Nov. 2021, doi: 10.46747/cfp.6711823.

[4] M. Andolfi, R. Potenza, R. Capozzi, V. Liparulo, F. Puma, and K. Yasufuku, "The role of bronchoscopy in the diagnosis of early lung cancer: A review," *J. Thoracic Disease*, vol. 8, no. 11, pp. 3329–3337, Nov. 2016, doi: 10.21037/jtd.2016.11.81.

[5] A. D. Lerner and D. Feller-Kopman, "Bronchoscopic techniques used in the diagnosis and staging of lung cancer," *J. Nat. Comprehensive Cancer Netw.*, vol. 15, no. 5, pp. 640–647, May 2017, doi: 10.6004/jnccn.2017.0065.

[6] H. Minami, Y. Ando, F. Nomura, S. Sakai, and K. Shimokata, "Interbronchoscopist variability in the diagnosis of lung cancer by flexible bronchoscopy," *Chest*, vol. 105, no. 6, pp. 1658–1662, Jun. 1994, doi: 10.1378/chest.105.6.1658.

[7] Y. Hadhoud et al., "A two-step hybrid CNN-ViT model for chest disease classification based on X-ray images," *Diagnosis*, vol. 14, no. 23, Nov. 2024, Art. no. 2754, doi: 10.3390/diagnostics14232754.

[8] C. Hu, N. Cao, H. Zhou, and B. Guo, "Medical image classification with a hybrid SSM model based on CNN and transformer," *Electronics*, vol. 13, no. 15, p. 3094, Aug. 2024, doi: 10.3390/electronics13153094.

[9] Y. Xie et al., "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science), vol. 12903. Cham, Switzerland: Springer, 2021, pp. 236–246, doi: 10.1007/978-3-030-87199-4_16.

[10] H. Lee and J. Yoo, "MetaFormer and CNN hybrid model for polyp image segmentation," *IEEE Access*, vol. 12, pp. 133694–133702, 2024.

[11] M. Hayat, "Endoscopic image super-resolution algorithm using edge and disparity awareness," Ph.D. dissertation, Dept. Electrical Eng., Chulalongkorn Univ., Bangkok, Thailand, 2023, doi: 10.58837/CHULA.THE.2023.901.

[12] M. Hayat, S. Aramvith, and T. Achakulvisut, "SEGSRNet for stereo-endoscopic image super-resolution and surgical instrument segmentation," in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2024, pp. 1–4, doi: 10.1109/embc53108.2024.10782794.

[13] Y. Deng, Y. Chen, L. Xie, L. Wang, and J. Zhan, "The investigation of construction and clinical application of image recognition technology assisted bronchoscopy diagnostic model of lung cancer," *Frontiers Oncol.*, vol. 12, Oct. 2022, Art. no. 1001840, doi: 10.3389/fonc.2022.1001840.

[14] P.-H. Feng, Y.-T. Lin, and C.-M. Lo, "A machine learning texture model for classifying lung cancer subtypes using preliminary bronchoscopic findings," *Med. Phys.*, vol. 45, no. 12, pp. 5509–5514, Dec. 2018, doi: 10.1002/mp.13241.

[15] R. Setayeshi, J. Vahidi, E. Kozegar, and T. Tan, "An end-to-end multitask deep learning framework for bronchoscopy image classification," *Multimedia Syst.*, vol. 30, no. 6, p. 361, Dec. 2024, doi: 10.1007/s00530-024-01579-3.

[16] A. Banach, F. King, F. Masaki, H. Tsukada, and N. Hata, "Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102164, doi: 10.1016/j.media.2021.102164.

[17] Q. Chang et al., "ESFPNet: Real-time lesion segmentation in autofluorescence bronchoscopic video," *Proc. SPIE*, vol. 12468, Mar. 2023, Art. no. 1246803, doi: 10.1117/12.2647897.

[18] J. Y. Yoo, S. Y. Kang, J. S. Park, Y.-J. Cho, S. Y. Park, H. I. Yoon, S. J. Park, H.-G. Jeong, and T. Kim, "Deep learning for anatomical interpretation of video bronchoscopy images," *Sci. Rep.*, vol. 11, no. 1, p. 23765, Dec. 2021, doi: 10.1038/s41598-021-03219-6.

[19] K. M. Cold, K. Agbontaen, A. O. Nielsen, C. S. Andersen, S. Singh, and L. Konge, "Artificial intelligence for automatic and objective assessment of competencies in flexible bronchoscopy," *J. Thoracic Disease*, vol. 16, no. 9, pp. 5718–5726, Sep. 2024, doi: 10.21037/jtd-24-841.

[20] K. M. Cold, K. Agbontaen, A. O. Nielsen, C. S. Andersen, S. Singh, and L. Konge, "Artificial intelligence improves bronchoscopy performance: A randomised crossover trial," *ERJ Open Res.*, vol. 11, no. 1, pp. 00395–2024, Jan. 2025, doi: 10.1183/23120541.00395-2024.

[21] V. G. Vu et al., "BM-BronchoLC–A rich bronchoscopy dataset for anatomical landmarks and lung cancer lesion recognition," *Sci. Data*, vol. 11, Mar. 2024, Art. no. 321, doi: 10.1038/s41597-024-03145-y.

[22] D. R. Duhamel and J. H. Harrell, *Atlas Clinique Des Maladies Des Voies Respiratoires: Bronchoscopie, Radiologie et Pathologie*. Amsterdam, The Netherlands: Elsevier Saunders, 2005.

[23] P. Shah, *Atlas De La Bronchoscopie Flexible*. Boca Raton, FL, USA: CRC Press, 2011, doi: 10.1201/b13458.

[24] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Comput. Biol. Med.*, vol. 157, May 2023, Art. no. 106791, doi: 10.1016/j.compbiomed.2023.106791.

[25] D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre, "Learning long-range spatial dependencies with horizontal gated-recurrent units," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, BC, Canada, Dec. 2018, pp. 1–11.

[26] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai, H. Machino, K. Kobayashi, K. Asada, M. Komatsu, S. Kaneko, M. Sugiyama, and R. Hamamoto, "Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review," *J. Med. Syst.*, vol. 48, no. 1, pp. 1–15, Sep. 2024, doi: 10.1007/s10916-024-02105-8.

[27] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. De Nadai, "Efficient training of visual transformers with small datasets," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 1–13.

[28] K. Fitzgerald, J. Bernal, A. Histace, and B. J. Matuszewski, "Polyp segmentation with the FCB-SwinV2 transformer," *IEEE Access*, vol. 12, pp. 38927–38943, 2024, doi: 10.1109/ACCESS.2024.3376228.

[29] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.

[30] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020, doi: 10.1109/TMI.2019.2959609.

[31] Z. Zhou et al., "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, pp. 3–11, 2018, doi: 10.1007/978-3-030-00889-5_1.

[32] W. Wu, T. Dai, Z. Chen, X. Huang, J. Xiao, F. Ma, and R. Ouyang, "Adaptive patch contrast for weakly supervised semantic segmentation," *Eng. Appl. Artif. Intell.*, vol. 139, Jan. 2025, Art. no. 109626.

[33] W. Wu, X. Chen, Z. Chen, J.-E. Jiang, K.-F. Tsang, X. Huang, F. Ma, and J. Xiao, "Tag-enriched multi-attention with large language models for cross-domain sequential recommendation," *IEEE Trans. Consum. Electron.*, vol. 71, pp. 1–10, Oct. 2025, doi: 10.1109/TCE.2025.3620527.

[34] S. R. Pavel, Y. D. Zhang, S. Sun, and A. L. F. de Almeida, "Tensor reconstruction-based sparse array 2-D DOA estimation of mixed coherent and uncorrelated signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 12876–12880, doi: 10.1109/icassp48485.2024.10447692.

[35] W. Wu, X. Qiu, S. Song, Z. Chen, X. Huang, F. Ma, and J. Xiao, "Prompt categories cluster for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2025, pp. 3189–3198.

ROLYPH ERWAN NTOUTOUME NGUEMA received the B.Sc. and M.Sc. degrees in biomedical engineering from the Université Polytechnique de l'Ouest Africain (UPOA), Dakar, Senegal, in 2021 and 2024, respectively. His research presented in this article was conducted during his Master's studies at UPOA. He has been admitted to the Ph.D. program in engineering at the École de Technologie Supérieure (ÉTS), Montreal, Canada, in 2026. His research interests include advanced analysis of physiological signals, particularly polysomnography recordings, as well as deep learning and computer vision models for medical image analysis, with a focus on pulmonary cancer lesion and bronchoscopic structure recognition.

MOHAMAD FOROUZANFAR (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa. Following his doctoral studies, he completed Postdoctoral fellowships with Stanford University and Harvard University. He previously a Research Scientist within the Human Sleep Research Laboratory, SRI International. He is currently an Associate Professor with the École de technologie supérieure, Université du Québec. His research interests include development of advanced instrumentation, signal and image processing, and machine learning, specifically as they apply to non-invasive physiological monitoring for sleep and cardiovascular health. He is an Associate Editor-in-Chief for IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.

ALI TRAORE received the master's degree in medical physics from University Paris Descartes, in 2010, and the M.S. degree in biomedical engineering from the University of Montréal, Canada, in 2013, and the Ph.D. degree in medical physics from the Complutense University of Madrid, Spain, as a Marie Curie Fellow, from 2014 to 2018. Since 2020, he has been the Director of biomedical engineering and physics with the Université Polytechnique de l'Ouest Africain, Dakar, Senegal. His current research interests include medical imaging and artificial intelligence, Monte Carlo simulations in medical physics for radiotherapy application and the development of advanced methods for adaptive radiotherapy. Since 2023, he has been pursuing a residency in medical physics with the Centre Hospitalier Universitaire of Brest, France. He is a member of French Society of Medical Physics (SFPM).