ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240657

Deep Learning for *in vivo* Bronchial Carinae Detection in Flexible Bronchoscopy

Robin Ghyselinck[a,*], Valentin Delchevalerie[a], Pierre Poitier[a], Benoît Frénay[a] and Bruno Dumas[a]

[a]Faculty of Computer Science, NaDI & NaXys institutes, University of Namur, Namur, Belgium

Abstract. Early lung cancer detection strongly increases survival rate. During a navigational bronchoscopy, pulmonologists perform tissue sampling for biopsies based on preoperative medical images. The bronchial carina is an airway structure that appears at each bronchus bifurcation. It is an important landmark to detect during navigations as it indicates the need to choose between multiple paths and keep track of the position in the lungs. In this paper, we assessed various deep learning pipelines including the use of semi-supervised, segmentation and recurrent methods under different setups to perform bronchial carinae detection. In contrast to most previous works that focus on phantoms, cadavers or virtual images, we exploit a large corpus of proprietary *in vivo* data captured during real endoscopic procedures using a mini probe. To the best of our knowledge, it is the first work that deals with this quantity of real and challenging data. After performing a comparison study, we conclude that the best performance to detect bronchial carinae is achieved by a semi-supervised pipeline that leverages the ability of nnU-Net to solve segmentation tasks, coupled with Gated Recurrent Units that extracts temporal contexts from image sequences.

1 Introduction

Lung cancer is the primary cause of cancer-related fatal outcomes in the world [18]. Currently, patients have a 5-year survival rate of 13.0% [20], which increases to 62.8% if diagnosed and treated early [9]. Hence, early diagnosis has a tremendous potential to save lives. In practice, pulmonologists detect potential cancers (i.e., pulmonary nodules) using medical imaging techniques such as Computed Tomography (CT) scans, and confirm or reject their diagnosis by performing tissue sampling for a biopsy. This process is known as navigational bronchoscopy [4], as they use an endoscope to navigate in the lungs, from the trachea to the target nodule(s).

Regrettably, over 60% of lung cancers are localised in the periphery of the lungs and cannot be reached with regular endoscopes because of the branching complexity of the bronchial tree and the size of these endoscopes [5]. Indeed, one must navigate many bronchial subdivisions, which become increasingly narrower, before reaching peripheral nodules. These subdivisions are harder to reach and the procedure is prone to miss-interpretations of the position of the endoscope by the pulmonologist. As a consequence, tissues are often taken blindly, resulting in a poor yield of biopsies [7]. While ultra-thin bronchoscopy [11] can help reaching small bronchial subdivisions, no previous research has been able to acquire a large amount of such precise data in order to assess the feasibility of using various deep learning techniques on it. Next to that, few solutions are proposed to provide accurate support for the navigation from the trachea to the pulmonary nodule(s). In the lung, there exists a structure known as the bronchial carina [14] (or carina, see Figure 1), which appears at each bifurcation in the bronchial tree as one moves from the trachea to the bronchioles (i.e., thin bronchus at the end of the tree, leading to pulmonary alveoli). Detecting a carina indicates that a bifurcation has been reached. At this location, the pulmonologist needs to select one of several bronchus (two in most cases) to navigate towards the target nodule(s).

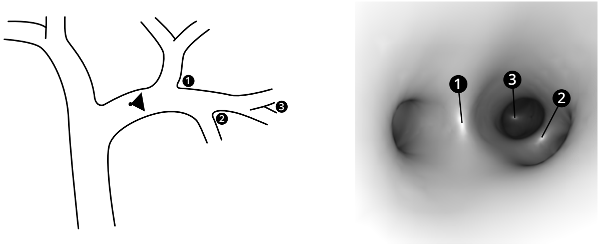

Figure 1: Inspired from Prakash et al. [14]. Three carinae are displayed with their position in the airways. Each of them indicates a bifurcation in the bronchial tree.

In order to bring an end-to-end solution to navigational bronchoscopy, three pieces can be considered: (i) a 3D map of the lungs, (ii) a Six Degrees Of Freedom (6DOF) Electromagnetic (EM) sensor, and (iii) a bronchial carinae detector. With these, it becomes conceivable to bring real-time support to help pulmonologists or a robotic system navigate to target nodule(s) in the lungs. Regarding the first point, the Airway Tree Modeling challenge (ATM22) [23] has yielded several pulmonary airway segmentation models that help create 3D representations of the lungs. One such example is Navi-Airway [21] and achieves accurate performances in constructing 3D models of the bronchial tree up to the 12th generation of bronchioles (see Figure 2) and preserves the airway topology. For the second piece, several 6DOF EM sensors have been developed and are able to give real-time position of the endoscope [6]. With the information they give, the voxels from the 3D model can be translated into real world positions. However, this information is not reliable enough since patient's breathing and movements add noise to those measurements [2]. Finally, a bronchial carinae detector, which is the subject of this paper, would give the signal that a bifurcation has been reached. Subsequently, the 6DOF information together with the 3D model could be used to indicate which bronchus has to be followed to move towards the target. Despite the apparent feasibility of a carina detector, it still constitutes a challenging task for many reasons such

[*] Corresponding Author. Email: robin.ghyselinck@unamur.be.

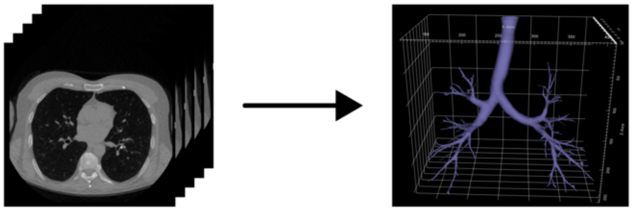

Figure 2: CT-scan slices from a real patient (left) and its corresponding 3D modeling of bronchial tree generated with NaviAirway (right).

as the high variability observed in images between different patients, and numerous artifacts that appears when dealing with real patients (e.g., breathing, movements, etc.).

In this work, a first step toward accurate *in vivo* navigational bronchoscopy support is proposed. It distinguishes from previous work by evaluating several deep learning techniques to detect bronchial carinae, such as binary classifiers (i.e., detector), segmentation, and recurrent models. Indeed, as opposed to some previous works that rely on geometry, deep learning has the potential to enhance robustness against variability and artifacts by leveraging a vast quantity of data available to us. Two pipelines are setup: a **supervised** pipeline and a **semi-supervised** pipeline. They benefit from a large amount of data coming from *in vivo* navigational bronchoscopy videos, which are partially annotated by biomedical experts. The supervised pipeline consists of two deep learning models trained on labeled images in order to detect carinae. In contrast, the semi-supervised pipeline generates pseudo-labels for all the frames of an unlabeled video dataset, and trains four deep learning models in a semi-supervised fashion. Two models use image embeddings and Recurrent Neural Networks (RNNs) (to assess the benefit from adding video temporality), while the two others directly use image embeddings with fully connected layers (to assess the benefit from the semi-supervised pipeline) instead. By evaluating those models on a third dataset of fully annotated videos, we show that the best performance is achieved by a semi-supervised pipeline that leverages the ability of nnU-Net [10] to solve segmentation tasks, coupled with the use of Gated Recurrent Units (GRU) [3] that accounts for the temporal context.

The main contributions of this work are the followings:

- detection of bronchial carinae with deep learning;
- ablation study of various deep learning techniques like nnU-Nets, recurrent and semi-supervised techniques in order to assess the result of such techniques on a large corpus of both labeled and unlabeled data;
- exploitation of images from *in vivo* navigational endoscopy recordings that include images of very narrow bronchus bronchi that traditional endoscopes cannot read (i.e., bronchioles and pulmopulmonary alveoli), as compared to other works that mostly use phantom, cadaver or virtual images of the first generations of bronchi.

This paper starts by reviewing related works, then it continues with a description of the data. Next, the proposed methodologies are described, followed by a section on experiments and results before concluding on the results and further works.

2 Related Works

Previous works can be divided into two different categories. The first one is mainly based on traditional computer vision approaches. In their work, Shen et al. [17] extract salient regions using a Maximally Stable Extremal Regions [13] detector on depth maps of video frames (obtained with shape from shading), and on depth maps of virtual images generated from CT scans. Sequentially, regions that indicate bifurcations are detected using a SVM classifier. Nonetheless, it appears that this work uses an Olympus BF-260 that has 5.5mm of outer diameter, which makes it impossible to reach smaller bronchi. Moreover, most of the work is validated on a phantom (i.e., an artificial model replicating the lungs' structure), with only 347 frames of *in vivo* bronchoscopic examination.

Sanchez et al. [15] detect lumen centers as the dark region of the image based on their appearance and geometry using gradient maps and k-means clustering. However, this work uses only 125 bronchoscopy images (with varying resolution of maximum 288×288 and minimum 114×144) which is very low to capture all the variance in the problem. Furthermore, it mainly uses geometry-based techniques. Most of this kind of approaches are likely to perform poorly in real applications due to a lack of robustness as stated in several works [16, 1]. Indeed, those methods generally make strong assumptions about geometrical properties of the airway, which diminishes its ability to generalize to new patients, and do not deal well with *in vivo* artifacts (breathing, movements, etc.).

The second group of approaches have shifted towards the use of deep learning techniques. This move was mainly motivated by the advances in the field, the availability of data, and the higher generalization power of such approaches compared to traditional ones. For example, Sganga et al. [16] use Convolutional Neural Networks (CNNs) to localize the endoscope in the lungs with simulated images from CT scans. However, this work is based on a robotic device that only operates in a restricted and controlled environment. It is also too expensive for most practitioners to acquire. Furthermore, the model is only evaluated on phantoms and cadavers (avoiding the hard task of dealing with artifacts), and the phantoms are limited to only 3-8 bronchus generations (avoiding the hard task of dealing with the periphery of the lung). A previous analysis by Borrego-Carazo et al. [1] uses RNNs and 3D Convolutions on synthetic data for vision-based bronchoscopy tracking, which tries to estimate the position of a virtual endoscope in the lungs. Unfortunately, the synthetic data are created from CT-scans, which is not realistic in comparison to *in vivo* images, both in terms of texture and conditions (e.g., no occlusion, breathing or artifacts). Under those circumstances, one cannot infer what the performance of the models would be in realistic conditions. Xu et al. adopted a lumen detection strategy for lung navigation [22] using encoder-decoder architectures. However, the data is a mix of phantom (3,818 frames) and *in vivo* images (2,871 frames) captured with an Olympus BF-P290 bronchoscope. This device has 4.2mm of diameter, which makes it physically impossible to reach smaller bronchi where nodules might be located. Moreover, this work does not use sequences of frames and only predicts the presence of lumens for a single frame without any context.

To summarize, every previous work faces at least one of the following issues: (i) a lack of generalization and robustness to real *in vivo* conditions, (ii) use of cadaver, phantom or virtual data, (iii) a need for specific and expensive equipment, such as robots, (iv) the treatment of frames as single, isolated images instead of sequences of data, and finally (v) the use of an endoscope that is too large to reach smaller bronchus, let alone bronchioles. Our work addresses those issues by assessing various deep learning techniques along with a large amount of real *in vivo* data that also involves bronchioles and pulmonary alveoli thanks to the use of a very specific medical hardware (a total of 777,522 frames from 252 videos featuring navigational endoscopy). Classic computer vision models like ResNet-50 are tested along with the use of segmentation, recurrent and semi-supervised techniques in order to assess which deep learning components can be useful for the task of carinae detection. On the one side, using segmentation technique may be useful as it gives a better feedback to the model while training. Indeed, model should not only perform a classification by highlight particular regions of the images, which forces the models to focus on the meaningful parts of the image. On the other side, recurrence may also be meaningful as a prediction on a frame at time t may also depend on the previous frames, especially if the camera punctually faces some challenging conditions (i.e., presence of artifacts), which is often the case in real endoscopic videos. Also, we consider in this work the use of a semi-supervised pipeline for leveraging a large amount of unlabeled data that can potentially still be useful for training. To the best of our knowledge, this is the first attempt to use unlabelled data for carinae detection.

3 Data

Through a collaboration with a company that employs biomedical experts and specializes in lung endoscopy, the authors had access to 252 proprietary videos recorded between 2019 and 2023 in 5 European hospitals. Those proprietary videos (recorded with a mini endoscope) show *in vivo* navigational bronchoscopies performed by a pulmonologist exploring a patient's airways. As the related works section points out, the datasets that have been explored in the past do not contain any footage from the smallest bronchi, due to the mechanical properties of the endoscopes that are used. Conversely, this dataset is captured with a mini probe (much smaller than regular endoscopes), and displays recordings of very small bronchi, up to the pulmonary alveoli (which is the farthest one can navigate in the lungs). The images extracted from those videos either contain one or several carinae or not. The videos are captured in RGB at a resolution of 400×400 pixels with 30 frames per second, and have a duration ranging from 30 seconds to 5 minutes. More precisely, the data used in the next sections can be divided into 3 distinct subsets as follows:

A labeled dataset of 7,000 images carefully selected and extracted from a subset of 100 videos. In particular, endoscopic procedure footages often contain a lot of artifacts due to (i) poor lighting conditions (too much light or not enough), (ii) physical obstructions or (iii) the presence of liquids such as blood, mucus, etc. Hence, the selection of frames is performed by experts at the company, with the idea to keep a set of meaningful, representative and diverse images that also feature challenging conditions. For each of those frames, carinae (if any) were highlighted using polygons. One should already mention that the proportion of frames that show one or several carinae, can differ significantly from one video to another, ranging from 0% to 100%. The resulting annotated set is imbalanced with 34% of video frames showing at least one carina. Figure 3 shows some examples of images along with their respective annotation mask.

A test dataset consisting of 5 videos (in full) that have been selected and annotated by experts for carinae detection. This means that for each frame, either the label 0 is assigned if there is no carina or the label 1 is assigned if there is at least one. Consequently, this test dataset results in a total of 24,119 labeled images (of which ∼18% display carinae). The selection criteria are video diversity (including challenging conditions), adequate length, and a sufficiently large amount of frames as compared to the labeled dataset (at least 3 times more in this case). These videos represent patients that are not featured in any of the other image sets. With this set, the intent is to simulate real navigational bronchoscopy and assess the performance of the models.

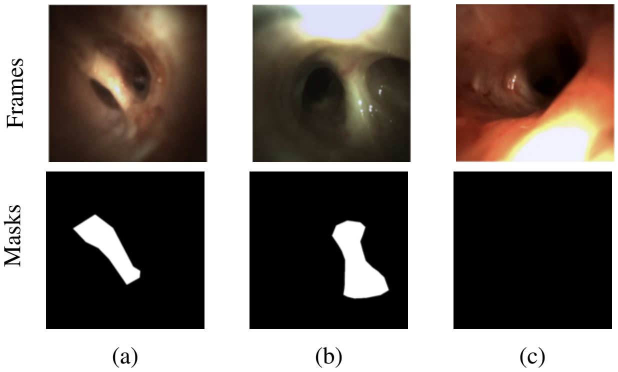

Figure 3: Example of images from the annotated set along with their corresponding mask provided by an expert. (a) and (b) contain a carina, while this is not the case in (c).

An unlabeled dataset of 147 videos (leading to 746,403 images or frames). The annotation process is difficult, as it is a time-consuming process that requires specific skills in medical image analysis. As a result, those 147 videos were not considered during manual annotation. Consequently, there is a large amount of unlabeled, yet potentially interesting images. A semi-supervised pipeline can capitalize on these by generating pseudo-labels (more on this in Section 4). Although precise quantification of this dataset's balancing is not possible under these circumstances, one can estimate it with a pseudo-labeling. This process is explained in the next section, and it shows that approximately 17% of images display carinae. This is yet another indication of the data imbalance property of this last dataset.

Both the labeled dataset and the unlabeled dataset exhibit data imbalance. In order to deal with such property, three different strategies are applied. First, the data loaders use a sampler, such that the data from the training come from 50% of each class. Second, the binary crossentropy loss function uses positive weights to apply an additional penalty for classification errors on the minority class. Third, the results are validated on the balanced accuracy metric (amongst other), which accounts for data imbalance. Those strategies are described in details in the subsequent sections.

4 Proposed Methodologies

This section presents our strategy for building bronchial carina detectors. On the one hand, segmentation models are able to capture masks from images, which can in turn indicate the presence of an object such as a bronchial carina. On the other hand, using Recurrent Neural Networks (RNNs) is a common choice for solving tasks involving sequences [19], such as video frames classification. Indeed, RNNs are able to learn meaningful context from the sequences by keeping track of previous events in their hidden units, and are therefore able to make predictions that are consistent in time. However, data available to us only feature sparse annotations in videos, which makes it impossible to run a straightforward training of RNN. Therefore, based on available data, we set up two deep learning pipelines. The first pipeline trains two bronchial carina detectors in a supervised way on the labeled set. The second one is a semi-supervised pipeline that uses these trained detectors on inference mode to generate pseudo-labels for all the frames of each videos in the unlabeled dataset. Subsequently, it trains recurrent and non-recurrent models on those pseudo-labels. The following sections gives more details about those two pipelines, and are summarized in Figure 4.

4.1 Carinae detection on labeled data

The supervised deep learning pipeline trains two different models on the labeled dataset in order to perform a binary classification task (i.e., a carinae detection, see Figure 4). Moreover, each of the two models are considered in the semi-supervised pipeline to generate pseudo-labels.

As it is a common baseline in deep learning, the first detector is made of a ResNet-50 [8] backbone pre-trained on ImageNet. The weights of the model are frozen, except for its classification layer. The latter is replaced by a linear layer with an output size of 1 for binary classification purposes. As the labels are initially segmentation masks, the positive label is beforehand assigned to images for which segmentation mask is present. The second model addresses the detection task by first considering a more general segmentation task. Starting from the segmentation masks generated by the model, one can infer the presence or absence of a bronchial carina in the image. The main motivation is that ground-truth masks give a much more informative feedback to the model. Indeed, instead of just checking if the binary output of the model is correct or not, the model is forced to highlight the meaningful regions of the images. While the ResNet-50 should figure out by itself the objects to detect in the images, it becomes much more straightforward when the problem is framed as a segmentation task. To do this, one can use nnU-Net [10], a well-known semantic segmentation method that can automatically configure U-Net architectures to a specific biomedical tasks. At the end, an average pooling with a sliding window of size 10×10 is performed on the output probabilities to smooth the predictions in the image, and a positive label is assigned if any probability is greater than 0.5.

4.2 Carinae detection on unlabeled data

As explained in Section 3, 147 videos lack expert annotation. The first step of this semi-supervised pipeline (see Figure 4) consists in generating pseudo-labels for each frame of these videos. To do so, the two models described in the previous section are used to generate pseudo-labels. For all the methods trained in the semi-supervised pipeline, the pseudo-labels used are those obtained thanks to the nnU-Net model in the supervised pipeline. This is motivated by the higher performances that are obtained compare to the same setup when using those from the ResNet-50. The nnU-Net model has been directly trained on the labeled dataset, whereas the ResNet-50 is pre-trained on ImageNet, similarly to the one from the previous section. The second step involves extracting the embeddings from each frames to serve as input features for the unlabeled data. In the case of the ResNet-50, the last 2048-dimensional embeddings are extracted. For the nnU-Net, the different embeddings from the encoding part of the U-Net are concatenated to build a 2016-dimensional embeddings, capturing features at different scales (more details about the architectures are given in the next section). After collecting pseudo-labels and embeddings for all the videos in the unlabeled set, four deep learning models are trained in the third step:

- **RNN nnU-Net**, which utilizes the embeddings generated by the nnU-Net as input for a simple Gated Recurrent Units (GRU) [3] network;
- **FC nnU-Net**, which inputs the embeddings generated by the nnU-Net into a simple Fully Connected (FC) layer;
- **RNN ResNet-50**, which channels the embeddings generated by the ResNet-50 as input for the GRU;
- **FC ResNet-50**, which passes the embeddings generated by the ResNet-50 to a FC layer.

The rationale for building such models is two folds, and is part of an ablation study. First, one can assess if the use of a semi-supervised pipeline that fine-tunes a FC layer improves the model's performance (by comparing the FC ResNet-50 and FC nnU-Net with the ResNet-50 and nnU-Net, respectively). In one case, a pre-trained ResNet-50 (on ImageNet) is used. This model is the same as the one from the supervised pipeline. In the other case, the nnU-Net model was trained on the labeled dataset (in the supervised pipeline). This means that the difference between the two pipelines is the underlying training set. While, the supervised pipeline uses the labeled dataset, the semi-supervised pipeline uses the unlabeled dataset (for which pseudo-labels were generated). Second, one can assess if processing sequences of frames (with a RNN) instead of individual frames (with a FC layer) enhances performance (by comparing the FC ResNet-50 and FC nnU-Net with the RNN ResNet-50 and RNN nnU-Net, respectively). Sequential models are good candidates for treating video data in theory [19] because experts in endoscopy reported that it is helpful to visualize previous frames when labeling a particular one, as context matters.

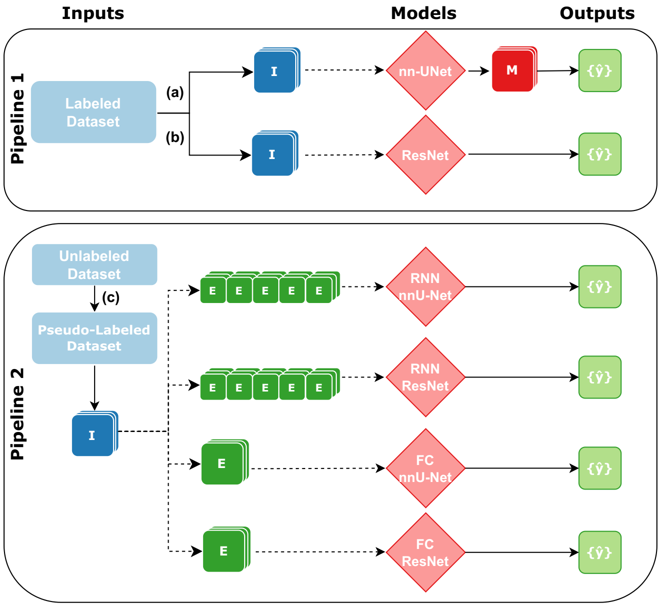

Figure 4: **Pipeline 1 (supervised)** shows the training process for two models on the labeled dataset: (a) takes images (I) as input to generate masks (M), and output the probabilities of detecting a carina ($\{\hat{y}\}$), while (b) directly outputs probabilities from the images. **Pipeline 2 (semi-supervised)** shows the training process for the four models on the unlabeled dataset. (c) represents pseudo-labels generation for the unlabeled dataset (with the two models from pipeline 1). Subsequently, two RNNs and two FC models are trained using these generated pseudo-labels. The RNNs take sequences of embeddings (E) as inputs whereas the FC models take a single embedding (E). Finally, the output is a probability.

5 Experiments and Results

In the first part of this section, more practical details are given about the data preprocessing steps, the architecture of the different models, the meta-parameter choices, the optimizer setup, and metrics selected for models evaluation. In the second part, results are presented before concluding this section with a discussion.

5.1 Experimental Setup

5.1.1 Data preprocessing

Section 4 presents two models in the supervised pipeline. The fine-tuning of the pre-trained ResNet-50 starts with data augmentation on the 400×400 RGB input images of the labeled dataset. First, a resize to 236×236 pixels is performed, then a random crop to 224×224, which is followed by a rotation of $0°, 90°, 180°$ or $270°$ with probability 0.25 each. After this, a horizontal flip or a vertical flip with probability 0.5 each is applied. Finally, the image is normalized. As already stated in Section 3, the labeled set is imbalanced regarding the two classes (34% of images show one or several carinae). Therefore, an imbalanced sampler is used for loading images into training batches. This sampler uses a multinomial sampling, based on the label frequencies in the data set. As a result, this sampler will force each training batch to contain ∼50% of images of each class. In order to validate the different models when performing grid-search for different meta-parameters, a stratified 5-fold cross validation is used. The stratification helps in dealing with the diversity of the videos, both in terms of carina ratio (from 0% to 100% of the frames may show carinae) and of labeled frames count (each video contains between 11 to 457 frames). Indeed, with the stratification, each of the 5 folds has the same proportion of any video. The second model presented in Section 4, nnU-Net, also takes the 400×400 RGB input images and is trained with its standard preprocessing [10].

The semi-supervised pipeline also uses a 5-fold cross validation scheme to train 4 models on the unlabeled dataset's videos. It applies the same preprocessing for both FC and recurrent models. In this process, an imbalanced sampler is used too, such that ∼50% of the frames considered display a (pseudo-labeled) carina. Regarding recurrent models, one must insist on the fact that full video sequences are not used to avoid imbalance and overfitting issues. The different images for each video are pre-processed into embeddings (see Figure 4) instead of directly using the 400×400 RGB images. These embeddings vary in dimensions based on the model used for extraction. The pre-trained ResNet-50 yields embeddings with 2048 dimensions, leading to sequences of 2048-dimensional embeddings. In contrast, the trained nnU-Net generates embeddings of 2016 dimensions, leading to sequences of 2016-dimensional embeddings. This brings two practical advantages. First, the training is faster by reducing the amount of I/O operation on the disk (data can fit in memory with 64GB of RAM), which benefits both FC and recurrent models. Second, it makes the training of the recurrent models more efficient by focusing on learning temporal relationships instead of having to extract features from the images. Specifically, the GRUs that are proposed take advantage of the local temporal relationships by adding contextual information to each of the video frames. In particular, the context contains the embeddings from the 5 previous frames. Then, this context is used in the GRUs to predict the label of the current frame. It is important to note that frames in the future are never used as contextual information. Indeed, as the model is used in an online fashion during inference, only an auto-regressive scheme can be utilized. Contexts of different sizes have also been evaluated without any improvement in performance.

5.1.2 Architecture and Meta-Parameters

Most of the meta-parameters given below are obtained thanks to a grid-search process with the 5-fold cross validation described previously. For each of the 5 folds, and for each model, the epoch showing the highest validation balanced accuracy is kept. One should note that to avoid overfitting, a higher validation accuracy can be considered only if there has been less than 10% of continuous training epochs without any improvement. This strategy allows for the implementation of early stopping for each of the 5 folds individually.

In the supervised pipeline, no modifications are applied to the ResNet-50 except the size of the last layer, which is resized to 1 to allow fine-tuning for binary classification. Training is performed through 200 epochs with a batch size of 256. The AdamW [12] optimizer is used along with a constant learning rate of 5×10^{-3}, a weight decay of 10^{-2}, early stopping after 20 epochs without any improvement and gradient clipping. The selected loss function is a Weighted Binary Cross-Entropy (WBCE) loss. Its weight is based on the positive label frequency in the dataset.

The architecture of the U-Net is obtained thanks to nnU-Net, and mainly consists of 7 encoder blocks and 6 decoder blocks, with 2 convolutional layers with a kernel size of 3×3 per block. Encoding blocks are followed by a 2×2 pooling operation, except for the first encoding layer. The number of feature maps for the first block is set to 32, and is multiplied by 2 after each block until it reaches the maximal value of 512. As nnU-Net aims to automatically infer the best architecture and training setup, neither the architecture, nor the training loop have been modified by the authors.

As already explained, FC models and RNNs from the semi-supervised pipeline directly use the embeddings either from the pre-trained ResNet-50 or the trained nnU-Net. The two FC models and the two GRUs were trained during 50 epochs each, with a learning rate of 10^{-5} and the AdamW optimizer. Similarly to the supervised pipeline, the loss function is a WBCE. This loss only differs from the one of the supervised pipeline by the value of the weight, since the datasets differ. Moreover, gradient clipping applies and the same early stopping as described above is used.

The architecture of the FC ResNet-50 and of the FC nnU-Net are straightforward. The embeddings of 2048-d (ResNet-50) or 2016-d (nnU-Net) are passed to a linear layer with 32 hidden units, followed by a ReLU activation function, and sent to a last linear layer with output size 1 for binary classification.

Regarding recurrent models, the best architecture among several options is a GRU with 32 hidden units, 1 recurrent layer, and a dropout rate of 0.3. As 5 frames of contextual information are added to the current one, the GRU takes sequences of 6 embeddings as input. Common values of each of those hyper parameters were tested and the one with the best results on the test dataset were kept. The LSTM architecture was also used and the performance was similar to that of the GRU. Since the primary interest was to understand the impact of RNNs against other techniques, LSTMs were not investigated in further details.

5.1.3 Evaluation and Metrics

In order to rigorously evaluate the performance of the 6 deep learning models, one can generate predictions on the test set made of 5 fully annotated videos featuring unseen patients (see Figure 5). As mentioned earlier, each model is trained with a 5-fold cross-validation approach, leading to 5 distinct trained versions for each of the 6 deep learning models. Each of these is used to calculate the frame-by-frame probability that a given image shows at least one carina. Then, a single consolidated prediction is computed for each model from the average of the output probabilities across its 5 folds. Finally, a threshold of 0.5 is applied to assign the final probability. A probability lower than 0.5 corresponds to no carina detected, and a probability higher than 0.5 corresponds to a detection of at least one carina.

Based on this inference, two metrics are assesses for each method on all the videos contained in the test dataset. Given the imbalance of the data, the balanced accuracy is used as a first evaluation metric. Indeed, this metric averages the recalls and can therefore show a potential poor performance on the minority class. Next to that, the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) is a common choice for assessing the quality of models dealing with medical data. In particular, it has the advantage of showing the impact of moving the classification threshold (from 0 to 1) on true positive and false positive predictions. Finally, for each frame of the test set, the predictions of each model against the ground truths is qualitatively assessed as a temporal sequence segmentation. This assessment presents the ability of each model to recognize the transitions of sequences with and without a carina.

5.2 Results

Table 1 shows the results of each model on the selected performance metrics for each video individually, as well as the mean computed over those 5 videos (see test dataset in Section 3). This table shows no comparison to other methodologies (see Section 2) as there is no prior similar work. Instead, our work is a first step towards an accurate classification of bronchial carina. For that purpose, we compare several architectures (described in Section 4) that allow us to draw conclusion on the relevance of classification, segmentation, and recurrent models. Since the result presented for each video is an average of the probabilities from the 5-folds, the 95% Confidence Interval (CI) is presented next to the average for each video. In many cases, the width of the CI is very small, even zero sometimes. The results of the supervised pipeline (top part of the table, methods related to Section 4.1) show that for models trained on the labeled set, the ResNet-50 has a higher balanced accuracy than the nnU-Net for 3 out of 5 videos, and performs better on average. In terms of AUC ROC, the nnU-Net has higher scores for all 5 videos. The results of the semi-supervised pipeline (bottom part of the table, methods related to Section 4.2) show that the RNN nnU-Net has a higher balanced accuracy for all the 5 videos. Among the models of this semi-supervised pipeline, RNN nnU-Net achieves the highest AUC ROC scores for 3 out of 5 videos. Finally, when comparing the results from the two pipelines altogether, one notices that RNN nnU-Net has the highest balanced accuracy for 5 out of 5 videos, while nnUnet has the highest AUC ROC for 3 out of 5 videos and on average.

Finally, Figure 6 shows a qualitative temporal segmentation for video 5 where the ground truth (from human annotator) is compared against the predictions obtained with each model. This video was selected as it exhibits the greatest number of transitions between temporal sequences featuring a carina and those without. Regions in white indicate the absence of a bronchial carina, whereas regions in gray indicate its presence. Figure 6 shows that all models are quite capable of dealing with the first half of the video. Between frames 3,250 and 5,000, all models are aligned and do not detect any carina, which is the expected prediction as shown by the ground truth. However, it seems that RNN nnU-Net, FC nnU-Net, and FC ResNet-50 hallucinate by detecting some carinae between frames 5,000 and 6,000, where there was in fact none. Near frame 6,000, only the RNN nnU-Net and the FC nnU-Net are able to recognize (some) carinae. No model was able to recognize the sequence near frame 6,500. Finally, all models deal well with the end of the sequence.

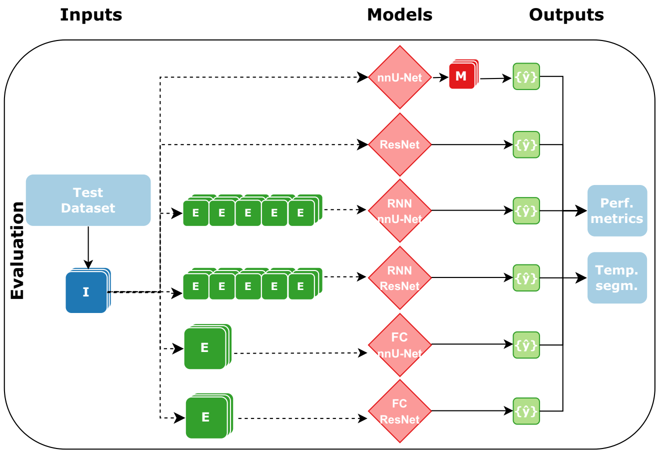

Figure 5: Process by which all models are evaluated on the test dataset. The two first models directly use the images (I) whereas the embeddings (E) are extracted from images (I) in the four other cases. The nnU-Net extracts a mask (M) before outputing a probability, where the five other models directly generate aprobability. Finally, the output ($\{\hat{y}\}$) of each model is quantitatively assessed.

5.3 Discussion

The models of the supervised pipeline perform differently depending on the considered evaluation metric. On average, the balanced accuracy is favorable to the ResNet-50, while the AUC ROC is higher for nnU-Net (see Table 1). Note that the very narrow confidence interval are most likely due to the unbalanced nature of the data and the sensitivity of AUC ROC to imbalance data.

The models of the semi-supervised pipeline provide information on two levels: datasets and temporal context. First, one can see the impact of a semi-supervised approach by comparing models trained on the labeled dataset against those trained on the unlabeled dataset (ResNet-50 with FC ResNet-50, and nnU-Net vs FC nnU-Net). In terms of balanced accuracy, FC models show higher scores than models from the supervised pipeline. However, in terms of AUC ROC, models from the supervised pipeline have higher scores than their FC version from the semi-supervised pipeline. Second, one can assess the impact of using the temporality of the context by comparing the performance of FC models (FC ResNet-50 and FC nnU-Net against RNN ResNet-50 and RNN nnU-Net) against recurrent models. While the RNN nnU-Net shows higher average balanced accuracy and AUC ROC than its FC counterpart, the reverse is true for the RNN ResNet-50. This indicates that the context given by the previous frames to the RNN nnU-Net help improve its performance while it is not the case for the ResNet-50.

Overall, through the lenses of the balanced accuracy, the RNN nnU-Net has the highest performance across all the videos (on average on the 5-folds). This seems to indicate that the embeddings from nnU-Net were able to capture meaningful features from the images. Besides, the addition of the GRU helps in dealing with the temporality. In the context of detecting bronchial carinae, a wrong prediction would lead to a wrong indication during navigational bronchoscopy, regardless of the error type. Indeed, either one misses a carina (false negative) or one wrongly detects a carina (false positive). Therefore, both indications would incorrectly lead to a movement in the 3D map of the lungs. In light of these elements, the RNN nnU-Net could be considered as the best of the proposed models. However, if one is interested in detecting true positive (detecting all the carinae) with minimum false positives, the AUC ROC becomes an interesting performance metric. Looking at the AUC ROC, the conclusion

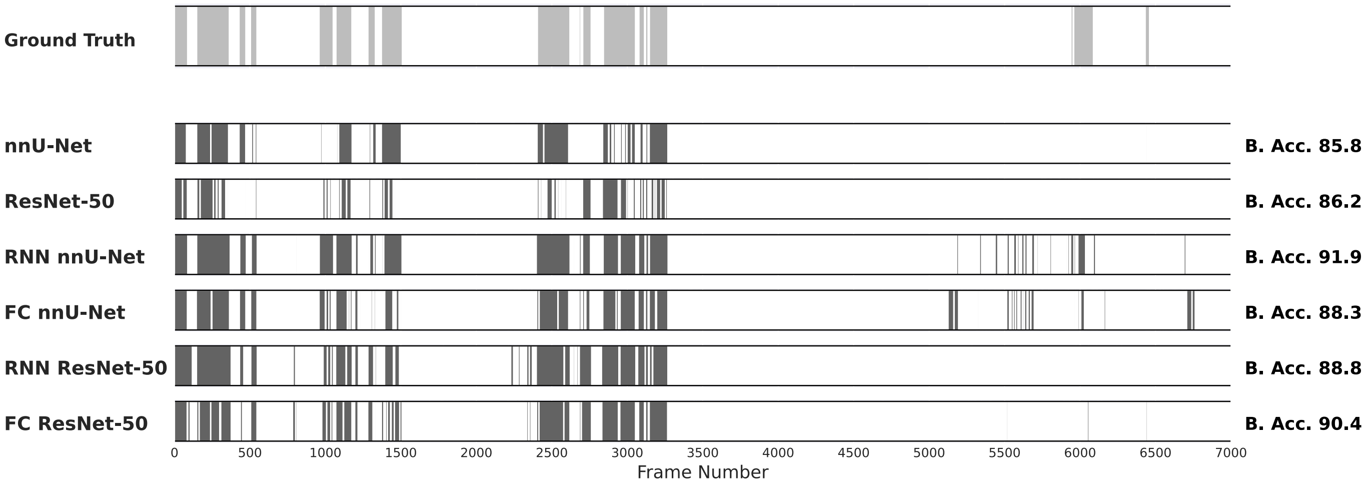

Figure 6: Temporal sequence testing in the particular case of video 5. Gray vertical bars correspond to frames with at least one carina, while white vertical bars correspond to frames without any carina. The first horizontal bar plot is the ground truth for each video frame, while others are the predictions made by specific models. B. Acc. stands for Balanced Accuracy. See Table 1 for a quantitative analysis on 5 videos.

Table 1: Balanced accuracy and AUC ROC for the different models on each video (5-fold average ±95% confidence interval) and on average (higher scores means better performance). The first two lines correspond to methods in the supervised pipeline, while the next four lines correspond to methods in the semi-supervised pipeline. Highest values for each pipeline are highlighted in bold.

	Balanced Accuracy (%)						*AUC ROC (%)*					
	Video 1	Video 2	Video 3	Video 4	Video 5	**Avg.**	Video 1	Video 2	Video 3	Video 4	Video 5	**Avg.**
nnU-Net	**69.9**±9.4	**75.8** ±0.6	71.8±1.7	66.5±1.3	85.8±2.3	74.0	**98.8**±0.0	**91.3**±0.0	**93.6**±0.0	**90.7**±0.0	**97.7**±0.0	**94.4**
ResNet-50	67.0±0.5	72.1±1.2	**80.1**±0.6	**69.0**±1.6	**86.2**±0.9	**74.9**	75.8±0.0	87.8±0.0	92.8±0.0	84.6±0.0	96.6±0.0	87.5
RNN nnU-Net	**73.7**±1.9	**78.6**±2.8	**84.7**±2.2	**89.2**±2.7	**91.9**±2.0	**83.6**	**84.2**±0.0	86.8±0.0	**92.5**±0.0	92.2±0.0	**97.8**±0.0	**90.7**
FC nnU-Net	71.1±3.1	75.4±0.7	83.9±0.9	88.7±1.1	88.3±1.4	81.5	80.3±0.0	85.0±0.0	92.1±0.0	**93.0**±0.0	96.4±0.0	89.4
RNN Resnet-50	58.8±6.1	73.8±1.6	82.6±1.7	74.7±8.3	88.8±2.5	75.7	69.3±0.1	84.0±0.0	92.3±0.0	83.4±0.1	96.4±0.0	85.1
FC ResNet-50	62.0±1.5	77.8±0.5	83.6±1.4	80.6±1.3	90.4±0.7	78.9	72.4±0.0	**87.0**±0.0	92.2±0.0	88.2±0.0	96.3±0.0	87.2

is different. Indeed, the nnU-Net from the supervised pipeline shows better results, except for videos 4 and 5. On average, the result of 94.4% is roughly 4% better than the second best performing model, the RNN nnU-Net. These observations show that the nnU-Net either standalone or combined with a RNN performs better than ResNet-50 models. Finally, one should note that although the 5 selected videos contain 24, 119 frames, the results of the analysis would likely be different should one use different videos as a test set.

A visual inspection of Figure 6 shows more errors after frame 5,200. An hypothesis to explain this lower performance is that we may have reached narrower regions where structures are smaller and harder to distinguish. However, this needs to be evaluated in full depth because the endoscope is going back and forth in the bronchial tree. In the current datasets, the depth in the bronchial tree is not available. With this information, one could gain insights on the change in accuracy depending on the depth in bronchial tree.

6 Conclusion

This paper exploits three different datasets featuring *in vivo* images from endoscopic procedures: the first one has 7, 000 labeled images from 100 videos, the second one has 746, 403 unlabeled images from 147 videos, for which pseudo-labels were generated using a deep learning model, and the third one has 24, 119 images for which frame-by-frame annotations are available. Six deep learning models are explored for detecting bronchial carinae during navigational bronchoscopies. These models are assessed on two performance metrics, which are the balanced accuracy and the AUC ROC. The experiments performed in this work show that both the use of nnU-Net, RNNs and a semi-supervised pipeline seem to be beneficial. More precisely, the standalone nnU-Net trained in a supervised pipeline and the RNN nnU-Net trained in the semi-supervised pipeline achieve the best balanced accuracy and AUC ROC, respectively. One can therefore conclude that those methods should be considered for navigational bronchoscopy.

In the future, one could spend additional effort in collecting more labeled data, either in form of mask annotations with polygons or binary annotations indicating the presence of carinae. Indeed, even if the pseudo-labels combined with RNNs increased the performance on some evaluation metrics, using ground truth could lead to an improvement in performance of the models. Moreover additional data augmentation techniques, such as GANs or VAEs could be explored. Indeed, these have already shown to improve training quality and help models to better generalize. In the medical field where labeled data are sparse, they seem like a natural candidate to explore.

This work is the first step towards building a robust navigational bronchoscopy system. In the future, one should investigate multimodal data fusion, such as combining (i) a 6DOF position from an electromagnetic system, (ii) a 3D map of the lungs that shows the path to follow to reach pulmonary nodules, and (iii) a bronchial carina detector that would indicate that a bifurcation was reached, and that the 3D map should be updated to continue the navigation.

Acknowledgements

R.G. benefits from the support of University of Namur, F.S.R grant. V.D. and P.P. both benefit from the support of the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS). The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247. The data used for the purpose of this study were captured with a CE compliant probe and patient's consent. The authors thank Benjamin Mertens, Charline Dardenne, Julien Albert, Sacha Corbugy and Simon Lejoly for their comments and the fruitful discussions on this paper.

References

[1] J. Borrego-Carazo, C. Sanchez, D. Castells-Rufas, J. Carrabina, and D. Gil. Bronchopose: an analysis of data and model configuration for vision-based bronchoscopy pose estimation. *Computer Methods and Programs in Biomedicine*, 228:107241, Jan. 2023.

[2] A. Chen, N. Pastis, B. Furukawa, and G. A. Silvestri. The effect of respiratory motion on pulmonary nodule location during electromagnetic navigation bronchoscopy. *Chest*, 147(5):1275–1281, May 2015.

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, Oct. 2014.

[4] J. Cicenia, S. K. Avasarala, and T. R. Gildea. Navigational bronchoscopy: a guide through history, current use, and developing technology. *Journal of Thoracic Disease*, 12(6):3263–3271, June 2020.

[5] X. Dhillon and Y. Harris. Bronchoscopy for the diagnosis of peripheral lung lesions. *Journal of Thoracic Disease*, 9(Suppl 10):S1047–S1058, Sept 2017.

[6] X. Duan, D. Xie, R. Zhang, X. Li, J. Sun, C. Qian, X. Song, and C. Li. A novel robotic bronchoscope system for navigation and biopsy of pulmonary lesions. *Cyborg and Bionic Systems*, 4:0013, Jan. 2023.

[7] Y. Han, H. J. Kim, K. A. Kong, S. J. Kim, S. H. Lee, Y. J. Ryu, J. H. Lee, Y. Kim, S. S. Shim, and J. H. Chang. Diagnosis of small pulmonary lesions by transbronchial lung biopsy with radial endobronchial ultrasound and virtual bronchoscopic navigation versus CT-guided transthoracic needle biopsy: A systematic review and meta-analysis. *PLOS ONE*, 13(1):e0191590, Jan. 2018.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, June 2016.

[9] N. C. Institute. Seer stat fact sheets: Lung and bronchus cancer, 2023. Consulted on: October 31, 2023.

[10] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec. 2020.

[11] S. H. Kim, J. Kim, K. Pak, and J. S. Eom. Ultrathin bronchoscopy for the diagnosis of peripheral pulmonary lesions: A meta-analysis. *Respiration*, 102(1):34–45, Nov. 2022.

[12] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, Sept. 2004.

[14] U. B. Prakash and R. S. Fontana. Functional classification of bronchial carinae. *Chest*, 86(5):770–772, Nov. 1984.

[15] C. Sánchez, J. Bernal, D. Gil, and F. J. Sánchez. On-line lumen centre detection in gastrointestinal and respiratory endoscopy. In *Clinical Image-Based Procedures. Translational Research in Medical Imaging*, pages 31–38. Springer International Publishing, 2014.

[16] J. Sganga, D. Eng, C. Graetzel, and D. B. Camarillo. Autonomous driving in the lung using deep learning for localization. *ArXiv:1907.08136*, 2019.

[17] M. Shen, S. Giannarou, P. L. Shah, and G.-Z. Yang. Branch: Bifurcation recognition for airway navigation based on structural characteristics. In *MICCAI 2017*, volume 10434, pages 182–189, 2017.

[18] R. Siegel, D. Naishadham, and A. Jemal. Cancer statistics. *A Cancer Journal for Clinicians*, 62(1):10–29, Jan. 2012.

[19] K. Smagulova and A. P. James. A survey on lstm memristive neural network architectures and applications. *The European Physical Journal Special Topics*, 228(10):2313–2324, Oct. 2019.

[20] J. P. van Meerbeeck and C. Franck. Lung cancer screening in europe: where are we in 2021? *Translational Lung Cancer Research*, 10(5): 2407–2417, May 2021.

[21] A. Wang, T. C. C. Tam, H. M. Poon, K.-C. Yu, and W.-N. Lee. Naviairway: a bronchiole-sensitive deep learning-based airway segmentation pipeline, 2022.

[22] J. Xu, T. Zhang, Y. Wu, J. Yang, G.-Z. Yang, and Y. Gu. Cdfi: Cross domain feature interaction for robust bronchi lumen detection, 2023.

[23] M. Zhang, Y. Wu, H. Zhang, Y. Qin, H. Zheng, W. Tang, C. Arnold, C. Pei, P. Yu, Y. Nan, G. Yang, S. Walsh, D. C. Marshall, M. Komorowski, P. Wang, D. Guo, D. Jin, Y. Wu, S. Zhao, R. Chang, B. Zhang, X. Lu, A. Qayyum, M. Mazher, Q. Su, Y. Wu, Y. Liu, Y. Zhu, J. Yang, A. Pakzad, B. Rangelov, R. S. J. Estepar, C. C. Espinosa, J. Sun, G.-Z. Yang, and Y. Gu. Multi-site, multi-domain airway tree modeling. *Medical Image Analysis*, 90:102957, Dec. 2023.