IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 November 2023, accepted 22 December 2023, date of publication 5 January 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350430

RESEARCH ARTICLE

Computer-Aided Diagnosis System for Lung Fibrosis: From the Effect of Radiomic Features and Multi-Layer-Perceptron Mixers to Pre-Clinical Evaluation

M. FONTANELLAZ[1], A. CHRISTE[2], S. CHRISTODOULIDIS[3], (Member, IEEE), E. DACK[1], J. ROOS[4], D. DRAKOPOULOS[2], D. SIERON[5], A. PETERS[2], T. GEISER[6], M. FUNKE-CHAMBOUR[6], J. HEVERHAGEN[2], H. HOPPE[7,8,9,10,11], A. K. EXADAKTYLOS[12], L. EBNER[2,13], AND S. MOUGIAKAKOU[1,2,12], (Member, IEEE)

[1]AI in Health and Nutrition, ARTORG Center for Biomedical Engineering Research, University of Bern, 3012 Bern, Switzerland
[2]Department of Diagnostic, Interventional, and Pediatric Radiology, Inselspital, Bern University Hospital, University of Bern, 3012 Bern, Switzerland
[3]Laboratory of Mathematics and Informatics (MICS), CentraleSupélec, Université Paris Saclay, 91190 Gif-sur-Yvette, France
[4]Institute of Radiology and Nuclear Medicine, Luzerner Kantonsspital, Lucerne, 6000 Canton of Lucerne, Switzerland
[5]Division of Magnetic Resonance Imaging, Silesian Center for Heart Diseases, 41800 Zabrze, Poland
[6]Department of Pulmonary Medicine, Inselspital, Bern University Hospital, University of Bern, 3012 Bern, Switzerland
[7]Department of Radiology, Lindenhofspital, 3001 Bern, Switzerland
[8]Campus Stiftung Lindenhof, 8090 Bern, Switzerland
[9]Swiss Institute for Transnational and Entrepreneurial Medicine, 3010 Bern, Switzerland
[10]University of Bern, 3012 Bern, Switzerland
[11]Department of Health Sciences and Medicine, University of Lucerne, 6002 Lucerne, Switzerland
[12]Department of Emergency Medicine, Inselspital, University Hospital Bern, University of Bern, 3012 Bern, Switzerland
[13]Department of Radiology, Cantonal Hospital Lucerne, 6000 Luzern, Switzerland

Corresponding author: S. Mougiakakou (stavroula.mougiakakou@unibe.ch)

This work was supported in part by the Emergency Department and the Department of Diagnostic, Interventional, and Pediatric Radiology of Inselspital Bern; and in part by the Campus Stiftung Lindenhof (Campus SLB) and the Hasler Foundation.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Medical image segmentation is a crucial element of computer-aided diagnosis (CAD) systems. Segmentation maps are used to calculate imaging features, such as quantitative disease distribution and radiomic features. Since their introduction in 2015, UNets have become the state-of-the-art segmentation tools. However, since that time, many new methods for image processing have been introduced, such as vision transformers and multi-layer-perceptron-mixers (MLP-Mixers). Alongside baseline UNets, we have now investigated the application of such MLP-Mixers for medical image segmentation, as part of a CAD system for the diagnosis of interstitial lung diseases (ILDs). Furthermore, we have investigated the effect of 2D and 3D data representations on segmentation and the final CAD results. We have evaluated the performance of the baseline segmentation methods and the MLP-Mixer primary on the overall diagnostic performance of the CAD system - as well as on the accuracy of segmentation as an intermediate step. In addition to network and data representation variations, we have investigated two different techniques for selecting features, an agnostic method and an alternative approach which selects features tailored to a specific segmentation map and diagnosis task. Finally, the CAD's performance was compared with that of four independent specialists in chest radiology. Among the 105 test cases, the diagnostic accuracy was 77.2±1.6% for the AI-approaches and 79.0±6.9% for the radiologists, indicating that the proposed systems perform comparably well to human readers in most of the cases. For the task of ILD pattern segmentation, similar results were obtained with 3D data and 2D tomography slices.

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

source: https://doi.org/10.48350/194109 | downloaded: 28.4.2024

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

INDEX TERMS Chest X-ray, chest computed tomography, interstitial lung diseases, radiomics, computer-aided diagnosis, segmentation, machine learning.

I. INTRODUCTION

The umbrella term of interstitial lung diseases (ILDs) encompasses a heterogeneous group of chronic lung disorders characterised by either fibrosis and/or inflammation of the interstitium. A large group of ILDs is of unknown aetiology with broadly variable fibrosis and inflammation [1]. As shown in Fig. 1, this group of ILDs is referred to as idiopathic interstitial pneumonias (IIPs). Idiopathic pulmonary fibrosis (IPF) is the most prevalent IIP and is associated with the poorest prognosis. Therefore, accurate diagnosis of IPF is essential for optimal treatment [2]. It is difficult to make a reliable diagnosis, as the clinical presentation may be similar or overlapping in different IIPs.

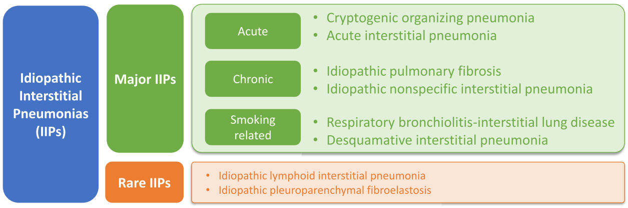

Idiopathic Interstitial Pneumonias (IIPs)
Major IIPs
Acute
Cryptogenic organizing pneumonia
Acute interstitial pneumonia
Chronic
Idiopathic pulmonary fibrosis
Idiopathic nonspecific interstitial pneumonia
Smoking related
Respiratory bronchiolitis-interstitial lung disease
Desquamative interstitial pneumonia
Rare IIPs
Idiopathic lymphoid interstitial pneumonia
Idiopathic pleuroparenchymal fibroelastosis

FIGURE 1. Overview over IIPs with similar radiological patterns, such as IPF, non-specific interstitial pneumonia (NSIP), respiratory bronchiolitis, cryptogenic organising pneumonia, and lymphocytic interstitial pneumonia as introduced in [3].

The differential diagnosis of IIP is largely based on recently proposed harmonised guidelines [4]. A preliminary diagnosis is made by radiologists in conjunction with a clinical board of pneumonologists and histopathologists, and employing the classification into usual interstitial pneumonia (UIP), typical UIP, probable UIP, indeterminate UIP, and non-IPF, [4], [5]. Computer Tomography (CT) scans exhibit a high positive predictive value for UIP and are assigned a pivotal role in the diagnostic process within the proposed Fleischner Society guidelines of 2018 [6]. Therefore, if a typical or probable UIP pattern is present on CT images, additional invasive procedures such as transbronchial or surgical biopsy can be avoided [4]. The most clinically relevant distinction among IIPs is between IPF and NSIP [3], [7]. Early distinction between these two diseases is crucial for making fundamental decision on treatment.

Due to the difficulty in the classification of IIPs and the benefits of CT scans, computer vision (CV) methods are an attractive approach for enabling faster and more efficient patient outcomes. Recent reviews [8], [9], [10] have underscored the benefits of utilising Artificial Intelligence (AI) techniques, such as Convolutional Neural Networks (CNNs), for tasks related to the diagnosis and prognosis of ILDs. CNNs can fuse local texture-based features with high-level global context features depending on their architecture. Features of the former type are generated by the first few layers in the CNN, whereas the latter feature type is generated by neurons with large receptive fields from deeper layers. The larger an input image is, the more layers are needed to be stacked to achieve a reasonable receptive field size. Therefore, CNNs are not always the ideal choice when working with imaging data. Modern techniques in the field of CV suggest that vision transformers(ViTs) could be used [11], as the global context is available from the first dedicated layer and throughout the network. Several implementations of ViTs have already been applied to medical images [12], [13], [14], [15]. In this research, our primary focus is to explore innovative AI methods as alternatives to those relying on attention mechanisms as those seen in ViTs. This approach aims to improve efficiency while achieving good generalisation on our datasets.

Throughout this study we design and evaluate algorithmic approaches for the segmentation of ILD patterns as well as for the diagnosis of ILDs. Within this context, we investigated i) whether novel image processing methods such as MLP-Mixers can be applied in future studies to medical image segmentation to overcome the limitations implied by CNNs (minimum network-depth for adequate visual field size), ii) the difference in performance between 2D and 3D data, and iii) the effect of feature selection methods in the identification of the most informative features. We applied various deep learning (DL) models to retrieve segmentation maps, which were used to assess the pathological pattern distribution and to calculate dedicated radiomic features for UIP patterns. Disease distribution, radiomic and demographic (age and gender) features were used to predict case affiliation for two different tasks; the first task was a 4-way diagnosis into a typical UIP CT pattern, a probable UIP CT pattern, a CT pattern indeterminate for UIP, or CT features that are most consistent with a non-IPF diagnosis, - all according to the guidelines published in [6]. The second task was a 2-way diagnosis with pooled classes where UIP patterns and the non-UIP patterns were joined together. The optimised system performance was compared with the performance of four chest radiologists.

II. RELATED WORKS

A. COMPUTER-AIDED DIAGNOSIS

The introduction of computer-aided diagnosis (CAD) tools marked the development of systems capable of imitating human experts in the field of diagnosis [18]. With the ongoing advances in CV and AI, especially in the field of Deep Learning (DL) [11], [12], [13], [14], [15], [16], [17], CAD systems are deriving substantial benefits ([19]. These systems offer the advantage of providing rapid and dependable information to assist human experts. Nevertheless, they have encountered challenges that require attention, particularly in advanced segmentation techniques for medical images [20]. Despite this, we have seen great success of CADs in radiology, with breast cancer diagnosis being comparable to human experts [21] and implemented as the second blind reader in many hospitals in Europe [22]. While comparable achievements have been seen in CAD's application to UIP pattern detection [5], [23], the broader implementation of such powerful tools as the norm is yet to be realized in this specialized field [24]. Reasons for the absence of production ready CAD systems is mainly the scarceness and quality of annotated data, as well as model generalization deterioration across different populations and datasets acquired by a diverse set of machines. The largest influence on annotation quality is given by considerable inter-observer variability. Inter-observer agreement depends on the specialists and the pattern itself [25], [26], [27]. To address these challenges, we propose a novel segmentation method that integrates into an existing CAD design [5], aiming to surmount these barriers and enhance the accuracy of pulmonary fibrosis diagnosis.

B. DL AND ILDS

An established method for ILD diagnosis published in [23] implements the CNN Inception-ResNet V2 (IRV2) [28] on montages created from random CT slices. Despite desirable, no intermediate segmentation maps for algorithm credibility are available. This study is solely based on imaging features and achieves human-like results. Shaish et al. [29] took an alternative approach, initially segmenting the lung image and subsequently generating numerous wedges from CT scans. They employed a CNN to address a UIP classification problem, and their methodology also involves integrating histopathology data into the final analysis. A weakness of this work is that the data originates from a single institution; this will limit the overall generalisability of the models learned. More recently, Mei et al. [30]. combines imaging and clinical data to distinguish individual ILDs. The majority of recent work focuses on the classification of ILD patterns. Anthimopoulos et al. [31] introduced one of the first attempts using a CNN to classify common ILD patterns. Furthermore, Christodoulidis et al. [32] built upon this work by pretraining CNNs on publicly available texture databases and then further fine-tuning the networks to ILD databases; the proposed method resulted in a 2% increase from previous baselines. The evaluation of ILDs at CT level still remains challenging due to the nature of the task and the inter-reader variability. An alternative approach uses content-based retrieval of similar chest CT images utilising a DL approach [33]. While this approach uses segmentation and pattern classification of known ILDs based on a 2D deep learning algorithm, we wish to compare 2D against 3D segmentation methods and furthermore extract radiomic features from ILD patterns and combine those with demographic data to make an overall UIP classification.

C. MULTI-LAYER PERCEPTRON MIXERS

A promising approach to retrieve imaging information is based on multi-layer perceptrons (MLPs) in the form of MLP-Mixers [17], where neither convolutional nor attention operations are necessary to fuse low-level texture and high-level global context features. However, both ViTs and MLP-Mixers have been developed for classification, thus not directly applicable to segmentation-based CAD systems. However, in [35], a segmentation mechanism on top of ViTs was introduced to generate densely annotated segmentation maps, giving promising results on bench-marking datasets. Recently, MLP-Mixers have been successfully applied for medical diagnosis [36], [37], where some of them exhibit more complex modifications [38]. Research in 3D data has shown that their usage improves the performance of segmentation algorithms for anatomical structures and focal lesions such as lungs, brains and tumours [39], [40]. However, their performance in the segmentation of diffused radiological patterns has not been systematically investigated. In any case, the segmentation maps are used to automatically generate imaging features applicable to diagnosis tasks. Among different types of imaging features, radiomic features have been successfully applied to images of lung diseases [41], [42] and have been proven to be vital in assessing the severity of ILDs [43], [44], [45].

III. MATERIALS AND METHODS

A. DATABASES

Within the scope of this study, we utilised two databases - the Lung Tissue Research Consortium Database (LTRC-DB), where written informed consent of each participant was required before any LTRC procedure was performed [46] and the Inselspital Interstitial Lung Diseases Database (INSEL-DB) [5], where depending on the annotation type, cases were either assigned to the segmentation database (INSEL-DB-Seg) or the the diagnosis database (INSEL-DB-Diag). Acquisition for the latter was based on written consent given by patients on hospital admission following ethics approval. The segmentation database encompasses 338 cases in total (278 cases from the LTRC-DB and the 60 cases from the INSEL-DB-Seg). From this database, 60% (173 LTRC-DB and 31 INSEL-DB-Seg cases) of the data was used to train the models, 20% (53 LTRC-DB and 14 INSEL-DB-Seg cases) used for the evaluation and 20% (52 LTRC-DB and 15 INSEL-DB-Seg cases) for testing. In order to ascertain the ground truth for both databases, two chest radiologists - with 21 and 15 years of experience - classified the cases into the 4 UIP CT patterns (typical UIP CT, probable UIP CT, indeterminate

TABLE 1. High level description of the two Databases used withing this work.

Database name	Annotation type	#cases
LTRC-DB	dense ILD pattern diagnosis lung/airway	278
INSEL-DB-Seg	sparse ILD paattern	60
INSEL-DB-Diag	diagnosis	105

for UIP, and CT features mostly consistent with a non-IPF diagnosis), according to the recommendations of the Fleischner Society [6]. The radiologists first reviewed and classified all cases independently and then met to discuss the ambiguous cases to determine the classification through consensus.

CT scans within the INSEL-DB were retrospectively collected with irreversible data anonymisation from October 2015 to June 2017; IRB approval could therefore be waived. Images were acquired on a third-generation dual-source CT (Somatom Definition Flash; Siemens Healthineers, Forchheim, Germany). CT scans were performed during the end-inspiratory phase using the breath-hold technique with patients in the supine position, from the apex of the lung to the costodiaphragmatic recess, with a slice thickness of 1 mm. A tube voltage from 100 to 120 kVp and reference mAs from 100 to 120 were applied. On the 128-detector scanner, collimation of 128 × 0.6 mm was used, with a pitch of 0.6. A slice thickness of 1 mm was reconstructed, with a SAFIRE (Sinogram Affirmed Iterative Reconstruction) level 3 and with a high frequency lung kernel (I70f). The LTRC-DB was originally created in 2005 by the National Institutes of Health and is made up of four clinical centres from around the United States: Mayo Clinic Rochester, University of Michigan–Ann Arbor, University of Pittsburgh, and Temple University. Each centre contributes CT images of various vendors with a slice thickness of 1 mm and hard kernel reconstruction.

B. CAD OUTLINE

As shown in Fig. 2a, the CAD support system for ILDs consists of four modules: 1) segmentation and lung tissue characterisation, 2) quantification, 3) radiomic feature extraction, and 4) diagnosis. Each module is described in the following paragraphs.

C. SEGMENTATION AND LUNG TISSUE CHARACTERISATION

1) LUNG/AIRWAY SEGMENTATION

To simultaneously extract lung and airway segmentation, we implemented a 3D CNN-MLP-Mixer hybrid network, referred to as “3D-Hybrid”, in a multi-task setup. We evaluated the performance of our novel MLP-Mixer against the 3D full-resolution nnUNet [47]. As the main operations in MLP-Mixers are matrix-multiplications, a fixed size input is required. Thus, each CT scan was re-scaled to a cube with a side length of 256 voxels. An MLP-Mixer is composed of two blocks: the token-mixing and the channel-mixing blocks, where the first token-mixing block accepts as input linearly projected image patches - the tokens. The token-mixing block allows communication between spatial locations, while the channel-mixing block provided a possibility for inter-token communication. Mathematically, the two blocks can be described as follows.

$$U_i = X_i^T + W_2 \cdot \sigma(W_1 \cdot \text{LayerNorm}(X)_i^T) \quad \text{for } i = 1 \ldots T \tag{1}$$

$$Y_j = U_j^T + W_4 \cdot \sigma(W_3 \cdot \text{LayerNorm}(U)_i^T) \quad \text{for } i = 1 \ldots C \tag{2}$$

where U_i and Y_j are the token and channel embedding, respectively, W_1 to W_4 are four different weight matrices, σ is an activation function, “$\cdot$” is the dot-product, and “^{T}” is the transposing operator. For the lung/airway segmentation, the token-mixing block was designated as “auto encoder” in order to reduce computational costs. The combination of the two blocks is called a “mixer-layer” and an MLP-Mixer can contain several of these layers. The networks architecture can be seen in Fig. 2b. The initial convolutional stage and the first 5 mixer layers contain task-shared weights. For each task, 3 consecutive mixer layers and convolutional up-sampling is implemented. The cubic segmentation map is finally reshaped to its original CT scan shape by utilising spline interpolation. In order to exclude tissue belonging to the main airways we subtracted the extracted airways from the lung segmentation. The nnUNet framework was used without applying the in-built post-processing tools; instead unified post-processing through morphological operations was applied for both, the 3D-Hybrid and the nnUNet. To train the network, we applied a region-based dice loss for foreground and background and a distribution-based weighted binary cross entropy (BCE) loss with dynamic loss weights; the class weights (foreground/background) were calculated per CT scan and utilised to weight the binary cross entropy case-by-case. Both losses can be calculated as follows.

$$\text{Dice} = \frac{2 \times \text{Area of Overlap}}{\text{Area of Region 1} + \text{Area of Region 2}} \tag{3}$$

$$\text{BCE loss} = -(y * \log(p) + (1 - y) * \log(1 - p)) \tag{4}$$

where y is the ground truth label and p is the network output.

2) ILD PATTERN SEGMENTATION

Seven different ILD patterns (ground glass opacity - GGO, reticulation, consolidation, honeycombing, reticulation + ground glass, bronchiectasis, and emphysema) plus healthy tissue were considered [48]. We utilised three different methods to retrieve different ILD pattern segmentation maps from CT scans. We evaluated the performance of those models against each other and against the 3D full-resolution nnUNet [47]. For all 2D methods, axial CT slices and for the 3D methods, cubic patches with a side length of 32 voxels were used. The first two implemented methods were custom made baseline U-nets for 2D and 3D data, referred to as “2D-UNet” and “3D-UNet”, with exactly the same network architecture (Fig. 2c). The third method is a 2D CNN-MLP-Mixer hybrid model, referred to as “Hybrid”, which is similar to the network used for the lung/airway segmentation (Fig. 2c). To counteract class imbalance, we applied a class weight, inversely proportional to the number of samples and adjusted the cross-entropy (CE) loss. The adjusted CE loss was used within a focal loss [49] as followed.

$$\text{Focal Loss} = -(1 - p)^{\gamma} * w * \log(p) \tag{5}$$

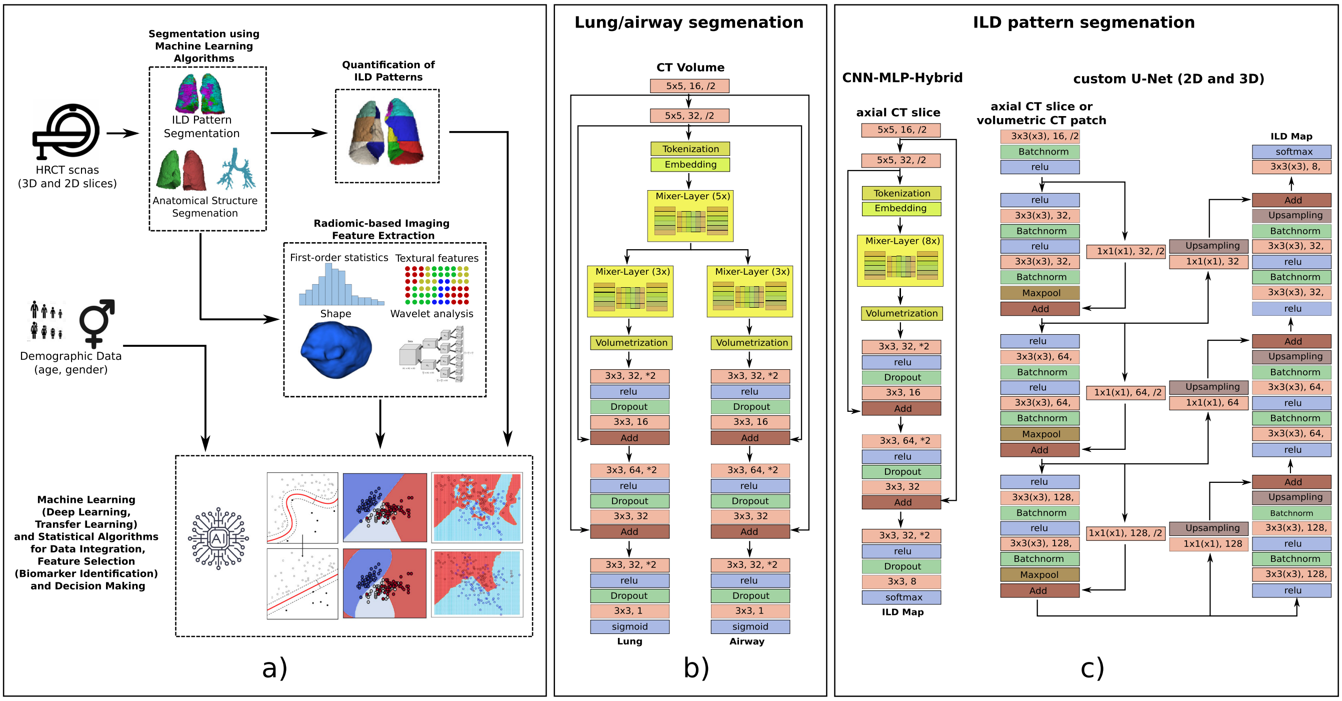

Segmentation using Machine Learning Algorithms
ILD Pattern Segmentation
Anatomical Structure Segmentation
Quantification of ILD Patterns
HRCT scnas (3D and 2D slices)
Radiomic-based Imaging Feature Extraction
First-order statistics
Textural features
Shape
Wavelet analysis
Demographic Data (age, gender)
Machine Learning (Deep Learning, Transfer Learning) and Statistical Algorithms for Data Integration, Feature Selection (Biomarker Identification) and Decision Making
a)
Lung/airway segmentation
CT Volume
5x5, 16, /2
5x5, 32, /2
Tokenization
Embedding
Mixer-Layer (5x)
Mixer-Layer (3x)
Volumetrization
3x3, 32, *2
relu
Dropout
3x3, 16
Add
3x3, 64, *2
3x3, 32
3x3, 1
sigmoid
Lung
Airway
b)
ILD pattern segmentation
CNN-MLP-Hybrid
axial CT slice
Mixer-Layer (8x)
3x3, 8
softmax
ILD Map
custom U-Net (2D and 3D)
axial CT slice or volumetric CT patch
3x3(x3), 16, /2
Batchnorm
3x3(x3), 32,
Maxpool
1x1(x1), 32, /2
Upsampling
1x1(x1), 32
3x3(x3), 64,
1x1(x1), 64, /2
1x1(x1), 64
3x3(x3), 128,
1x1(x1), 128, /2
1x1(x1), 128
3x3(x3), 8,
c)

FIGURE 2. **a) Outline of the CAD system for ILD diagnosis and, b) high-level network architecture for lung/airway segmentation, c) high-level network architecture for the ILD pattern segmentation (Hybrid and U-nets (2D and 3D)). As for the convolutional layers (salmon-colored boxes) in b) and c), the first parameter is the kernel-size for 2D and 3D models, the parameter in the middle is the number of kernels, and the last parameter is the down-sampling stride (/2) or the up-sampling stride (*2).**

TABLE 2. **Different radiomic features extracted for n.**

Feature type	#features
first order statistic	19
3D shape based	16
grey-level co-occurrence matrix (GLCM)	23
grey-level run length matrix (GLRLM)	16
grey-level size zone matrix (GLSZM)	16
neighbouring grey tone difference matrix (NGTDM)	5
grey level dependence matrix (GLDM)	14
total radiomic features	109

where $(1-p)^{\gamma}$ is the focal loss' modulating factor aiming to improve the class-weighting, $-\log(p)$ is the CE loss, and w is the class weight. The sparse ILD pattern annotations in INSEL-DB-Seg provide unlabelled information that is exploited in an unsupervised manner by minimising the entropy of the output probabilities for non-labelled voxels [50].

D. IMAGING FEATURES

1) DISTRIBUTION-BASED ILD

Disease distribution features are calculated on the basis of the occurrence frequency of a specified lung tissue pattern (e.g. 25% of the overall lung tissue is affected by consolidation). Distribution-based features can be divided into those calculated over the whole area of the lung parenchyma and those calculated for designated areas of the lung, as described in [5]. We extracted a total of 63 disease distribution features.

2) RADIOMIC FEATURES

Hand-crafted features were estimated for each pathological tissue type, as well as for healthy tissue. We extracted a total of 872 features, 109 for each pattern. An overview over the extracted features can be found in Table 2. Additionally, we calculated the Minkowski-Bouligand dimension to supplement the pool of radiomic features with eight fractal dimension features: one for each type of lung tissue considered.

3) FEATURE SELECTION

To identify the most important features among the distribution based and radiomic features, two different feature selection approaches (FSA1 and FSA2) were applied and assessed, as based on the predictive performance on the ILD diagnosis task.

4) FSA1: COMBINED UNIVARIATE AND MULTIVARIATE FEATURE ANALYSIS

Firstly, each feature is tested for normality using the Shapiro-Wilk test, and, depending on the result, either, ANOVA, or a Kruskal-Wallis H-Test was conducted to test for significance. All features were eliminated that exhibit a p-value above 0.05. The multivariate analysis was conducted separately for the distribution-based and the radiomic features. The maximum relevance minimum redundancy (MRMR) algorithm [51] was used to select the most important features. In Algorithm 1, we denote a pseudo code for our implementation of FSA1 up to this point. The procedure is repeated for all three different ILD segmentation maps generated by our custom algorithms, resulting in three different sets of selected features. A feature was selected for diagnosis if it was among the most important features for at least two of the three ILD segmentation maps. For the diagnosis, two clinical parameters, age and gender, were appended to the features set.

Algorithm 1 FSA1

```
Input: Feature set $X$
Output: Selected feature set $X_{\text{selected}}$

Step 1: Feature Normality Test
for each feature $f$ in $X$ do
    Perform Shapiro-Wilk test to check for normality.
    if feature $f$ is normally distributed then
        Conduct ANOVA to test for significance.
    else
        Conduct Kruskal-Wallis H-Test to test for significance.
    end if
    if p-value of the test $> 0.05$ then
        Remove feature $f$ from $X$.
    end if
end for

Step 2: Multivariate Analysis and MRMR Feature Selection
Perform multivariate analysis on distribution-based features.
Perform multivariate analysis on radiomic features.
Apply MRMR algorithm to select most important features for each analysis.

Output $X_{\text{selected}}$
```

5) FSA2: FEATURE REDUCTION

The method is based on eliminating redundant information. To this end, the correlation between all features is calculated. The correlation values above a certain threshold T are counted. The feature with the highest count is removed. Next, the correlation values between the remaining features are calculated and the features are eliminated until no correlation value above T is observed. Thresholds of 0.1, 0.3, 0.5, 0.7, 0.9, and 0.99 were used. In the second step, features relevant to classification were selected according to a feature ranking provided by a random forest classifier on the combined feature set (diseases distribution + radiomic features + clinical). In Algorithm 2, we denote a pseudo code for our implementation of FSA2. The resulting feature sets with all different Ts were then used for the classification on the test set. We report the performance measures only for the best performing combination. Thus, FSA2 provides different feature sets for each of the three different ILD segmentation maps and both diagnosis classification tasks (4-way and 2-way diagnosis), resulting in a total of six different feature sets.

Algorithm 2 FSA2

```
Input: Feature set $X$
Output: Selected feature set $X_{\text{selected}}$

Step 1: Feature Correlation Elimination
Calculate correlation matrix $C$ between all features in $X$.
Count the number of correlation values above threshold $T$.

Remove the feature with the highest count.
while correlation value above $T$ is observed do
    Calculate correlation matrix $C$ between remaining features in $X$.
    Remove features with correlation values above $T$.
end while
$X_{\text{selected}} \leftarrow X$

Step 2: Relevant Feature Selection
Rank features in $X_{\text{selected}}$ using a random forest classifier.
Select features relevant to classification on the combined feature set (diseases distribution + radiomic features + clinical).
```

E. DIAGNOSIS

Diagnosis classification was performed on two different tasks. Specifically, the random forest classifiers were implemented to i) classify the lung fibrosis for each case into four UIP patterns according to the guidelines (Fleischner) and ii) classify pooled classes where UIP patterns and the non-UIP patterns were joined together. The number of cases per class with and without the class re-balancing is described in Table 3.

TABLE 3. Class-distribution for testing set with and without data augmentation.

Pattern	#Cases	#Cases with augmentation
Typical UIP CT	14	37
Probable UIP CT	9	29
Indeterminate for UIP	18	35
CT features most consistent with a non-IPF diagnosis	64	64

F. RADIOLOGICAL IMAGE ASSESSMENT

A chest imaging specialist, a senior general consultant, a general consultant, and a chest fellow radiologist with 20-, 9-, 5- and 4-years' experience, respectively, in chest CT imaging scored the images on a clinical Picture Archiving and Communication System (PACS Sectra, Linköping, Sweden).

Lung windows settings were used to read the hard kernel reconstructions (I70f). Both radiologists were blinded to the ground truth and had to classify the patient's images into the four UIP categories (Table 3).

G. EXPERIMENTAL SETUP

To assess the performance of our methods for segmentation of the lung anatomy, we calculated sensitivity, specificity, balanced accuracy, and the soft dice-coefficient (Equation 3) on the testing set. Sensitivity, specificity, and balanced accuracy are defined as follows

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{8}$$

where TP stands for true positive, FN for false negative, TN for true negative, FP for false positive, and balanced accuracy computes the average of sensitivity and specificity, giving equal weight to both metrics, to provide a balanced assessment of the model's overall performance.

To compare the different ILD pattern segmentation pipelines, we calculated sensitivity, specificity, and the balanced accuracy on the testing set. To assess the difference between individual segmentation models based on the dice metric for the lung segmentation and for the balanced accuracy metric for the ILD pattern segmentation, we applied the paired sample t-test. The significance level was set to 0.05. For the diagnosis classification, we utilised random forest classifiers and 105 cases from the validation (53 cases) and testing (52 cases) sets of the LTRC-DB. For each of the 105 diagnosis cases, an individual classifier was trained on the remaining 104 cases and additional cases from the INSEL-DB-Diag, in order to reduce the class imbalances (Table 3). In addition to our segmentation based pipeline, we implemented the algorithm proposed in [23], trained it on our data, and finally compare the results against our method. Moreover, and different to [23], we did not only used random weight initialization of the deep CNN, but used pretrained RadImageNet [52] weights as type of transfer learning for network initialization. Finally, balanced accuracy (Equation 8) a.nd F-score were calculated for the readers and the proposed CAD systems, using the independent chest radiology experts' classification (1–4) as the ground truth. F-score, the harmonic mean of precision (fraction of TP out of all positive predictions and recall (sensitivity), is defined as follows

$$\text{F-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{9}$$

The McNemar test was used to compare accuracy between the readers and the CAD system. The significance level was set to 0.05. Statsmodels (v0.13.2), was used for the McNemar test and sklearn (v0.23.2) was used for Cohen-Kappa tests. The 4 categories were first analysed unpooled for all entries. Then the 4 groups were pooled into two categories: i) cases needing a biopsy for further diagnostic testing, according to the white paper of the Fleischner Society [53] (groups 3 and 4), and ii) cases without further need for diagnostic testing (groups 1 and 2). Interobserver agreement between radiologists, as well as between radiologists and the CAD system, were calculated individually from the weighted Kappa as follows: slight (0–0.2), fair (0.21–0.4), moderate (0.41–0.6), substantial (0.61–0.8), and almost perfect agreement (0.81–1).

TABLE 4. Comparison between 3D-Hybrid and nnUNet for lung segmentation.

Method	Sens	Spec	B-acc (%)	Dice
3D-Hybrid	0.979	0.987	98.3	0.981
nnUNet	0.983	0.985	98.4	0.984

"Sens is short for sensitivity, "Spec" is short for specificity, and "B-acc" is short for balanced accuracy.

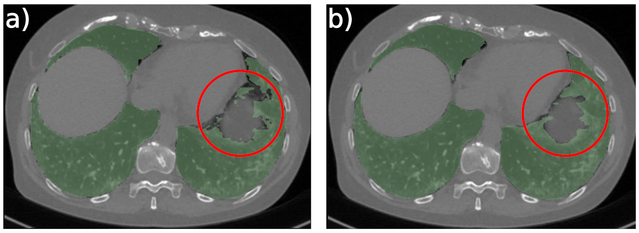


FIGURE 3. Comparison between segmentation maps retrieved by MLP-Mixers with and without the presence of convolutional layers. a) segmentation map, up-scaled with spline interpolation. Artifacts at small recesses are preserved. b) segmentation map up-scaled with transposed convolution. Artifacts could be partially cleaned out.

IV. RESULTS

A. LUNG/AIRWAY SEGMENTATION

With a dice-coefficient of 0.981, our lung segmentation method utilising MLPs, performs comparably well to the nnUNet baseline (dice: 0.984, p-value: 0.62, Table 4). In Fig. 3, we present different outputs for the segmented lung parenchyma - once for a pure MLP-Mixer architecture and once for the architecture as introduced in Fig. 2b. We observe that the pure MLP-Mixer architecture generates pixelated outputs, as they internally down-sample the provided images by the factor of the chosen token-size. Adding a convolutional up-sampling stage at the end greatly improves the quality of the segmentation map.

B. ILD PATTERN SEGMENTATION

Table 5 presents the performance for the four different ILD segmentation methods, as measured in terms of sensitivity,

TABLE 5. Comparison between ILD segmentation methods.

Method	Sens	Spec	B-acc (%)
Hybrid	0.59	0.90	73.94
2D-UNet	0.51	0.84	68.35
3D-UNet	0.56	0.90	72.89
nnUNet	0.53	0.92	73.21

"Sens is short for sensitivity, "Spec" is short for specificity, and "B-acc" is short for balanced accuracy.

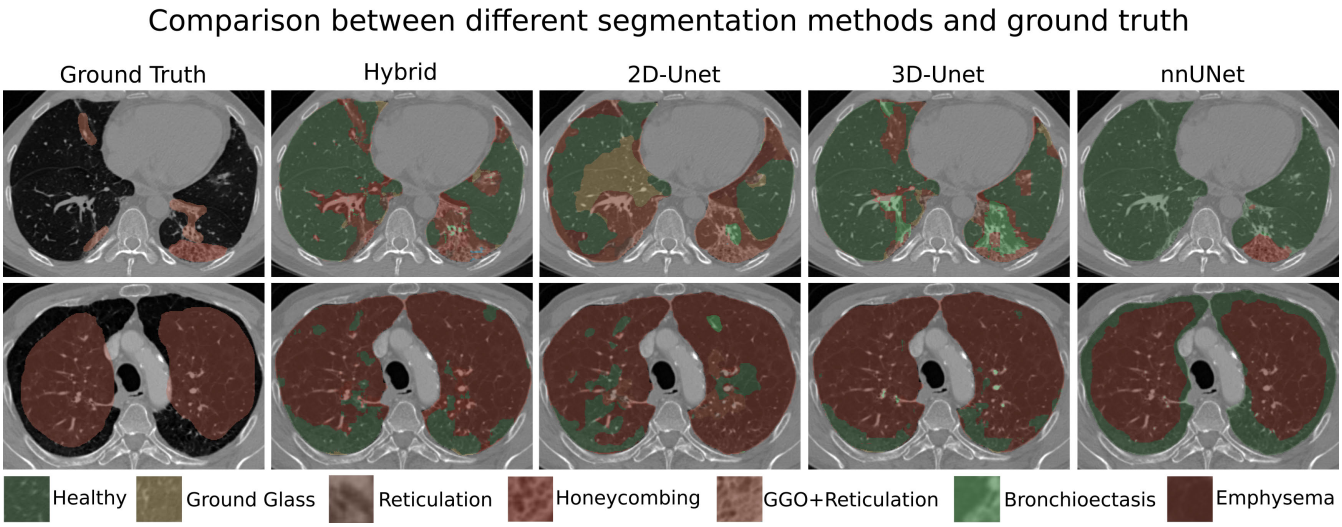
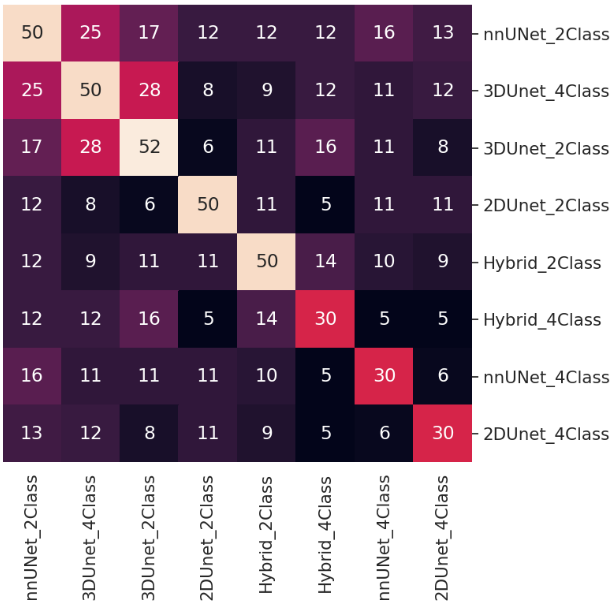


FIGURE 4. **Comparison between ground truth and the four proposed segmentation methods. For all three custom methods (Hybrid, 2D-UNet, and 3D-UNet) we observe over-segmentation, while for the baseline nnUNet, we observe under-segmentation in the presented cases.**



	nnUNet_2Class	3DUnet_4Class	3DUnet_2Class	2DUnet_2Class	Hybrid_2Class	Hybrid_4Class	nnUNet_4Class	2DUnet_4Class
nnUNet_2Class	50	25	17	12	12	12	16	13
3DUnet_4Class	25	50	28	8	9	12	11	12
3DUnet_2Class	17	28	52	6	11	16	11	8
2DUnet_2Class	12	8	6	50	11	5	11	11
Hybrid_2Class	12	9	11	11	50	14	10	9
Hybrid_4Class	12	12	16	5	14	30	5	5
nnUNet_4Class	16	11	11	11	10	5	30	6
2DUnet_4Class	13	12	8	11	9	5	6	30

FIGURE 5. **Dependent on the underlying segmentation method and the goal task (4-way or 2-way classification) the selected features (by FSA2) could differ significantly. The number of shared features among different combinations of underlying segmentation methods and goal tasks can be found on the non-diagnoal entries, whereas the total number of features contained in one set can be found on the diagonal.**

specificity, and balanced accuracy. We display the average over all 8 tissue pasterns (health + 7 ILD patterns) and all cases within the test set. We did not observe any statistically significant difference between nnUNet, the 2D-UNet and the Hybrid (p-values of 0.14, 0.28). However, there was significant difference between nnUNet and the 3D-Unet ($p = 0.03$). Furthermore, the Hybrid performed comparably well to the 2D-UNet ($p = 0.65$), but differently from the 3D-UNet ($p = 0.0007$). 2D-UNet and 3D-UNet with the same network structures perform differently ($p = 0.0002$). In Fig. 4 we present output samples for each of the four segmentation methods and the corresponding ground truth. In the top row, we have a mix of honeycombing and a mixed pattern (GGO + reticulation) and in the middle row a predominant case of emphysema - all pathological patterns present in ILDs.

C. FEATURE SELECTION

FSA1 returns 33 features (19 radiomic, 12 disease distribution, and 2 clinical features), while FSA2 returns between 30 and 52 features for the 4-way and 2-way diagnosis, respectively. More specifically, features originating from 3D-UNet segmentation maps were selected by applying a threshold of 0.5 with subsequent selection of the 50 features and by applying a threshold of 0.3 with subsequent selection of the 52 features for the 4-way and 2-way diagnosis, respectively. Feature sets generated by the 2D-UNet and the Hybrid were selected with thresholds of 0.7 and 0.99 with subsequent selections of the 30 and 50 most informative features for the 4-way and 2-way diagnosis, respectively. In spite of this, having the same threshold and number of features does not mean that the features for those methods are identical. Lastly, feature sets generated by the nnUNet were also selected with thresholds of 0.99 and 0.9 with subsequent selections of the 30 and 50 most informative features for the 4-way and 2-way diagnosis, respectively. As seen in Fig. 5, many of the selected features differ between the segmentation methods and within the segmentation methods for different tasks (4-way and 2-way diagnosis). Lists of the selected features can be seen in the supplementary material.

D. DIAGNOSIS

As seen for all three custom segmentation algorithms, the feature set selected by FSA2 led to better performance with respect to both metrics for the 4-way (FSA1 vs FSA2: Accuracy 49.6±0.71 vs 58.1±2.6 and F-score 0.50±0.01 vs 0.62±0.03) and the 2-way (FSA1 vs FSA2: Accuracy 70.2±0.95 vs 77.2±1.6 and F-score (0.70±0.01) vs 0.77±0.01) for the diagnosis problem (Table 6). To assess whether the system can provide diagnosis at least equal to that of human readers, we calculated the corresponding p-values utilising the McNemar test. We only analysed the best performing feature set (FSA2 feature sets). We observed no significant difference between the 3 custom models and the nnUNet baseline for either the 4-way or the 2-way diagnosis; (p-values ranging from 0.24 to 0.80). For both diagnosis tasks, all four models performed similarly to readers 1, 3, and 4, with p-values ranging from 0.077 to 0.86, with the single exception of the 4-way diagnosis, where Hybrid vs. Reader 4 performed differently (p-values: 0.048) and one exception for the 2-way diagnosis, where 3D-UNet vs. reader 1 performed differently (p-values: 0.045). For the 4-way diagnosis, reader 2 performed significantly differently from all other readers and all four models (p-values ranging from 0.00002 to 0.00046). For the same diagnosis task, reader 1 performed differently from all other readers (p-values: reader 1 vs. reader 2: 0.00002, reader1 vs. reader 3: 0.05, and reader 1 vs. reader 4: 0.033). For the 2-way diagnosis task, reader 1 performed differently from all other readers with p-values ranging from 0.00067 to 0.012. For the same diagnosis task, there was no significant difference between readers 2, 3, and 4. Furthermore, Hybrid, nnUNet and 2D-UNet performed differently from reader 2 (p-values: 0.013, 0.034, 0.0093), whereas 3D-UNet performed comparably to reader 2 (p-value: 0.097). Readers 3, 4, and all four models performed comparably with p-values ranging from 0.072 to 0.80. When comparing the performance of our proposed segmentation based models to the IRV2 methods [23], we observe a different behaviour in performance between the IRV2 with random initialized weights and our 4 models (max p-value of 0.034 for the 4-way classification and 0.001 for the 2-way classification). In comparison, the IRV2 initialized with RadImageNet weights performs more in line with our models (p-values between 0.02 and 0.18 as well as 0.05 and 0.19 for the 4-way and 2-way classification, respectively. To assess interobserver agreements, the Cohen-Kappa test was applied. For the 4-way diagnosis, reader 2 exhibits the weakest agreement with all other readers and all four models, with values between 0.1 and 0.32 - corresponding to slight to fair interobserver agreement. For the same diagnosis task, the interobserver agreement between readers 1, 3, 4, and all four models, is fair to moderate, with values ranging from 0.29 to 0.59. For the 2-way diagnosis, only the Hybrid vs. reader 3 exhibit a Cohen-Kappa of 0.19 (slight agreement). Most other reader-reader, reader-model and model-model combinations exhibit fair to moderate agreement. Reader 1 vs. reader 3, reader 3 vs. reader 4, reader 1 vs 3D-UNet, and nnUNet vs 2D-UNet showed substantial agreement with values of 0.61, 0.63, 0.63, and 0.64, respectively. Furthermore, the Kappa Fleiss-test was performed on the groups (4-way, 2-way diagnosis) - all readers (0.37, 0.50), all readers and the Hybrid (0.35, 0.42), all readers and nnUNet (0.36, 0.48), all readers and 2D-UNet (0.34, 0.46), all readers and 3D-UNet (0.34, 0.49), and all readers in combination with all four algorithms (0.36, 0.45).

E. IMPLEMENTATION

For both lung/airway and ILD segmentation we used early stopping to avoid overfitting. For parameter optimisation, we utilised Adam that was initialised with a beta_1 of 0.9, a beta_2 of 0.999, and a learning-rate of 0.001. No learning rate decay was applied. The best hyper parameters for the loss weighting within the lung/airway segmentation were 0.7 and 0.3 for dice loss and cross-entropy loss, respectively, where 90% of the dice loss is associated to the foreground and 10% to the background. Radiomic features were retrieved utilising the PyRadiomcs library [54], a standardised python software-package [55]. With an empirical search, we found that selecting the 15 most important features among the distribution-based features and the 35 most important features among the radiomic features provides the best features set for FSA1. For FSA2, different feature sets were chosen for the different segmentation methods and classification tasks. The random forests for the 4-way and 2-way diagnosis were configured to use 100 decision trees and 15 features per leaf with balanced class-weights. Calculations were partially performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

V. DISCUSSION

The motivation behind this study was to design and evaluate algorithmic approaches for the detection and diagnosis of ILDs. We investigated the applicability of novel image processing methods such as MLP-Mixers. We additionally investigated how the performance in terms of diagnostic accuracy is influenced i) when 2D data and 3D data are used to retrieve ILD pattern segmentation maps, and ii) feature-reduction methods are applied for selecting the most informative radiomic and diseases distribution features.

While our proposed segmentation based CAD system is more complex than pure CNN based approaches and does not out-perform established systems like [23], our solution provides radiologists with auxiliary segmentation maps, empowering the human expert to verify the algorithm's behaviour. Additionally, we utilised random forests for the classification of our transparent segmentation-map-based features. Random forests, unlike deep neural networks, belong to the family of explainable algorithms; hence our approach aims to increase trust in the algorithms decision making and encourage practicing radiologists to use an AI-powered tool in their clinical practice. Moreover, the original study in IRV2 reports no significant differences in results when initialising their models with pre-trained ImageNet [56] weights. In contrast, we observe the IRV2 model with random initialisation performs differently when initialised with domain-specif RadImageNet weights, while

TABLE 6. Diagnostic performance for CAD systems and human experts.

	Reader	4-way diagnosis		2-way diagnosis	
		Balanced accuracy (%)	Average F-Score	Balanced accuracy (%)	Average F-Score
Custom System*	Hybrid	**61.7** (50.3)	**0.66** (0.51)	**77.4** (69.0)	**0.79** (0.70)
	3D-UNet	**56.7** (48.6)	**0.61** (0.49)	**79.0** (70.4)	**0.76** (0.69)
	2D-UNet	**55.8** (49.8)	**0.60** (0.51)	**75.2** (71.3)	**0.77** (0.72)
	Average	**58.1±2.6** (49.6±0.71)	**0.62±0.03** (0.50±0.01)	**77.2±1.6** (70.2±0.95)	**0.77±0.01** (0.70±0.01)
Baseline System**	nnUNet	59.2	0.61	77.1	0.77
Method by [24] (random weights)	IRV2	55.7	0.43	74.9	0.63
Method by [24] (RadImageNet weights)	IRV2	59.0	0.53	76.3	0.70
Human	Reader 1	76.5	0.71	90.2	0.86
	Reader 2	39.1	0.36	73.4	0.68
	Reader 3	55.8	0.54	73.1	0.72
	Reader 4	59.4	0.53	79.1	0.73
	Average	57.7±13.3	0.54±0.12	79.0±6.9	0.75±0.07

*Performance measures are in the format of feature set from **FSA2** (FSA1). ** for the baseline, only features selected by FSA2 were used.

the overall scores are similar (Table 6). This may indicate the introduction of some bias towards other imaging modalities like Magnetic Resonance Imaging (MRI) or CT patterns with high frequency within the RadImageNet dataset, while random initialization allows the network to focus on the current data distribution. The optimal case might be an ensemble of pre-trained and random initialised models.

We found that MLP-Mixer in combination with convolutional pre- and post-processing can be applied to medical image segmentation within the framework of detecting diffused pathological patterns in the lungs. These techniques not only show strong results in this area but also have the potential to be applied to other medical and non-medical image-processing domains [57], [58]. In general, incorporating CNN components into MLP-Mixers necessitates a shallower neural network overall, thanks to the distinctive architecture of the MLP-Mixer. This design efficiently merges low-level texture and high-level global context features. However, according to Fig. 4, nnUNet seems to be more robust than all three custom models. This might be caused by the different data preparation used for the custom models (voxel value clipping to -1000 and 200 Hounsfield Units and normalisation to a range between 0 and 1) and nnUNet (voxel value clipping using the 0.5 and 99.5 percentiles of the foreground voxels and image normalisation utilising the global foreground mean and standard deviation). Furthermore, we have to evaluate whether additional auxiliary segmentation losses at down-sampled stages - as used by nnUNet - would also lead to more robust outputs for our three custom models, as with such intermediate auxiliary losses, the network learns how to gradually fine-tune a proposed segmentation map. Segmentation maps appear to be sensitive to the overall CAD pipeline (data pre- and postprocessing as well as DL model) as segmentation maps produced by our proposed CAD system appear more similar under each other (despite quite different network architectures) than compared to those generated by the nnUNet pipeline.

For diffused patterns, such as abnormal tissue alterations in ILDs, we found that 2D slice-based segmentation did not perform worse than 3D patch-based segmentation. Even though, the segmentation was based on 2D data, diagnosis relevant distribution-based and radiomic features were calculated on the reconstructed 3D segmentation maps. This is in line with other proposed systems for medical diagnosis [59], where 2D slices were used to retrieve 3D predictive imaging features for computer aided diagnostic support. One reason that the 3D segmentation method does not outperform the 2D segmentation may be that the third dimensions does not provide additional information for diffused radiological patterns, since those patterns appear to be similar in all directions. By showing that 2D methods are readily applicable for ILD segmentation we can benefit from the plethora of well-established and pretrained models that can be applied for transfer learning and to augment our segmentation pipelines not only with volumetric CT data, but also with 2D chest X-rays.

To achieve precise lung and airway segmentation, we adopted a 3D data processing network approach, enabling the integration of shape features across spatial dimensions. We also employed a multi-task setup with shared first-layer weights to extract shared CT-specific patterns. In task-specific layers, which combine information through subtraction, addition, and more complex functions, the final segmentation output is generated. The challenge lay in deciding when to transition from shared-weights to task-specific weights. While lung and airway segmentation share some similarities, they demand distinct finesse. Lung parenchyma extraction involves blob-shaped structures, while airways are tubular and gradually decrease in diameter as they merge into lung tissues. To accommodate these nuances, we introduced residual connections to the initial convolutional down-sampling stages. This empowers the task-specific layers to derive characteristic combinations of information needed to fulfil their respective objectives.

The radiomic- and disease-distribution features selected for each segmentation method and classification tasks (4-way or 2-way) individually (FSA2) outperformed the approach of standardised feature selection (FSA1) that utilises selection techniques based on univariate and multivariate features. However, with p-values for the 4-way diagnosis of 0.02, 0.06, and 0.25 and with p-values for the 2-way diagnosis of 0.16, 0.73, and 0.80 for the Hybrid, the 2D-UNet and the

3D-UNet, respectively, most of the differences are not significant (0.05 significance level). Nonetheless, we conclude that standardisation to select features across methods and tasks (4-way and 2-way diagnosis) is sub optimal and features should be individually selected for each segmentation method and task, even though we are at risk of selecting features that do not represent a characteristic biomarker [60]. From Fig. 5, we observe that our statistically similar models can identify different features, as is in line with recent findings [60]. With those experiments, we want to highlight the sensitivity of feature selection. Even within the same dataset, different features are perceived to be important for related but not identical machine learning methods and goal tasks (4-way and 2-way classification of ILDs), hence accenting the difficulty of CAD systems to adopt to unseen data.

The accuracy of the radiologists depended on their experience. The most experienced radiologist with 20 years of chest CT practice demonstrated significantly better performance in UIP pattern classification than the other readers and most algorithms. With ranges from fair over moderate up to substantial, interobserver agreement was different for many radiologist-, radiologist- and machine-, and machine-pairs. Pairs below the substantial level might be used as complement to a human readers in a second reader setup or to question the human expert (similar as for DL assisted nodule detection [61]), whereas those with substantial agreement, classified more or less the same cases correctly or falsely.

Processing CT scans requires reduction in data volume if computational resources are not to be exceeded. While for the IRV2 model, down-sampled montages of 4 randomly selected axial CT-slices were chosen to select a whole CT volume [23], we decided to heavily down-sample all CT scans for the lung/airway segmentation, since the overall shapes of the lung/airways are of major importance. Prior to the down-sampling, we accounted for the different voxel spacing across the three axis through interpolation. The scans have subsequently been down-sampled to a cube with a fixed size of 256 voxels per side. For the ILD pattern segmentation we either split each CT scans into axial slices (2D) or into cubic patches (3D). As MLP-Mixers require a fixed size input, we re-scaled the axial slices by sizing the longer side of the CT-slice to 352 pixels and padded the smaller side to 336 pixels. With padding, distortions in the CT could be avoided. Due to the moving-window framework within CNNs (convolutional down-sampling stage within the Hybrid), the padded region does not affect the network performance. The 3D-UNet uses 3D patches that are sampled from a CT scan within the region of the detected lung parenchyma. To avoid artefacts at the boarders of the patches, we decided to sample the patches with 25% overlap and average out multiple class votes at the border regions.

VI. LIMITATIONS

Although we have shown that MLP-Mixers may be used in the segmentation of medical images, we have to mention some limitations. Firstly, MLP-Mixers require the input to have the same size for all samples. Secondly, images need to be tokenised and the tokens need to be flattened into an array of channels. The size of the tokens determines the resolution of the outputted segmentation map. Therefore, if the token is too large, the output will be pixelated and in the best case, demands post processing. The token size however cannot be arbitrarily decreased, as the number of neurons in the token-mixing block correlates directly with the number of tokens within an image and an excessive number of neurons increases the complexity of the training procedure. We can mitigate the effect of the token size on output resolution by choosing an adequate token-size. For the lung/airway segmentation, a token size of $4 \times 4 \times 4$ was chosen for feature map resolution of $64 \times 64 \times 64$; for ILD segmentation a token size of 4×4 was chosen for a feature map resolution of 88×84. The outputted down-sampled segmentation map is subsequently up-sampled by learnable convolutions. As seen in Fig. 3, MLP-Mixers utilised as segmentation models can greatly benefit from combination with learnable up-sample functions. A third drawback of MLP-mixers is the inability to process unevenly padded regions across different cases. In our experiments with ILD-mixers on the ILD segmentation task, we observed poor performance with a weighted accuracy of 55.59%. By adding a convolutional stage prior to the MLP-Mixer layers, we empowered the system to carefully select the information that should be applied to the powerful mixer layers. Combination of convolutional down-sampling and up-sampling with mixer layers and intermediate dense processors boosted the performance by 18.35% to 73.94%.

A limiting factor in our reported performance (segmentation and diagnosis) could be the mix of two different databases what is challenging for any deep learning based system. Recently published studies have mentioned a drastic decrease in performance if the system was evaluated on a population other than that used for training [62] and that different cohorts result in different feature sets [63].

VII. CONCLUSION

We found that MLP-Mixers in combination with convolutions are applicable for medical image segmentation. The Hybrid performs similarly to all other networks with respect to balanced accuracy on the segmentation task and on the diagnosis. In future experiments, we will add additional auxiliary segmentation losses at different resolutions to achieve gradual fine-tuning of the final segmentation map. We hypothesise that investing further research into MLP-Mixer application for medical image segmentation could boost their performance beyond the level of state-of-the-art UNets. Furthermore, utilising 3D data neither improved the segmentation of ILD patterns nor the diagnosis. Lastly, we found that features selected for a specific classification task (4-way or 2-way diagnosis) outperform standardised methods of agnostic feature-selection. Overall, CAD systems in the framework of ILDs perform similarly to human experts, with mostly moderate interobserver agreement.

ACKNOWLEDGMENT

(M. Fontanellaz, A. Christe, L. Ebner, and S. Mougiakakou contributed equally to this work.)

REFERENCES

[1] G. Raghu, D. Weycker, J. Edelsberg, W. Z. Bradford, and G. Oster, "Incidence and prevalence of idiopathic pulmonary fibrosis," *Amer. J. Respiratory Crit. Care Med.*, vol. 174, no. 7, pp. 810–816, 2006.

[2] A. U. Wells, "Managing diagnostic procedures in idiopathic pulmonary fibrosis," *Eur. Respiratory Rev.*, vol. 22, no. 128, pp. 158–162, Jun. 2013.

[3] W. D. Travis et al., "An official American thoracic society/European respiratory society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias," *Amer. J. Respiratory Crit. Care Med.*, vol. 188, no. 6, pp. 733–748, Sep. 2013, doi: 10.1164/rccm.201308-1483ST.

[4] G. Raghu et al., "Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: An official ATS/ERS/JRS/ALAT clinical practice guideline," *Amer. J. Respiratory Crit. Care Med.*, vol. 205, no. 9, pp. e18–e47, May 2022.

[5] A. Christe, A. A. Peters, D. Drakopoulos, J. T. Heverhagen, T. Geiser, T. Stathopoulou, S. Christodoulidis, M. Anthimopoulos, S. G. Mougiakakou, and L. Ebner, "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images," *Investigative Radiol.*, vol. 54, no. 10, pp. 627–632, 2019.

[6] D. A. Lynch, N. Sverzellati, W. D. Travis, K. K. Brown, T. V. Colby, J. R. Galvin, J. G. Goldin, D. M. Hansell, Y. Inoue, T. Johkoh, A. G. Nicholson, S. L. Knight, S. Raoof, L. Richeldi, C. J. Ryerson, J. H. Ryu, and A. U. Wells, "Diagnostic criteria for idiopathic pulmonary fibrosis: A Fleischner society white paper," *Lancet Respiratory Med.*, vol. 6, no. 2, pp. 138–153, Feb. 2018.

[7] R. L. Riha, E. E. Duhig, B. E. Clarke, R. H. Steele, R. E. Slaughter, and P. V. Zimmerman, "Survival of patients with biopsy-proven usual interstitial pneumonia and nonspecific interstitial pneumonia," *Eur. Respiratory J.*, vol. 19, no. 6, pp. 1114–1118, Jun. 2002.

[8] S. Soffer, A. S. Morgenthau, O. Shimon, Y. Barash, E. Konen, B. S. Glicksberg, and E. Klang, "Artificial intelligence for interstitial lung disease analysis on chest computed tomography: A systematic review," *Academic Radiol.*, vol. 29, pp. S226–S235, Feb. 2022.

[9] H. Barnes, S. M. Humphries, and P. M. George, "Machine learning in radiology: The new frontier in interstitial lung diseases," *Lancet Digit Health*, vol. 5, no. 1, pp. e41–e50, Jan. 2023, doi: 10.1016/S2589-7500(22)00230-8.

[10] E. Dack, A. Christe, M. Fontanellaz, L. Brigato, J. T. Heverhagen, A. A. Peters, A. T. Huber, H. Hoppe, S. Mougiakakou, and L. Ebner, "Artificial intelligence and interstitial lung disease: Diagnosis and prognosis," *Investigative Radiol.*, vol. 58, no. 8, pp. 602–609, Aug. 2023, doi: 10.1097/RLI.0000000000000974.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[12] F. Shamshad, S. Khan, and S. W. Zamir, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802, doi: 10.1016/j.media.2023.102802.

[13] E. U. Henry, O. Emebob, and C. A. Omonhinmin, "Vision transformers in medical imaging: A review," 2022, *arXiv:2211.10043*.

[14] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2023, doi: 10.1016/j.imed.2022.07.002.

[15] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," 2022, *arXiv:2203.15269*.

[16] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," 2015, *arXiv:1512.07108*.

[17] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," 2021, *arXiv:2105.01601*.

[18] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112821.

[19] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Med. Phys.*, vol. 47, no. 5, pp. e218–e227, 2020.

[20] J. Yanase and E. Triantaphyllou, "The seven key challenges for the future of computer-aided diagnosis in medicine," *Int. J. Med. Informat.*, vol. 129, pp. 413–422, Sep. 2019, doi: 10.1016/j.ijmedinf.2019.06.017.

[21] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, Jan. 2020.

[22] Y. A. Ng, B. Glocker, C. Oberije, G. Fox, J. Nash, E. Karpati, S. Kerruish, and P. D. Kecskemethy, "A novel workflow for the safe and effective integration of AI as supporting reader in double reading breast cancer screening: A large-scale retrospective evaluation," *MedRxiv*, Jan. 2022.

[23] S. L. F. Walsh, L. Calandriello, M. Silva, and N. Sverzellati, "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study," *Lancet Respiratory Med.*, vol. 6, no. 11, pp. 837–845, Nov. 2018.

[24] A. A. Trusculescu, D. Manolescu, E. Tudorache, and C. Oancea, "Deep learning in interstitial lung disease—How long until daily practice," *Eur. Radiol.*, vol. 30, no. 11, pp. 6285–6292, Nov. 2020, doi: 10.1007/s00330-020-06986-4.

[25] S. L. F. Walsh, L. Calandriello, N. Sverzellati, A. U. Wells, and D. M. Hansell, "Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT," *Thorax*, vol. 71, no. 1, pp. 45–51, Jan. 2016.

[26] S. L. F. Walsh and M. Kolb, "Radiological diagnosis of interstitial lung disease: Is it all about pattern recognition?" *Eur. Respiratory J.*, vol. 52, no. 2, Aug. 2018, Art. no. 1801321.

[27] J. Widell and M. Lidén, "Interobserver variability in high-resolution CT of the lungs," *Eur. J. Radiol. Open*, vol. 7, Jan. 2020, Art. no. 100228.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, Feb. 2017, vol. 31, no. 1.

[29] H. Shaish, F. S. Ahmed, D. Lederer, B. D'Souza, P. Armenta, M. Salvatore, A. Saqi, S. Huang, S. Jambawalikar, and S. Mutasa, "Deep learning of computed tomography virtual wedge resection for prediction of histologic usual interstitial pneumonitis," *Ann. Amer. Thoracic Soc.*, vol. 18, no. 1, pp. 51–59, Jan. 2021, doi: 10.1513/annalsats.202001-068oc.

[30] X. Mei et al., "Interstitial lung disease diagnosis and prognosis using an AI system integrating longitudinal data," *Nature Commun.*, vol. 14, no. 1, p. 2272, Apr. 2023, doi: 10.1038/s41467-023-37720-5.

[31] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[32] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 76–84, Jan. 2017.

[33] J. Choe, H. J. Hwang, J. B. Seo, S. M. Lee, J. Yun, M.-J. Kim, J. Jeong, Y. Lee, K. Jin, R. Park, J. Kim, H. Jeon, N. Kim, J. Yi, D. Yu, and B. Kim, "Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT," *Radiology*, vol. 302, no. 1, pp. 187–197, Jan. 2022, doi: 10.1148/radiol.2021204164.

[34] S. L. F. Walsh et al., "Deep learning–based outcome prediction in progressive fibrotic lung disease using high-resolution computed tomography," *Amer. J. Respiratory Crit. Care Med.*, vol. 206, no. 7, pp. 883–891, Oct. 2022, doi: 10.1164/rccm.202112-2684oc.

[35] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," 2021, *arXiv:2105.05633*.

[36] S. Pan, C. Chang, T. Wang, J. Wynne, M. Hu, Y. Lei, T. Liu, P. Patel, J. Roper, and X. Yang, "Abdomen CT multi-organ segmentation using token-based MLP-mixer," *Med. Phys.*, vol. 50, no. 5, pp. 3027–3038, May 2023.

[37] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Proc. MICCAI*, 2022, pp. 23–33.

[38] J. Yan, X. Wang, J. Cai, Q. Qin, H. Yang, Q. Wang, Y. Cheng, T. Gan, H. Jiang, J. Deng, and B. Chen, "Medical image segmentation model based on triple gate MultiLayer perceptron," *Sci. Rep.*, vol. 12, no. 1, p. 6103, Apr. 2022.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," 2016, *arXiv:1606.04797*.

[41] A.-N. Frix, F. Cousin, T. Refaee, F. Bottari, A. Vaidyanathan, C. Desir, W. Vos, S. Walsh, M. Occhipinti, P. Lovinfosse, R. Leijenaar, R. Hustinx, P. Meunier, R. Louis, P. Lambin, and J. Guiot, "Radiomics in lung diseases imaging: State-of-the-art for clinicians," *J. Personalized Med.*, vol. 11, no. 7, p. 602, Jun. 2021.

[42] S. Ather, T. Kadir, and F. Gleeson, "Artificial intelligence and radiomics in pulmonary nodule management: Current status and future applications," *Clin. Radiol.*, vol. 75, no. 1, pp. 13–19, Jan. 2020.

[43] L. Yu, G. Tao, L. Zhu, G. Wang, Z. Li, J. Ye, and Q. Chen, "Prediction of pathologic stage in non-small cell lung cancer using machine learning algorithm based on CT image feature analysis," *BMC Cancer*, vol. 19, no. 1, p. 464, Dec. 2019.

[44] J. Schniering, H. Gabrys, M. Brunner, O. Distler, M. Guckenberger, M. Bogowicz, and D. Vuong, "Computed-tomography-based radiomics features for staging of interstitial lung disease—Transferability from experimental to human lung fibrosis—A proof-of-concept study," in *Imaging*, vol. 54, 2019, Art. no. PA4806.

[45] A. Stefano, M. Gioè, G. Russo, S. Palmucci, S. E. Torrisi, S. Bignardi, A. Basile, A. Comelli, V. Benfante, G. Sambataro, D. Falsaperla, A. G. Torcitto, M. Attanasio, A. Yezzi, and C. Vancheri, "Performance of radiomics features in the quantification of idiopathic pulmonary fibrosis from HRCT," *Diagnostics*, vol. 10, no. 5, p. 306, May 2020.

[46] U.S. Department of Health and Human Services. *BIOLINCC: Lung Tissue Research Consortium (LTRC) (no Date) National Heart Lung and Blood Institute*. Accessed: Jan. 18, 2023. [Online]. Available: https://biolincc.nhlbi.nih.gov/studies/ltrc/

[47] 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*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[48] D. M. Hansell, A. A. Bankier, H. Macmahon, T. C. McLoud, N. L. Müller, and J. Remy, "Fleischner society: Glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, Mar. 2008.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[50] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[51] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[52] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," *Radiol., Artif. Intell.*, vol. 4, no. 5, Sep. 2022.

[53] D. S. Kim, "Classification and natural history of the idiopathic interstitial pneumonias," *Proc. Amer. Thoracic Soc.*, vol. 3, no. 4, pp. 285–292, Jun. 2006.

[54] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017.

[55] A. Zwanenburg et al., "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[57] H.-P. Lai, T.-T. Tran, and V.-T. Pham, "Axial attention MLP-mixer: A new architecture for image segmentation," in *Proc. IEEE 9th Int. Conf. Commun. Electron. (ICCE)*, Nha Trang, Vietnam, Jul. 2022, pp. 381–386, doi: 10.1109/ICCE55644.2022.9852066.

[58] H.-M.-Q. Le, T.-K. Le, V.-T. Pham, and T.-T. Tran, "AMG-Mixer: A multi-axis attention MLP-mixer architecture for biomedical image segmentation," in *Proc. Conf. Inf. Technol. Appl.*, in Lecture notes in networks and systems, Jan. 2023, pp. 169–180.

[59] Z. Chen, J. Liu, M. Zhu, P. Y. M. Woo, and Y. Yuan, "Instance importance-aware graph convolutional network for 3D medical diagnosis," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102421.

[60] A. Demircioğlu, "Evaluation of the dependence of radiomic features on the machine learning model," *Insights Imag.*, vol. 13, no. 1, p. 28, Dec. 2022.

[61] A. A. Peters, A. Decasper, J. Munz, J. Klaus, L. I. Loebelenz, M. K. M. Hoffner, C. Hourscht, J. T. Heverhagen, A. Christe, and L. Ebner, "Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists," *J. Thoracic Disease*, vol. 13, no. 5, pp. 2728–2737, May 2021.

[62] C. Fang, S. Bai, Q. Chen, Y. Zhou, L. Xia, L. Qin, S. Gong, X. Xie, C. Zhou, D. Tu, C. Zhang, X. Liu, W. Chen, X. Bai, and P. H. S. Torr, "Deep learning for predicting COVID-19 malignant progression," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102096.

[63] K. Gong et al., "A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records," *Eur. J. Radiol.*, vol. 139, Jun. 2021, Art. no. 109583.

M. FONTANELLAZ received the B.S. degree in mechanical engineering from the Bern University of Applied Science, Bern, Switzerland, in 2017, and the M.S. degree in biomechanical engineering from the University of Bern, Bern, in 2019, where he is currently pursuing the Ph.D. degree in AI in medicine with the AIHN, ARTORG Center for Biomedical Engineering Research.

From 2014 to 2017, he was an Assistant Professor with the Department of Engineering and Information Technology, Bern University of Applied Science. His research interests include the development of AI methods to solve open issues in medicine, thereunder AI assisted diagnosis and disease progression prediction for interstitial lung diseases, and methods to facilitate diabetes self-management, such as glucose prediction and early alarm generation for hypo- and hyperglycemia.

A. CHRISTE received the Medical degree in pathology and radiology from the University of Bern, Switzerland. He completed his Postdoctoral Fellowship in CT technique with Stanford University, CA, USA, from 2008 to 2010. Thereafter, he was the Chief of Body CT with the Department of Radiology, Inselspital, University of Bern. He received grants to investigate CT-parameters for lung nodule detection concerning lung cancer screening. In 2012, he received the Venia Docendi in radiology (Habilitation) and focused on computer aided diagnosis in radiology and he was awarded with Swiss National Science Foundation grants to apply artificial intelligence and machine learning to CT-images. Besides thoracic radiology his expertise includes abdominal imaging, where he studies liver diseases on magnetic resonance elastography. In 2015, he was elected as the Head of the Department of Radiology (City and County Hospitals of the INSELGROUP, Bern), and he was appointed as an Associate Professor in radiology with the University of Bern.

S. CHRISTODOULIDIS (Member, IEEE) received the degree in electrical and computer engineering and the Ph.D. degree in biomedical engineering from the University of Bern. Currently, he is an Assistant Professor with the Department of Mathematics, CentraleSupélec, where he is involved in research and teaching activities related with artificial intelligence. His research interest includes applied machine learning for healthcare. He has been awarded with a fellowship by the Swiss National Science Foundation with which he worked with Gustave Roussy as a Postdoctoral Researcher focusing on multi-modal breast cancer data while he is also a part of the PRISM Center for precision medicine in oncology.

E. DACK received the B.Sc. degree (Hons.) in mathematics from the University of York, the P.G.C.E. degree in secondary mathematics from the Institute of Education, UCL, and the master's degree in data science from the Birkbeck College, University of London. After a career change he smoothly transitioned to working in the IT industry, he has experience in software application support and engineering in the form of building web applications, designing/developing APIs, and database schemas. He was a Software Engineer. His thesis involved developing a deep learning framework to accurately predict mortgage product selection in the U.K. finance industry, a first attempt of any such research in the U.K. His current work involves applying computer vision and machine learning techniques to datasets of medical images and their respective clinical data.

J. ROOS received the M.D. degree from the University of Berne, Switzerland, in 1997. He completed his radiology residency from the University Hospital of Zürich, Switzerland, until 2003. With his board exam (FMH) in radiology, he was an Attending Radiologist with the University Hospital of Zürich, Switzerland, from 2003 to 2005, before he was appointed as a Research Associate and later as an Assistant Professor with the Stanford University, CA, USA, from 2005 to 2010. In 2012, he was promoted to an Associate Professor with Duke University, NC, USA, before he became the Chairman of the Department of Radiology and Nuclear Medicine, University Hospital Lucerne, Switzerland, in 2015. Since 2022, he has been a Full Professor with the University of Zurich, Switzerland. His research interests include computed tomography (CT) imaging and the post processing of cardiovascular and pulmonary diseases and the functional MR imaging of pulmonary diseases.

D. DRAKOPOULOS received the B.S. degree in medicine from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the Ph.D. degree in medicine from the University of Athens, in 2019. From 1997 to 2016, he was a Military Officer in the Greek army and after that since the present day he has been a Consultant Radiologist with the Radiology Department, INSEL Hospital, University of Bern. His research interests include artificial intelligence and health informatics/digital health.

D. SIERON received the degree in medical from the Medical University of Silesia, Katowice, the degree in economic and in public relations from the University of Economics (Karol Adamiecki in Katowice), the Managerial degree in health care units management from Jagiellonian University, Kraków, and the Medicine degree from Christian-Albrechts-Medical University, Kiel, Germany. He is currently a Specialist in EDIR radiology and imagine diagnostics. He is the author of many domestic and foreign publications in the field of medicine and PR, and one of the first monographs dealing with the use of PR in health care. He is the Creator of a handbook on diagnostic imaging addressed to physiotherapists and physiotherapists and a series of handbooks for doctors "Lepetytorium" and "LEK Last Minute." He worked and practiced at many medical universities, e.g., Carl Gustav Carus University Hospital Dresden, University of Aberdeen, Medical University of Graz, and Medical Ruhr-University Bochum. He is also with the Insel Group Hospital, Medical University of Bern, Switzerland, dealing with diagnostic imaging, interventional radiology and broadly understood didactics on a daily basis. He is an Organizer of trainings, both Polish and foreign, in the field of imaging diagnostics for doctors and physiotherapists, e.g., on imaging diagnostics of the musculoskeletal system.

A. PETERS received the Medicine degree from the University of Marburg, Germany, and the M.D. degree, in 2014. He defended his doctoral thesis in the following year. From 2015 to 2020, he completed his residency with the Department for Diagnostic, Interventional and Pediatric Radiology (DIPR), Inselspital, University Hospital Bern, University of Bern, Switzerland. Since 2020, he has been an attending Radiologist with DIPR with special focus on cardiothoracic imaging. His research interests include the clinical validation and comparison of AI algorithms in the field of CT lung cancer screening for detection, volumetry and the risk stratification of pulmonary nodules, and the evaluation of new imaging techniques in cardiac MRI.

T. GEISER is currently the Head of the Department of Pulmonary Medicine and Allergology and the Director for Research and Teaching with the University Hospital Bern, Switzerland. He trained in internal, pulmonary, and experimental medicine with the University of Bern and University of Zurich, and he completed his training with the University of California at San Francisco, San Francisco, CA, USA. His main clinical interest is interstitial lung diseases (ILD), and he is involved in clinical, translational, and experimental research projects in ILD. His research work, which is supported by the Swiss National Science Foundation, focuses on the cellular and molecular mechanisms of pulmonary fibrosis and the development of novel personalized in vitro disease modeling systems, like lung-on-chip and organoids to study precision medicine approaches in lung fibrosis. He is a member of the Swiss Academy of Medical Sciences (SAMW). He was the President of the Swiss Respiratory Society (SGP), the Head of the Cellular and Molecular Biology Assembly, and the Treasurer of the European Respiratory Society (ERS).

M. FUNKE-CHAMBOUR received the M.D. degree from Rheinische Friedrich-Wilhelms University, Bonn. She performed her internal medicine training in Lausanne, followed by her fellowship for pulmonary medicine in Bern. From 2007 to 2011, she was a Postdoctoral Research Fellow with the MGH/Harvard Medical School, investigating the pathomechanisms for pulmonary fibrosis. Since 2011, she has been a Research Group Leader and conducts basic, translational and clinical research in the field of lung fibrosis. Currently, she is an Associate Professor in pulmonary medicine with the University of Bern. She is also a Chief Physician and the Deputy Director of the Department for Pulmonary Medicine, Allergologie and Clinical Immunology, Inselspital, University Hospital Bern, Bern.

J. HEVERHAGEN received the Medical degree from the University of Marburg, Germany. He completed his residency in radiology from The Ohio State University, Columbus, OH, USA. He completed his Postdoctoral Fellowship in advanced MRI applications with The Ohio State University, from 2002 to 2006. Thereafter, he became a Professor in medical physics and the Vice Chair of Radiology with the University of Marburg. He received grants to for advanced imaging techniques, including MRI. In 2012 he became the Chair of Radiology with Inselspital, University Hospital Bern, and a Full Professor in radiology with the University of Bern. His research interests include MRI applications and abdominal radiology, especially quantitative applications. Besides abdominal radiology his expertise includes vascular, pelvic, and senologic imaging.

H. HOPPE received the M.D. degree. He is currently an Associate Professor with the University of Bern, Switzerland. He is a board certified diagnostic and an interventional radiologist. He completed a clinical fellowship in interventional radiology with the Dotter Interventional Institute, Oregon Health & Science University, Portland, OR, USA, from 2006 to 2007. Before that, among other places, he was trained in cardiovascular MRI with the Royal Brompton Hospital, London, in 2002. His research interests include the development and application of AI in diagnostic and interventional radiology and to assess, evaluate, and advance the development of novel endovascular percutaneous interventional procedures.

A. K. EXADAKTYLOS received the Graduate degree from the MLU Medical School, in 1996, the M.D. degree from MLU Germany, and the M.Sc. degree in translational medicine in Edinburgh. He completed FRCEM in emergency medicine in London, SGNOR in emergency medicine in Switzerland, and FMH on anaesthesia in Switzerland. He is currently the Chairman, the Clinical Director, and a Professor with the Department of Emergency Medicine, Inselspital, University Hospital Bern. As a Professor, he is a recipient of many awards and grants for his valuable contributions and discoveries in major area of research. He is an editorial board member of many peer-reviewed journals and his area of expertise, as a Professor credits him with many publications in national and international journals. He is committed to highest standards of excellence, and it proves through his authorship of many books.

L. EBNER received the M.D. degree from the University of Graz, Austria, and the Habilitation degree in chest radiology, in 2017. He completed his residency with Inselspital, University Hospital Bern, Bern. During his residency, he was a Research Fellow with the University of Zurich, Virtopsy Group, in 2014. From 2015 to 2016, he was appointed as a Research Fellow with the Chest-Imaging Department, Duke University, NC, USA. In 2017, he complemented his subspecialized training in chest radiology with the Radiology Department, University Hospital Vienna. He was appointed as an Associate Professor, in 2019. He is currently a Radiologist with Inselspital, University Hospital Bern, where he is also the Section Chief of the Thoracic Imaging Section. His research interests include the development and clinical translation of AI methods, specifically for the diagnosis, the quantification and prognostication of diffuse lung diseases, lungs cancer screening, and the functional MRI of the lungs.

S. MOUGIAKAKOU (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the National Technical University of Athens. She is currently an Associate Professor with the Faculty of Medicine, University of Bern. She also leads the Laboratory for AI in Health and Nutrition, ARTORG Center for Biomedical Engineering Research Bern, Switzerland. Her research interests include artificial intelligence, machine learning, computer vision and their application in prevention, diagnosis, and the treatment optimization of acute and chronic diseases. In recent years, she has successfully participated with her research group in many competitive national, European, and internationally funded research and development projects, resulting in numerous publications, a number of patents, and several technology transfer activities.