Accelerating research on 3D medical image classification and regression

Luuk Boulogne

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Accelerating research on 3D medical image classification and regression

Luuk Boulogne

Author: Luuk Boulogne
Title: Accelerating research on 3D medical image classification and regression

Radboud Dissertations Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Luuk Boulogne
Cover: Proefschrift AIO | Guntra Laivacuma
Printing: DPN Rikken/Pumbo

ISBN: 9789465150277
DOI: 10.54195/ 9789465150277
Free download at: www.boekenbestellen.nl/radboud-university-press/dissertations

RADBOUD
UNIVERSITY
PRESS

Accelerating research on 3D medical image classification and regression

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 12 februari 2025
om 10.30 uur precies

door

Lucas Hendrikus Boulogne

geboren op 12 mei 1994
te Enschede

Promotoren:	**Prof. dr. B. van Ginneken**
	Prof. dr. H.F.M. van der Heijden
Copromotor:	**Dr. ir. C. Jacobs**
Manuscriptcommissie:	Prof. dr. G.J.S. Litjens
	Dr. A. Reinke (Deutsches Krebsforschungszentrum, Duitsland)
	Prof. dr. F.M.E. Franssen (Maastricht University)

Accelerating research on 3D medical image classification and regression

Dissertation

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Wednesday, February 12, 2025
at 10.30 am

by

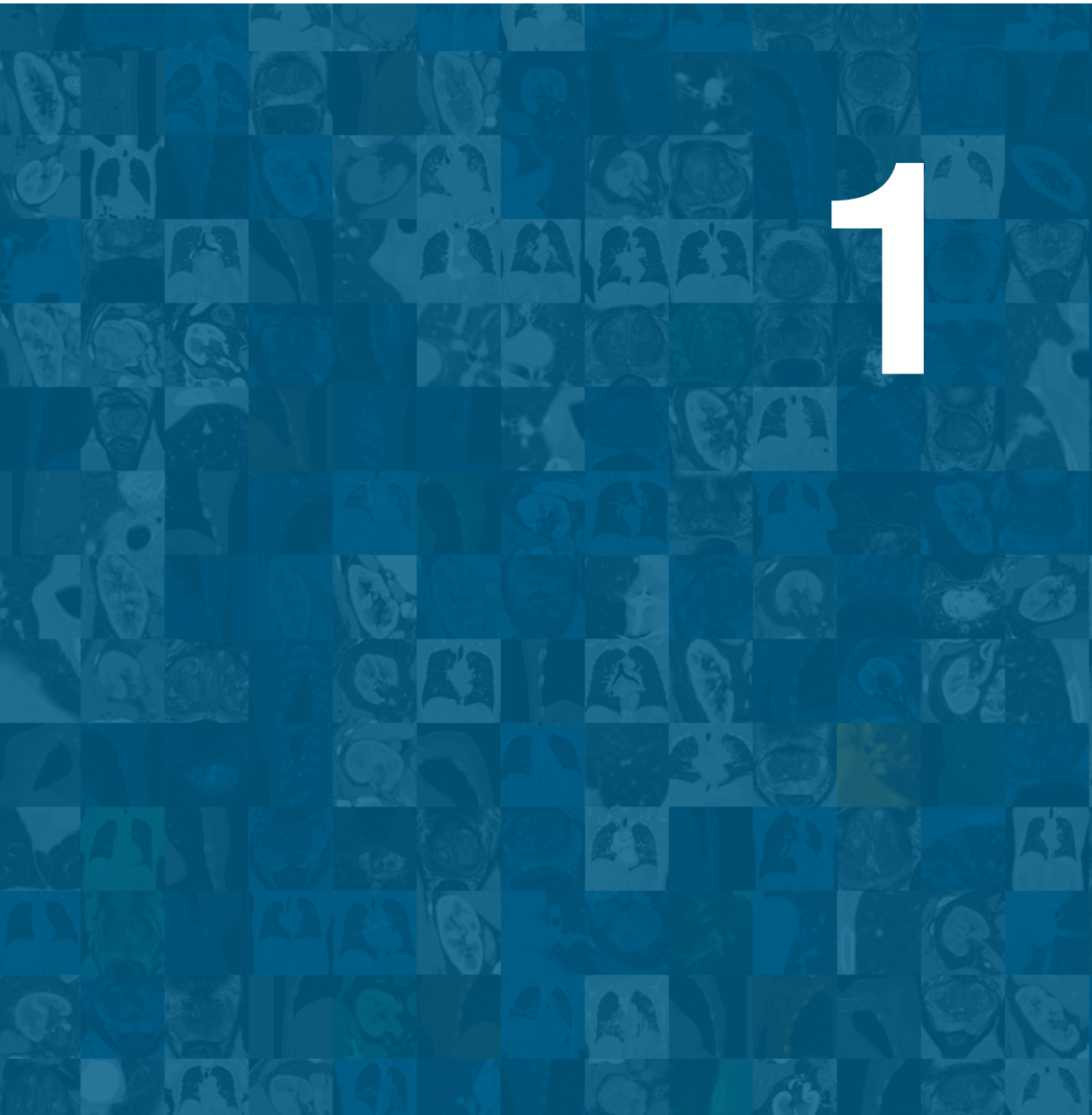
Lucas Hendrikus Boulogne

born on May 12, 1994
in Enschede (the Netherlands)

Supervisors:	**Prof. dr. B. van Ginneken**
	Prof. dr. H.F.M. van der Heijden
Co-supervisor:	**Dr. ir. C. Jacobs**
Manuscript Committee:	Prof. dr. G.J.S. Litjens
	Dr. A. Reinke (German Cancer Research Center, Germany)
	Prof. dr. F.M.E. Franssen (Maastricht University)

TABLE OF CONTENTS

Introduction



1

Medical images enable medical experts to spot abnormalities in the body that are otherwise invisible to the naked eye. Interpreting these images involves classification, i.e. sorting them into distinct categories, and regression, i.e. estimating or predicting continuous metrics from them. Three-dimensional (3D) medical images, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), are especially detailed and time-consuming to interpret. The production of medical imaging data is increasing. In the Netherlands for example, this trend can be observed from the number of CT and MRI studies conducted [1, 2]. With the increased production of imaging data, a shortage of medical experts to interpret medical images has emerged. This shortage is only expected to keep growing in the coming years [3]. The aid of automated systems may help reduce the workload and improve the accuracy of medical professionals for several medical image analysis processing tasks.

This thesis focuses on the development of automated classification and regression methods from 3D medical images. The developed methods may be useful for various purposes, including diagnosis, monitoring, and assessing the risk associated with interventions and surgical procedures.

1.1 3D image classification and regression

1.1.1 Deep learning

Deep Learning (DL) [4] models have achieved human-level performance for various image processing tasks [5–7]. Applying a DL model to an image consists of performing a sequence of simple processing steps, which are often referred to as layers. Each of these processing steps involves a collection of parameters.

Changing the parameters of a layer changes the operation that it performs. Training a DL model involves adjusting the parameters in such a way that the model performs a specific task. This training process requires data that describe the task of interest to guide how to adjust the parameters. Each individual layer in a DL model can only learn to extract simple features from its input. However, by optimizing the parameters of all the layers together, the model can learn complicated concepts, built up from the simple features learned by the individual layers.

Different training methods can be used to train DL models. A DL training method defines which model configuration, also referred to as architecture, is used, and how the training data is used to adjust the model parameters.

1.1.2 Two dimensional image processing

Most applications that require natural images to be processed are aimed at processing two-dimensional (2D) photographs. Because of this, most of the existing computer vision systems are designed to handle 2D natural images only. Research to build such systems well has resulted in a large body of literature focused on DL systems for processing these 2D images [5, 8–11]. This research has led to easy-to-use DL platforms for computer vision, the widespread adoption of convolutional neural networks (CNNs) [8], and more recently to the adoption of Visual Transformers (ViTs) for image processing as well [11, 12].

CNNs and transformers are both types of DL models, each defined by the different types of layers that they consist of. In CNNs, most of the processing power is allocated to convolutional layers, which extract features from their input by applying filters. Filters in early layers detect simple features such as edges, while later layers combine simpler features to detect complex shapes and structures. CNNs are used in every chapter of this thesis.

The heavy lifting in ViTs is performed by self-attention layers that weigh the importance of different parts of the input data, regardless of their positions in the image. In order to focus on any relevant parts of the input, self-attention layers require a large memory footprint. ViTs were employed in Chapter 5 of this thesis.

1.1.3 3D image processing

Most of the research done in the 2D image domain transfers well to 3D image processing. However, with respect to 2D, 3D volumes do pose some additional challenges for effectively processing images with DL, namely handling the additional spatial dimension and with this, dealing with the larger memory footprint that 3D DL models have with respect to their 2D counterparts.

The first approach for handling the extra spatial dimension involves viewing the 3D image as a collection of 2D images. This is a common approach in the multiple instance learning paradigm [13]. With the multiple instance learning approach, 2D DL models are used to extract features from each of these images separately. The extracted 2D features are aggregated along the third dimension to obtain a 3D representation, which in turn is used to make predictions about the complete image volume. The aggregation step can be implemented in multiple ways. One straightforward implementation is to simply obtain the 3D image representation by taking a sum or average of the extracted 2D features [14]. This approach has a downside in that it regards the 2D images that make up the 3D volume as an unordered set, resulting in the loss of structural information along the third dimension. This 3D structure

can be helpful for evaluating and classifying 3D medical images [15]. DL systems that can utilize 3D information well may thus ultimately result in better patient care. Regardless of how the aggregation step is implemented, first extracting features in 2D and subsequently aggregating them to get a 3D representation may be inefficient for obtaining a representation that captures information about 3D structures well and may be suboptimal for model accuracy and throughput.

Another way to handle the extra spatial dimension is to view the 3D volume as a sequence of 2D images. Methods that are well-suited for processing sequences of data, such as recurrent neural networks, can therefore also be used for processing 3D image volumes. Contrary to the approach described above where 2D models first extract features that are then aggregated along the third dimension, sequence processing models can aggregate features along the third dimension throughout the feature extraction process. This allows the extracted features to capture and represent structural information in the third dimension.

3D volumes can also be processed with approaches that handle all three spatial dimensions equally. Many medical image volumes, such as spiral CT and MRI images are acquired as a collection of 2D slices. Sequence modeling approaches treat the first two dimensions that make up these 2D slices differently than the third dimension along which the volume is divided into slices. Such an approach does not recognize that the imaged structures are inherently three-dimensional by treating the three dimensions equally. An alternative approach that can treat all three dimensions equally is the 3D CNNs [16]. A 3D CNN is a conventional 2D CNN where the convolution kernel which is usually two-dimensional has been extended into the third dimension.

Some applications of 3D medical image processing require large volumes high resolution 3D images to be processed. In such applications, the file sizes of the 3D images are significantly larger than the file sizes of 2D images typically encountered for main-stream AI applications. This in turn increases the memory requirements of the processing methods. ViTs have gained much popularity during the preparation of this thesis, especially for 2D image processing. Their larger memory footprint with respect to CNNs magnifies their memory requirements for 3D image processing with respect to 2D image processing even more.

1.2 Applications

The applications of regression and classification that this thesis mainly focuses on are processing thoracic CT scans for automatic Pulmonary Function Test (PFT) estimation and COVID-19 classification. CT scans are three-dimensional images of the

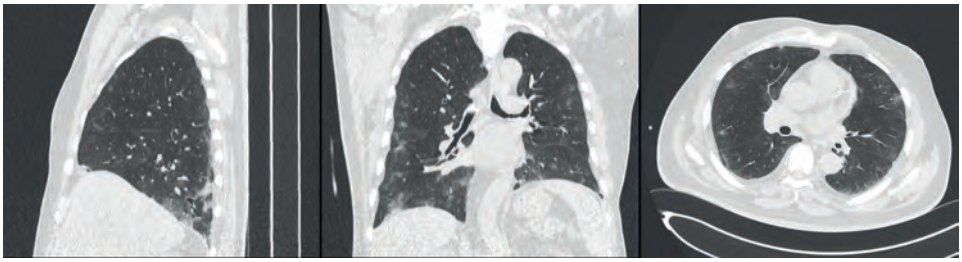


Figure 1.1: Sagittal, coronal, and axial cross-sections of a CT scan imaging a patient with RT-PCR confirmed COVID-19 from the STOIC2021 public training dataset [17].

inside of the body. They consist of a series of axial cross-sectional images, each of which is reconstructed from multiple X-ray measurements taken from different angles. Figure 1.1 shows an example of a thoracic CT scan.

1.2.1 Pulmonary function test result estimation

PFTs produce measurements of how well the lungs are functioning. They are useful for gaining valuable insights relevant for many purposes, including the diagnosis and staging of respiratory diseases [18–20] and risk assessment of bronchoscopic and surgical lung volume reduction [21, 22]. Two types of PFT that are often performed for these purposes are spirometry testing and single breath Diffusion Capacity for carbon monoxide (DLCO).

Spirometry

Spirometry tests measure the volume of air a patient exhales or inhales as a function of time [19]. Two key measurements performed during spirometry testing are the volume of air the patient can expel from their lungs in a set time, denoted as Forced Expiratory Volume in one second (FEV1), and the volume of air that the patient can expel from their lungs in total, denoted as the Forced Vital Capacity (FVC).

DLCO

DLCO tests measure how well the lungs can transfer oxygen from the alveolar gas into the bloodstream, by using the uptake of carbon monoxide (CO) as a proxy [20]. During a DLCO test, a patient fully exhales, then inhales a mixture of CO and tracer gas (usually helium) to full inflation, holds their breath for ten seconds, and then exhales. DLCO is computed by analyzing the exhaled CO and tracer gas concentrations. The DLCO measurement depends on the the rate of uptake of CO by the lungs

from the alveolar gas and as well as the total volume of the alveoli.

PFT variability

Regarding PFT estimation from CT, it may be noted that the reliability of PFT results depends on the consistency of the testing procedure, which relies on effective patient coaching. Test quality is controlled by assessing whether repeated tests fall within guideline boundaries [19, 20]. During the acquisition of CT scans, patients are instructed to perform full inspiration and expiration, but they are generally not coached during this process. Because of this, PFTs involve more extreme levels of inspiration and expiration than inspiratory and expiratory CT scans.

1.2.2 COVID-19 classification

During the COVID-19 pandemic that began in late 2019, Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests for the detection of COVID-19 were not yet widely available. Researchers around the globe investigated the use of CT for combating the disease. This led to the development of the CO-RADS scoring system [15], which is a standardized reporting system that indicates the level of suspicion of a COVID-19 infection. Research to combat the pandemic also included the development of automatic COVID-19 classification from CT [14, 23–44], often aimed at patient triaging and reducing the workload of medical professionals. Challenges that remained included how to leverage these methods in such a way that they can be used for effective communication within hospitals, and performing fair comparisons between different methodologies.

1.3 Grand challenges

Competitions for developing systems that automatically analyze medical images have shown to be useful for obtaining high-performance solutions to medical image analysis problems. Although medical image analysis challenges are widely applicable to many tasks, individual challenges often focus on narrow tasks [6, 17, 45–53]. They often revolve around obtaining a high-performance solution for the specific task at hand with a data set specifically collected for the event. It is often not possible to retrain the resulting solutions on new data and the possibility of translating lessons learned from these challenges to new problems is limited. Recently, challenges have emerged that focus on the development of general algorithms that can solve a wide variety of similar tasks [54]. Solutions to a more general set of problems can provide out-of-the-box methods that are applicable to many tasks [55–57]. Such

solutions may furthermore accelerate the progress of the field by providing baselines to compare novel methodologies to.

1.4 Outline

This thesis discusses methods for automated classification and regression from 3D medical images, as well as ways to accelerate the development of such methods.

- Chapter 2 describes a method for estimating PFT results from inspiration CT scans. This method was developed to also estimate the contribution of each lobe to the total patient-level lung function. It is aimed at improving the assessments of restrictive pulmonary diseases as well as risk assessments of bronchoscopic and surgical lung volume reduction.
- Chapter 3 provides a systematic comparison of automatic methods for COVID-19 classification from CT scans and provides insights into the added value of individual algorithm components to accelerate the creation of tools for accurate COVID-19 grading. It furthermore proposes adherence of automated systems to the CO-RADS reporting format to increase compatibility with clinical workflow.
- Chapter 4 describes a challenge format for training solutions on private data that guarantees reusable training methodologies of challenge solutions. It applies this format to a medical image analysis challenge aimed at identifying severe COVID-19 infections from thoracic CT scans. Severe COVID-19 was defined as death or intubation within one month after the CT scan was made.
- Chapter 5 describes a database that can be used for the development of a general-purpose automatic 3D medical image classifier to accelerate future research on 3D medical image classification and regression.

Finally, Chapter 6 summarizes the methodologies, results, and findings presented in this thesis. It furthermore indicates possible directions for future research.

Estimating lung function at the patient and lobe level

2

Authors: **Luuk H. Boulogne, Jean-Paul Charbonnier, Colin Jacobs, Erik H. F. M. van der Heijden, and Bram van Ginneken**

Original title: **Estimating Lung Function From Computed Tomography at the Patient and Lobe Level using Machine Learning**

Published in: **Medical Physics, volume 51, number 4, pp. 2834-2845, 2024**

Abstract

Background: Automated estimation of Pulmonary function test (PFT) results from Computed Tomography (CT) could advance the use of CT in screening, diagnosis, and staging of restrictive pulmonary diseases. Estimating lung function per lobe, which cannot be done with PFTs, would be helpful for risk assessment for pulmonary resection surgery and bronchoscopic lung volume reduction.
Purpose: To automatically estimate PFT results from CT and furthermore disentangle the individual contribution of pulmonary lobes to a patient's lung function.
Methods: We propose I3Dr, a deep learning architecture for estimating global measures from an image that can also estimate the contributions of individual parts of the image to this global measure. We apply it to estimate the separate contributions of each pulmonary lobe to a patient's total lung function from CT, while requiring only CT scans and patient level lung function measurements for training. I3Dr consists of a lobe-level and a patient-level model. The lobe-level model extracts all anatomical pulmonary lobes from a CT scan and processes them in parallel to produce lobe level lung function estimates that sum up to a patient level estimate. The patient-level model directly estimates patient level lung function from a CT scan and is used to re-scale the output of the lobe-level model to increase performance. After demonstrating the viability of the proposed approach, the I3Dr model is trained and evaluated for PFT result estimation using a large data set of 8 433 CT volumes for training, 1 775 CT volumes for validation, and 1 873 CT volumes for testing.
Results: First, we demonstrate the viability of our approach by showing that a model trained with a collection of digit images to estimate their sum implicitly learns to assign correct values to individual digits. Next, we show that our models can estimate lobe-level quantities, such as COVID-19 severity scores, pulmonary volume, and functional pulmonary volume from CT while only provided with patient-level quantities during training. Lastly, we train and evaluate models for producing spirometry and diffusion capacity of carbon mono-oxide estimates at the patient and lobe level. For producing FEV1, FVC, and DLCO estimates, I3Dr obtains mean absolute errors of 0.377 L, 0.297 L, and 2.800 mL/min/mm Hg respectively. We release the resulting algorithms for lung function estimation to the research community at `https://grand-challenge.org/algorithms/lobe-wise-lung-function-estimation/`
Conclusions: I3Dr can estimate global measures from an image, as well as the contributions of individual parts of the image to this global measure. It offers a promising approach for estimating PFT results from CT scans and disentangling the individual contribution of pulmonary lobes to a patient's lung function. The findings presented in this work may advance the use of CT in screening, diagnosis, and staging of re-

strictive pulmonary diseases as well as in risk assessment for pulmonary resection surgery and bronchoscopic lung volume reduction.

2.1 Introduction

Pulmonary function tests (PFTs) are widely used to assess the respiratory health of a patient, with spirometry and Diffusion Capacity of Carbon mono-Oxide (DLCO) measurements being particularly useful for this purpose [18–20]. Spirometry assesses how a person inhales or exhales a given volume of air over time [19], while DLCO measures the rate of oxygen uptake in the lungs [20].

In this pilot study, we present a machine learning method that automatically estimates spirometry and DLCO test results from Computed Tomography (CT). Our method also estimates the contribution of each pulmonary lobe to a patient's lung function, providing clinically relevant information that is impossible to obtain from PFTs.

Methods for accurate automatic lung function prediction from CT have the potential to replace PFTs when CT scans are already available. This would make them useful for the many purposes that PFTs are used for today, such as diagnosing [18, 19] and determining the efficacy of treatment [18] of astma, and the diagnosis [19] and staging [18, 58] of Chronic Obstructive Pulmonary Disease (COPD).

In order to determine operability of patients with an early stage lung cancer, clinicians largely depend on the prediction of post-operative pulmonary function (ppo) based on calculation of the effect of the number of segments that will be removed upon surgery [59] . These estimations are relatively crude, are based on a homogeneous perfusion of each lobe and do not compensate for the changes in volume of the remaining pulmonary tissue in the operated hemi-thorax [60]. More precise lobe level lung function estimates could improve the risk assessment of pulmonary resection surgery.

Both spirometry and DLCO measurements are currently used for the risk assessment for [21] and advocated to be used in the inclusion criteria for bronchoscopic lung volume reduction [22]. Access to lobe level lung function estimates could improve the accuracy of these assessments.

In this work, we utilized an Inflated 3D ConvNet (I3D) [16] backbone to construct the regional I3D (I3Dr) model for estimating lung function at the patient level and at the lobe level. I3Dr was trained and evaluated for estimating spirometry, namely Forced Expiratory Volume in one second (FEV1) and Forced Vital Capacity (FVC), as well as DLCO measurement outcomes using a large dataset of 12 045 CT volumes. It was designed to estimate the separate contributions of each pulmonary lobe to a patient's total lung function, while requiring only CT scans and patient level lung function measurements for training.

We validated I3Dr through several experiments. We first explored the viability

of the I3Dr model in a toy experiment by showing that, when only presented with a collection of digit images and their sum, it learned to assign correct values to individual digits. Next, we validated its ability to estimate lobe level quantities from CT while only training with patient level labels through estimating lobe level COVID-19 severity scores, Pulmonary Volume (PV), and Functional PV (FPV). Finally, we used the I3Dr model to estimate lobe level lung function quantities that sum up to patient level PFT estimates. The presented methods outperformed using FPV for PFT result estimation.

The main contributions of this work are as follows:

- We show that FEV1, FVC, and, as a first to the best of our knowledge, DLCO measurement outcomes can be estimated from CT using machine learning.
- We proposed and validated the I3Dr model for disentangling regional contributions from a patient level label and applied this model to estimate lobe level lung function.
- We made our models for PFT result estimation at the patient and lobe level publicly available to the research community.

2.2 Related work

2.2.1 Estimating regional contributions using machine learning

The presented method for lobe level lung function estimation was inspired by class activation maps (CAMs) [61] and in particular by the Bag-of-local-Features (BagNet) model [62]. Like most deep learning classification models, BagNet consists of a series of feature-extracting convolution layers followed by global average pooling and a linear layer that outputs the logits used for classification. By swapping the final pooling layer and the final linear layer of the model at test time, which does not change the model output [61, 62], BagNet can produce interpretable heat maps. After swapping, each activation value produced by the linear layer can be interpreted as a regional output for the receptive field patch of the corresponding neuron.

BagNet has been used for a variety of medical image processing tasks in various modalities. It was deployed for sex and age prediction to generate heat maps for brain MRI volumes [63] and retinal images [64, 65]. In other work, BagNet was extended with a MIL branch and trained to generate interpretable heat maps for histology images describing malignant and benign regions [66].

In our work, we exploited the principle of swapping the final pooling layer and linear layer of a model to produce lung function estimations at the pulmonary lobe

level. Instead of producing heat maps, we designed a model in which the receptive fields of the output neurons of the final linear layer each contain one pulmonary lobe. Our method assigns lung function estimates to each pulmonary lobe, such that they sum up to an image level estimate.

2.2.2 PFT prediction

Earlier work has shown promising results for producing PFT results at a patient level using convolutional neural networks. For example, total lung volume has been estimated from chest radiographs [67] and CT scans have been used for estimating spirometry test results [68]. These methods did not produce lobe level estimates.

Various methods for lobe level lung function estimation that do not make use of the complex features in a CT scan have been introduced in previous works. The risk assessment guidelines for pulmonary resection surgery advise to estimating the residual lung function of a patient with a simple calculation using pulmonary segment counting [59]. Lobe level lung function has furthermore been estimated using Pulmonary Volume (PV) as obtained from a lung segmentation [69], as well as through Functional Pulmonary Volume (FPV) [70–74]. Here, after obtaining the lobe level PV, dual thresholding operations on the CT scan are applied to acquire an estimation of the volume of functional parenchyma. These methods have been shown to outperform the segment counting method for the task of predicting PFT results after lung resection surgery [69, 72]. Post-operative PFT results have also been predicted using an estimation of lobar collapsibility [71], which is computed as the fraction of change in FPV between an inspiration and an expiration CT scan. Lastly, SPECT/CT images have been used to determine the lobe level FEV1 of a patient [75] by computing the proportion of radioactivity in the lobe of interest with respect to the total radioactivity.

To the best of our knowledge, we are the first to propose a method for lobe level lung function estimation that makes use of the detailed information available in a CT scan and that does not require the patient to undergo additional imaging.

2.3 Methods

2.3.1 I3Dr

Fig. 2.1 shows the training and inference pipelines of the regional I3D (I3Dr) model proposed in this work, which consists of patient level model A and a lobe level model B.

Patient level model

First, we trained model A to map a CT scan x to a corresponding PFT result vector y. Model A consisted of a feature extractor $f^A(\cdot)$ followed by a linear fully connected layer $r^A(\cdot)$:

$$\hat{y}^A = W^A f^A(x) + b^A.$$

Here, the linear fully connected layer $r^A(\cdot)$ is parameterized by weight matrix W^A and bias b^A. Fig. 2.1a schematically presents this conventional approach.

Lobe level model

Next, we trained a model B that processed images of all anatomical pulmonary lobes of a patient in parallel to produce a patient level PFT result estimate. The model was designed so that it could produce meaningful lobe level lung function estimates during inference. More specifically, the training and evaluation pipelines were altered to force the model to output lobe level lung function estimates that sum up to a patient level estimate. Fig. 2.1b and Fig. 2.1c shows these modifications.

Training To extract feature vectors per lobe, the lobes were segmented from the input CT scan and presented individually to feature extractor $f^B(\cdot)$. The resulting feature vectors were aggregated and presented to the final linear layer $r^B(\cdot)$. This process can be described as:

$$\hat{y}^B = W^B \left(\sum_{l=1}^{L} f^B(x_l) \right) + b^B.$$

Here, a linear fully connected layer $r^B(\cdot)$ is parameterized by a weight matrix W^B and bias b^B. L is the number of regions for which the model produced estimates at test time. For the task of PFT result estimation, $L = 5$, as it represents the number of anatomical pulmonary lobes. Each x_l is a separate input region for which the model produces estimates. For PFT result estimation, x_l is the image of a pulmonary lobe l of CT scan x.

Only a single feature extractor $f^B(\cdot)$ was used for processing all five lobes. This sharing of weights allows $f^B(\cdot)$ to take advantage of the common features of interest across different anatomical lobes.

Inference At test time, the summation operation and final linear layer $r^B(\cdot)$ of model B are swapped, which does not change the model's output [61, 62]. After

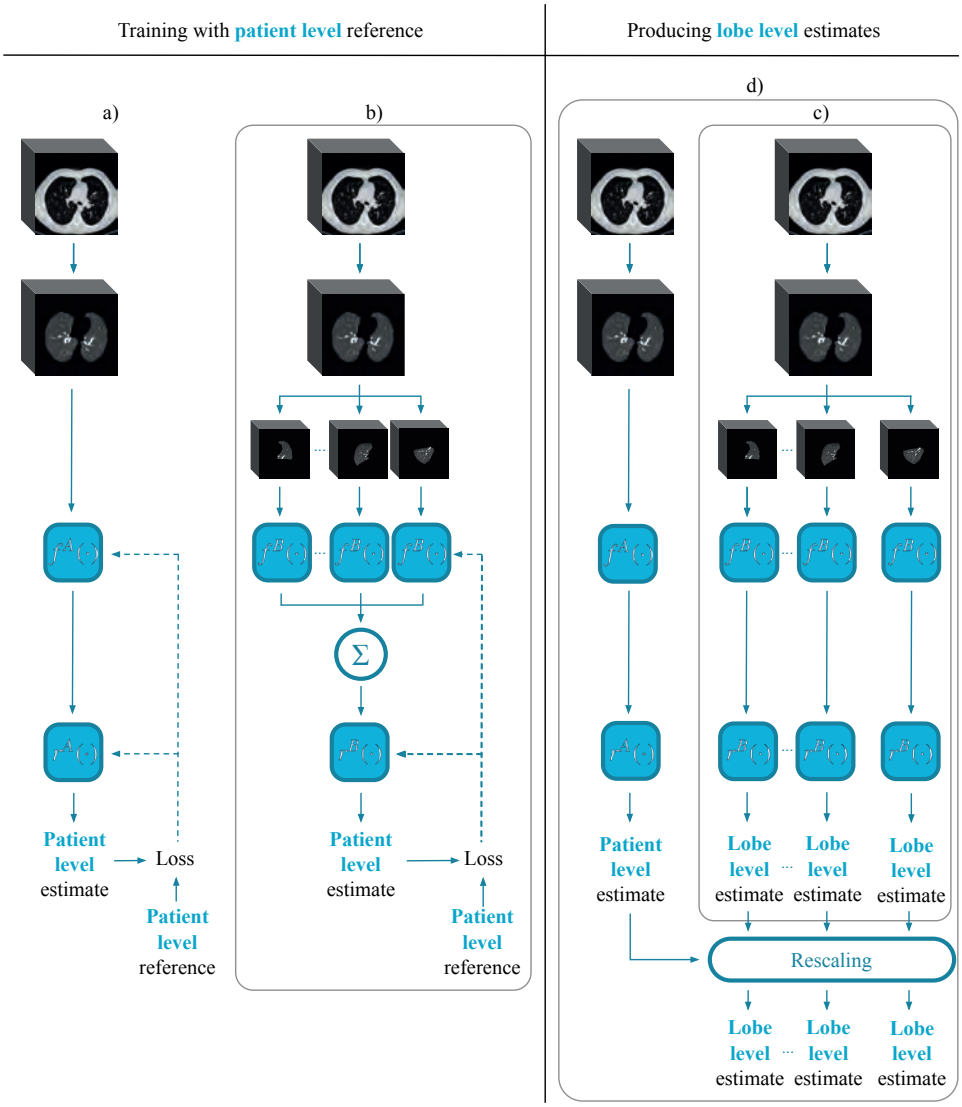

Training with patient level reference
Producing lobe level estimates
a)
b)
c)
d)
Patient level estimate
Loss
Patient level reference
Patient level estimate
Loss
Patient level reference
Patient level estimate
Lobe level estimate
Lobe level estimate
Lobe level estimate
Rescaling
Lobe level estimate
Lobe level estimate
Lobe level estimate

Figure 2.1: Schematic representations of the training and evaluation pipelines for the I3Dr model, which combines two distinct machine learning models. a) Training procedure for the model that learned to produce patient level lung function estimates from a CT scan (model A). First, a lung mask was applied to the CT scan. Next, features were extracted by the feature extractor $f^A(\cdot)$. The linear layer $r^A(\cdot)$ was applied to obtain a patient level estimate. Loss was computed by comparing this estimate with the patient level reference to update $f^A(\cdot)$ and $r^A(\cdot)$. b) Training procedure for the model that learned to produce lobe level estimates (model B). First, a lobe mask was applied to the input CT scan. The segmented lobes were processed by the feature extractor $f^B(\cdot)$ individually. During training, the feature vectors corresponding to each lobe were summed together. The resulting feature vector was presented to the linear layer $r^B(\cdot)$ to produce a patient level estimate. Only a patient level reference was used to compute the loss and update $f^B(\cdot)$ and $r^B(\cdot)$. c) Inference pipeline for model B. During inference, the feature vectors corresponding to each lobe were not aggregated. Instead, they were presented to linear layer $r^B(\cdot)$ individually to produce lobe level estimates that sum up to a patient level estimate. d) To increase performance, the output of model A, that learned to produce patient level lung function estimates from a CT scan, was used to rescale the lobe level estimates produced by model B.

swapping, each of the activation values produced by the final linear layer had a receptive field containing one pulmonary lobe. These values could therefore be interpreted as lobe level lung function estimates that sum up to a patient level PFT result estimate. Fig. 2.1c shows this inference pipeline.

For an image of pulmonary lobe l of CT scan x, a lobe level estimate was obtained as follows:

$$\hat{y_l}^B = W^B f^B(x_l) + \frac{b^B}{L}.$$

These lobe level estimates could be aggregated again by the summation operation to produce the patient level estimate that the model produced before swapping the summation operation and final linear layer $r^B(\cdot)$:

$$\hat{y}^B = \sum_{l=1}^{L} \hat{y_l}^B.$$

Combined model

We found in our experiments that model A (see section 2.3.1) regularly outperformed model B (see section 2.3.1) on the patient level. We therefore combined models A and B into one model that we refer to throughout this work as the I3Dr model. The output of this model was obtained by simply rescaling the output of model A with the output of model B as follows:

$$\hat{y_l} = \hat{y_l}^B \frac{\hat{y}^A}{\hat{y}^B}.$$

Fig. 2.1d shows the complete inference pipeline for the I3Dr model.

2.4 Experiments

Since PFTs do not describe lobe level lung function, the approach described in section 2.3.1 for lobe level lung function estimation cannot be validated directly. We therefore designed and performed several experiments where we do have access to local and global measurements to validate whether our models can actually produce meaningful regional (lobe level) output when only receiving feedback based on global (patient level) measurements during training. In our final and most comprehensive experiment, we trained a model with patient level PFT results to produce lobe level lung function estimates.

2.4.1 Summing digits

In this proof-of-concept experiment, we used 2D images of digits instead of images of pulmonary lobes, and sums of digits instead of patient level PFT results.

Dataset

We conducted this experiment with the MNIST dataset [8], which contains a training set of 60 000 and a test set of 10 000 images of digits of 28×28 pixels each.

Experiment design

For each of the models, the training set of digit images was randomly divided into collections or 'bags' of images prior to training. We used the bag sizes of 2^i with $i \in [0..8]$. Each bag was labeled with the sum of all the digits it contained. Fig. 2.2a shows some input-label example pairs for a bag size of 8.

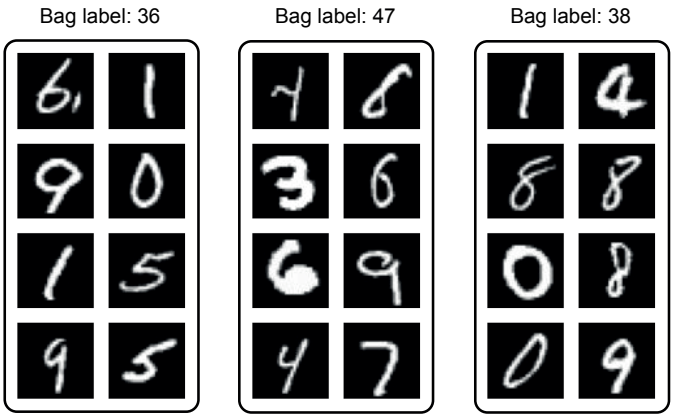
The models were trained with the methodology described in section 2.3.1 to produce these bag level sums in a regression setting using 85% of the 60 000 MNIST training images. A validation set to monitor the performance on this task was constructed with the remaining 15%. No individual digit labels were presented to provide feedback to the model during training, except when training the model with a bag size of one. After training, the models were evaluated on their performance for estimating individual digit labels using the 10 000 test images. For this experiment, the rescaling described in section 2.3.1 was not performed.

2.4.2 COVID-19 severity scoring

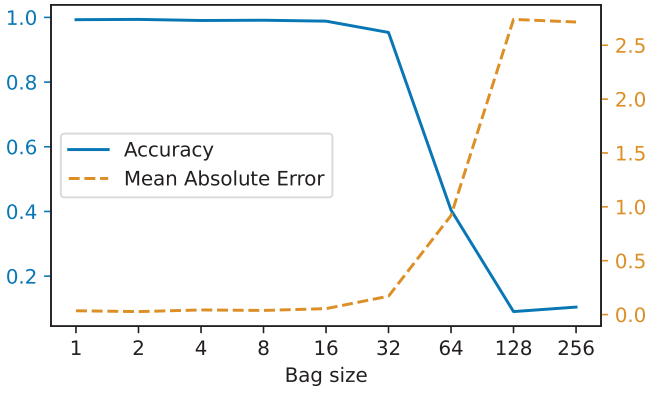
Subsequently, to validate whether our approach could be used for estimating meaningful lobe level quantities using only patient level labels, we trained two models to estimate COVID-19 CT Severity Scores (CTSSs) [76]. The CTSS indicates the severity of a COVID-19 infection for individual lobes. In clinical practice, these lobe-wise scores are summed up to a patient-level CTSS. This summing is analogous to how many patient-level Pulmonary Function Test (PFT) results are the sum of the contributions of lobe level quantities.

Dataset

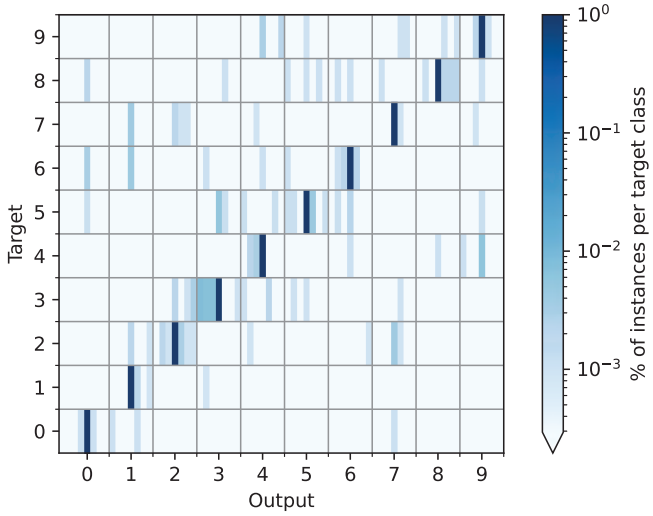
To develop and test models for CTSS estimation, we adopted the internal dataset from [33]. This dataset contains 482 CT scans for which each of the pulmonary lobes were scored with a CTSS by a radiologist. The dataset was split at the patient level


Bag label: 36
Bag label: 47
Bag label: 38

(a)


1.0
0.8
0.6
0.4
0.2
2.5
2.0
1.5
1.0
0.5
0.0
Accuracy
Mean Absolute Error
1
2
4
8
16
32
64
128
256
Bag size

(b)



(c)

Figure 2.2: MNIST experiment to validate whether a model can implicitly learn the contribution of individual images to a collection level label. a) Three examples of input-label pairs for this experiment. During training, the machine learning model is presented with a collection (bag) of images and a bag level label. Each bag is labeled with the sum of all the labels of the digits in that bag. The model only receives feedback based on these bag labels during training, and not based on the individual digit labels. b) Performance on the test set of models trained to predict the sum of a bag of digits, evaluated on the task of predicting the value of a single digit for different bag sizes. c) Individual digits from the test set were presented to the model trained for estimating the sum of a bag of 8 MNIST digits. The model was evaluated for the task of predicting the individual label of each digit. Each row shows a histogram of the continuous model output for all images depicting a single target digit.

into a training set (293 scans) validation set (84 scans) and test set (105 scans). The test set used in this work was identical to the test set used by N. Lessmann *et al* [33].

Experiment design

First, a baseline model was trained for the task of producing lobe level CTSS estimates from images of individual pulmonary lobes to obtain a performance limit for lobe level CTSS estimation. This model was trained in a fully supervised fashion using the conventional methodology described in section 2.3.1.

Furthermore, we trained an I3Dr model using the methodology described in section 2.3.1. Only patient level CTSS labels were used for training this model.

Both models were evaluated on the task of producing lobe level CTSS labels.

2.4.3 Lung function estimation

Lastly, we trained an I3Dr model to jointly produce patient level Pulmonary Volume (PV), Functional Pulmonary Volume (FPV), and PFT results from CT scans.

The PV was estimated by taking the volume of a lung mask. The FPV is the PV from which emphysematous regions, large blood vessels and airways, and dense lesions such as fibrosis are excluded, so that only the functional parenchyma remains. It was computed as the volume of a CT scan within the lung mask where only voxels with HU in the range [-950, -500] HU were included. The resulting subject level PV and FPV measures were used as reference targets for training the I3Dr model.

Lobe level PV and FPV values were computed in a similar fashion, but using segmentations of individual lobes instead of lung masks. These lobe level PV and FPV measures were not used during training. They were only used as ground truth for evaluating the model.

The PFT results that the model was trained to estimate were Diffusion capacity of the Lungs for Carbon monOxide (DLCO) as well as spirometry measurements, namely Forced Expiratory Volume in one second (FEV1) and Forced Vital Capacity (FVC). In this work, spirometry measurements that were performed pre- and post-bronchodilator are indicated with superscripts ‘pre’ and ‘post’ respectively.

After training, the performance of the I3Dr model for predicting PFT results was compared to using FPV for PFT result estimation.

Dataset

For this experiment, we used inspiration CT scans from the COPDGene study [77]. Data for this clinical trial was collected from 21 imaging centers in the United States.

Table 2.1: Number of CT Images (and Subjects) in the COPDGene Dataset used for the lung function estimation. In this dataset, DLCO measurements were not available as often as spirometry measurements.

	Train	Val	Test	Total
Total	8 433 (6 023)	1 775 (1 271)	1 837 (1 304)	12 045 (8 598)
$FEV1^{pre}$	8 415 (6 020)	1 772 (1 271)	1 831 (1 304)	12 018 (8 595)
$FEV1^{post}$	8 356 (5 987)	1 760 (1 262)	1 819 (1 298)	11 935 (8 547)
FVC^{pre}	8 414 (6 020)	1 772 (1 271)	1 831 (1 304)	12 017 (8 595)
FVC^{post}	8 355 (5 987)	1 760 (1 262)	1 819 (1 298)	11 934 (8 547)
DLCO	2 350 (2 350)	496 (496)	500 (500)	3 346 (3 346)

The COPDGene dataset was divided randomly into separate sets for training, validation, and testing. This split was performed at a subject level to ensure that the model performance on the test set is not tainted due to overfitting on subject-specific information. An overview of the numbers of CT scans, patients, and measurements used in this work can be found in Table 2.1.

The PV and FPV reference were computed using the lung masks available in this dataset.

Experiment design

An I3Dr model was trained using all CT scans in the COPDGene training set (see Table 2.1) to produce patient level PV, FPV, $FEV1^{pre}$, $FEV1^{post}$, FVC^{pre}, FVC^{post}, and DLCO estimates. The performance of the model for producing these patient level labels was evaluated, as well as its performance for producing lobe level PV and FPV.

2.4.4 CT scan preprocessing

Prior to presenting the CT scans to a model, they were clipped between -1100 and 300 HU and the voxel values were scaled to the range [0,1]. After this, the CT scans were isotropically resampled to 1.6 mm^3 with linear interpolation. A lobe segmentation was used to mask out all voxels outside of the region of interest. This region was either a lung mask or a mask of a single lobe, depending on the model being trained. The resulting volumes were centered around the region of interest and cropped to $240 \times 240 \times 240$ voxels.

For the COPDGene dataset, the lobe segmentations that were provided with this

dataset were used for masking and cropping the CT scans. These lobe segmentations had been automatically generated by a commercialized software (LungQ, Thirona, Nijmegen, The Netherlands), and were manually corrected by trained analysts that had at least one year experience in annotating pulmonary structures on CT.

RTSU-Net [78], was used to obtain lobe segmentations for masking and cropping the CT scans of the CTSS dataset from N. Lessmann *et al* [33].

2.4.5 Training details

Training models for this work was done on a single GPU, using NVIDIA GeForce GTX TITAN X, GeForce GTX 1080, GeForce GTX 1080 Ti, GeForce RTX 2080 Ti, and TITAN Xp graphics cards.

Summing digits

For the experiments conducted with the MNIST dataset, the images were resized to 33×33 pixels, and the BagNet-33 architecture [62] was used as a feature extractor. The models were trained to minimize the L1 loss between their bag level predictions and the bag level labels using the Adam optimizer (β_1=0.9, β_2=0.999) with a learning rate of 0.001 and a batch size of 512 divided by the bag size. The models were trained for 1 000 epochs and performance on the validation set was monitored after each epoch. The model weights with the best performance on the validation set were used for evaluation.

CTSS and PFT result estimation

For CTSS and PFT result estimation, we modeled the feature extractors $f^A(\cdot)$ and $f^B(\cdot)$ with the Inflated 3D ConvNet (I3D) [16], without its final layer. Because pre-training has been shown to be advantageous for processing medical images with convolutional neural networks [79], we initialized our model with publicly available weights trained for RGB video classification [16].

All models were trained using the Adam optimizer (β_1=0.9, β_2=0.999) to minimize the L1 loss with a learning rate of 1×10^{-4}, weight decay of 1×10^{-2}, and a batch size of 2. Early stopping was used with a patience of 15 epochs.

Data augmentation was used to increase training data diversity without altering relevant image features. The augmentations consisted of rotation, translation, and shearing in the axial plane, elastic deformation, and adding Gaussian noise. To accommodate the RGB input format of the I3D model, the CT scans were copied along the channel dimension after applying data augmentation.

All CT scan labels were divided by their mean standard deviation in the training set before computing loss. This was done to avoid unbalanced penalties due to the differences in magnitude distributions of different measurement types. For input CT scans for which not all labels were available, the loss was computed using only the available labels.

To compute the loss for a single CT scan when training a model with the methodology presented in section 2.3.1, the feature extractor $f^B(\cdot)$ would conventionally process a minimal batch size equal to the number of lobes in a CT scan. This minimum batch size of five was too large to fit on a 12GB GPU. To circumvent this problem, we applied the gradient checkpointing method described in Algorithm 1. We did not pass a full batch of lobe images through $f^B(\cdot)$ when training these models. Instead, we further divided the training batch X comprised of lobe images x_i into a set of virtual mini-batches M, such that each virtual mini-batch $m \in M$ contained two lobe images. These virtual mini-batches were passed through the feature extractor separately without gradient computation (lines 1-3). Subsequently, for each CT scan s from the set S of CT scans with which the training batch X was constructed, the features extracted from the lobes of that CT scan were aggregated and cached (line 4). The aggregated features were processed by final linear layer $r^B(\cdot)$ (line 5) and the gradient for $r^B(\cdot)$ was computed (line 6). Lastly, the gradient $f^B(\cdot)$ was aggregated by computing it for each virtual mini-batch separately (lines 7-12). More specifically, for each virtual mini-batch, the features extracted by $f^B(\cdot)$ were re-computed, this time storing the computational graph (line 9). The features for this virtual mini-batch were aggregated with the features cached in the original forward pass (line 10) in order to continue back-propagation through $f^B(\cdot)$ (line 11).

2.4.6 Evaluation

The correlation between FPV and PFT measurements and between model output and PFT measurements was computed on the test set using the Pearson correlation coefficient.

Following [33], the agreement between the CTSS labels and the output of the trained models was evaluated in terms of linearly weighted κ.

The performance of the models trained in this work were additionally evaluated in terms of Mean Absolute Error (MAE). 95% confidence intervals for performance measures were computed as the interval between the 2.5% and 97.5% percentiles of a bootstrap distribution generated with 1 000 iterations [80].

All p-values were obtained using standard permutation tests for matched pairs [80] with 10 000 iterations.

Algorithm 1 Gradient computation with checkpointing

Input: Set of lobe images in training batch $X := \{x_i\}_1^N$, set of virtual mini-batches M that is a partition of the indices of X, set of CT scans S that is a partition of the indices of X, loss function $\mathcal{L}(\cdot)$, feature extractor $f^B(\cdot)$ with learnable parameters ϕ, linear function $r^B(\cdot)$ with learnable parameters θ, mapping $Q(i)$ that maps an index i of a lobe image to the CT scan $s \in S$ that contains it.

Output: ∇r, ∇f

for all $m \in M$ **do**
  $h_i \leftarrow f^B(x_i)$ **for all** $i \in m$ {Without gradient computation}
end for
$z_s \leftarrow \sum_{i \in s} h_i$ **for all** $s \in S$
$o_s \leftarrow r^B(z_s)$ **for all** $s \in S$
$\nabla r^B \leftarrow \sum_{s \in S} \frac{\partial \mathcal{L}(o_s)}{\partial \theta}$
$\nabla f^B \leftarrow 0$
for all $m \in M$ **do**
  $\tilde{h}_i \leftarrow f^B(x_i)$ **for all** $i \in m$
  $\tilde{z}_i \leftarrow \tilde{h}_i + \sum_{j \in Q(i), j \neq i} h_i$ **for all** $i \in m$
  $\nabla f^B \leftarrow \nabla f^B + \sum_{i \in m} \frac{\partial \mathcal{L}(o_{Q(i)})}{\partial z_{Q(i)}} * \frac{\partial \tilde{z}_i}{\partial \phi}$
end for
return $\nabla r^B, \nabla f^B$

2.5 Results

2.5.1 Summing digits

The models trained to predict the sum of a bag of MNIST digits using the methodology described in section 2.3.1 were evaluated for the task of predicting the value of a single digit. Fig. 2.2b shows that when increasing the bag size, performance decreased slowly in terms of MAE and accuracy for bag sizes up to 16. For larger bag sizes, performance decreased more rapidly. For small bag sizes, the models generally produced output within a narrow band around the target quantity. Fig. 2.2c shows this for a bag size of 8.

2.5.2 CTSS scoring

The I3D model that was trained to produce lobe level CTSS labels from individual lobe images was evaluated for this task. It obtained a linearly weighted κ of 0.565, 95% CI: (0.520, 0.609), indicating moderate agreement, which was similar to the lin-

early weighted κ of 0.54 reported by N. Lessmann *et al* [33].

The I3Dr model had a more challenging task. It was evaluated for predicting these same lobe-level CTSS labels, but was trained using only patient level CTSS scores. It obtained a linearly weighted κ of 0.491, 95% CI: (0.448, 0.534), which indicates moderate agreement with the CTSS labels from the test set.

2.5.3 Lung function estimation

Table 2.2 shows the performance of the I3Dr model with and without the rescaling step (see section 2.3.1) at the patient level and for the different anatomical lobes, namely the left upper lobe (LUL), left lower lobe (LLL), right upper lobe (RUL), right lower lobe (RLL) and the right middle lobe (RML).

For both PV and FPV, the I3Dr model was able to produce meaningful lobe level estimates, even though it was only trained with patient level labels. The rescaling step in the I3Dr model increased performance.

Table 2.2: Model performance in Mean Absolute Error (in mL/min/mm Hg for DLCO and in L for other measures) at the patient level and for each of the anatomical pulmonary lobes: Left Upper Lobe (LUL), Left Lower Lobe (LLL), Right Upper Lobe (RUL), Right Lower Lobe (RLL), Right Middle Lobe (RML).

Model	Level	PV	FPV
I3Dr (no rescaling)	LUL	0.055	0.063
	LLL	0.059	0.055
	RUL	0.057	0.065
	RLL	0.060	0.058
	RML	0.037	0.041
	Patient	0.124	0.133
I3Dr	LUL	0.053	0.053
	LLL	0.045	0.051
	RUL	0.065	0.078
	RLL	0.046	0.054
	RML	0.032	0.037
	Patient	0.053	0.118

Table 2.3 shows the MAE of the I3D model for the task of PFT result estimation. Fig. 2.3 shows the correlation between the I3Dr output and PFT results for the COPDGene test set. In terms of the Pearson correlation coefficient, the correlation between the I3Dr output and PFT test results was substantially better for all PFT

Table 2.3: Model performance in Mean Absolute Error (in mL/min/mm Hg for DLCO and in L for other measures) at the patient level.

Model	FVC^{pre}	FVC^{post}	$FEV1^{pre}$	$FEV1^{post}$	DLCO
I3Dr (no rescaling)	0.423	0.411	0.357	0.347	3.336
I3Dr	0.388	0.377	0.307	0.297	2.800

types than the correlation between functional lung volume and PFT test results. The PFT measurements were better correlated with the rescaled I3Dr output than with the I3Dr output before the rescaling step ($p < 0.001$ for all PFT types). Fig. 2.4 shows representative qualitative PFT estimation results of the I3Dr model.

2.6 Discussion

In this paper, we introduced a method that can estimate global measures (lung function) and can also estimate the contributions of individual parts to this global measure (in this case lobar contributions to lung function). Especially for DLCO, the results are promising. Recommendations state that when measuring DLCO, there should be at least two acceptable tests that meet the repeatability requirement of either being within 3 mL CO (STPD)/min/mm Hg (or 1 mmol/min-1/kPa) of each other or within 10% of the highest value [20]. When regarding the ground truth PFT measurement and the corresponding output of the I3Dr model as the two acceptable tests, the I3Dr model meets this repeatability requirement in 64% of the cases in the test set.

In our evaluation of the I3Dr model, we found that the PFT results correlated substantially better with the output of the I3Dr model than with PFV. The latter has been used as a proxy for lobe level PFT results in previous research [70–74]. The I3Dr model may be viable for directly producing lobe level lung function estimations from CT.

The I3Dr model presented in this work has some limitations. Firstly, the I3Dr model was both trained and evaluated with CT scans and PFT results from the COPDGene study [77]. Despite the scope of the COPDGene study, encompassing data from 21 imaging centers in the United States, this training data may not be representative for patients with different pathologies and/or demographics. Consequently, this could affect the efficacy of the I3Dr model when applied to populations outside the study's demographic or those with different lung pathologies. The generalization ability of our model may be quantified in future work through external

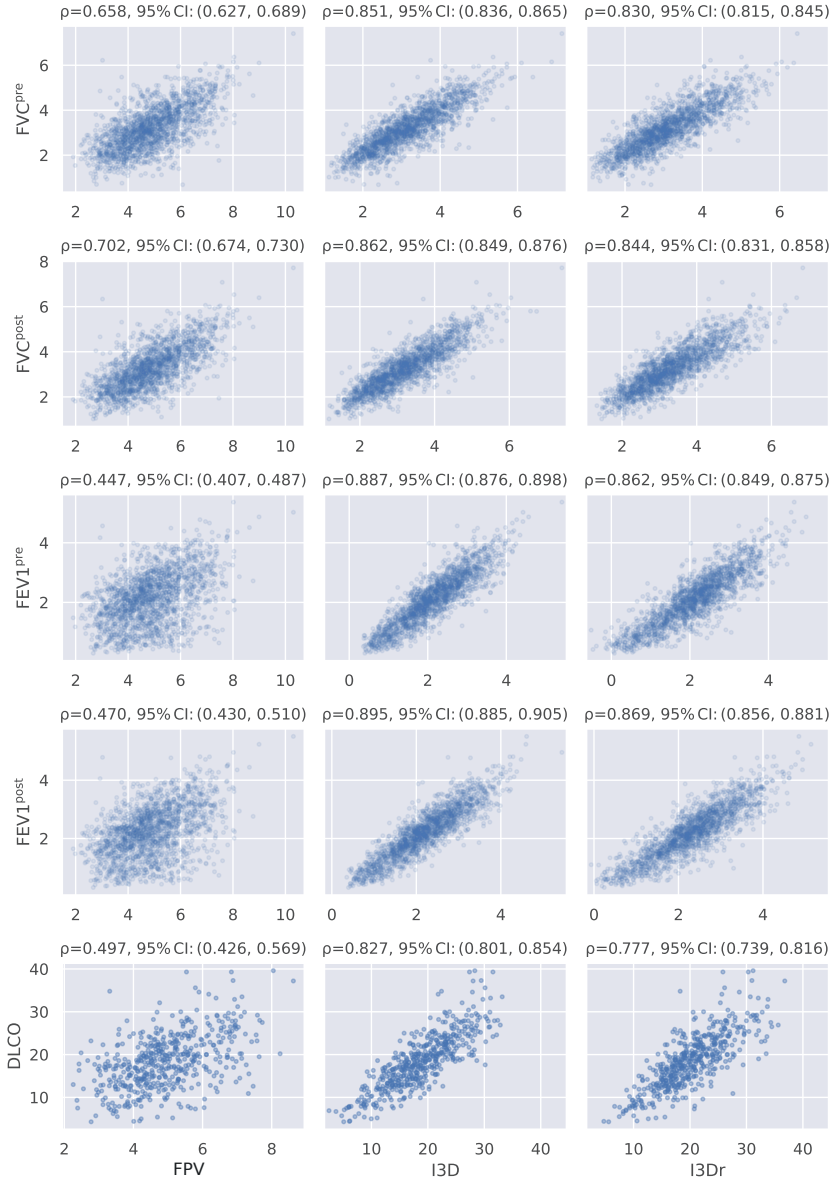


Figure 2.3: Correlation between PFT results from the test set and functional pulmonary volume (top row), the corresponding output of the I3D model (middle row), and the corresponding output of the I3Dr model (bottom row). Pearson correlation coefficients with 95% confidence intervals are shown above each plot. DLCO is shown in mL/min/mm Hg. All other measures are shown in L.

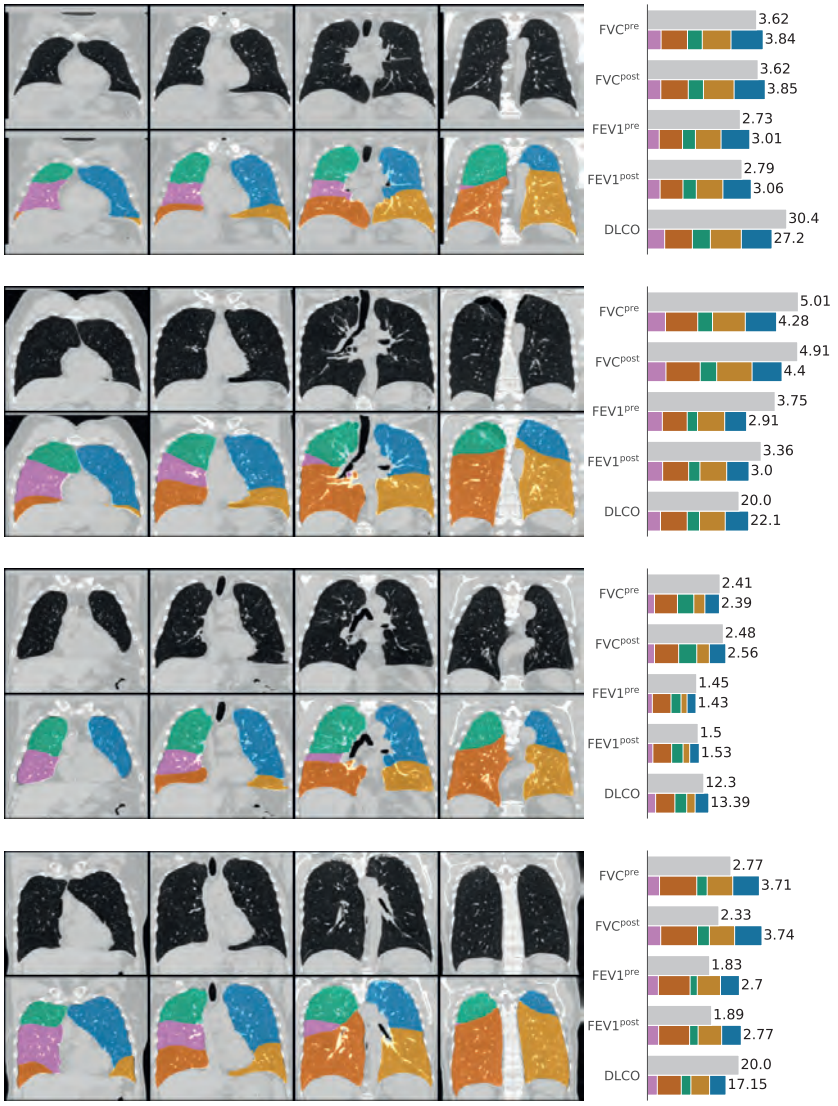


Figure 2.4: Qualitative results of the I3Dr model for four randomly selected CT scans from the test set. The left images show central, evenly spaced coronal slices of the input CT scan cropped to the lobes, as well as a voxel-wise annotation of the pulmonary lobes. The bar plots on the right show the corresponding PFT measurement values in gray and the lobar lung function estimations in the colors corresponding to the lobe segmentation. DLCO is shown in mL/min/mm Hg. All other PFT measurement values are shown in L.

validation of the publicly available algorithm.

Furthermore, a direct lobe level evaluation for spirometry and DLCO results was not possible, since these measurements can only be performed at a subject level. The I3Dr's ability to produce lobe-level CTSS scores, PV, and FPV, coupled with its performance in predicting subject level spirometry and DLCO results, suggests that it is a promising approach for estimating lobe-level lung function measures.

Lastly, we only trained the I3Dr model with inspiratory CT scans, since expiratory CT scans are not always available in clinical practice. Presenting both inspiratory and expiratory CT scans as input may increase performance for spirometry test result estimation. This might also allow the I3Dr model to indirectly capture interplay between lobes, such as the decreased inflation of one lobe due to the hyperinflation of another.

Incorporating a patient level model in the I3Dr inference pipeline as described in section 2.3.1 increased patient level performance. A reason for this increase could be that the patient level model can model interactions between lobes, which is not possible when processing the images of each anatomical lobe separately.

In our work, the I3Dr model was applied for lobe level lung function estimation. It could trivially be extended to estimate lung function per pulmonary segment when a segment segmentation is available. The I3Dr model might also be applied for determining functional measures of other organs that can be divided into separate regions with similar functionality such as the liver.

2.7 Conclusion

In this work, we conducted several experiments to validate the ability of the presented I3Dr model to produce meaningful regional labels, while being trained with only patient level labels.

Firstly, we performed a proof of concept to test whether a model could accurately predict the label of individual 2D digit images when trained with a set of images and its sum. We showed that up to sets of 32 digits, the results were nearly flawless.

Our next experiments showed that our methodology for implicitly learning regional quantities can also be applied to regression from CT scans. We trained I3Dr models to produce meaningful lobe level quantities from CT, while only using patient level labels during training. In this setting, the I3Dr model was able to estimate a lobe level CT Severity Score (CTSS), Pulmonary Volume (PV), and Functional Pulmonary Volume (FPV) from CT.

After validating that the I3Dr model could estimate meaningful lobar quantities from CT, we found that it is able to also estimate patient level PFT results. Over-

all, we found that I3Dr can estimate global measures from an image, as well as the contributions of individual parts of the image to this global measure. I3Dr offers a promising approach for estimating PFT results from CT scans and disentangling the individual contribution of pulmonary lobes to a patient's lung function.

We hope that the findings presented in this work may advance the use of CT in screening, diagnosis, and staging of restrictive pulmonary diseases as well as in risk assessment for pulmonary resection surgery and bronchoscopic lung volume reduction.

2.8 Acknowledgments

This work was supported by the European Regional Development Fund, as well as by NHLBI U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer and Sunovion.

A systematic comparison of automated COVID-19 grading algorithms

3

Authors: **Coen de Vente*, Luuk H. Boulogne*, Kiran Vaidhya Venkadesh, Cheryl Sital, Nikolas Lessmann, Colin Jacobs, Clara I. Sánchez, and Bram van Ginneken**

Original title: **Automated COVID-19 Grading With Convolutional Neural Networks in Computed Tomography Scans: A Systematic Comparison**

Published in: **IEEE Transactions on Artificial Intelligence, volume 3, number 2, pp. 129-138, 2021**

** Shared first author*

Abstract

Applied artificial intelligence (AI) research focuses disproportionately on novel architecture modifications that do not necessarily generalize to other datasets, while neglecting systematic comparisons between commonly used algorithm components. This inhibits the deployment of AI for real-world applications. For automatic COVID-19 grading specifically, attention for compatibility of AI with clinical workflow is lacking. This paper presents a systematic investigation of COVID-19 grading algorithm components using a large publicly available dataset. The results are published in an online challenge. These contributions speed up the development of AI applications for COVID-19 grading by establishing insights into the components of such applications and by allowing applications produced by future research to be compared in a fair manner. The adherence to a standardized COVID-19 grading system may increase the compatibility between AI and clinical workflow. Altogether, this work may increase the efficiency and accuracy of radiologists when reading CT scans during this pandemic.

3.1 Introduction

Imaging of COVID-19 with chest computed tomography (CT) has been found to be helpful for diagnosis of this disease in the current pandemic [81]. With the aim to reduce the workload of radiologists, various machine learning techniques have been proposed to automatically grade and classify the presence of COVID-19 in CT images [14, 23–38, 40, 82–85]. Automatic COVID-19 classification methods have already been deployed in several medical centers [24].

By far the most common technique for automatic COVID-19 classification from CT images is the Convolutional Neural Network (CNN) [86, 87], which is the current state-of-the-art for image classification [88]. The works that use this approach can be divided into those that use 2D CNNs [14, 23, 25–28, 30, 40, 83, 84] and those that use 3D CNNs [31–38, 84]. While 3D CNNs are directly capable of exploiting 3D information present in CT volumes, 2D CNNs can only indirectly use 3D information by aggregating their output for individual slices of the image to produce an image level prediction. 3D CNNs are typically more memory intensive than 2D CNNs, but Graphics Processing Units (GPUs) with sufficient memory to train 3D models are becoming increasingly available. Moreover, radiologists are specifically instructed to take 3D information into account by inspecting different orthogonal views for assessing the suspicion of COVID-19 in CT scans [15]. This indicates that 3D information is essential for radiologists in assessing the patterns indicative for COVID-19. Additionally, the slice thickness of CT scans are increasingly becoming smaller [89] so that the scans contain more detailed 3D information. We therefore hypothesize that 3D CNNs are more suitable for COVID-19 classification from CT scans than 2D CNNs.

A major issue that inhibits the utilization of artificial intelligence in real-world applications, such as COVID-19 diagnosis from CT, is the excessive focus of research on novel architectures, while scientifically sound comparisons and proper evaluations on external datasets are lacking. Often, small additions and adaptations to model architectures for incremental improvements on specific datasets are proposed that do not generalize well to other datasets. This issue is increasingly being recognized and simple baselines have been proposed which perform comparably to or better than over-engineered solutions [55, 90].

The goal of this paper is therefore not to introduce novel architectural tweaks, but instead to perform a comparative study that evaluates existing approaches. To indicate the generalization capabilities of automatic COVID-19 classification systems, some methods have been validated on data from different centers than the data that were used for training [31, 33]. Also, the same validation methods, such as receiver operating characteristic (ROC) curves and the area under the ROC curve

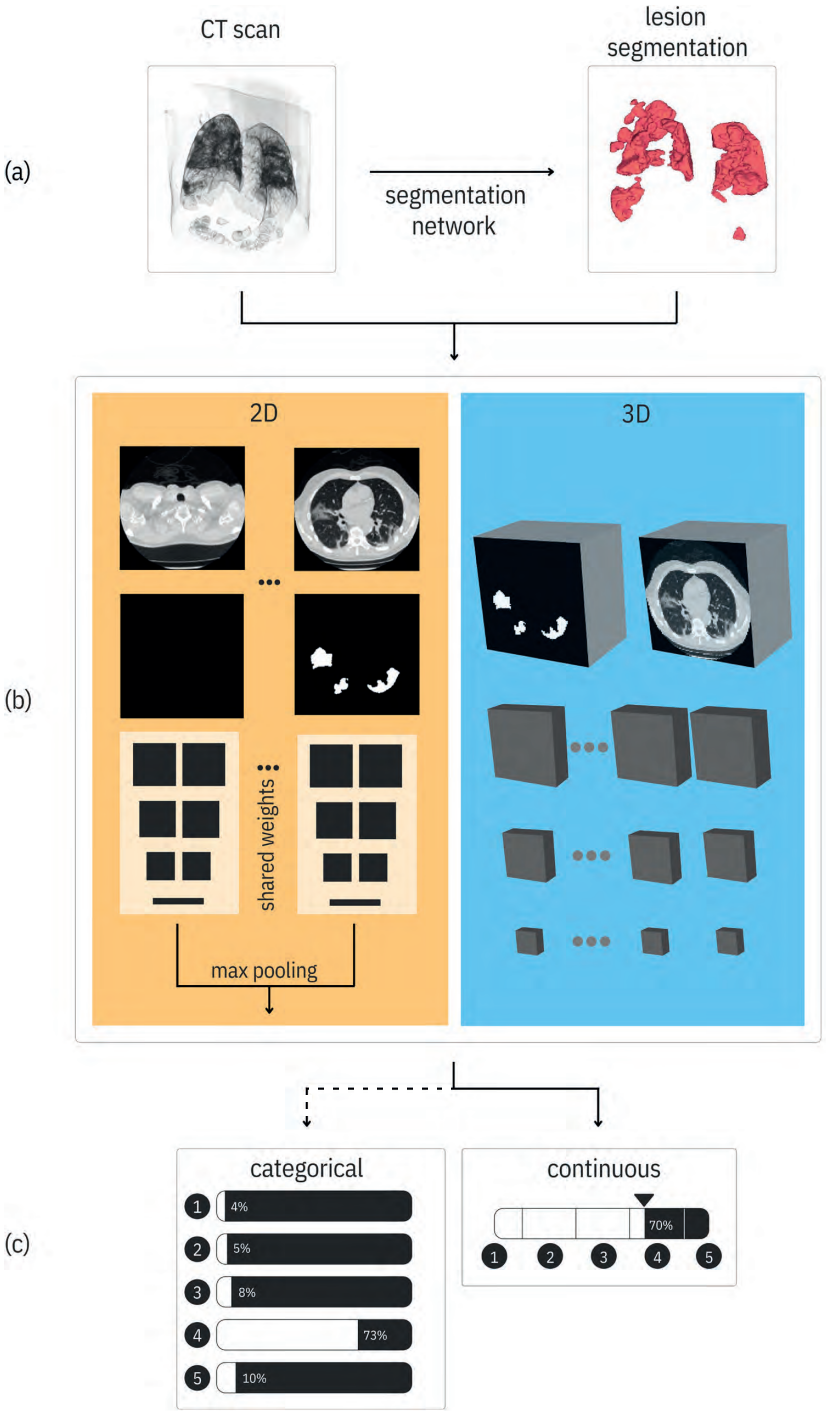

(a)
CT scan
lesion segmentation
segmentation network
(b)
2D
3D
shared weights
max pooling
(c)
categorical
1 4%
2 5%
3 8%
4 73%
5 10%
continuous
70%
1 2 3 4 5

Figure 3.1: Schematic representation of the different components used for CO-RADS grading from CT scans using convolutional neural networks in patients suspected with COVID-19. This processing pipeline was used in all experiments of this work. (a) The input CT scan is fed into a lesion segmentation network. The CT and the lesion segmentation are used as separate input channels to the classification network as described in Section 3.3.3. In one of the ablation study experiments, this lesion segmentation input was left out. (b) We compared a variety of 3D (top) and 2D architectures (bottom) as described in Section 3.3.3. The 3D architectures take as input the full volume. The 2D architectures use individual slices as input. (c) We compared a continuous output to a categorical output in the ablation study. Section 3.3.3 describes the continuous output in detail. The dashed line indicates that the categorical output replaces the continuous output in one of the models in the ablation study and all models in the architecture search, but it is not incorporated in the main approach.

(AUC), have been reported across different studies [14, 23–28, 30–36, 40, 84]. However, since each study used different datasets for training and for validation, the need for fair, direct comparisons of the performance of these algorithms remains unsatisfied. Recently, the “CT images and clinical features for COVID-19" (iCTCF) dataset was made publicly available [91], enabling a fair comparison of COVID-19 classification methods.

This paper compares a variety of 2D and 3D CNN architectures for COVID-19 classification. We trained and evaluated the approaches on the same internal dataset. Moreover, in an ablation study, we investigated performance changes due to 1) using transfer learning for 2D and 3D COVID-19 classification models, 2) using prior information in the form of COVID-19 related lesion segmentations as additional input to the network, 3) replacing the categorical output with a continuous output.

We furthermore created a public challenge [92] for evaluating and comparing different COVID-19 classification algorithms. Algorithms can be submitted to the challenge as Docker containers and are evaluated on the iCTCF dataset that we used in this paper. This allows their performance to be compared to the methods presented in this paper, as well as to other COVID-19 grading and classification algorithms that are submitted to the challenge.

3.2 Background

3D CNNs were initially proposed for processing video data [87], where the third dimension of the convolutional layers dealt with the temporal dimension. In later works, 3D CNN architectures were derived from 2D CNN architectures by expand-

ing the 2D filters into 3D [16]. Methods based on these inflated 3D CNNs, in particular the Inflated Inception-v1 (I3D) model, have recently been successfully employed for lung nodule detection and scan-level classification tasks from thorax CT scans [93, 94].

The large majority of the architectures used for COVID-19 classification from CT scans in previous works [14, 23–40] are heavily or completely based on the ResNet [95], DenseNet [96], or Inception [9] architecture families. Especially ResNet architectures have been used frequently [14, 23–28, 35–39]. Some works did not use a full ResNet architecture, but did incorporate residual blocks into their model [29, 30]. Architectures from the DenseNet [27, 31, 32] and Inception [33, 40] families have been used less frequently. Other architectures such as VGG-19 [97], Inception-ResNet-v2 [98], NASNet [99], and EfficientNet [100] have also been used in research for COVID-19 classification from CT scans [39, 41–44]. Due to the lack of standardized data for testing across different works, previous research does not identify which architecture produces the best performance for COVID-19 classification from CT.

Fine-tuning is a widely used technique in research on deep learning in medical imaging [101] and COVID-19 classification specifically [102]. With fine-tuning, models are initialized with pre-trained weights from models trained on a different task or dataset. They are commonly pre-trained on the ImageNet [103] dataset that contains a large variety of 2D natural images. Afterwards, the models are trained for the task at hand. Pre-training speeds up training and can offer performance gains for large models [101]. It has been used in several 2D CNN COVID-19 classification methods [14, 23, 26, 28, 40]. Pre-trained weights have also been used for 3D CNN-based methods. Wang *et al.* [31] pre-trained their model for COVID-19 classification on a large number of CT scans from lung cancer patients. Inflated 3D CNNs can conveniently be initialized by inflating 2D weights. 2D weights have been used to pre-train I3D models for video classification [16] and chest CT classification [93] tasks.

Before presenting CT images to the CNN, they are often pre-processed by extracting the lung region using lung or lobe segmentation algorithms. These lung regions are then used either for cropping around and centering to the lungs [23, 26, 31, 33, 37] and/or by suppressing non-lung tissue [14, 23–25, 31, 34–36, 38]. Yang *et al.* [27] used a lung segmentation as an additional input channel and used lesion masks as extra information by training their model to perform lesion segmentation and COVID-19 classification simultaneously. Lessmann *et al.* [33] also added a lesion segmentation to the input of their model.

Most studies on automated detection of COVID-19 employ a categorical classification output format that uses a softmax or sigmoid activation [102]. Previous works have trained models to discern between COVID-19 positive and negative patients

[23, 25–30, 32, 34, 37, 82], COVID-19 positive patients and patients with other types of pneumonia [31, 35, 40], and between all three [14, 36, 38]. In this work, we followed Lessmann *et al.* [33] and trained our models to produce CO-RADS [15] scores on chest CT scans of suspected COVID-19 patients. The CO-RADS score denotes the suspicion of COVID-19 on a scale from 1 to 5 and was developed to standardize reporting of CT scans of patients suspected with COVID-19 [15]. Scoring systems, like CO-RADS, have been advocated for better communication between radiologists and other healthcare providers [15, 33].

3.3 Methodology

3.3.1 Data

Training and internal test data

The internal dataset contained CT scans from consecutive patients who presented at the emergency wards of the Radboud University Medical Center, the Netherlands in March, April and May 2020 and were referred for CT imaging because of moderate to severe COVID-19 suspicion. The retrospective and anonymous collection of this data was approved by the ethical review board of Radboudumc (CMO2016-3045, Project 20027) prior to the study. Further details such as imaging parameters can be found elsewhere [33].

CO-RADS scores were reported by a radiologist as part of routine interpretation of the scans. CO-RADS 1 was used for normal or non-infectious etiologies, having a very low level of suspicion. CO-RADS 2 was used if the CT-scan was typical for other infections than COVID-19, indicating a low level of COVID-19 suspicion. CO-RADS 3 implies equivocal findings and features compatible with COVID-19, but characteristics of other diseases are also found. CO-RADS 4 and 5 indicate a high and very high level of COVID-19 suspicion, respectively.

We randomly split the dataset into a development set with 616 patients and an internal test set of 105 patients. The patients in the development set were split into 75% for training and 25% for validation using data stratification based on the CO-RADS scores. The distribution of CO-RADS scores over the different splits is displayed in Table 3.1. All data splits were made such that all scans from a patient with multiple visits ended up in the same split.

Table 3.1: Number of CT Images in Internal Dataset.

	CO-RADS							
	1	2	3	4	5	Total	Neg	Pos
Development set								
Training	253	71	78	37	73	512	324	188
Validation	81	24	26	11	23	165	105	60
Internal test set	20	10	19	17	39	105	30	75
Total	354	105	123	65	135	782	459	323

Table 3.2: Number of CT Images in External Dataset.

Grade [84]							
Control	Mild	Regular	Severe	Critically ill	Total	Neg	Pos
207	23	363	117	32	742	207	535

External test data

For external evaluation, we used the publicly available CT images and clinical features for COVID-19 dataset (iCTCF) dataset [84, 91]. Since we focused on comparing architectures for CT image processing for COVID-19 classification, we did not incorporate the clinical features from this dataset into the input for our models. In iCTCF, patients were categorized with a Chinese grading system that distinguishes the classes as Control, Mild, Regular, Severe, Critically ill and Suspected. Since there was no etiological evidence available for the presence of COVID-19 in Suspected cases [84], we did not use them for testing our models. The distribution of the other classes is displayed in Table 3.2. The grading system uses etiological laboratory confirmation and other factors such as clinical features and CT imaging [84]. The control cases include both healthy patients and patients with community acquired pneumonia. Most of the iCTCF data has been made publicly available, but some CT scans were not available at the time of conducting this study. We validated our models with all available data from the first iCTCF cohort for which etiological evidence for the presence of COVID-19 was available [91].

3.3.2 2D and 3D architectures

We compared the performance of a variety of popular 2D and 3D CNN architectures for the task of COVID-19 classification from CT. More specifically, we compared vanilla 2D and 3D versions of DenseNet-121, DenseNet-169, DenseNet-201, Inception-v1, ResNet-18, ResNet-34, and ResNet-50. Section 3.2 describes previous works that have used many of these architectures.

Since we used scan-level labels for training and testing these models, the 2D architectures required the integration of a slice-wise reduction step, while the 3D architecture did not. For the 2D architectures, we therefore integrated the slice-wise reduction step presented by Li *et al.* [14]. First, the 2D CNN extracts features of individual axial slices. A global max pooling step reduces these features to a 1D vector, to which a fully connected layer is applied with an output size equal to the number of classes.

3.3.3 Ablation study

We investigated whether additional model components had an effect on COVID-19 classification performance in an ablation study. Fig. 3.1 shows a summary of the processing pipeline that was used.

Since performing the ablation study for all 2D and 3D architectures would require a large quantity of computational resources, the ablation study was instead performed with only the best performing architecture in terms of quadratic weighted kappa (QWK).

Lesion map as prior information

To aid the model in localizing COVID-19 related parenchymal lesions, we provided a lesion segmentation map as additional input in a separate input channel. More specifically, the CT image was fed into the first input channel, the lesion segmentation into the second channel, and the third channel was presented with zeros. When training models without the additional lesion segmentation input, the CT image was fed into all three input channels.

A 3D nnU-Net [55] trained by Lessmann *et al.* [33], which segments ground-glass opacities (GGOs) and consolidations, provided the lesion segmentations. GGOs and consolidations are biomarkers with major importance in diagnosing COVID-19 [15].

Dimensionality

Since various components were added to the models in the ablation study, we trained both the 2D and 3D variants of the best performing architecture. This allows for an analysis of the performance difference solely due to the dimensionality of the model in our complete processing pipeline.

Pre-training

We investigated the performance changes due to pre-training on a natural image classification task. The 2D models were initialized with weights pre-trained on ImageNet. The 3D models were initialized with the same weights by inflating the pre-trained 2D convolution kernels to 3D.

Continuous output

The standard output format of CNNs used for categorical classification does not capture the ordinal nature of the CO-RADS scoring system. Furthermore, although the CO-RADS scoring system allows for a higher level of interpretability than a binary system, the fact that a CO-RADS suspicion score of 3 indicates that it is unclear whether COVID-19 is present makes it difficult to decide on the onset of the positive class for the predicted scores in ROC analyses. For these reasons, we considered the CO-RADS classification to be a regression task. Hence, the model had one output node that was forced to the range $(0, 1)$ using the sigmoid function. CO-RADS scores were mapped to target values in the range $[0, 1]$ with a uniform spacing between CO-RADS classes such that CO-RADS scores of 1 and 5 were assigned target values of 0 and 1, respectively. As the network had one output node, binary cross-entropy was used as loss function. With this method, unlike a standard categorical approach with a softmax layer and categorical cross-entropy loss, predictions that are further off from the target are penalized more heavily than predictions that are closer. To obtain a CO-RADS score during inference, the sigmoid output was multiplied by 4, rounded to the nearest integer and added to 1. De Vente *et al.* [104] explored this approach for prostate cancer grading and found that it outperformed other regression and categorical output methods.

3.3.4 Pre-processing

The CT scans were clipped between -1100 and 300 Hounsfield units, normalized between 0 and 1, and resampled to a voxel spacing of 1.5 mm^3 using linear interpolation. The scans were further pre-processed using a lung segmentation algorithm

that was trained on data from patients with and without COVID-19 [78]. More specifically, any slices with a distance of 10 mm or more to the lung mask were discarded and the remaining slices were cropped to 240 $\times$ 240 pixels around the center of the mask. Following previous research with I3D models [16, 93, 94], we trained our models with a fixed 3D input size. To achieve this without adding extra slices that do not contain information regarding the presence of COVID-19, we uniformly sampled 128 axial slices along the z-axis.

3.3.5 Training

We trained all networks with a batch size of 2, the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 10^{-4}. Data augmentation consisted of random zooming between -20% and +20%, rotation between -15% and +15%, shearing between -10% and +10% and elastic deformations in the axial plane, translation between -2 and +2 voxels in the z-direction, -20 and +20 voxels in both the x- and y-direction, and additive Gaussian noise with a mean of 0 and a standard deviation between between 0 and 0.01 (after intensity normalization between 0 and 1). To correct for the class imbalance, we monitored the performance on the validation data in the development set during training with balanced samples based on the distribution of CO-RADS classes in the training set. We used early stopping with a patience of 10 000 training batches and the QWK on the validation set for the stopping criterion. Gradient checkpointing [105] reduces GPU memory requirements for training deep neural networks without affecting performance. This technique was used when necessary to enable a batch size of 2 for the 2D models.

To rule out the possibility that performance differences between the 3D and 2D approach were due to other factors such as pre-processing or data augmentation, we kept all hyperparameters the same during training.

Each model was trained on a single GPU, using NVIDIA GeForce GTX TITAN X, GeForce GTX 1080, GeForce GTX 1080 Ti, GeForce RTX 2080 Ti, TITAN Xp, and A100 SXM4 cards.

3.3.6 Ensembling

The models were sensitive to the randomness of the training process introduced by initialization of weights without pre-training, sample selection, and data augmentation. In order to enable stable comparisons, we obtained ensembles by training 10 instances of the same model with different random seeds. The ensemble output was obtained by simply taking the mean of the individual model outputs. For categorical model ensembles, the output was the mean of the probability output vectors

of the individual models. All results presented in Section 3.4 were obtained from ensembles unless stated otherwise.

3.3.7 Evaluation

We evaluated the CO-RADS scoring performance using the QWK score. This measure accounts for the ordinal nature of the CO-RADS score by weighting mismatches between true and predicted labels differently based on the magnitude of the error. Following previous works on COVID-19 classification and grading [14, 23, 24, 31, 33–36, 40, 84], diagnostic performance was evaluated using the AUC and ROC curves.

We calculated 95% confidence intervals (CIs) with non-parametric bootstrapping and 1000 iterations [106]. Statistical significance was computed with the same bootstrapping method [107].

The AUCs that our models achieved on the external test set are additionally listed on the Grand Challenge platform [92] to allow for a direct comparison between our and future COVID-19 grading and classification solutions.

Inference duration was calculated on the same machine for each architecture, using a GeForce RTX 2080 Ti card. The reported durations were averaged over 50 forward passes of a batch with one sample.

3.4 Results

3.4.1 Architecture selection

Fig. 3.2 shows the QWK and AUC for the different 2D and 3D architectures. Table 3.3 shows the number of trainable parameters, single-model inference time for one sample and FLOP count for each architecture. All 2D architectures were outperformed by their 3D counterparts both in terms of QWK and AUC. The 3D DenseNet-201 architecture performed best in terms of QWK, followed by the 3D Inception-v1 architecture. In terms of AUC, the Densenet-169 obtained the best performance, again followed by the 3D Inception-v1 architecture.

In the architecture selection, on average, training of the individual 3D models required approximately 26 700 iterations, while it required about 29 800 iterations for the 2D models.

Since the QWK takes into account the ordinal nature of the CO-RADS score, this metric was used to select the architecture to execute the ablation study with. In the rest of this section, we refer to the 3D DenseNet-201 ensemble as the 3D model and to the 2D Densenet-201 ensemble as the 2D model.

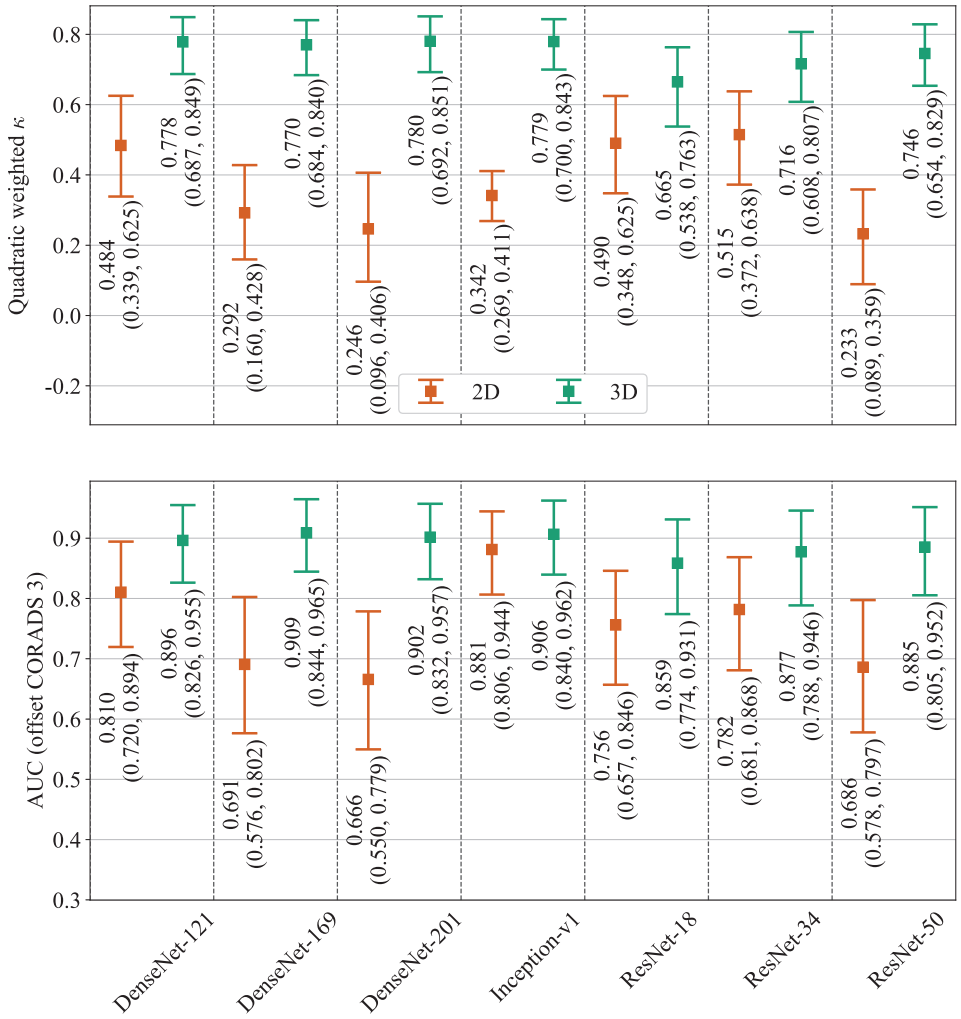


Figure 3.2: Performance of 2D and 3D CNN architectures on the internal test set for the task of CO-RADS grading from CT images is shown in QWK and AUC, respectively. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans).

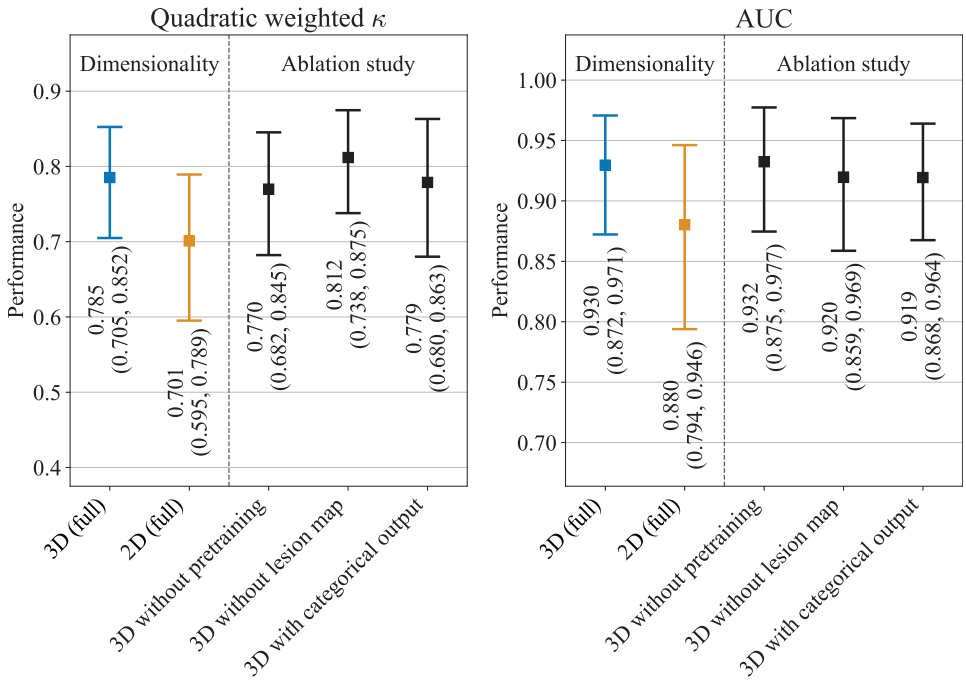


Figure 3.3: Comparison of 2D and 3D Densenet-201 models and ablation study with this architecture for the task of CO-RADS grading from CT images. The analysis was performed on the internal test set. The error bars indicate the 95% CIs. The AUC was computed with CO-RADS 1-2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans).

Table 3.3: Architecture properties

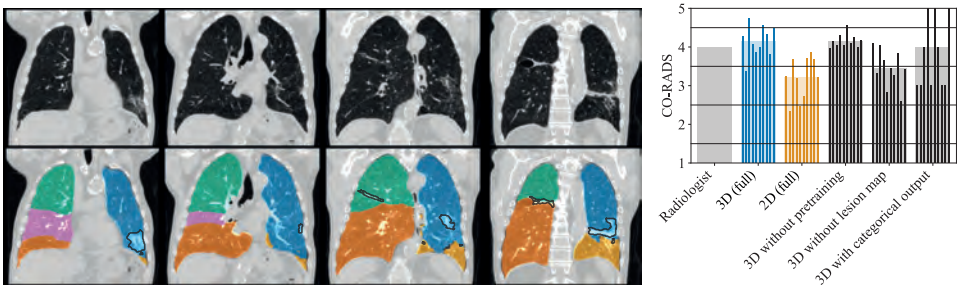
Dim.	Architecture	Parameter count ($\times 10^6$)	Inference time (ms)	FLOP count ($\times 10^{11}$)
2D	DenseNet-121	6.88	151.09 ± 8.76	8.43
	DenseNet-169	12.33	255.21 ± 17.47	9.93
	DenseNet-201	17.87	326.30 ± 3.81	12.70
	Inception-v1	5.59	40.92 ± 10.16	4.47
	ResNet-18	11.17	8.95 ± 1.25	5.52
	ResNet-34	21.27	13.53 ± 1.42	11.06
	ResNet-50	23.47	35.91 ± 10.50	12.39
3D	DenseNet-121	11.24	25.07 ± 8.49	10.88
	DenseNet-169	18.54	31.49 ± 11.48	11.30
	DenseNet-201	25.33	38.48 ± 15.65	12.14
	Inception-v1	12.29	36.74 ± 16.75	5.13
	ResNet-18	33.21	28.33 ± 31.31	6.08
	ResNet-34	63.52	22.56 ± 14.37	9.29
	ResNet-50	46.21	31.09 ± 8.49	7.39

3.4.2 2D vs. 3D CNNs

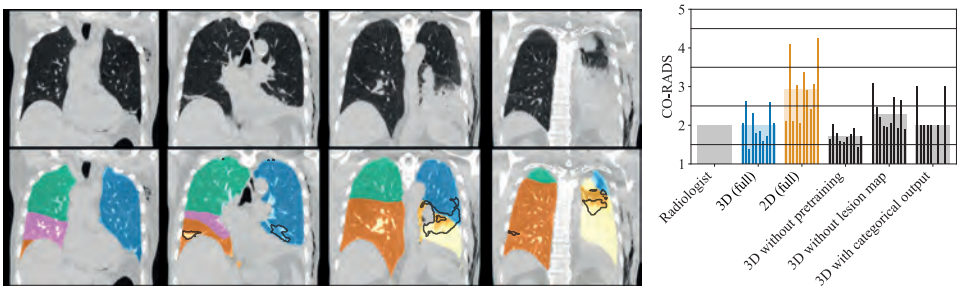
On the internal dataset, both the AUC and the QWK scores were significantly higher for the full 3D model (with transfer learning, lesion maps and continuous output) than for the full 2D model ($p = .006$ for AUC and $p = .007$ for QWK). Figures 3.3 and 3.6 show the corresponding CIs and ROC analyses respectively. Fig. 3.4 shows prediction examples from the full 3D, full 2D and ablated 3D models in blue, yellow, and black respectively.

We also trained an ensemble with the COVNet pipeline from Li *et al.* [14], which contains a ResNet-50 backbone that was pre-trained on ImageNet. With COVNet, we obtained a lower performance on the internal test set than when we applied the 3D model in our own pipeline. COVNet obtained a QWK of 0.567 (95% CI: 0.411-0.703, $p = .004$) and a lower AUC of 0.828 (95% CI: 0.741-0.906, $p = .017$) Our 2D model also outperformed COVNet in terms of both the QWK ($p = .074$) and AUC ($p = .179$).

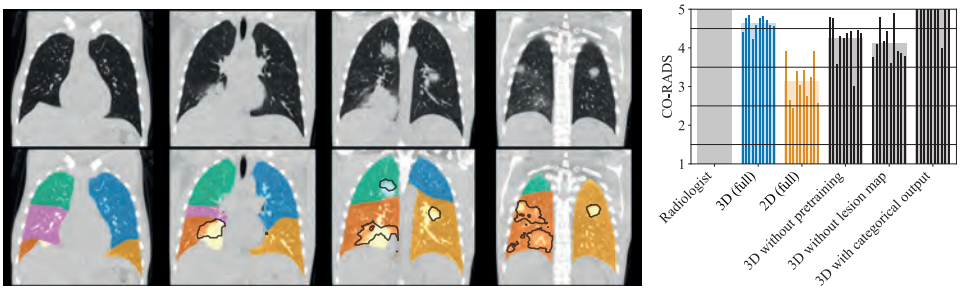
Fig. 3.5 shows confusion matrices for the two dimensionalities. For 13 scans, the full 3D approach had predictions that were more than one CO-RADS category off. For the full 2D approach this was the case for 19 scans. Furthermore, the full 3D approach and 2D approach both had two cases that were further off than 2 categories.


CO-RADS
5
4
3
2
1
Radiologist
3D (full)
2D (full)
3D without pretraining
3D without lesion map
3D with categorical output

(a)


CO-RADS
5
4
3
2
1
Radiologist
3D (full)
2D (full)
3D without pretraining
3D without lesion map
3D with categorical output

(b)


CO-RADS
5
4
3
2
1
Radiologist
3D (full)
2D (full)
3D without pretraining
3D without lesion map
3D with categorical output

(c)

Figure 3.4: Example input-output pairs for the task of CO-RADS grading on the internal test set for the trained DenseNet-201 ensembles. Input examples are shown on the left. Top row: Coronal slices of an input CT scan. Bottom row: Lung segmentation used for centering and cropping are displayed with colored overlays. Delineations of the lesion masks that were used as a separate input channel are depicted as black lines. Output examples of the ensembles (wide, light bars) as well as the individual models these ensembles are composed of (narrow, dark bars) are shown on the right. (a) Radiology report: "GGO and consolidations especially lower lobes and posterior. Has had prior lung carcinoma. COVID-19 is probable, but other infection intrapulmonal is also possible." (b) Radiology report: "COVID-19 not probable, but also not ruled out. Known post-traumatic thorax, persistent pleura fluid, slice pneumothorax. Small amount of GGO and consolidation (left). Some pneumonia at thorax trauma, post-traumatic deviations." (c) Consolidation and GGO in all lobes. According to radiologist: "Very suggestive for COVID. Also positive PCR. Proven comorbidity."

3.4.3 Ablation study

The results of an ablation study to investigate the effect of each of the additional components added to the 3D CNN are shown in Fig. 3.3. The 3D model without ablations obtained an AUC of 0.930 (95% CI: 0.872-0.971) and a QWK of 0.785 (95% CI: 0.705-0.852). Removing any of the additions had a smaller effect on these perfor-

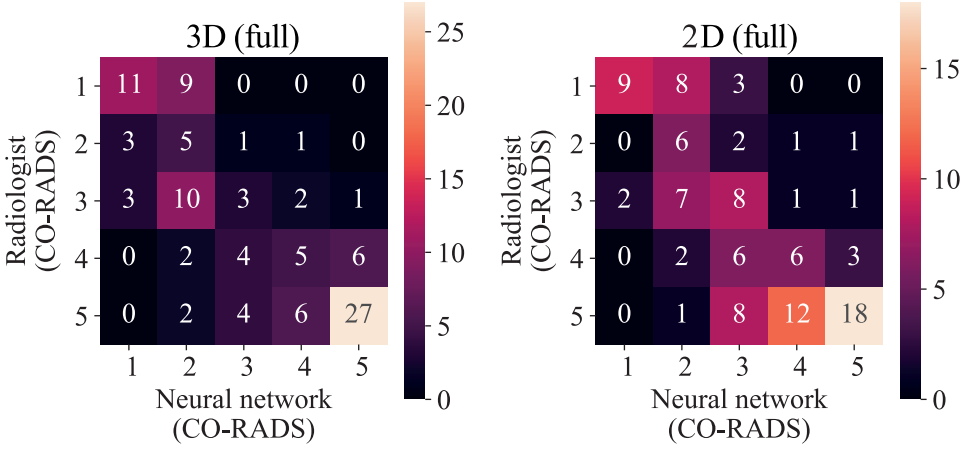


Figure 3.5: Confusion matrices for CO-RADS grading of the 2D and 3D DenseNet-201 model predictions on the internal test set. These models were trained with transfer learning, lesion maps and produced continuous output. The true label reference is from the radiology report. Cells contain the number of CT scans.

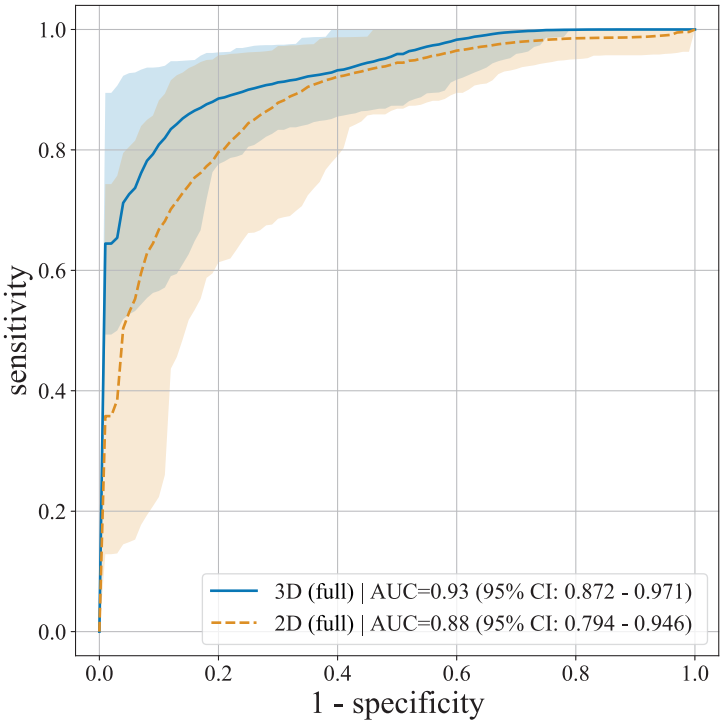


Figure 3.6: ROC analysis for the 2D and 3D Densenet-201 models on the internal test set from Radboudumc (105 CT scans) for the task of CO-RADS grading. The analysis was performed with CO-RADS 1 and 2 as the negative class (30 scans) and CO-RADS 3-5 as the positive class (75 scans). It was performed for the full 2D and 3D models trained with transfer learning, lesion maps and continuous output.

mance metrics than changing the dimensionality of the architecture to 2D. Removing pre-training reduced the QWK to 0.770 (95% CI: 0.682-0.789, $p = .278$), but increased the AUC to 0.932 (95% CI: 0.857-0.977, $p = .428$). When the lesion segmentation input was removed from the model, the QWK was increased to 0.812 (95% CI: 0.738-0.875, $p = .091$) and the AUC was reduced to 0.920 (95% CI: 0.859-0.969, $p = .292$). Replacing the regression approach with a categorical target had a negative effect on both metrics, reducing the QWK to 0.799 (95% CI: 0.680-0.863, $p = .421$) and the AUC to 0.919 (95% CI: 0.868-0.964, $p = .324$). Fig. 3.4 shows prediction examples from the ablation study models in black.

The 3D model required 31 550 iterations for training on average. The 2D model, the network without pre-training, and the model without categorical output all required less iterations (25 650, 31 000 and 22 450, respectively). The model without

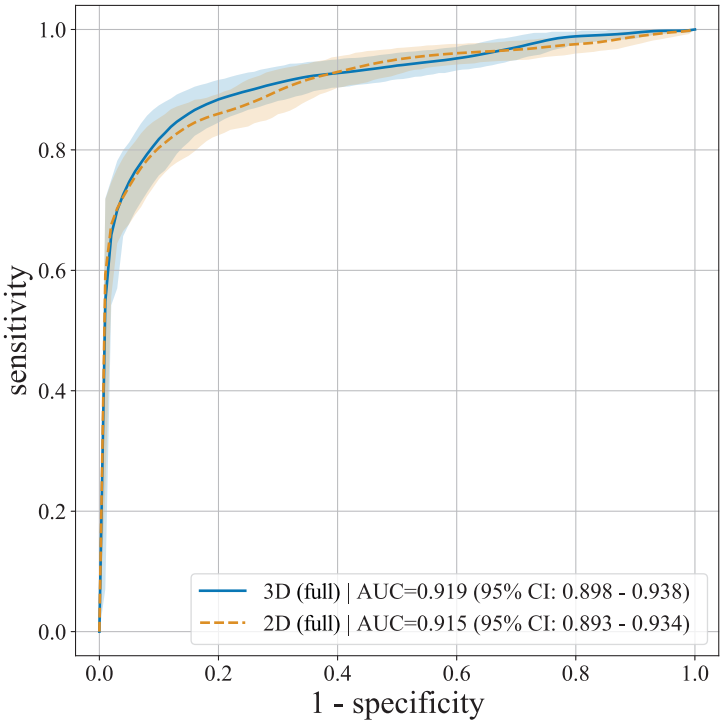


Figure 3.7: ROC analysis for the 2D and 3D Densenet-201 models on the external iCTCF test set (742 CT scans) for the task of COVID-19 classification. The analysis was performed with 207 COVID-19 negative (Control) cases and 535 positive (Mild, Regular, Severe, Critically ill) cases.

lesion input required more iterations (32 750).

3.4.4 External evaluation

Fig. 3.7 shows the ROC curves of the full 3D and the full 2D model for the external iCTCF test set.

The 3D approach obtained an AUC of 0.919 (95% CI: 0.898-0.938) and outperformed the 2D approach that obtained an AUC of 0.915 (95% CI: 0.893-0.934, $p = .215$).

3.4.5 Lesion segmentation model

For a single patch the lesion segmentation model inference time was 178.66 ms $\pm$ 14.56 ms, using 9.41×10^{11} FLOPs. The CT scans in the test set contained 12.8 patches on average. The model had 29.69×10^{6} parameters. Performance metrics for this model were reported by Lessmann *et al.* [33].

3.5 Discussion

In this paper, we identified and tested components of CNN based automated COVID-19 grading models. More specifically, we investigated how the performance of such models is affected by using different 2D and 3D CNN architectures, adopting pre-trained weights, using automatically computed lesion maps as additional network input, and predicting a continuous output instead of a categorical output. We evaluated all models with the same datasets to allow for a fair comparison between models.

Based on the architectures used in earlier automated COVID-19 classification research, we selected and compared the performance of the 2D and 3D variants of 7 CNN architectures for this task. We found that for all architecture types, the 2D models were outperformed by their 3D counterparts. The best performing model was a 3D DenseNet-201. In the rest of this section, we refer to the 3D DenseNet-201 as the 3D model and to the 2D Densenet-201 as the 2D model.

The full 3D model (with transfer learning, lesion maps and continuous output) outperformed the full 2D model in terms of AUC and QWK score on the internal test set for COVID-19 classification and CO-RADS grading.

We compared our 2D model with COVNet, an architecture previously used in a similar COVID-19 classification task in CT [14], for which the authors reported an AUC of 0.96 for differentiating between COVID-19 positive and negative patients. The substantial difference between this result and our observations with COVNet illustrates the importance of using the same dataset when comparing different approaches.

We also observed a better diagnostic performance for COVID-19 classification by the 3D model on the external test set, although this performance increase was not statistically significant for a significance level of 0.05. AUC was 0.919 for the full 3D model, while it was 0.915 for the full 2D model. Ning *et al.* [84] developed a 2D model with slice-level annotations indicating if the slice was COVID-19 positive, negative or non-informative. Using a superset of the external set used in this paper for evaluation an AUC of 0.919 was obtained, which is the same as the AUC of our

3D model, even though our 3D model was trained with weaker labels and on data from a different population. This further emphasizes the importance of using 3D rather than 2D models.

The internal test set was comprised of data from the same population as the data the model was trained on, while the external test set was comprised of data from a different population. For the full 2D model, a lower AUC was obtained on the internal test set than on the external test set. This difference might be due to population differences between the internal and external test set, or due to the different definitions of the positive class, which were presence of COVID-19 and high suspicion of COVID-19 for the internal and external test sets respectively.

On the external test set, the full 3D model outperformed the full 2D model by a smaller margin in terms of AUC than on the internal dataset. This difference could be partly due to the different definitions of the positive class. However, we also found that it partly arises from the larger overall slice thickness in the external test set. All scans in the internal test set had a slice thickness of 0.5 mm. In contrast, 207 scans (40 COVID-19 positive, 167 negative scans) in the external test set had a slice thickness larger than 1.5 mm, which was the input resolution in our training and testing pipeline. When evaluating only on these scans, we obtained an AUC of 0.885 (95% CI: 0.835-0.931) for the full 3D model and an AUC of 0.891 (95% CI: 0.843-0.932) for the full 2D model. The external test set contained 535 scans (167 COVID-19 positive, 368 negative) with a slice thickness smaller than or equal to 1.5 mm. On these scans we obtained an AUC of 0.926 (95% CI: 0.902-0.947) for the full 3D model and an AUC of 0.918 (95% CI: 0.892-0.941) for the full 2D model. The performance of both models is lower for scans with a large slice thickness, but this effect is more apparent for the 3D model. Taking into account the increasingly smaller slice thickness of CT scans [89], this observation further supports our hypothesis that 3D models are better suited for COVID-19 grading applications than 2D models.

A possible explanation for why adding the extra dimension to the convolutions improves the performance is that it allows the CNN to take into account the 3D structure and full volume of individual lesions. This explanation is in line with the fact that radiologists typically use both the axial and coronal views to visualize the spread of COVID-19 related lesions across the lungs in CT scans, such as GGOs [15].

We could not directly compare the CO-RADS classification performance on the external set, since CO-RADS labels were not available. Moreover, the CO-RADS grading cannot be directly translated to the system used in the iCTCF dataset, since the former measures the probability of COVID-19 presence, while the latter quantifies the severity of the disease.

The ablation study on the internal test set showed that the further additions to

the network and training procedure did not have a significant effect on the performance. Regardless of performance increases, using a continuous output removes the disadvantage of having to decide on the onset of the positive class for the predicted CO-RADS scores. Adding lesion maps as input and using inflated ImageNet weights for pre-training might both be ineffective for 3D automated CNN based COVID-19 grading methods.

The full 2D DenseNet-201 model managed to obtain a better performance than the 2D DenseNet-201 model without pre-training, additional lesion map input, and continuous output. This indicates that some of these additional components positively affected the performance of the 2D model. However, even with all additional components, it was still outperformed by the vanilla 3D DenseNet-201.

We did not use clinical features available for the external dataset as input to the models trained in this work, since the main goal of this paper was to demonstrate the effect on performance of different COVID-19 grading and classification algorithm components.

3.6 Conclusion and Future Work

We compared a variety of 2D and 3D Convolutional Neural Network (CNN) architectures for COVID-19 classification from computed tomography scans and found that for all architectures considered, the 3D variants outperformed their 2D counterparts. We investigated how the performances of the best performing architecture and its 2D counterpart were affected by including COVID-19 related lesion segmentations as additional input, using pre-trained weights, and replacing the categorical output with a scalar continuous output.

We intentionally did not develop novel non-trivial architectural tweaks for small performance improvements, as many of them have been shown to be unnecessary and to not generalize well to other datasets and tasks [55, 90]. We leave systematic comparisons that explore other transfer learning schemes, make use of slice-level annotations, and use clinical features as model input for future work.

Radiologists can be aided in assessing CT scans on the presence of COVID-19 by automatic COVID-19 grading systems. This paper advances and speeds up the development of such systems in the following ways. Firstly, our findings aid in advancing the performance of automated COVID-19 grading systems and provide insight into the performance benefits of several of their components. These insights primarily indicate that future research and clinical applications should move towards using 3D CNNs for COVID-19 grading in CT scans. Secondly, the models and the automatic evaluation method used in this paper have been made available on the on-

line Grand Challenge platform [92]. This allows researchers to obtain and compare the performance of their COVID-19 grading and classification solutions to other solutions on the platform. Thirdly, the output of all models used in this paper adheres to the standardized CO-RADS reporting system to facilitate easier integration into clinical workflow.

Reusable methods for training COVID-19 classifiers

4

Authors: **Luuk H. Boulogne, Julian Lorenz, Daniel Kienzle, Robin Schön, Katja Ludwig, Rainer Lienhart, Simon Jégou, Guang Li, Cong Chen, Qi Wang, Derik Shi, Mayug Maniparambil, Dominik Müller, Silvan Mertes, Niklas Schröter, Fabio Hellmann, Miriam Elia, Ine Dirks, Matías Nicolás Bossa, Abel Díaz Berenguer, Tanmoy Mukherjee, Jef Vandemeulebroucke, Hichem Sahli, Nikos Deligiannis, Panagiotis Gonidakis, Ngoc Dung Huynh, Imran Razzak, Reda Bouadjenek, Mario Verdicchio, Pasquale Borrelli, Marco Aiello, James A. Meakin, Alexander Lemm, Christoph Russ, Razvan Ionasec, Nikos Paragios, Bram van Ginneken, Marie-Pierre Revel**

Original title: **The STOIC2021 COVID-19 AI challenge: applying reusable training methodologies to private data**

Published in: **Medical Image Analysis, volume 97, p. 103230, October 2024**

Abstract

Challenges drive the state-of-the-art of automated medical image analysis. The quantity of public training data that they provide can limit the performance of their solutions. Public access to the training methodology for these solutions remains absent. This study implements the Type Three (T3) challenge format, which allows for training solutions on private data and guarantees reusable training methodologies. With T3, challenge organizers train a codebase provided by the participants on sequestered training data. T3 was implemented in the STOIC2021 challenge, with the goal of predicting from a computed tomography (CT) scan whether subjects had a severe COVID-19 infection, defined as intubation or death within one month. STOIC2021 consisted of a Qualification phase, where participants developed challenge solutions using 2 000 publicly available CT scans, and a Final phase, where participants submitted their training methodologies with which solutions were trained on CT scans of 9 724 subjects. The organizers successfully trained six of the eight Final phase submissions. The submitted codebases for training and running inference were released publicly. The winning solution obtained an area under the receiver operating characteristic curve for discerning between severe and non-severe COVID-19 of 0.815. The Final phase solutions of all finalists improved upon their Qualification phase solutions.

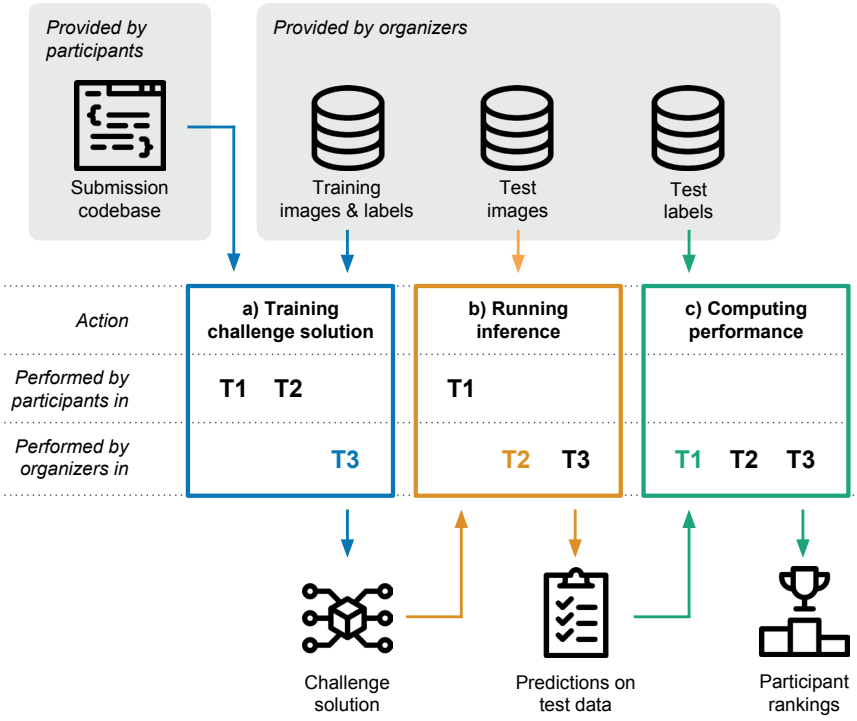


Figure 4.1: Schematic representation of the submission pipeline of challenges of Type One (T1), Type Two (T2), and Type Three (T3). a) A challenge solution is trained by applying a participants' codebase to images and labels provided by the challenge organizers. With T1 and T2, participants perform this step. With T3, the challenge organizers perform training. b) The solution is applied to test images, producing predictions. The introduction of the T2 format allowed challenge organizers to perform this step. c) The resulting predictions are compared with test labels to compute the submission's performance. Participants are ranked based on their performance. In all challenge types, the performance is computed by the organizers.

4.1 Introduction

Grand challenges for medical image analysis aim to provide the best solutions to clinical problems that the field of artificial intelligence has to offer. The sensitive nature of medical images can limit the quantity of data for model development that challenge organizers release publicly, which can in turn limit the performance of challenge solutions. Although some recent challenges ensured that the winning solu-

tions were readily available after the challenge had completed, [6, 108–111] reusability of the methods with which these solutions were trained was not enforced.

This work implements a challenge format that allows for training submissions on private data. This ensures that the winning solutions can easily be retrained on new datasets after the challenge has concluded. We aim to demonstrate the effectiveness of this challenge format in the STOIC2021 challenge, available at `https://stoic2021.grand-challenge.org`.

CT scans of COVID-19 patients can be used in the diagnostic process, as they can show clear indicators of the disease, including ground-grass opacities, typically distributed bilaterally, with or without consolidations [15]. Automatic algorithms that analyze CT scans of COVID-19 patients have the potential to aid healthcare professionals in the diagnostic process [112]. The focus of STOIC2021 was to produce fully automatic methods for discriminating between severe and non-severe COVID-19 subjects, with severe COVID-19 defined as death or intubation after one month. The challenge was organized with data from the STOIC project, [48] a multi-center dataset that comprises CT scans of 10 735 subjects. The STOIC project protocol can be accessed via ClinicalTrials.gov with identifier NCT04355507.

Through STOIC2021, this study provides the public release of CT scans of 2 000 subjects suspected for COVID-19, along with RT-PCR results, disease severity at one month follow-up, age, and sex labels under a CC-BY-NC 4.0 licence.

The submission pipeline of a challenge generally consists of training a challenge solution, running inference with it on a test set, and using the resulting predictions to compute the submission's performance. In this work, we define different challenge types by considering which steps are performed by challenge participants, and which steps are performed by challenge organizers. Figure 4.1 describes the challenge submission pipeline, previously used challenge formats that are referred to in this work as Type One (T1) and Type Two (T2), as well as the Type Three (T3) challenge format.

In T1 challenges, [54, 111, 113–149] participants perform inference on a publicly released test set themselves, which does not preclude them from meddling with their predictions, compromising the integrity of their submission's performance. T2 challenges [6, 108–110, 150, 151] solve this issue by requiring participants to submit functional algorithms. These can be made easily accessible to third parties [6, 108–110], and generate reproducible results [6, 150].

We implement the Type Three (T3) challenge structure, which has only seen limited use in medical image analysis research [152]. With T3, participants do not submit an algorithm for inference, but they instead submit an uncompiled codebase for training and inference. The challenge organizers apply the codebase to the training

set, generating the corresponding challenge solution. This allows for training on a combination of public and sensitive private training data. It guarantees that not only inference methods, but also training methods work out-of-the-box for third parties.

4.2 Materials and methods

4.2.1 Materials

Data from the STOIC study [48] was used to construct the database used for the STOIC2021 challenge. For each subject in the database, the initial CT examination, performed at presentation, was selected. The subjects were represented by one thoracic CT scan when available, or otherwise by one CT scan that imaged more of the body. Slices more than 80 mm above and 110 mm below the lungs were discarded based on corrected lung masks produced by RTSU-Net [78], as they were considered outside the typical scope of a thoracic CT scan. For all subjects, sex and age labels, binned into ten year ranges, were provided as optional additional model input. RT-PCR results, and outcome, defined as death or intubation at one month, were used as ground truth for COVID-19 infection and severity respectively. Figure 4.2a depicts how the preprocessed database was split into training and evaluation sets for the Qualification and Final phases of STOIC2021.

4.2.2 Performance metric

Performance on all leaderboards was measured in terms of Area Under the receiver operating characteristics Curve (AUC) to reflect class imbalance [153]. Participants were ranked based on AUC for classifying COVID-19 severity, computed over cases with a positive COVID-19 RT-PCR result. AUC for COVID-19 presence, computed over all cases, was used solely as additional feedback for participants and did not directly influence ranking. Submissions with missing results on any of the test cases were regarded as invalid.

4.2.3 Study design

STOIC2021 was organized on the grand-challenge.org platform. It consisted of a Qualification phase followed by a Final phase as shown in Figure 4.2. These phases respectively followed the T2 and T3 format illustrated in Figure 4.1. Anyone with a verified, authentic user account on grand-challenge.org platform could join the challenge. Participants had the option to collaborate by forming non-overlapping teams.

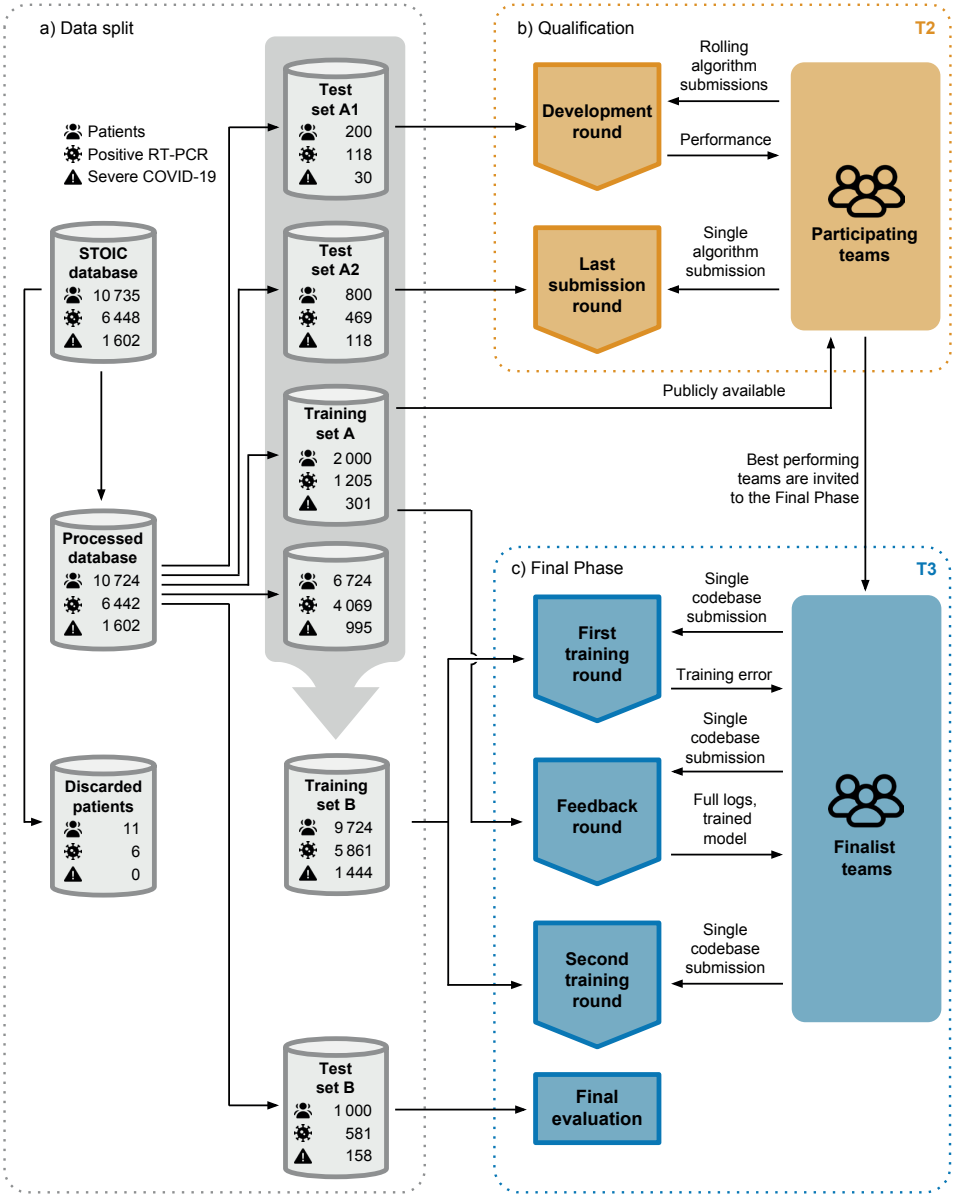

a) Data split
Patients
Positive RT-PCR
Severe COVID-19
STOIC database
10 735
6 448
1 602
Processed database
10 724
6 442
1 602
Discarded patients
11
6
0
Test set A1
200
118
30
Test set A2
800
469
118
Training set A
2 000
1 205
301
6 724
4 069
995
Training set B
9 724
5 861
1 444
Test set B
1 000
581
158
b) Qualification
T2
Development round
Rolling algorithm submissions
Performance
Last submission round
Single algorithm submission
Participating teams
Publicly available
Best performing teams are invited to the Final Phase
c) Final Phase
T3
First training round
Single codebase submission
Training error
Feedback round
Single codebase submission
Full logs, trained model
Second training round
Single codebase submission
Final evaluation
Finalist teams

Figure 4.2: Schematic overview of the STOIC2021 challenge. Each patient was represented by a single CT scan. a) Schematic overview showing how many CT scans were used for what purpose, how many of them showed patients with a positive RT-PCR result, and how many of those patients suffered from severe COVID-19. The CT scans in the STOIC database were discarded when severe motion artifacts that affected the entire scan were present, and preprocessed otherwise. From this processed database, training set A, and test sets A1, A2 and B were randomly sampled without replacement. Training set A, and test sets A1 and A2 were used in the Qualification phase. All processed data not present in test set B, including the 6 724 CT scans not used in the Qualification phase, were used to form training set B. Training set B and test set B were used in the Final phase. All data except for the public training set A was kept secure on the grand-challenge.org platform at all times and could not be downloaded by participants at any point. The large sizes of test sets A2 and B were chosen to obtain accurate performance measures despite the class imbalance. Test set A1 was deliberately chosen to be smaller to lower the challenge organization costs of rolling submissions. b) In the T2 Qualification phase, participating teams trained challenge solutions on training set A and submitted them in a rolling fashion. They could view their performance on test set A1 through a public leaderboard. At the end of the Qualification phase a single submission for evaluation on test set A2 determined which teams were invited to join the Final phase. c) The T3 Final phase started with a first training round in which participants made a single codebase submission. The challenge organizers applied these codebases to training set B. The submitting teams received any training errors that their codebase generated. Subsequently, the finalist teams could make a Feedback codebase submission to resolve these errors. This codebase was applied to public training set A so that each finalist could inspect all results of their Feedback run. Lastly, finalists could submit their revised codebases to training set B, forfeiting their first training round submission. The models trained in the Final phase on training set B were evaluated on test set B.

Qualification phase

During the Qualification phase, participating teams submitted solutions in the form of containerized algorithms trained on the publicly available training set A (see Figure 4.2a), which was publicly released on December 6th, 2021.

Rolling submissions On December 23rd, a submission tutorial accompanied by a baseline system was released and rolling submissions were opened. The rolling submissions were evaluated on test set A1 (see Figure 4.2a). This tutorial and source code is available on `https://github.com/luukboulogne/stoic2021-baseline`. Test set A1 consisted of only 200 subjects to limit the computational costs of the rolling submissions. Teams could view their performance on a public leaderboard. A count-down time between submissions of seven days was enforced. Violating this rule resulted in a submission time-out with a duration equal to the ignored count-down time.

Last submission Teams submitted to test set A2 to qualify for the Final phase. To prevent the performance on the corresponding leaderboard to be tainted by overfitting, there existed no overlap between test set A1 and A2, and each team could submit their solution to be evaluated on test set A2 only once. Participants had a total of four months for developing their solutions. Submissions to both test set A1 and A2 were closed on April 13th, 2022.

Final phase

The finalists were the 10 best performing teams that accepted an invitation to the Final Phase. Of these teams, the teams that ranked 1st, 2nd, 4th to 8th, and 14th in the Qualification phase submitted code bases for performing training and inference with their solution. A codebase for training and performing inference with the baseline system along with submission instructions for the Final phase was released on February 23rd, 2022. This tutorial and source code is available on `https://github.com/luukboulogne/stoic2021-baseline-finalphase`. These instructions ensured that the winning solutions could be used out-of-the-box by the challenge organizers and by third parties after the challenge had completed.

The Final phase initially consisted of a single round in which the challenge organizers used the finalists' training code bases to train solutions. Since not all submissions completed training successfully during this first training round, the Final phase was extended with a feedback round and a second training round.

Participating teams' members qualified as author when submitting a codebase for training their solution to the Final Phase. Participating teams could publish their own results separately, without embargo.

Training environment The training environment for the Final phase was drafted on March 17th based on resource requests and discussion with the Qualification phase participants, and was finalized on April 29th. Final phase training was performed on an Amazon EC2 p3dn.24xlarge instance. Each submission was allowed training for a maximum of 120 hours with access to two Tesla V100 GPUs with 32 GB vRAM each, 16 cpus with a total of 128G RAM, and 2 000 GB of Elastic Block Storage for storing intermediate results such as preprocessed data.

First training round Finalists could submit a single code base for training and inference with their solution in the form of a GitHub repository until May 12th. The challenge organizers generated training algorithms in the form of Docker [154] container images from the submitted code bases and applied these to training set B (see Figure 4.2). Each finalist obtained any error messages that their training algorithm generated in the first training round. These error messages were first scrutinized by the challenge organizers to ensure no leakage of sensitive information from training set B and to confirm the absence of indications of model performance.

Feedback round To acquire additional feedback about running their code base in the training environment, finalists could submit any code base before July 17th following the final submission guidelines. These codebases were applied to the training environment and participants received the complete training logs and the resulting trained model. For the Feedback round only, two modifications were made to the training environment. Firstly, to ensure that training set B was kept secure, training set B was swapped out for the public training set A. Secondly, run time was limited to 24 hours to keep down computational costs.

Second training round Finalists were given the opportunity to make a second submission to the Final phase until July 27th. They could update their codebases to make their resulting training and inference containers run and complete successfully. For this update, methodological changes with respect to the first training round submission were not allowed. The codebases were checked for adherence to this rule by the challenge organizers and no violations were found. Finalists that chose to submit to the second training round were required to renounce their first training round submission.

Prizes

Prizes in Amazon Web Services (AWS) credits were awarded to the best performing teams of the Final phase with values of $10 000, $6 000, and $4 000 for 1st, 2nd, and 3rd place respectively. The winners were announced during a public webinar on October 18th, 2022.

Future submissions

After STOIC2021 had concluded, rolling submissions to test set A1 were re-opened. Submissions to the leaderboard corresponding to test set A2 have been made available for submission upon request to the challenge organizers.

4.2.4 Statistical tests

The DeLong [155, 156] test is widely used for comparing AUCs and was also adopted for the statistical analysis in this work. 95% confidence intervals were computed as the interval between the 2.5% and 97.5% percentiles of a bootstrap distribution generated with 1 000 iterations [157].

4.2.5 Baseline method

The baseline for STOIC2021 implemented a simple training and evaluation pipeline for an Inflated 3D convnet (I3D) [16].

Preprocessing strategy The input CT scans were resampled to an isotropic spacing of 1.6 mm3. A center crop of 240*240*240 voxels was extracted from the CT, using zero padding when necessary. The voxel values were clipped between -1100 and 300 HU and rescaled to the range [0,1].

Training strategy A single I3D model [16], initialized with publicly available weights trained for RGB video classification, was trained to estimate both COVID-19 presence and severity. The model was trained on all training data for 40 epochs using the AdamW optimizer [158] with a learning rate of 10, momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 0.01. Data augmentation was employed in the form of zoom, rotation, translation, and adding gaussian noise. Patient age and sex information were not incorporated as input to the model.

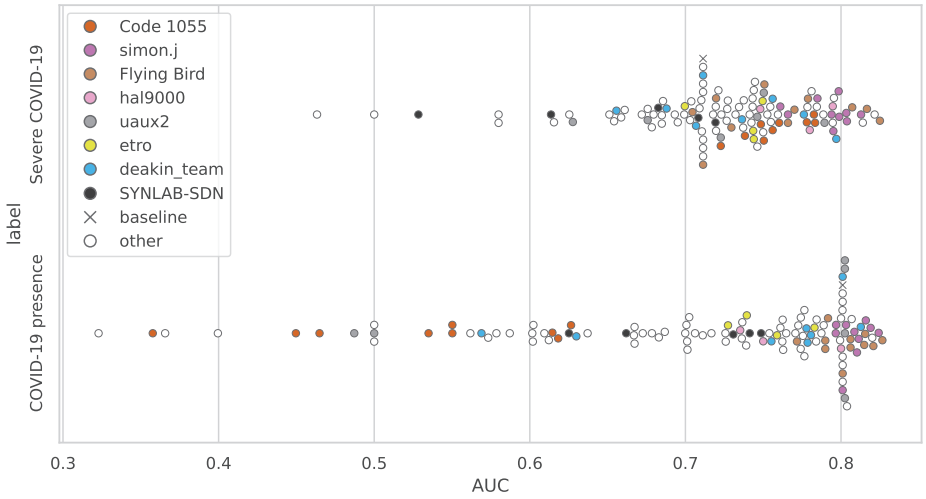


Figure 4.3: Performance distribution of the rolling submissions to test set A1 during the Qualification phase. The performance of the baseline is represented by an 'x'. Submissions by the eight finalist teams are represented by colored circles. All other submissions are represented by white circles. Details about the metrics used are described in section 4.2.2.

Table 4.1: Performance on test set B. Solutions trained on training set A and B respectively are printed in regular and bold text. The top three ensemble was obtained by averaging the predictions of the best performing solutions, the AUCs of which are marked with '*'. Details about the metrics used are described in section 4.2.2.

Team name	AUC severe COVID-19	AUC COVID-19 presence
Top three ensemble	**0.817**	**0.849**
Code 1055	**0.815***	**0.616**
simon.j	**0.810***	**0.845***
Flying Bird	**0.794***	**0.838***
hal9000	**0.788**	**0.829***
uaux2	**0.787**	**0.825**
baseline	**0.775**	**0.818**
etro	**0.763**	**0.677**
deakin_team	0.741	0.820
SYNLAB-SDN	0.722	0.789

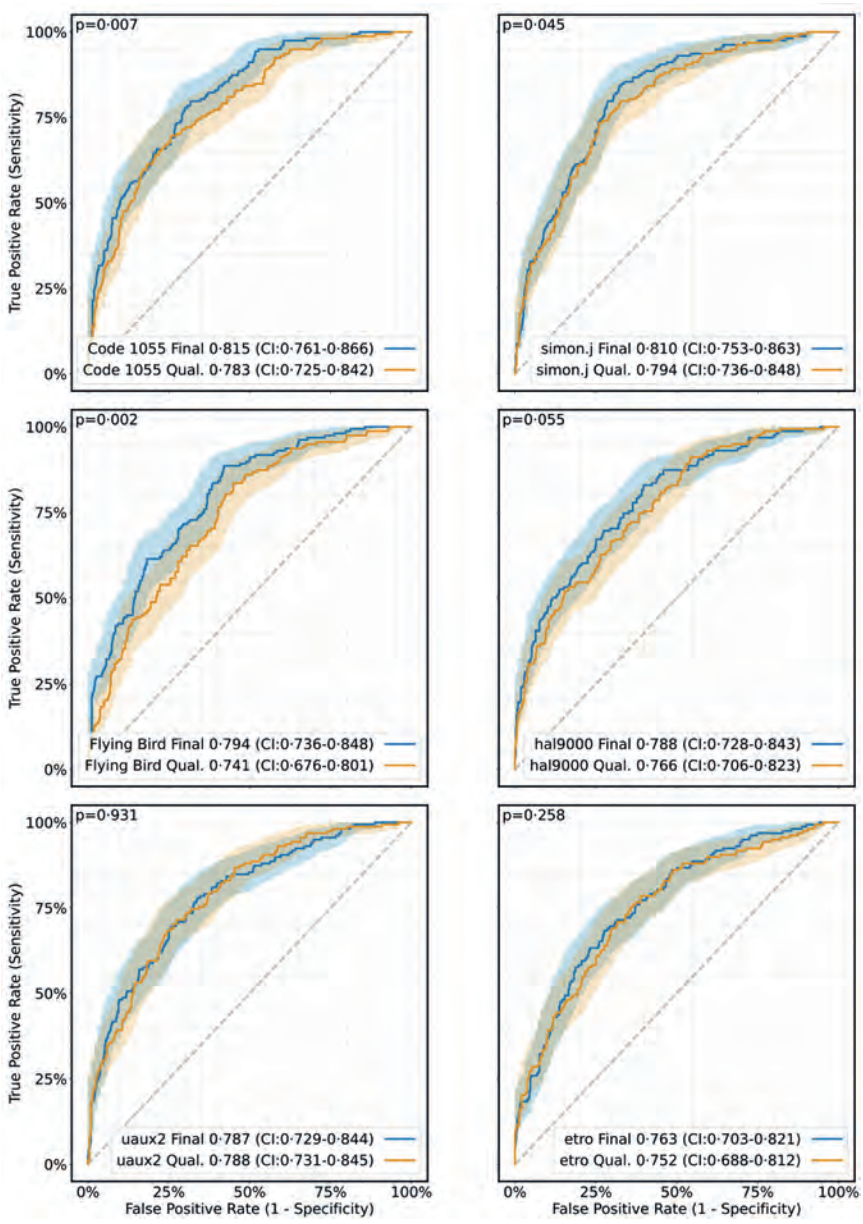


Figure 4.4: ROC curves with confidence intervals (CIs) for discriminating between severe and non-severe COVID-19 on test set B. The curves for the codebase submissions in the Final phase that completed training on training set B successfully are shown in blue. The ROC curves of the submissions that represented these teams in the Qualification phase, trained on training set A, are shown in orange. DeLong p-values are shown in the top left. AUCs with CIs are shown in the legends.

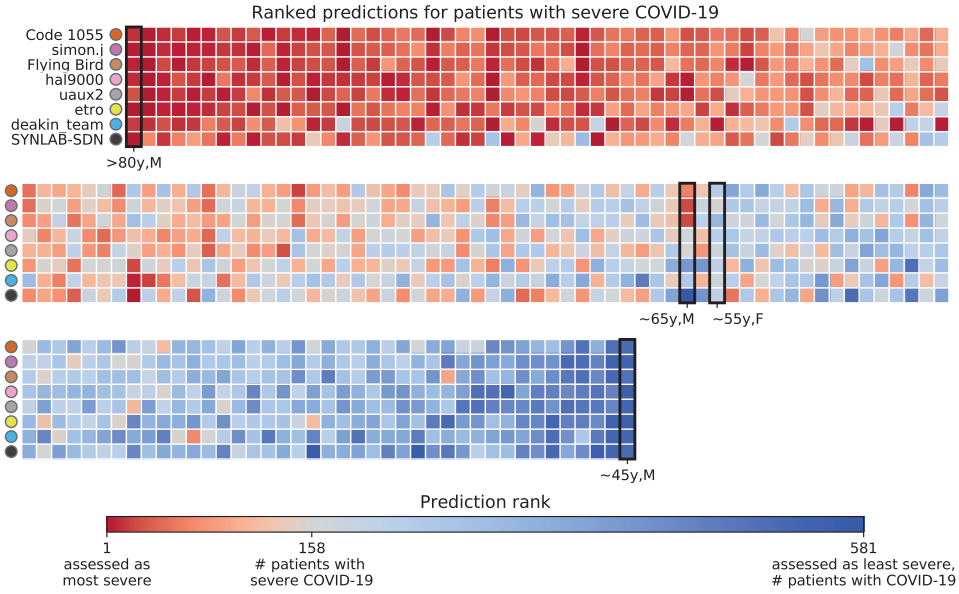


Figure 4.5: Ranked predictions for the subjects with severe COVID-19. Ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19. Each column shows the ranked predictions of all finalist teams for one subject. The subjects are ordered by the average rank of all corresponding finalist predictions. Figure 4.7 shows the CT scans corresponding to the columns that are outlined in black and annotated with age and sex.

4.3 Results

4.3.1 Qualification phase

413 participants registered to STOIC2021. During the rolling submissions, 30 teams, comprising 68 participants developed and successfully submitted 119 solutions to test set A1. Figure 4.3 shows an overview of the performance of these submissions. 20 teams competed for admission to the Final phase by successfully submitting to test set A2. The best performing teams on test set A2 were selected to advance to the Final phase, with invitations extended to the top ten teams that accepted.

4.3.2 Final phase

First training round

Eight of the ten Finalist teams submitted a codebase for training their solution on training set B. These eight teams are highlighted with unique colors in Figure 4.3. In

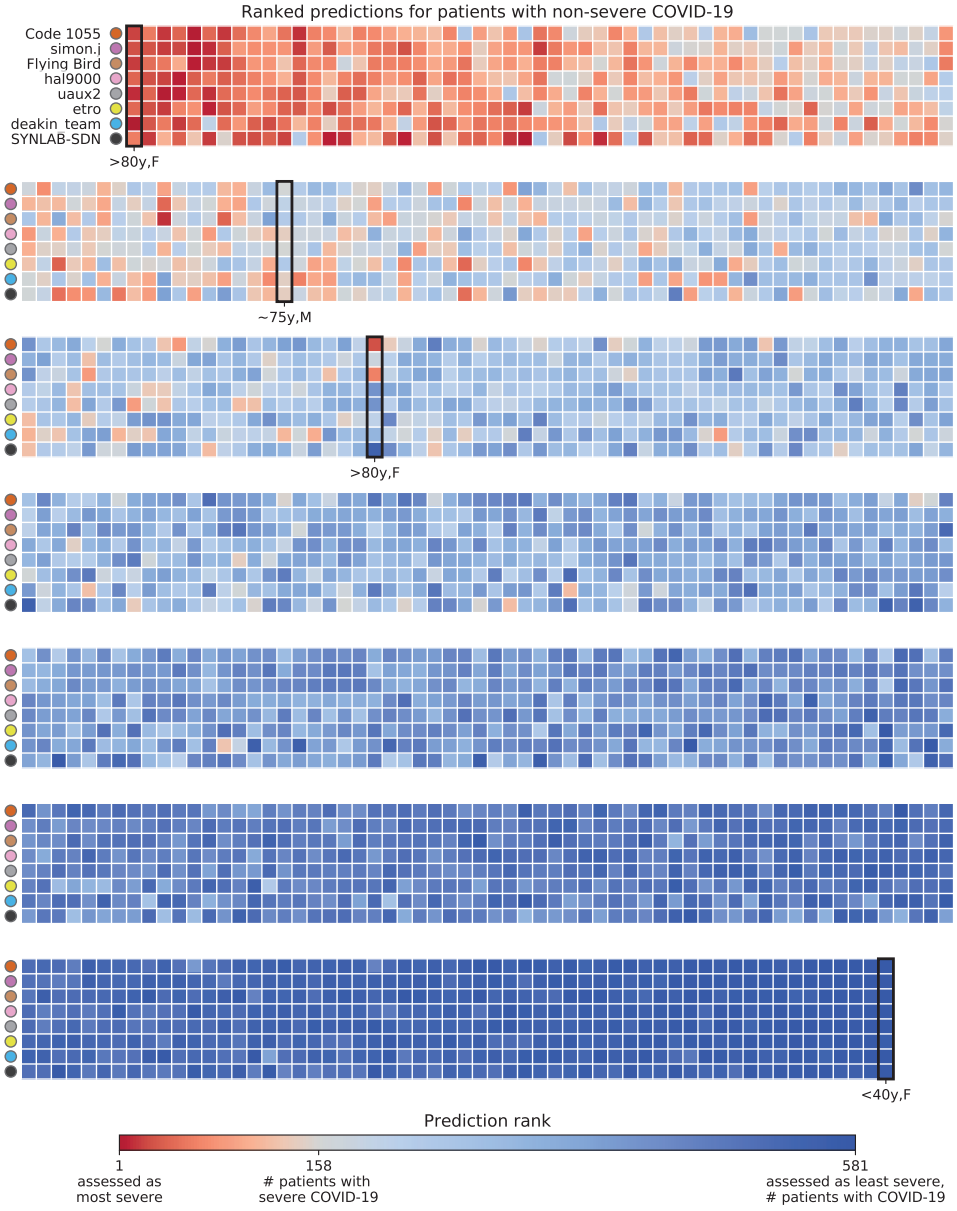
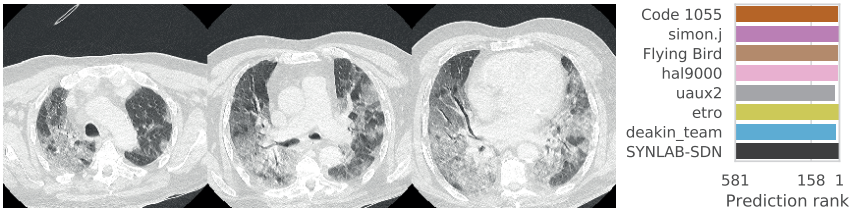
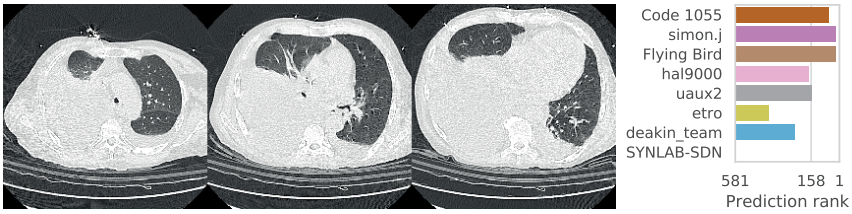


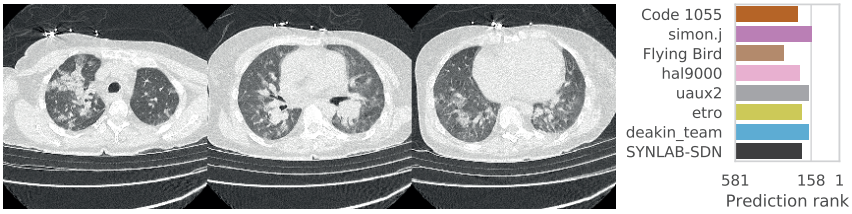
Figure 4.6: Ranked predictions for severe COVID-19 for subjects with non-severe COVID-19. Ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19. Each column shows the ranked predictions of all finalist teams for one subject. The subjects are ordered by the average rank of all corresponding finalist predictions. Figure 4.8 shows the CT scans corresponding to the columns that are outlined in black and annotated with age and sex.



(a) 80+ years old, male



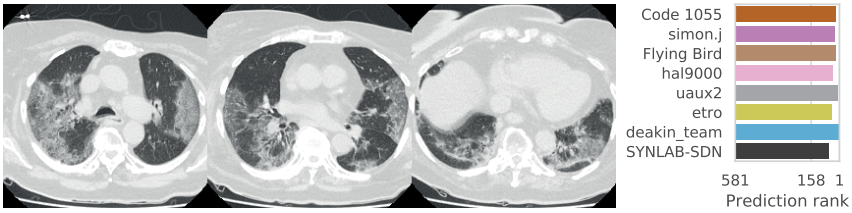
(b) 60-70 years old, male



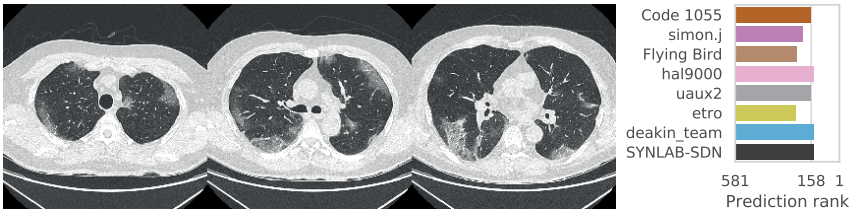
(c) 50-60 years old, female

(d) 40-50 years old, male

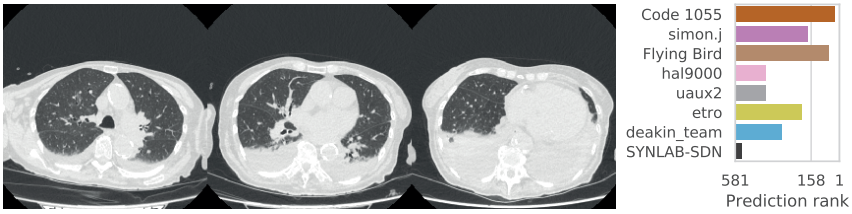
Figure 4.7: Subjects from test set B with severe COVID-19 that were highlighted in Figure 4.5. For each subject, three axial slices of a CT scan are shown on the left. The right shows how each finalist ranked the subject for presence of severe COVID-19. These ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19.



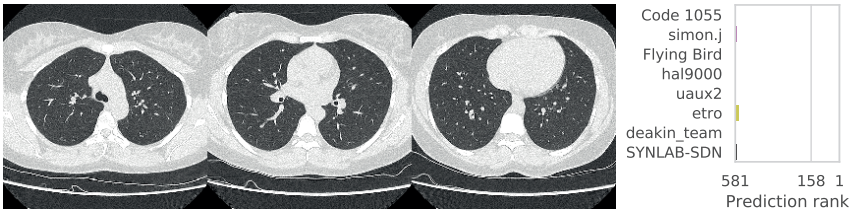
(a) 80+ years old, female



(b) 70-80 years old, male



(c) 80+ years old, female



(d) 40- years old, female

Figure 4.8: Subjects from test set B with non-severe COVID-19 that were highlighted in Figure 4.6. For each subject, three axial slices of a CT scan are shown on the left. The right shows how each finalist ranked the subject for presence of severe COVID-19. These ranks were computed over all subjects from test set B with a positive RT-PCR test for COVID-19.

the first training round, the codebases submitted by the teams simon.j, Flying Bird, and etro completed successfully. All other codebases exited training with an error.

Feedback round and second training round

The teams Code1055, uaux2, and hal9000 submitted codebases to the feedback round and to the second training round. All three submissions to the second training round completed successfully, resulting in a total of six successful Final phase submissions.

Performance

Table 4.1 shows the AUC on test set B for COVID-19 presence and severity of the teams that submitted to the Final phase. Figure 4.4 shows Receiving Operating Characteristics (ROC) curves of the six successful Final phase submissions for discriminating between severe and non-severe COVID-19 subjects from test set B. Figures 4.5 and 4.6 show how the finalists ranked the subjects from test set B with severe and non-severe COVID-19 respectively for presence of severe COVID-19. Figures 4.7 and 4.8 highlight some individual cases from test set B. During the original STOIC project [48], a logistic regression model was developed to predict severe COVID-19 using clinical variables and CT annotations by radiologists. It was developed and evaluated using the patients from the STOIC who were COVID-19 positive for both RT-PCR and CT, and had unenhanced CT. Of these 4238 patients, 1000 developed severe COVID-19. Revel and colleagues 6 reported an AUC for this model of 0.69 (CI: 0.67-0.71). To compare this model against the results from STOIC2021, an ensemble of the top three solutions for severe COVID-19 prediction was evaluated on the 367 patients from test set B who were COVID-19 positive for both RT-PCR and CT, and had unenhanced CT. 97 of these patients developed severe COVID-19. The top three ensemble achieved an AUC of 0.783 (CI: 0.706-0.848).

4.3.3 Solution methodology overview

Most finalists used lung and/or lesion segmentation methods [159, 160] to extract relevant features or to preprocess the input CT scan. Other preprocessing methods used were combinations of resampling, cropping, clipping, and normalizing or standardizing the image. End-to-end deep learning was the most common approach. The teams trained 2D or 3D versions of varying convolutional neural network architectures, [10, 96, 161, 162] often starting from pre-trained weights, and using varying data augmentation methods. The finalists that did not employ end-to-end learning employed logistic regression on top of either processed features extracted by vision

transformers [11] (simon.j) or features designed based on generated lung [159] and lesion masks [163] (etro and SYNLAB-SDN). Compared to the end-to-end deep learning methods, these methods consumed less time and memory during training. Most teams used an ensemble of classifiers. The rest of this section contains a detailed overview of the methods that were successfully submitted to the Final Phase. Table 4.2 presents URLs to the finalist codebases and the corresponding licences.

Team	Repository on `https://github.com/DIAGNijmegen/`	Licence
Code 1055	`stoic2021-finalphase-submission-code1055`	GPL-3.0
simon.j	`stoic2021-finalphase-submission-simonj`	MIT
Flying Bird	`stoic2021-finalphase-submission-flyingbird`	MIT
hal9000	`stoic2021-finalphase-submission-hal9000`	MIT
uaux2	`stoic2021-finalphase-submission-uaux2`	GPL-3.0
etro	`stoic2021-finalphase-submission-etro`	Apache-2.0

Table 4.2: Codebases and licences of finalist teams with successful submissions to the Final Phase.

Code 1055

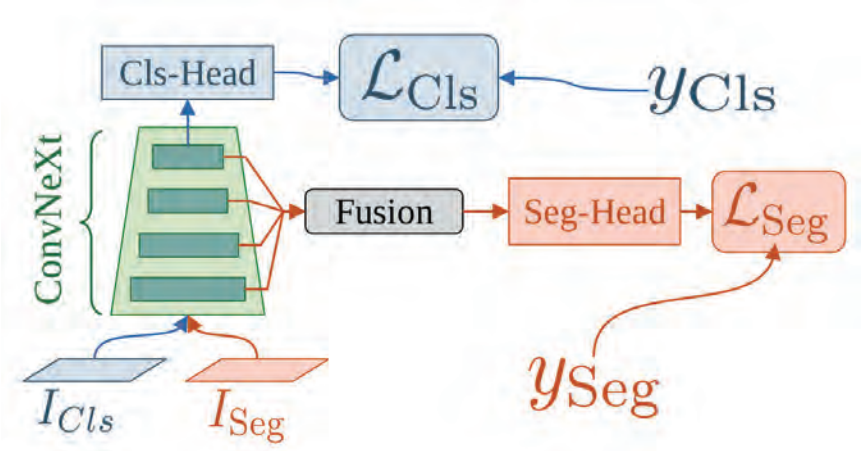


Figure 4.9: The pretraining pipeline is depicted. If segmentation data (I_{Seg}) is used as input, the features of each stage are upsampled, concatenated and the segmentation map is calculated with a segmentation head. If the classification data (I_{Cls}) is used as input, the severity prediction is obtained with a classification head using the features of the last stage. The overall loss is calculated as $L = L_{Cls} + L_{Seg}$

.

Severity classification using CT data is very similar to classical image classification apart from dealing with 3D tensors instead of 2D images. This allows us to employ the pre-existing techniques used in image classification. The ConvNeXt model [161] combines the benefits of the modern Vision Transformers [11] with Convolutional Neural Networks (CNN) and thus reaches state-of-the-art ImageNet results. We implement – to the best of our knowledge – the first 3D version of this architecture and, thus, boost the performance for severity classification in contrast to conventional CNNs.

Preprocessing strategy The input CT scans were resized to $256 \times 256 \times 256 voxels$. Their intensity values were clipped between -1100 and 300 HU and normalized around zero with a standard deviation of one.

Training strategy Even though the STOIC project [48] is a comparably large database of CT scans, it is exceedingly small in contrast to ImageNet [164]. Nevertheless, we are able to use a network with a large number of parameters and still prevent overfitting. For that purpose, we employ pretrained weights, a cosine learning rate scheduler, an early stopping strategy, an exponential moving average of the network parameters and efficient online data augmentation. Moreover, we balance our dataset in order to avoid learning a bias in the label distribution induced by the small number of severe cases.

In order to initialize our model with useful weights, we pretrain our network on two additional datasets. First, we train a 2D ConvNeXt on grayscale images from ImageNet. We calculate a superposition of gaussian inflated 2D weights to obtain 3D ImageNet weights. To further adjust these inflated ImageNet weights to our three dimensional task, we perform an additional multitask-pretraining using a segmentation [165–167] and classification [168] dataset. We use an architecture inspired by UPerNet [169] to concurrently perform segmentation of the lung region showing signs of COVID-19 infection for the segmentation data and prediction of severity for the classification data. This pre-training scheme is depicted in Figure 4.9. We are able to increase the performance of our model significantly with this additional pretraining in contrast to randomly initialized weights or inflated ImageNet weights.

In order to prevent overfitting and achieve greater generalization we use online data augmentation to virtually increase the dataset size. Besides using standard transforms like flipping, rotation or cropping, we apply a novel implementation of elastic deformations. By separating the gaussian kernels and utilizing GPU hardware, we are able to perform extremely fast elastic deformations. Consequently, we can augment our data with almost no additional cost. Furthermore, we perform

5-fold cross-validation during training.

Follow-up work is published by Kienzle *et al*. [170].

Inference strategy We average the outputs of the 5 networks trained in the cross validation. Therefore, we are able to train with the complete dataset and still generalize very well.

Public access Code for training and inference publicly available at `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-code1055`. Algorithm available for public use at `https://grand-challenge.org/algorithms/code-1055-second-final-phase-submission/`.

simon.j

Balaitous is an updated version of the AI-severity algorithm [171] implemented in the scancovia repository [172]. Given an input CT scan, the model outputs a probability for COVID-19 disease and for severe outcome (intubation or death within one month).

Preprocessing strategy The CT scan was rescaled to a resolution of 1.5 mm $\times$ 1.5 mm $\times$ 5 mm and reshaped to a shape of 224 $\times$ 224 $\times$ D, where D is the original dimension of the rescaled image along the axis orthogonal to the axial plane. A lung segmentation mask was computed using a 2D U-Net [159] and cleaned. The scan was cropped to the slices containing the lungs. For each slice, a first feature vector X_{full} was extracted using a ViT-L model [173]. This model was pretrained on ImageNet-22k using iBOT [173] and fine-tuned for 35 epochs on 165k CT scan images from 4k patients and 7 datasets. Next, the lung mask was applied so that only the lungs were visible and a second feature vector X_{lung} was extracted using the same ViT-L model without fine-tuning. For both VIT-L models, the extracted features of the individual slices were combined through pixel-wise average pooling.

Training strategy For the severe outcome two logistic regressions were applied to [X_{full}, age, sex] and [X_{lung}, age, sex]. The two predictions were aggregated through a learned weighted average. For the COVID-19 presence two logistic regressions were applied to X_{full} and X_{lung} and the two predictions were aggregated through a learned weighted average. Training was performed in 32 folds in the form of four different eight-fold cross validations.

Inference strategy The predictions were combined linearly with weights optimized that maximize the performance on the 32 training folds.

Methods altered from Qualification phase to Final phase None.

Public access Code for training and inference publicly available at `https://github.com/SimJeg/balaitous` and `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-simonj`. Algorithm available for public use at `https://grand-challenge.org/algorithms/simonj-first-final-phase-submission/`.

Flying Bird

The method employed was end-to-end deep learning with ResNet18 [10] models.

Preprocessing strategy In order to minimize image size and eliminate irrelevant regions, an open source lung segmentation model [159] was employed. The lung masks were used to crop the images, and were expanded by 6 mm to ensure complete coverage. The resulting cropped images were rescaled to $256 \times 256 \times 256$ voxels using trilinear interpolation. The voxel values were then clipped to the range (-1024, 512), and standardized with a mean of -237 and a standard deviation of 404.

Training strategy Due to the substantial volume of data, training a 3D network from scratch without a pre-trained model would be time-consuming. Regrettably, there is no all-purpose pre-trained model suitable for 3D networks. As a result, our approach involves initially training a pre-trained model via self-supervision [174], followed by conducting classification tasks built upon the pre-trained model. We used 5-fold cross validation. For training each fold, we appended a decoder to the ResNet18 network. Then, following the method described in [10], we applied some transformations to the input image and fed the transformed image into the network. We trained the network to enable it to recover the original image from the transformed image. After training, we obtained a pre-trained ResNet18 model. In the subsequent COVID-19 classification task and severity task, we initialized our models using pre-trained ResNet18. For both the COVID-19 classification task and severity task, we employed the same data augmentation techniques, including rotation, scaling, flipping, elastic transformation, Gaussian noise, and Gaussian smoothing. We used cross-entropy loss function and AdamW [158] optimizer, along with a one-cycle learning rate policy. For the severity task, we also incorporated age information by concatenating the age, which was divided by 100, with the output of ResNet18,

thereby taking into account the influence of age on severity. Furthermore, the data used in this task only consisted of COVID-19 positive cases.

Inference strategy For each model obtained through the cross-validation, test time augmentations are applied. The original input image is passed through the model, as well as variants of it obtained by flipping along each of the three axes, obtaining four outputs per model. Finally, the outputs of all models are averaged to obtain the final output.

Methods altered from Qualification phase to Final phase The data augmentation methods underwent minor modifications. The severity model was trained using both COVID-19 negative and positive images during the qualification phase, whereas only COVID-19 positive images were utilized in the final phase. Combinatorial image flipping was applied for test time augmentation during the qualification phase, along each of the three axes, resulting in a total of 8 outputs per model ($2 \times 2 \times 2$). In the final phase, only 4 outputs were generated, including the original image and those flipped along the x, y, and z axes.

Public access Code for training and inference publicly available at `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-flyingbird`. Algorithm available for public use at `https://grand-challenge.org/algorithms/flying-bird-first-final-phase-submission/`.

hal9000

We employed an ensemble of ResNet18 [10], and MoblieNetV3-Large [162] models trained end to end to predict COVID-19 disease and severity. In each model, embeddings of all slices were averaged and passed through a classifier to get the disease and severity probabilities. The ensemble of multiple models was used by averaging the probabilities of each model.

Preprocessing strategy 32 equidistant slices were sampled from the input CT scan. These slices were resampled to 224×224 pixels. The pixel values were clipped between -1350 and 150 HU. The images were normalized to a mean of 0.5 and a standard deviation of 0.5.

Training strategy The data for model development was split ten times into a training and validation set, such that the training set contained 85% of the data. A ResNet18 [10] was trained on five of these splits, and a MobileNetV3-Large [162]

was trained on the other five. Before presenting input data to a model, data augmentations were applied in the form of resizing, horizontal flipping, random cropping, gamma correction, color jitter, rotation, and blurring. The embeddings of all 32 slices were averaged and passed through a classifier to get the disease and severity probabilities. All models were trained using the Adam optimizer, with a learning rate of 0.0001 and weight regularization of 0.0005. The learning rate decayed by a factor of 0.1 every 40 epochs.

Inference strategy All model predictions were combined through averaging. We employed extensive test time augmentations involving five different crops (four corner crops and the center crop), and three different rotations (minus five degrees, plus five and plus ten degrees), and averaged the predictions for each augmentation. This was done for all five models for each model class. The ensemble prediction was obtained by averaging the probabilities.

Methods altered from Qualification phase to Final phase In the Qualification phase, we trained an ensemble of only MobileNet V3 Large models.

Public access Code for training and inference publicly available at `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-hal9000`. Algorithm available for public use at `https://grand-challenge.org/algorithms/hal9000-second-final-phase-submission/`.

uaux2

To assess the severity of SARS-CoV-2 (COVID-19) based on Computed Tomography (CT) scans of the lung, we apply an ensemble method approach, where we combine meta-data and 3D-CNN predictions. In addition to the information on patient age and sex already present in the data set, we rely on the respective Infection-Lung-Ratio (ILR) to generate our predictions. For implementation, we used our in-house developed framework AUCMEDI which is built on TensorFlow [175].

Preprocessing strategy For preprocessing, first, all data samples were resampled to a voxel spacing of 1.48 $\times$ 1.48 $\times$ 2.10 and clipped to the range [-1024, 100] to exclude irrelevant Hounsfield Unit areas [176]. Subsequently, the data was standardized to grayscale. Training samples that might exceed the accepted input image size of 148 $\times$ 224 $\times$ 224 were either randomly cropped or zero-padded to match the required size. For inference, center cropping was applied. To enable transfer learning, the grayscale images were converted to RGB. The intensities were scaled to the range

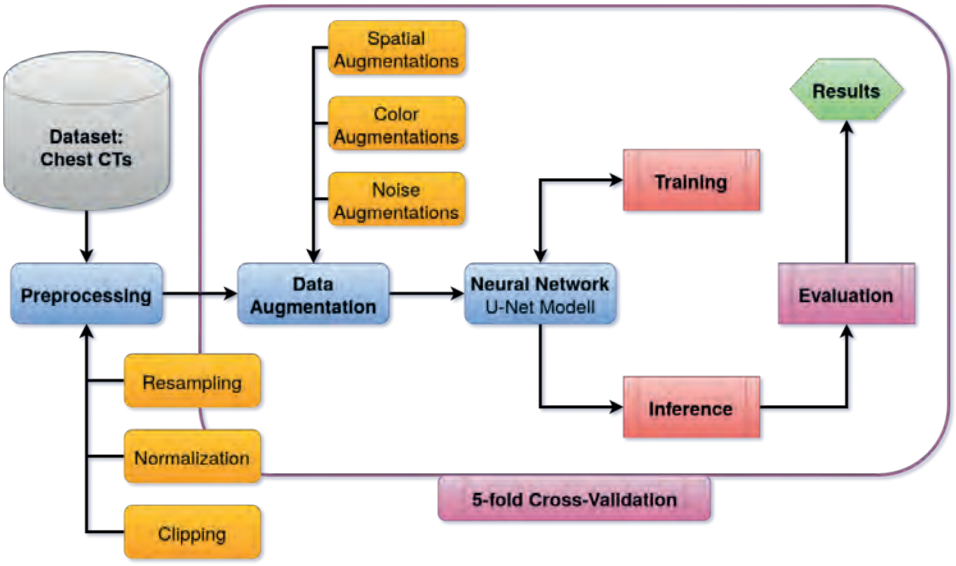


Figure 4.10: The MIScnn pipeline for SARS-CoV-2 segmentation to calculate the Infection-Lung-Ratio [160]

of [0, 1]. Then, normalization was applied via the Z-Score normalization approach based on the mean and standard deviation computed on the ImageNet dataset [103].

Training strategy In line with current state-of-the-art approaches, we applied several augmentation methods on the dataset, including rotation, flipping, scaling, gamma modification, and elastic deformations. Our main model for COVID-19 Severity prediction is based on a custom 3D version of the DenseNet121 architecture. We modified the classification head to additionally take metadata into account, which is described later on. For the training process, we applied transfer learning on the classification head and a fine-tuning strategy on all layers. The transfer learning on the classification head is done for 10 epochs, using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4. The fine-tuning runs for a maximum of 240 epochs, using a dynamic learning rate starting from 1×10^{-5} to a maximum decrease to 1×10^{-7} (decreasing factor of 0.1 after 8 epochs without improvement on the monitored validation loss). Furthermore, an early stopping technique was utilized, stopping after 36 epochs without improvement. As a loss function, we utilized the weighted Focal loss [177]. For inference, the model with the best validation loss is used. For COVID-19 presence prediction, we utilize a model based on the 3D ResNet34 architecture with the same hyperparameter settings as described above, that predicts 3 classes (negative/positive/severe). COVID-19 pres-

ence equals the sum of positive and severe cases. The metadata consists of three parts: Patient age, sex, and the ILR of each sample. The latter describes the ratio between infected parts of the lung and healthy tissue. We calculate the ILR by feeding the data into the MIScnn segmentation framework [160, 178], which utilizes a standard U-Net to predict infected areas Figure 4.10. For COVID-19 severity prediction, we applied cross-validation with a dynamic number of folds as a bagging approach for ensemble learning and monitored the outputs on the validation loss. We aimed to create a variety of models which were trained on different subsets of the training data.

Inference strategy Our final COVID-19 severity prediction comprises the averaged sum of all predictions from the ensemble. This approach not only allows for a more efficient usage of the available training data but also increases the reliability of the prediction.

Methods altered from Qualification phase to Final phase In the Qualification phase, cross-validation was done with five folds.

Public access Code for training and inference publicly available at `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-uaux2`. Algorithm available for public use at `https://grand-challenge.org/algorithms/uaux2-second-final-phase-submission/`.

etro

A short-term COVID-19 severity classifier was developed through logistic regression considering age, sex, and several image-derived features. A previously trained lung lesion segmentation model was used to extract volume fractions for ground glass opacities and consolidations. The segmentations were used in combination with the CT scan to derive mean intensities, kurtosis, and skewness for healthy lung parenchyma and lesion tissue. The final severity prediction was made by an ensemble of 20 models, trained on covid-positive samples selected through bootstrapping with replacement.

Preprocessing strategy The lungs were segmented using an open-source segmentation model [159]. A postprocessing step was added retaining only the 2 largest components and setting a minimum size for the components to exclude any regions outside the lungs that may have been segmented. CT scans were cropped to the lung mask and resampled to an isotropic spacing of 1 mm. The intensities were clipped

to [-1000 HU, 100 HU] and scaled to [-1, 1]. Ground glass opacity and consolidation patterns were segmented using a previously trained lung lesion segmentation model. The nnU-Net implementation in Monai [179] was used. The hyperparameters for this deep learning pipeline were determined automatically using the heuristics developed in nnU-Net [180]. The network was trained using the sum of the mean dice loss and the cross entropy, and deep supervision. Training data included 199 CT scans of the COVID-19 lesion segmentation challenge [181], 69 scans and manual lung lesion segmentations from the icovid consortium [182], 70 scans from the COPLENet public dataset [183] and 10 scans from the publicly available COVID-19 CT Lung and Infection Segmentation Dataset [184]. From these lung and lesion segmentations, the lesion volume fractions were calculated by dividing the lesion volume by the total lung volume. Additionally, the mean intensity, kurtosis and skewness were derived for each type of lesion and the healthy lung tissue.

Training strategy A logistic regression was trained for severity. Patient age and sex categories were assigned numerical values and were complemented with several image-derived features. Volume fractions of ground glass opacity and consolidation were included, as well as the mean intensity, kurtosis and skewness for healthy lung parenchyma and both lesion classes separately. For patients that were considered lesion free, the intensities and textural features of the ground glass opacity and consolidation were given the values of the healthy tissue. All intensity features were rescaled to [-1, 1]. To improve robustness, the severity classifier was built up by bagging 20 models where each training set was composed using bootstrapping with replacement on the covid-positive samples.

Inference strategy For inference, the intensity features were rescaled using the corresponding extrema from the training set. Final probabilities for severe COVID-19 were obtained by averaging the predictions of the 20 models. The probability of COVID-19 was predicted by a previously trained 3D ConvNext [161] model.

Methods altered from Qualification phase to Final phase For the Qualification phase, the model for severity was trained on both COVID-19 positive and negative patients versus only positives for the Final phase. For COVID-19 presence detection, the ConvNext model was added in the Final phase while a regression model similar to the severity classifier was used for the Qualification.

Public access Code for training and inference publicly available at `https://github.com/DIAGNijmegen/stoic2021-finalphase-submission-etro`. Algorithm

available for public use at https://grand-challenge.org/algorithms/etro-first-final-phase-submission/.

deakin_team

The method employed was end-to-end deep learning with DenseNet-201 [96].

Preprocessing strategy The input CT scans were resampled to an isotropic spacing of 1.6 mm3. A center crop of 240 × 240× 240 voxels was extracted from the CT, using zero padding when necessary. The voxel values were clipped between -1100 and 300 HU and rescaled to the range [0,1].

Training strategy A 3D DenseNet-201 [96], initialized with weights trained on the public STOIC2021 training set, was trained using the Adam optimizer with a learning rate of 0.00004, and a batch size of two for 15 epochs.

Inference strategy Inference was performed by a forward pass through the trained DenseNet-201 model.

Methods altered from Qualification phase to Final phase An ensemble approach incorporating multiple models, specifically DenseNet-201, -169, and -121, was initially proposed for this study. However, due to constraints related to computational resources and time in the training environment, we were ultimately only able to train a DenseNet-201 model.

Public access Algorithm available for public use at https://grand-challenge.org/algorithms/baseline-13/ (Qualification phase submission).

SYNLAB-SDN

The method was based on logistic regression using patient age, sex and features extracted from lesion masks.

Preprocessing strategy The CT voxel intensity values were clipped to the range [-1000, 500]. Afterward, a pre-trained model for COVID-19 lesion segmentation by Nvidia Clara (2) was used to obtain suitable masks representative of COVID-19 lesion burden. Furthermore, a lung mask was segmented from the input CT scan using a U-Net [159]. From the lesion masks, the following features were extracted:

- Mean HU value,

- Standard deviation intensity,
- Percent of lesion volume, computed as lesion volume divided by lung volume,
- Number of connected components in the lesion mask.

In addition, patient age and sex were included as features.

Training strategy For the classification, the dataset was randomly split into a training/validation (80%) and testing set (20%). Z-normalization was applied to the features constituting the training set, and the mean and standard deviation values calculated on the training set were used on the validation and test set. A downsampling strategy was applied to balance the dataset. We have trained logistic regression to solve the tasks. K-Fold cross-validation with K = 5 was applied to the training dataset for model selection in the form of hyperparameter tuning.

Inference strategy The trained logistic regression model was applied to perform inference.

Methods altered from Qualification phase to Final phase None.

Public access Algorithm available for public use at `https://grand-challenge.org/algorithms/2steps-2/` (Qualification phase submission).

4.4 Discussion

The Type Three (T3) medical image analysis challenge format presented in this study allows solutions to be trained on private data and that guarantees that their training methodologies are reusable. T3 was implemented in the STOIC2021 challenge, in which participants predicted from an initial CT scan, whether a COVID-19 patient would be intubated or would die within one month.

To evaluate their solutions, challenges typically release test set images to enable participants to run inference on them [54, 111, 113–149]. STOIC2021 consisted of a Qualification phase that instead followed the structure implemented of some recent challenges [6, 108–110, 150, 151] where participants submit solutions trained on public data, and of a T3 Final phase. The Final phase solutions consistently outperformed the solutions submitted to the Qualification phase by the same participants. This indicates that T3 may improve challenge solution performance through training on a combination of public and private data.

STOIC2021 resulted in six publicly available codebases through which the training and inference methods for the top performing solutions can be accessed. The challenge organizers tested these codebases by training the corresponding solutions without manual intervention by the participating teams. This guaranteed the reusability by third parties of these publicly released training methodologies. Links to these codebases can be found in section 4.3.3. Most finalists used sex and age information as additional input to their model. Advanced age and male sex are risk factors for severe outcome of a COVID-19 infection [48]

The released codebases may be useful for the development of tools to assist in the diagnostic process of COVID-19 infections in patients with suspected COVID-19. The methods developed for the STOIC2021 challenge may be useful for triaging patients based on the severity of their infection, which could help with optimizing the allocation of healthcare resources. This could be especially helpful in high-demand situations, and/or in medical centers where access to specialized readers is limited. Additionally, the released training methods may be useful for any 3D medical image classification tasks. This versitality stems from the fact that, besides employing a pre-trained segmentation model, most of the submitted solutions use 3D image processing methods that are not specific to one task or image modality.

This work demonstrated through the STOIC2021 challenge that the T3 challenge format allowed for training on private data and for the developed training methods to be re-usable. This suggests that future challenges that implement the T3 format may also reap these benefits. Future challenges may also benefit from incorporating a T2 Qualification phase before a T3 Final phase. In STOIC2021, this set-up minimized overhead during method development for the participating teams and kept down costs for the challenge organizers.

STOIC2021 participants were not incentivized to focus on the confirmation of COVID-19 presence, since this is possible with high sensitivity through RT-PCR testing [185]. The absence of this incentive explains why team Code 1055, which achieved the highest AUC for discriminating between severe and non-severe COVID-19 in the Final phase, achieved the lowest AUC for detecting COVID-19 presence of all finalists. It also explains why, overall, the finalists' performances on the auxiliary metric of detecting COVID-19 presence did not align with the finalists' ranks in the Final phase.

This study has limitations. Participants of STOIC2021 were not incentivized to focus on the calibration or interpretability of their solutions. Also, datasets for externally validating solutions on their ability of predicting intubation or death within one month were not publicly available. This also prohibited directly comparing the presented performances to the algorithms trained to predict severe COVID-19 out-

come by [171]. However, the solution by simon.j was heavily based on this work. Furthermore, T3 challenges are limited by the computational budget of the challenge organizers. STOIC2021 therefore implemented a limit to the compute resources for training the Final phase solutions, as detailed in section 4.2.3, and allowed for a limited number of finalists. Lastly, the maximum obtainable performance is limited by imperfections in the COVID-19 severity and presence labels. Death at one month follow-up could have resulted from any cause. RT-PCR is an imperfect ground truth for infection. For the STOIC study, 39% of initially negative RT-PCR tests were found to be positive when repeated in patients with typical clinical signs of COVID-19 [48].

Conclusion

This work showed the efficacy of the T3 medical image analysis challenge format. T3 has two benefits with respect to previous challenge formats. Firstly, it allows challenge solutions to be trained on private data. This results in training on bigger data, which can increase the performance of the resulting challenge solutions. Secondly, it ensures that the training methods developed for the challenge can be used out-of-the box by third parties.

Acknowledgments

The European Regional Development Fund had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript. Amazon Web Services funded algorithm evaluation, algorithm training for the Final phase, and prizes to the best performing teams. This study was endorsed by The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society. The STOIC study [48] was sponsored by Assistance Publique Hôpitaux de Paris and was funded by Fondation APHP pour la Recherche, Guerbet, Innothera, Fondation CentraleSupélec. For the STOIC study, General Electric Healthcare provided a 3D image visualization web application and Orange Healthcare a data repository.

Role of the funding resource

The European Regional Development Fund had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript. Amazon Web Services funded algorithm evaluation, algorithm training for the Final phase, and prizes to the best performing teams.

Data statement

Any collaborative research project led by an academic partner who requires access to the STOIC data shall be analyzed, validated, and authorized by the Steering Committee of STOIC. To this end, the academic partner shall send a document describing the research project to the Stoic Steering Committee at the following email address: `marie-pierre.revel@aphp.fr`, with the following subject: STOIC DATA ACCESS PERMISSION. After acceptance by the Steering Committee, the academic partner shall sign a specific agreement (Data Transfer Agreement - DTA) with AP-HP, who is legally responsible for the STOIC data as Sponsor of the STOIC research. Refer to the STOIC princeps paper [48] for more information.

A diverse multi-task database for universal 3D medical image classification

5

Authors: **Luuk H. Boulogne, Natália Alves, Luc Builtjes, Khrystyna Faryna, Alessa Hering, Colin Jacobs, Anindo Saha, Clarisa Sánchez, Kiran Vaidhya Venkadesh, Coen de Vente, and Bram van Ginneken**

Original title: **A Diverse Multi-Task Database for Universal 3-Dimensional Medical Image Classification**

Submitted

Abstract

Automatic medical image classification methods are valuable in various clinical applications, but application-specific solutions are often outperformed by generally applicable methods. This work presents a comprehensive database for the development of accessible and universally applicable 3D medical image classification algorithms. These datasets cover a broad spectrum of classification tasks containing images from multiple modalities and body parts. Each dataset is divided into a training and test set. The training sets are publicly released in a standardized format to foster algorithm development. Automatic evaluation of developed solutions on the test data is offered on the grand-challenge.org platform to ensure fair comparisons between methods. To validate the readiness of the database for algorithm development, a baseline system was trained and tested successfully across all datasets. The baseline codebase is made publicly available along with a detailed tutorial on using the database for development and utilizing the automatic evaluation. This database release is aimed to improve healthcare outcomes by spurring the development of universally applicable and publicly accessible medical image classification tools.

5.1 Background & Summary

Public benchmarks have played a crucial role in the progress of automated medical image analysis by establishing clear goals and providing shared research agendas.[6, 49, 51–54] These benchmarks offer standardized databases, allowing researchers to compare their methods' performance with others in the field, fostering collaboration, and driving the development of robust and reusable algorithms.

In response to the growing demand for universally applicable and publicly accessible methods for medical image analysis, we have created a comprehensive database for the development of universal 3D medical image classification systems. The database comprises a diverse range of 3D medical imaging datasets encompassing various characteristics and classification tasks. Figure 5.1 depicts an overview of these tasks.

The database includes seven datasets based on already publicly available data, and features a unified structure accross these datasets. Each dataset describes a different medical image classification task, including primary open-angle glaucoma (POAG) classification from optical coherence tomography (OCT),[50] nodule false positive reduction from computed tomography (CT),[49] molecular breast cancer subtyping from subtraction MRI,[47] COVID-19 presence and severity classification from CT,[17, 48] rib fracture type classification from CT,[46] and identifying clinically significant prostate cancer (csPCa) from multi-parametric MRI.[45] For the task of identifying abnormal kidneys, we publicly release CT images and abnormality labels depicting kidneys [186] and combine these with other publicly available kidney CT images.[51, 186]

Existing databases for 3D medical image classification typically focus on narrowly defined tasks,[6, 17, 45–53] encouraging specialized solutions tailored to each problem. Nevertheless, top-performing solutions on these benchmarks are frequently based on generally applicable deep learning methods. In many domains, task-specific customization of general deep learning methods is unnecessary to obtain good performance.[12, 187–189] Our study aims to encourage the development of approaches that are applicable and reusable for a broad range of applications in the medical image analysis domain.

Current databases for 3D universal medical image classification are insufficient for the development of real-world applications. For example, MedMNIST [190] is a light-weight database that describes many different medical image classification tasks involving a variety of imaging modalities. This database contains 2D images of 28×28 pixels and 3D images of $28 \times 28 \times 28$ voxels. Although this database could support numerous research and educational purposes by quickly illustrating

Figure 5.1: We have compiled seven medical image datasets, each describing a classification task where one or more labels are predicted from 3D input volumes [17, 45–51, 186]. This figure shows a cross-sectional view of example input images for each dataset, with the respective classification task noted above. Each image is marked with the corresponding image acquisition method in white. For tasks that involve input images of multiple acquisition techniques, only a part of each acquired image from a single case is depicted. The tasks include primary open angle glaucoma detection from an optical coherence tomography (OCT) image of the retina, lung nodule false positive reduction (FPR) from computed tomography (CT), molecular breast cancer subtyping from subtraction MRI, COVID-19 classification from CT, rib fracture type classification from CT, kidney abnormality detection from CT, and clinically significant prostate cancer (csPCa) grading from multi-parametric MRI.

the behavior of different algorithms, its small image size makes MedMNIST not representative for real-world medical image analysis applications. The RadImageNet [191] database contains CT, MRI, and ultra-sound images and labels for a wide variety of tasks. There is, however, no clearly structured set of tasks and access to the data is only available upon request and limited to images with reduced resolution and restricted labeling.

Although public databases for universal 3D medical image classification are lacking, a database for universal medical image segmentation has successfully established by the Medical Segmentation Decathlon,[54] a medical image analysis challenge. The winning solution of this challenge, nnU-Net,[189] is both publicly accessible and easy to use, addressing the growing demand for efficient and effective medical image analysis tools. In the format we propose for our dataset, we were inspired by design choices made by the nnU-Net [189].

We hope that the database released with this study will accelerate the development of robust and universally applicable algorithms for 3D medical image classification by providing a clear and objective way to evaluate and compare their performance across multiple tasks. This may contribute to the accessibility of universally applicable tools for developing medical image analysis solutions, resulting in enhanced patient care and improved healthcare outcomes.

5.2 Methods

Seven 3D medical imaging datasets were collected, each describing a unique task with varying characteristics, such as the image structure, imaging modality, number of images, image size, spacing, and isotropy as shown in table 5.2. The primary objective of our data collection was to maximize diversity between datasets so that machine learning classifiers that can learn to perform well across all these datasets would also generalize well to learning new classification tasks. Methods developed using these data may thus be useful for a vast array of clinical scenarios, including those that are not represented in our collected datasets.

Medical image classification algorithms are typically not utilized in isolation, but rather as components of an image processing pipeline. To ensure their applicability in real-world scenarios, the datasets in our database are designed to facilitate the development of algorithms that can be employed not only in pure classification tasks but also as part of such processing pipelines. More specifically, image classification is frequently deployed to categorize structures identified by a detection algorithm [49, 93] and can receive additional input next to a medical image in the form of a semantic segmentation of relevant structures [33]. Consequently, some datasets within the

database contain image crops that simulate processing pipelines incorporating both detection and classification components. To simulate the integration of classification after semantic segmentation of relevant structures, region of interest (ROI) masks are provided as supplementary input for certain datasets in our database.

To fairly compare automatic universal classifiers with each other, it is crucial that each system is evaluated on identical data. To ensure this consistency, an evaluation database was prepared separately from the database described in this Data Descriptor. The sections below describe in detail which data were included in the training database described in this Data Descriptor, and which data were reserved for the purpose of automatic evaluation on `https://auc23.grand-challenge.org/`.

Detailed descriptions of each dataset are provided in the following sections. We have followed all relevant ethical regulations for each dataset.

Primary open angle glaucoma classification from optical coherence tomography

Motivation and task

Glaucoma is a chronic eye disease that affects the optic nerve and is one of the leading causes of irreversible blindness worldwide. Estimates indicate that by the year 2040, the number of individuals affected by glaucoma and experiencing different levels of visual impairment will exceed 110 million [192]. Among this population, approximately 10% are expected to suffer from complete blindness in both eyes, while around 25% may experience blindness in one eye [193]. These statistics highlight the significant impact of glaucoma on global visual health and emphasize the urgent need for effective diagnosis, intervention, and management strategies to mitigate the burden of this disease. OCT imaging provides high-resolution cross-sectional images of the retina, making it a valuable tool for glaucoma diagnosis and monitoring. Using artificial intelligence to automatically classify glaucoma, may be key to achieve cost-effective of the screening of this disease. The feasibility of such approaches has been demonstrated before in literature [194]. The task associated with this dataset is to develop an algorithm capable of classifying OCT images as either normal or indicative of primary open-angle glaucoma (POAG).

Provenance

Our dataset for POAG classification is based on data described by Maetschke *et al.* [195], which was made publicly available on Zenodo[1]. This dataset contained OCT

[1]https://zenodo.org/record/1481223

scans centered on the optic nerve head (ONH) acquired from a total of 624 patients using a Cirrus SD-OCT Scanner (Zeiss, Dublin, CA, USA) [195]. To ensure data quality, scans with a signal strength less than 7 were discarded, resulting in a total of 1,110 scans. The scans were maintained in their original laterality without flipping left into right eye orientation. Among the 1,110 scans, 263 were diagnosed as healthy, while 847 were diagnosed with primary open angle glaucoma. Eyes diagnosed as glaucomatous were identified based on the presence of glaucomatous visual field defects and the requirement of at least 2 consecutive abnormal test results. The original scans had physical dimensions of 6 mm $\times$ 6 mm $\times$ 2 mm, consisting of 200 $\times$ 200 $\times$ 1024 voxels per volume. However, for network training purposes, they were downsampled to a size of 64 $\times$ 64 $\times$ 128 voxels.

On this dataset, the performance of a logistic regression model was compared to that of a 3D convolutional neural network (CNN). The logistic regression model used 22 outcome measures that were computed by the OCT scanner that was used for acquisition. The CNN achieved an AUC of 0.94, which was higher than the performance of the logistic regression model, which achieved an AUC of 0.89 [195].

Dataset preparation

The data described by Maetschke *et al.* [195] was used without any further preprocessing than described in the previous section.

Training/test split

To our knowledge, there was no public training/test split available for this dataset. Therefore, we split the dataset ourselves into 80% training and 20% testing on patient-level. This resulted in a training set with 884 OCT volumes from 499 patients, of which 677 OCT volumes from 390 patients were positive for POAG and 207 OCT volumes from 109 were negative. The test set contained 226 OCT volumes from 125 patients in total, of which 170 OCT volumes from 97 patients were positive and 56 OCT volumes from 28 patients were negative.

Pulmonary nodule false positive reduction from computed tomography

Motivation and task

Lung cancer is the most deadly cancer worldwide. Mortality can be reduced when high-risk individuals undergo low-dose computed tomography (CT) screening, with reported reductions of up to 26% [196, 197]. Early-stage lung cancer is visible on CT

as small pulmonary nodules, often measuring less than 1 cm in diameter. While CT imaging exhibits high sensitivity in detecting true nodule candidates, it presents a significant challenge in the form of also detecting false positive candidates. The dataset presented here reflects the task of classifying a CT region around a candidate as a nodule or a false positive, which was also a component of the LUNA16 challenge [143].

Provenance

The dataset comprised 888 CT scans from the LUNA16 challenge dataset[143]. From these scans, 1 186 candidates were identified as true nodules, based on consensus from at least 3 out of 4 radiologists who annotated the nodules. To acquire false-positive candidates, we utilized the CUMedVis system [143] with an operating threshold of 8 false positives per scan, resulting in 7 156 false-positive candidates. Together, the dataset comprises 8 342 candidates.

Dataset preparation

Following the approach of Venkadesh et al. [198], we extracted 3D cubes around each candidate. These cubes were standardized to a size of 50 mm, encompassing 64 pixels in each direction.

Training/test split

We split the dataset into 80% training and 20% testing at a patient-level. This resulted in 6 694 candidates (941 true nodules) for training and 1 648 candidates (245 true nodules) for testing.

Molecular breast cancer type classification from subtraction magnetic resonance imaging

Motivation and task

Radiogenomic analysis of breast cancer aims to establish a relationship between tumor imaging phenotypes and molecular markers, which could potentially provide non-invasive genomic analysis methods [47]. The task that we included in our database concerns classifying molecular breast cancer type from a breast subtraction MRI image, based on the human epidermal growth factor receptor 2 (HER2), estrogen receptor (ER), and progesterone receptor (PR) status. More specifically, it involved

discerning between the subtypes luminal A (ER and/or PR positive, HER2 negative), luminal B (ER and/or PR positive, HER2 positive), HER2 (ER and PR negative, HER2 positive), and triple negative (ER, PR, and HER2 negative).

Provenance

The dataset comprised 922 subtraction breast MRI scans from a previous study by Saha, A. *et al.*.[47] In this study, multivariate models were developed based on radiomics features, describing global, as well as localized image characteristics, including features from the breast, fibroglandular tissue, and tumour. These models were predictive of Luminal A subtype (AUC = 0.697) and triple negative breast cancer (AUC = 0.654).

Dataset preparation

The breast MRI images from Saha, A. *et al.*[47] were used without additional preprocessing.

Training/test split

The dataset was split randomly into a set of 737 cases for training (481 luminal A, 76 luminal B, 46 HER2, 134 tripple negative) and a set of 185 for testing (114 luminal A, 28 luminal B, 13 HER2, 30 triple negative).

COVID-19 presence and severity classification from computed tomography

Motivation and task

In response to the COVID-19 pandemic, extensive research has been conducted on use of CT imaging to automatically diagnose and classify the disease [112]. CT scans, providing high-resolution, three-dimensional views of the lungs, are able to capture key indicators of a COVID-19 infection, including diffuse patterns of ground-glass opacities in the lungs, with or without consolidations [15]. Leveraging artificial intelligence with CT imaging has the potential to assist in automating the diagnostic process, possibly enhancing the workflow and practices of doctors and radiologists [112]. Our database includes a public dataset that reflects the task of classifying from a CT scan whether a patient has a COVID-19 infection, as well as the severity of this infection.

Provenance

We base our dataset for COVID-19 presence and severity classification from CT on data from the Study of Thoracic CT in COVID-19 (STOIC) project [48]. This study collected data from patients of 20 French university hospitals between March 1 and April 30th, 2020. The STOIC study collected CT scans and reverse transcription-polymerase chain reaction (RT-PCR) test results from over 10 000 patients. Additionally, a one-month follow up was conducted to assess the development of severe COVID-19, which was defined as either death or intubation within the follow-up period. The STOIC project protocol can be accessed via ClinicalTrials.gov with identifier NCT04355507.

The STOIC2021 COVID-19 challenge [17] was held using data from the STOIC project. During this challenge, participants were tasked to develop fully automatic algorithms to severe COVID-19 from CT. 2 000 CT sans with COVID-19 presence and severity labels were publicly released on the Amazon Web Services (AWS) Registry of Open Data [17] to serve as a public training set for STOIC2021. An ensemble of the best performing solutions achieved an AUC of 0.849 for discerning between positive RT-PCR result from CT for patients included in the STOIC study and and AUC of 0.817 for discerning between severe and non-severe COVID-19.

Dataset preparation

The thorax CT scan at presentation utilized for the STOIC2021 challenge were reused without additional processing. As additional input, pulmonary lobe and COVID-19 lesion segmentations were generated using a relational two-stage U-Net [78]. The existing RT-PCR test results for each patient and one-month follow-up outcome were used as-is as COVID-19 presence and severity label respectively.

Training/test split

In accordance with the data split utilized in the STOIC2021 challenge [17], the training dataset comprises the CT scans of 2 000 patients that were publicly released [17]. The 200 CT scans for evaluating rolling submissions in the STOIC2021 challenge were employed as the test set.

Rib fracture type classification from computed tomography

Motivation and task

Rib fractures are a crucial indicator of trauma severity and degree of disability [46]. While CT offers a more accurate assessment than standard chest radiographs, iden-

tifying and diagnosing fractures in CT is a more complex and time-consuming task [46]. The task associated with this dataset involves the classification of individual rib fractures from CT into distinct types: buckle, displaced, non-displaced, or segmental.

Provenance

This dataset repurposes the 420 CT scans with rib fracture annotations that were publicly released for the Rib Fracture Detection and Classification challenge (RibFrac) [46]. The RibFrac dataset was originally collected for the development and validation of the FracNet model [46]. The annotation process of the rib fractures in the training cohort of this dataset involved five radiologists with varying experience levels. Two junior radiologists delineated the volume of interest of the traumatic rib fractures using the diagnosis reports of two other radiologists, and 3D Slicer software. A senior radiologist confirmed the volumes of interest.

Dataset preparation

This dataset has been derived from the publicly available RibFrac training dataset [46]. Blocks of dimensions $150 \times 150 \times 150$ mm, centered around each fracture, were extracted. To emulate this classification as part of a processing pipeline that also contains a segmentation step, the dataset also includes a segmentation mask of each fracture of interest. These masks were also provided in the RibFrac dataset.

Training and evaluation split

We split the dataset into 80% training and 20% testing. This resulted 1 325 fractures (of which 493 displaced, 460 non-displaced, 230 buckle, and 142 segmental fractures) for training and 332 fractures (of which 125 displaced, 108 non-displaced, 61 buckle, and 38 segmental fractures) for testing.

Kidney abnormality classification from computed tomography

Motivation and task

Kidney and renal pelvis cancer rank as the 6th most common and the 12th most deadly form of cancer in the United States [199]. Partial or radical nephrectomy provide a solution to kidney tumors with a relatively high survival rate and quality of life for a patient, but this requires the cancer to be found before metastasis can occur. CT scans allow for the differentiation of cysts, angiomyolipoma (benign kidney

tumors) and renal cell carcinoma from normal tissue. The utilization of artificial intelligence to differentiate healthy from unhealthy kidneys in such images could both the improve radiologist workflow and allow for potential higher detection rates of renal cancer when the two parties work in tandem.

Provenance

The dataset for kidney abnormality classification can be subdivided into two parts.

KiTS21 Firstly, data from the KiTS21 challenge[51] was used. This study includes a cohort of 300 patients who underwent nephrectomy for suspected renal malignancy between 2010 and 2020 in one of two medical centers in the US. Their dataset includes both abdomen CT-scans and segmentation masks for kidneys and, when applicable, renal tumors and cysts. These annotations were made by a team of both expert, trainees and laypeople in an iterative process. The KiTS challenges are biennial challenges where participants are asked to develop an automatic segmentation algorithm for kidney tumors using this dataset. We used the KiTS21 dataset in full for this project.

Kidney Abnormality Segmentation project Secondly, data was used from the "Kidney Abnormality Segmentation in Thorax-Abdomen CT Scans" project [186]. This study collected CT scans of patients at the Radboud University Medical Center (Nijmegen, The Netherlands) who underwent a thorax-abdomen CT scan in 2015. Approval of the medical ethical committee of the Radboud University Medical Center was obtained prior to the study. The need for written informed consent was waived, and data were collected and anonymized in accordance with local guidelines.

All CT scans were acquired with Toshiba (Aquilion One) or Siemens (Sensation 16, Sensation 64, and Somatom Definition AS) scanners, with FC09, FC09-H, B30f, B30fs, or I30f reconstruction kernels. The slice thickness ranged from 0.5 to 3mm.

An initial selection of 1905 studies from 929 patients were retrieved. The corresponding radiology reports were analyzed for sentences mentioning both the Dutch word for kidney ('nier' or 'nieren', excluding 'bijnier') and one or more abnormalities, including cysts ('cyste' or 'cysten'), lesions ('laesie' or 'lesies'), masses ('massa'), metastases ('metastase' or 'metastasen'), or tumors ('tumor'). This analysis split the cohort in 138 patients with one or more kidney abnormalities and 791 patients without.

Six of the 138 patients with kidney abnormalities were excluded due to the presence of infrequent abnormalities. Specifically, three patients had undergone kidney

transplants, two presented with irregular kidney sizes, and one had a horseshoe kidney. From the initial 791 patients without kidney abnormalities, 133 patients were selected, resulting in a total of 265 patients. For each of these patients, a single CT scan was selected to represent them.

Segmentations for both kidneys and abnormalities were annotated by a team of four medical students under supervision of an expert radiologist.

Dataset preparation

The data was further processed by extracting crops from the full CT images around the kidneys using the provided kidney segmentation masks with an additional border of 2 cm in each direction. Some patients had already undergone a radical nephrectomy at the moment that a scan was made, resulting in a total of 591 crops from the KiTS data, and 486 from the Kidney Abnormality Segmentation project.

Training/test split

The KiTS dataset and 215 patients from the Kidney Abnormality Segmentation project were included into our training database. This resulted in a training set with a total of 986 kidneys. 91 kidneys of 50 patients from the Kidney Abnormality Segmentation project were set aside for evaluation purposes and not included in the training database described in this Data Descriptor.

Clinically significant prostate cancer classification from magnetic resonance imaging

Motivation and task

Prostate cancer is one of the most prevalent cancers in men. One million men receive a diagnosis and 375,000 die from clinically significant cancer, each year, worldwide[200]. MRI is playing an increasingly important role in the early diagnosis of prostate cancer, and has been recommended by recent clinical guidelines in the European Union, United Kingdom and the United States. Radiologists follow a semi-quantitative assessment to read prostate MRI that mandates substantial expertise for proper usage, leading to high inter-reader variability. Artificial intelligence-assisted triaging or secondary reading can address the rising demand in prostate imaging, improve diagnostic accuracy and reduce inter-reader variability.

Provenance

This dataset repurposes 2600 biparametric MRI scans with case-level annotations for the presence or absence of clinically significant prostate cancer, from the PI-CAI challenge[45]. As part of the PI-CAI study protocol (ClinicalTrials.gov identifier NCT05489341), patient data from January 2012 through December 2021 were retrospectively collected and deidentified from four European tertiary care centers. All patients were adult men (23 to 92 years of age) suspected of harboring prostate cancer, with elevated levels ($\geq$ 3 ng/mL) of prostate-specific antigen, or an abnormal digital rectal examination, or both. Insignificant cancer was defined as Gleason grade group 1 (Gleason score 6; low risk), while clinically significant cancer was defined as Gleason grade group 2 (Gleason score 3+4=7; favourable intermediate risk), 3 (Gleason score, 4+3=7; unfavourable intermediate risk), 4 (Gleason score 8; high risk) or 5 (Gleason score 9 or 10; very high risk). Case-level outcomes were by trained investigators, under the supervision of expert radiologists at each participating data center. Annotations were independently reviewed by the steering committee at the central coordinating center (Radboud University Medical Center) for quality control.

Dataset preparation

For each exam, all biparametric prostate MRI sequences (axial T2-weighted imaging, high b-value diffusion weighted imaging, apparent diffusion coefficient maps) were resampled to a common spatial resolution of $0.5 \times 0.5 \times 3.0$ mm^3 and center-cropped to $256 \times 256 \times 20$ voxels.

Training/test split

In accordance with the data split utilized in the PI-CAI challenge[45], the training dataset comprises of the 1500 MRI scans (1476 patients) that were publicly released. Similarly, 1100 unseen MRI scans (1100 patients) from the challenge are used for validation/tuning (n=100) and testing (n=1000).

5.3 Data Records

The data associated with this work is available on Zenodo [201–212]. The record consists of the training data for the eight tasks. The datasets' URLs on Zenodo are specified in Table 5.1. Each of these training datasets is uploaded to a separate Zenodo repository and follows a standardized structure to allow for easy development of universally applicable classification methods. This format was heavily based on

those introduced by the Medical Image Segmentation Decathlon [54] and nnUnet [189].

Each repository includes a `LICENSE.txt` file that contains the license of the corresponding dataset, a `dataset.json` file that provides a detailed description of the dataset, and input images which optionally include ROI segmentations.

The `dataset.json` file contains the following information:

- A short textual description of the dataset;
- References to earlier uses and releases of the data;
- The name of the license with which the data is released;
- References to earlier uses and/or releases of the data;
- The number of training images in the dataset;
- Descriptions of the different data acquisition methods used for collecting the dataset that are used in the input image filenames and identification numbers for each of these aquisition methods;
- Descriptions of the classification labels present in the dataset;
- Descriptions of Region of Interest (ROI) segmentation labels, when available;
- A list of all cases in the dataset, where each item describes relative paths to the input image files and their related classification labels.

All images and ROI segmentations are stored in the MetaImage format, with the .mha file extension, which is widely used in medical imaging for the storage and transfer of 2D and 3D images. Many of the datasets involved individual cases corresponding to multiple input images obtained with different acquisition methods. Within each such case, input images were co-registered and processed to ensure consistent spacing. Filenames of the ROI segmentations follow the format `CaseID.mha` and all otehr images follow the format `CaseID_AcquisitionNumber.mha`, where `CaseID` is an identifier that is consistent among each input image that corresponds to a single case, and `AcquisitionNumber` is the identification number described in the JSON file that corresponds to the data acquisition method with which the image was obtained.

A summary of the datasets is available in Table 5.2.

Classification Task	Dataset URL	Licence
Breast MRI molecular cancer subtype	[201]	CC BY-NC 4.0
Retina OCT glaucoma	[202]	CC BY-NC 4.0
Lung nodule CT false positive reduction	[203]	CC BY 4.0
Rib CT fracture	[204]	CC BY-NC 4.0
Clinically significant prostate cancer	[205]	CC BY-NC 4.0
	[206]	CC BY-NC 4.0
	[207]	CC BY-NC 4.0
Lung CT	[208]	CC BY-NC 4.0
COVID-19	[209]	CC BY-NC 4.0
	[210]	CC BY-NC 4.0
	[211]	CC BY-NC 4.0
Kidney CT Abnormality	[212]	CC BY-NC-SA 4.0

Table 5.1: Zenodo URL for each task's publicly available data set. The full licences can be found in these repositories. We strived to assemble a database from datasets that allow distribution under lenient licenses to promote the creation of widely accessible algorithms.

5.4 Technical Validation

To validate that all datasets are ready to be used for the development of general medical image classification methods, we trained a baseline system on the seven datasets and subsequently tested it against the corresponding test sets. The baseline system and its comprehensive source code were made publicly available at `https://github.com/DIAGNijmegen/auc23-baseline`. By publicly releasing this example codebase, we intent to reduce the technical complexity of using the released database with the creation of new universal medical image classification systems.

The baseline codebase implements a data preprocessing step and a model training step, based heavily on the nnU-Net codebase [189]. Because the nnU-Net framework was originally designed for semantic medical image segmentation, it was adapted to facilitate training a classifier. Most notably it was altered to train an Inflated 3D ConvNet model [16] instead of a U-Net [213]. To lower the memory footprint of training the baseline system, preprocessing was extended to include cropping of input image volumes to ROI segmentations when such segmentations are available.

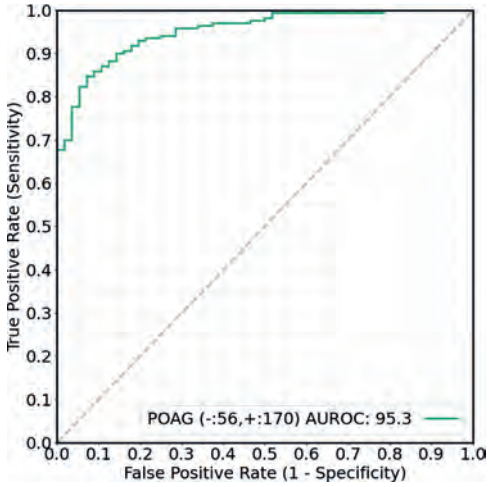
For each of the seven training datasets, a single I3D model was trained for 50 000 iterations. Training completed successfully for all tasks. Figure 5.2 shows the perfor-

Organ		Kidney	Prostate	Retina	Breast
Classification task		Abnormality (cyst or tumor)	Clinically significant cancer	POAG	Molecular cancer type
Provenance		[51, 186]	[45]	[50]	[47]
Total cases		586	1 500	884	737
Cases per class		176 Normal 410 Abnormal	1075 no csPCa 425 csPCa	207 Normal 677 POAG	46 HER2 481 Luminal A 76 Luminal B 134 TN
Acquisition methods		CT	T2 MRI, HBV MRI, ADC MRI	OCT	Subtraction MRI
Shape (voxels)	Axial	143.22 ± 54.76	20.00 ± 0.00	64.00 ± 0.00	168.26 ± 23.00
	Sagittal	146.24 ± 12.59	256.00 ± 0.00	64.00 ± 0.00	486.75 ± 45.18
	Coronal	150.27 ± 17.87	256.00 ± 0.00	128.00 ± 0.00	486.75 ± 45.18
Spacing (mm)	Axial	1.17 ± 0.38	3.00 ± 0.00	0.09 ± 0.00	1.07 ± 0.14
	Sagittal	0.76 ± 0.06	0.50 ± 0.00	0.09 ± 0.00	0.72 ± 0.11
	Coronal	0.76 ± 0.06	0.50 ± 0.00	0.02 ± 0.00	0.72 ± 0.11
ROI mask input		No	No	No	No

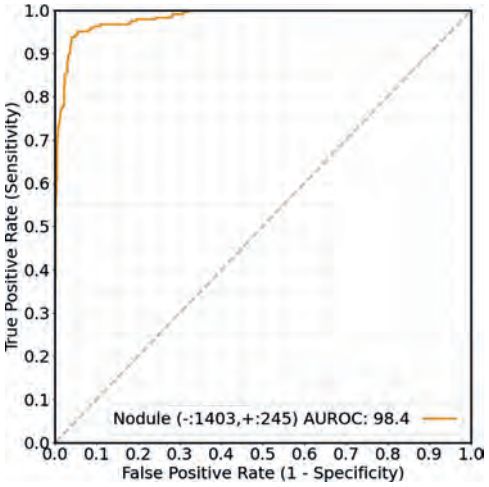
Organ		Rib	Lung	Lung
Classification task		Fracture type	Nodule false positive reduction	COVID-19
Provenance		[46]	[49]	[17, 48]
Total cases		1 324	6 694	2 000
Cases per class		230 Buckle 493 Displaced 460 Non-displaced 142 Segmental	941 Nodule 5753 No nodule	795 Negative 904 Non-severe 301 Severe
Acquisition methods		CT	CT	CT
Shape (voxels)	Axial	133.60 ± 15.99	64.00 ± 0.00	442.62 ± 124.48
	Sagittal	200.60 ± 20.17	64.00 ± 0.00	512.00 ± 0.00
	Coronal	200.60 ± 20.17	64.00 ± 0.00	512.00 ± 0.00
Spacing (mm)	Axial	1.14 ± 0.13	0.78 ± 0.00	0.80 ± 0.32
	Sagittal	0.75 ± 0.08	0.78 ± 0.00	0.73 ± 0.09
	Coronal	0.75 ± 0.08	0.78 ± 0.00	0.73 ± 0.09
ROI mask input		Yes	No	Yes

Table 5.2: Characteristics of the training datasets.

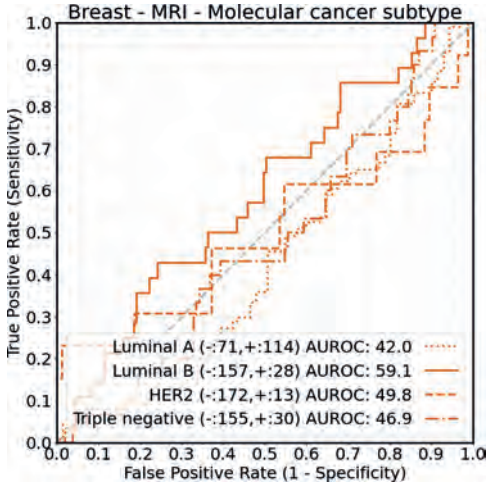
mance of the resulting systems on the corresponding test sets.



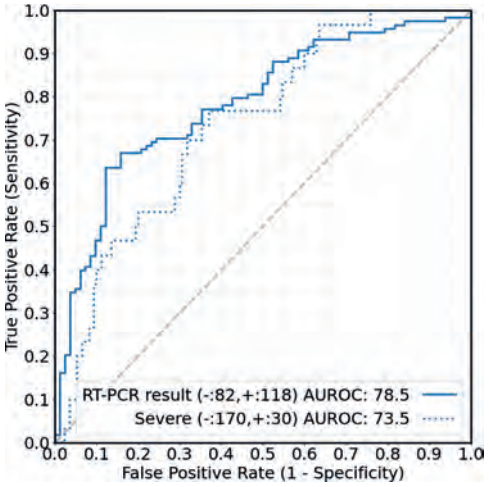
(a) Retina - OCT - Primary Open Angle Glaucoma (POAG)



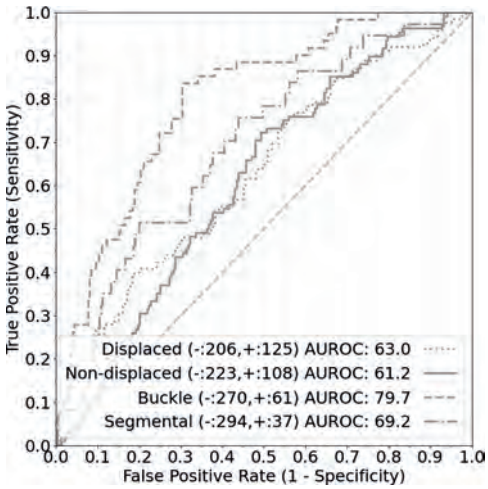
(b) Lung - CT - Nodule false positive reduction



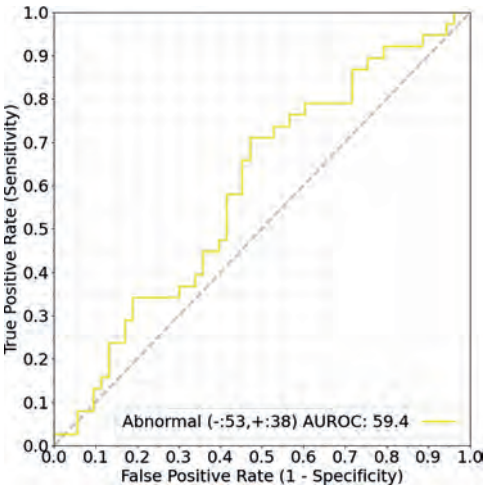
(c) Breast - MRI - Molecular cancer type



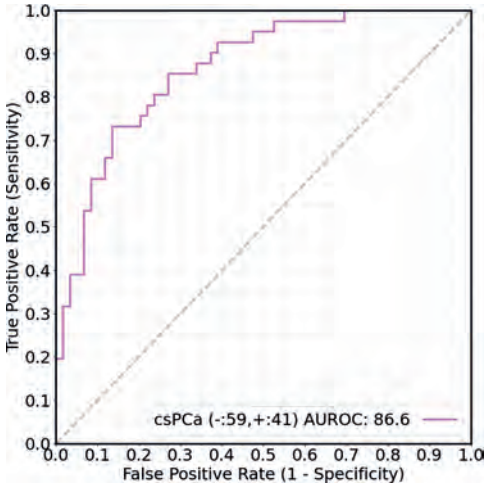
(d) Lung - CT - COVID-19 presence and severity



(e) Rib - CT - Fracture type



(f) Kidney - CT - Abnormality (tumor or cyst)



(g) Prostate - MRI - clinically significant prostate cancer (csPCa)

Figure 5.2: Receiver Operating Characteristics (ROC) curves of the baseline system for each of the tasks. Area under ROC curve (AUROC) values are computed separately for each class label. This was done by treating the class under consideration as the positive class and all other classes as negative.

5.5 Usage Notes

A tutorial for using the released database was released together with the baseline system. This tutorial includes an example of how to preprocess the training datasets and how to train a universal classification systems with them. At `https://auc23.grand-challenge.org/`, universal 3D medical image classification methods trained on each of the seven tasks can be evaluated.

For a fair comparison of universal 3D medical image classification systems, it is essential that they are evaluated using the same data and under identical conditions. With the release of our database, we therefore open a grand challenge with the goal of maintaining an up-to-date overview of the state-of-the art of universal classifiers.

At `https://auc23.grand-challenge.org/`, universal 3D medical image classification methods trained on each of the seven tasks can be freely evaluated. Trained solutions can be submitted to each of the seven tasks by submitting GitHub repositories containing codebases. To submit a solution codebase, developers can initially submit their trained models for evaluation on a select number of test cases from each test set, serving to verify model validity. Once successful verification across all seven tasks is achieved, users are then able to request the submission of their method for evaluation on the full test sets for all tasks. To encourage collaborative progress in the field, public release of the source code, under a permissive license, is a prerequisite before the final evaluation takes place.

For the grand challenge, metrics for each of the seven tasks are computed separately for each of the seven tasks. The metrics computed are Area Under the Receiver Operating Characteristic curve (AUROC), Cohen's kappa, and also quadratic wsevened kappa for classification tasks that contain ordinal output labels. For multi-class and multi-label tasks, AUROC values are computed separately for each class label. This is done by treating the class under consideration as the positive class and all other classes as negative. To obtain a single AUROC score per task, class-specific AUROC scores are aggregated through macro-averaging.

Code availability

Source code for reproducing all experiments in this data descriptor can be accessed at `https://github.com/DIAGNijmegen/auc23-baseline`. This codebase additionally contains a tutorial for submitting new universal 3D medical image classification methods to `https://auc23.grand-challenge.org`.

Discussion

This thesis presents several new machine learning approaches and datasets for the classification and regression from 3D medical image volumes. The corresponding tasks include patient and lobe-level lung function estimation, grading and classification of the likelihood and severity of COVID-19 infections, both using chest CT scans as input. Throughout this thesis, machine learning methods that are not restricted to specific 3D image modality and classification or regression tasks proved most successful.

This final chapter first reflects on the preceding chapters by discussing the different automation strategies, their clinical relevance, and implemented practices for accelerating research, including the use of different medical image analysis challenges. It concludes by discussing future research opportunities.

6.1 Automating classification and regression

This section describes different ways in which classification and regression were applied throughout this thesis, illustrated with examples from previous chapters. It describes several advantages and disadvantages of the strategies we employed.

6.1.1 As part of a pipeline

Automatic classification and regression tasks in medical image analysis are often components of a larger algorithm pipeline. More specifically, segmentation or detection steps were used prior to classification or regression to improve the efficiency and effectiveness of the presented methods in this thesis. Segmentation and detection require denser annotations than classification or regression tasks. Classification or regression models can only benefit from detection and segmentation models when relevant datasets with annotations are available to train such models, or when pretrained segmentation or detection models are available.

Segmentation and detection steps were employed for two purposes throughout this thesis, namely for limiting the size of the input volume, and providing additional information as input.

Limiting input volume size

Firstly, segmentation and detection were used to limit the input of the models to smaller regions of interest. This was done by cropping and/or masking the input volume. Presenting the memory-hungry 3D models with smaller inputs allows for more computationally efficient training and inference. Furthermore, excluding image regions that are not of interest from the input of the classification or regression

model may lead to better performance. This is firstly because of the increased signal-to-noise ratio in the input. Secondly, it also ensures that the model does not take into account nor becomes dependent on image features that the model should not focus on outside the region of interest.

All chapters of this thesis describe methods that utilize segmentation or detection to limit the input. In Chapter 2, input volumes were cropped to the lungs, and voxels outside of the lungs were masked out before presenting them to the models for lung function prediction. In Chapter 3, the models for COVID-19 grading relied on a lung segmentation for pre-processing. In Chapter 4, many challenge participants leveraged lung and segmentation algorithms in their training and inference pipelines for the same purpose. Chapter 5 formulated many classification tasks as part of a pipeline. The input volume for some tasks had already been cropped to the region of interest through the use of preceding detection or segmentation steps.

It is not a priori clear what the best size is of the region in the input volume to crop or mask to. This depends on how localized the information relevant for performing the classification or regression task is. For example, the information lung nodule malignancy classification task in Chapter 5 is highly localized. Cropping is performed to only a small area around the nodule. In contrast, the methodologies presented in Chapter 3 and 4 classification focus on the complete lungs because COVID-19 grading and classification requires the analysis of diffuse patterns.

Focusing only on specific parts of the image is an integral part of the methodology presented in Chapter 2. Here, the lobe level model relies heavily on a high-quality segmentation of the individual lobes. When analyzing an individual lobe to compute lobe-level lung function estimates, this allows the lobe-level model to ignore any image features present in other lobes.

Additional input

Secondly, segmentation and detection were used to extract relevant information from the input volumes. In Chapter 3, the added value of additional lesion segmentation input was investigated. In Chapter 4, the methodology of many participants relied on the output of a lesion segmentation. In Chapter 5, region of interest segmentations were provided as additional input for many of the tasks.

6.1.2 End-to-end learning

Instead of preceding the classification or regression model with preprocessing steps that extract the region of interest to be fed to the model, it is also possible to supply the model directly with the complete image. An example of this is the molecular

breast cancer type classification from Chapter 5. This approach is not possible when the methodology heavily relies on the partitioning of the image in regions of interest, as is the case with the lobe-level model presented in Chapter 2. Not using segmentation or detection models earlier in the processing pipeline also does not make the classification or regression model dependent on these preceding models. However, this approach is less computationally efficient and does not withhold the model from focusing on image features that are outside the region of interest.

6.1.3 Based on interpretable features

Classification or regression can also be performed by directly correlating the target output with features that are interpretable by humans. Taking a more classic approach to computer vision, these features can be handcrafted. Functional Pulmonary Volume (FPV) is an example of a highly interpretable hand-crafted feature that is predictive for lung function, as discussed in Chapter 2.

Another approach is to use interpretable image features extracted by (deep) learning algorithms. CORADS-AI [33] produced a COVID-19 CT severity score that was computed directly as the percentage of abnormal lung tissue within automatically generated lobe segmentation masks. Deep learning models can also be specifically engineered to provide such features. An example of this is BagNet [62] and the lobe level lung function prediction model presented in Chapter 2, where the architecture of the model forces it to produce lobe level lung function estimates that sum up to a patient level prediction.

A final approach for producing image-level output based on interpretable features is to use a combination of both handcrafted and automatically extracted features. In Chapter 4, the etro and uaux2 teams crafted features based on automatically extracted lesion segmentations.

6.1.4 Which strategy to use?

Methods based on hand-crafted features and/or segmentation masks alone are more interpretable than the output of current state-of-art machine learning models [214]. Current methods for improving the interpretability of machine learning models [214–216] do not yet alleviate this issue. Because of this, machine learning based 3D image classification and regression should be reserved for image assessment tasks where the classification or regression output cannot be directly extracted with sufficient accuracy from hand-crafted features or segmentation masks.

In general, when models for segmenting structures related to the task at hand are available, using them as part of the classification or regression pipeline is worth

considering. Automatically generated segmentation masks can be useful for cropping away or masking out image features that are outside the region of interest. This allows the classification or regression model to focus on the region of interest only. Using the segmentation mask or and-crafted features as additional input to the classification or regression model may also be useful, but does not always provide performance benefits.

Lastly, as also observed in preceding literature [55, 56, 217], a trend can be observed throughout this thesis that more general methods can provide strong performance. Task-specific algorithm improvements may in many cases only offer diminishing returns.

6.2 Clinical relevance

The methodology presented in this thesis may help healthcare professionals by saving them time and effort, as well as by providing them with additional insights that could be useful for clinical decision making.

6.2.1 Extending restricted healthcare capacity

The methods presented in this thesis may be particularly beneficial in situations that overload healthcare personnel and in a high-demand emergency situation like the COVID-19 pandemic. Automated medical image analysis might also prove useful for combating future large-scale disease outbreaks. Chapter 3 observed a lack of coordination in computer vision research on COVID-19 classification. Following the research practices described in section 6.3 may be vital for accelerating AI research in such situations.

Chapters 3 and 4 present methods that may be useful for computer-aided triaging of CT scans in the COVID-19 pandemic. These methods may be especially beneficial for alleviating personnel shortages in high-demand situations. Fortunately, COVID-19 infections have become much less prevalent and RT-PCR tests are available, providing an efficient and widespread means of diagnosing COVID-19.

Chapter 2 introduces a methodology for predicting lung function from CT scans, which, with further refinement, could potentially save time by eliminating the need for PFTs when a CT scan is readily available for a patient. Chapter 5 furthermore includes glaucoma and lung nodule malignancy prediction tasks that could provide aid in screening settings.

6.2.2 Providing additional insights

Automated classification and regression might be useful for tasks that are clinically relevant, but that medical professionals are not trained to perform. For such tasks, computers may aid clinical decision-making by offering additional perspectives. As for any machine learning problem, sufficient data that explicitly or implicitly describes the task is necessary for training such an algorithm.

In Chapter 2, machine learning models were tasked to extract information about lobe-level lung function that was implicitly available in the training set. Medical professionals are not trained to visually extract patient-level lung function estimates from CT, and even less so to estimate lobe-level lung function. When trained with sufficient data, machine learning models may be better suited than healthcare professionals to perform such a task. In such cases, a well-performing model could provide healthcare professionals with additional insights.

In Chapter 4, models were trained to predict one-month mortality or intubation risk for patients infected with COVID-19. Medical professionals are not explicitly trained to perform this task. Machine learning models could offer a way to make informed decisions in such a setting, such as prioritizing patients in saturated hospitals.

6.3 Accelerating research

This section describes methodologies utilized in this thesis that speed up progress in the field of computer-aided medical image analysis. These methodologies generalize to other fields for which machine learning applications are being developed, both related and unrelated to healthcare.

6.3.1 Open access

Providing open access to research artifacts increases transparency and accelerates research by allowing other researchers to replicate results and build upon these artifacts. The preceding chapters describe the public sharing of various algorithms, datasets, and codebases for these purposes.

All chapters of this thesis are linked to algorithms on grand-challenge.org that are available for public use. This includes the lobe level lung function prediction algorithm of Chapter 2, COVID-19 grading algorithms from Chapter 3, the baseline and finalist solutions of the STOIC2021 challenge described in Chapter 4, and the baseline algorithm from Chapter 5.

With the publication of the research presented in chapters 4, a database containing 2000 CT scans of patients suspected of COVID-19 was publicly released. Chapter 5 describes the public release of a structured database, designed to facilitate future research for developing universal classification methods. Furthermore, Chapter 3 reports performance on a public dataset, which allows other research to directly compare against the presented results

All code for training and running inference with the baseline and finalist solutions presented in 4 and the baseline presented in Chapter 5 was made publicly available.

6.3.2 Validation

Proper validation is crucial for the development of useful medical image analysis algorithms. Some good practices for validating and comparing models that were implemented in the preceding chapters are highlighted here.

Comparing methods on the same test set

For comparisons between different methods to be fair, they should be performed using the same metrics acquired through evaluation on the same test sets. Chapter 3 highlights that the research community cannot evaluate the performance of methods or model components when their evaluation on public benchmarks is not available. Chapters 3, 4, and 5 release or present public benchmarks with which the performance of the presented machine learning models is computed. This allows future research to compare newly developed methods directly against the methods presented in those chapters.

Evaluation on external or multi-center data

The generalization capability of machine learning models to unseen data should be evaluated as well for such models to be useful. Avoiding overlap of data from the same scans or patients in development and testing data is a first step for doing so. Overfitting models on other data specifics, such as scanner types, reconstruction kernels, and patient population, should also be avoided. Two strategies for measuring the generalization capabilities of machine learning models are evaluation on external data, and development and evaluation with large, heterogeneous data sets. These strategies were applied throughout this thesis where possible.

In Chapter 3, models were developed using data from the Radboud University Medical Center and evaluated using the external iCTCF dataset from China. Several

large databases containing data acquired from different scanners and medical centers were used in this thesis. COPDGene [77] were used for training and evaluating the models in Chapters 2. The STOIC database [48] was used for the research performed in Chapter 4, as well as for the COVID-19 classification task in Chapter 5. Furthermore, the PI-CAI database [45] was used for the prostate cancer classification task in Chapter 5. In all these applications, the databases were split into separate sets for development and testing.

6.3.3 Grand Challenges

Medical image analysis challenges have the potential of encouraging participants to follow the methodologies presented in the preceding sections 6.3.1 and 6.3.2. They provide important benchmarks for the community, allowing future publications to compare their approaches with other methodologies. Challenges furthermore provide clear directions through set objectives. Well-organized challenges can therefore be a powerful accelerator of research.

Medical image analysis challenges, in particular of Type Two (T2) and Type Three (T3), were used throughout this thesis to provide these benefits.

Type Two

In Chapter 3, a T2 challenge was opened for the systematic comparison between different methods, as well as to allow other researchers to compare their methods to the results in this chapter. In Chapter 4, the T2 challenge structure was used in the Qualification phase of the STOIC2021 challenge instead of the T3 structure to reduce the initial overhead for participants and to limit costs for the challenge organizers.

Type Three

This thesis does not only systematically compare different machine learning models but also discusses the evaluation of the methodologies for training such models through the T3 challenge format. In Chapter 4, reusable methodologies for training and inference are evaluated on their effectiveness in producing a well-performing machine learning model, given a dataset with a predetermined structure. More specifically, the finalists of the STOIC2021 challenge were evaluated on how well their methodology was able to produce a machine learning model for the task of COVID-19 severity classification. The T3 format requires that the codebase of participants is reusable, in the sense that it can be applied to any new dataset that follows the same predetermined dataset structure. This guarantees that the codebases can

be used to train models on any dataset that follows this structure, including private data sets and datasets constructed after the challenge has concluded.

The fact that the T3 challenge format requires participants to produce codebases for training models on any dataset of a predefined structure, makes the T3 challenge format well-suited for a head-to-head comparison of methods designed to work well for a variety of tasks. Chapter 5 presents a benchmark and challenge to obtain a universally applicable 3D classification method that performs well across varying medical image classification tasks. Following the Medical Image Segmentation Decathlon (MSD) [54], this chapter describes a diverse range of tasks, represented by medical imaging datasets, emphasizing the variety in modalities, anatomical regions, and diseases. A model that performs well on all of these tasks may also generalize to new unseen tasks. To evaluate this generalization capability, the MSD contained an evaluation phase after the development phase, in which datasets describing unseen tasks were released. The participants applied their methods to these datasets, and the predictions of the resulting models were evaluated.

Which challenge type to use?

Both T2 and T3 challenges were used throughout this thesis. The different challenge types each have their own advantages and disadvantages.

T1 challenges are the cheapest type of challenge to perform for challenge organizers. However, they are the least transparent and produce the fewest reusable research artifacts. T2 challenges produce more reusable research artifacts. They are more transparent and a fair way to compare different machine learning solutions. Because of this, it may benefit the research community when researchers adopt the T2 challenge structure where T1 is still being used.

When specifically searching for the most effective model for a specific task, T2 challenges may be preferred over T3 challenges. This is because T3 challenges provide much additional overhead and substantially raise costs for challenge organizers. The latter is especially true when considering compute-hungry models and large training datasets. To reduce costs, challenges may follow the hybrid format presented in Chapter 5, where the T2 structure is used for model development and the T3 structure is used to find the best-performing method.

T3 challenges should be used when the goal is to compare different training methodologies. They are especially useful for finding training methods that can be applied to a wide variety of tasks, including but not restricted to classification and regression tasks. T3 challenges may accelerate the development of universally applicable methodologies for computer-aided medical image analysis. T1 and T2 challenges have limitations for evaluating universally applicable methods because

they allow participants to potentially manually intervene in the training process.

6.4 Future outlook

6.4.1 Clinical adoption

Before the models presented in this thesis can be used for clinical adoption, several additional steps need to be taken.

Advancements in lobe level lung function prediction

The lung function and lobe level lung function prediction methods described in Chapter 2 may need to be more accurate before they become a viable alternative, or complementary to performing PFTs.

The methodology presented in Chapter 2 was initially developed for improved estimation of the risk of pulmonary resection, which is currently often estimated based on crude segment counting methods. This may be improved upon by utilizing the detailed information present in CT scans as done in this chapter.

In their presented form, the models trained in Chapter 2 only receive inspiration scans as input. As noted in Section 1.2.1, CT scans are often acquired at different inspiration and expiration levels than the levels achieved during PFTs. Providing models with CT scans taken at full inspiration with coaching similar to that applied during PFT measurements, and a combination of inspiratory and expiratory CT scans may also improve the accuracy of PFT estimation. A disadvantage of making the model reliant on both inspiratory and expiratory CT is that this will make the model inapplicable for patients where no expiratory CT scan is available.

PFTs are not perfectly reproducible, especially DLCO measurements [20]. To complement PFTs, CT features might be useful for relevant diffusion capacity estimates, and automatic medical image analysis could play a role in extracting such features.

It should be noted that post-operative PFT prediction may also never become perfectly accurate. This is firstly because PFTs themselves are not perfectly reproducible, but also because the risk estimation needs to be performed without knowledge about intra-operative decisions, complications during surgery, and patient recovery after surgery.

Clinical validation

None of the methods presented in this thesis have been sufficiently validated to be directly adopted in hospitals. Depending on the clinical setting in which the algo-

rithms are to be used, CT scans may image different patient populations and pathologies, and have different characteristics than those encountered in the training sets. For example, the COVID-19 models presented in Chapters 4 and 5 were developed on CT scans from patients suspected of an infection of COVID-19. If these models were to be used for analyzing standard radiology CT scans, their performance might potentially decline in such a different clinical context. Before any method can be used for clinical adoption, it needs to be certified for clinical use.

6.4.2 Type three challenges

The main drawbacks of T3 challenges are that they add additional overhead for challenge participants and that they substantially raise costs for challenge organizers. They furthermore separate the training data preparation process from the model development process. The latter limits the refinement of the training data through the use of model output. T2 challenges do not have these drawbacks and allow participants to provide feedback about the training data with challenge organizers.

This limits the widespread adoption of T3 challenges for task-specific challenges. In future endeavors, T3 challenges will shine where their benefits outweigh these downsides. They will be used for obtaining task-specific models on large private datasets, and for benchmarking the generalizability of universally applicable models to unseen tasks.

6.4.3 Universally applicable methods

This thesis highlights a trend toward universally applicable methods for training 3D medical image classification and regression models. The trend towards universally applicable methods is a general trend in AI research and can also be observed for medical image segmentation [54, 55]. The rise of universally applicable methods for classification and regression will make task-specific algorithm improvements less prevalent. Obtaining and refining datasets will remain important and may receive more attention from researchers.

Current developments in AI research also show a trend toward the use of base models [188, 218]. Base models are DL models that have been pretrained on vast datasets that encompass a wide range of information and context. Future base models trained on medical images will have a broad understanding of anatomy and disease patterns, enabling them to assist in accurate and efficient interpretation across a wide range of medical conditions and imaging modalities. Future universal methods built upon such base models will need much less task-specific data than state-of-the-art methods need now to achieve high performance on medical image classification

and regression tasks.

Overall, research in medical image analysis may be accelerated by prioritizing the development and benchmarking of universally applicable methods.

Summary

Chapter 1 introduces the most important concepts related to this thesis. It first provides a background on Deep Learning (DL) and Three-dimensional (3D) medical image processing. It then introduces the regression and classification applications this thesis mainly focuses on, namely thoracic computed tomography (CT) scans for automatic Pulmonary Function Test (PFT) estimation and COVID-19 classification. Lastly, it briefly explains the role of grand challenges in medical image analysis research.

In **Chapter 2**, a DL method is described for estimating global measures from an image that can also estimate the contributions of individual parts of the image to this global measure.

In an initial proof-of-concept, this chapter shows that a model trained with a collection of digit images to estimate their sum implicitly learns to assign correct values to individual digits. Next, it shows it is possible to estimate lobe-level quantities, such as COVID-19 severity scores, pulmonary volume, and functional pulmonary volume from CT while only provided with patient-level quantities during training.

Lastly, it is shown that the introduced DL approach can be used for estimating spirometry and diffusion capacity of carbon monoxide (DLCO) results from CT scans and disentangling the individual contribution of pulmonary lobes to a patient's lung function. The findings presented in this work may advance the use of CT in screening, diagnosis, and staging of restrictive pulmonary diseases as well as in risk assessment for pulmonary resection surgery and bronchoscopic lung volume reduction.

In **Chapter 3**, it is observed that applied artificial intelligence (AI) research focuses disproportionately on novel architecture modifications that do not necessarily generalize to other datasets, while neglecting systematic comparisons between commonly used algorithm components. This issue was especially prevalent in research on automated applications of AI for COVID-19 classification and grading from CT images with convolutional neural networks (CNNs).

Chapter 3 addresses this issue through a systematic investigation of COVID-19 grading algorithm components using a large publicly available dataset. The results are published in an online challenge. These contributions speed up the development of AI applications for COVID-19 grading by establishing insights into the components of such applications and by allowing applications produced by future research to be compared in a fair manner. The adherence to a standardized COVID-19 grading system may increase the compatibility between AI and clinical workflow.

Chapter 4 implements the Type Three (T3) challenge format for medical image analysis challenges. This format allows for training solutions on private data and guarantees that the training methods developed for the challenge can be used out-of-the box by third parties.

The T3 format was implemented in the STOIC2021 challenge, to predict from a CT scan whether subjects had a severe COVID-19 infection, defined as intubation or death within one month. STOIC2021 was implemented in two phases. It consisted of a Qualification phase, where participants developed challenge solutions using 2 000 publicly available CT scans, and a Final phase, where participants submitted their training methodologies with which solutions were trained on CT scans of 9 724 subjects.

The challenge organizers successfully trained six of the eight Final phase submissions. The submitted codebases for training and running inference were released publicly. The best-performing solutions were generally applicable deep learning approaches that made use of automatically generated segmentation masks.

Chapter 5 builds upon the observation made in the previous chapters that application-specific medical image classification solutions are often outperformed by generally applicable methods. To spur the development of universally applicable and publicly accessible medical image classification tools, the chapter presents a comprehensive database for the development of accessible and universally applicable 3D medical image classification algorithms. These datasets cover a broad spectrum of classification tasks containing images from multiple modalities and body parts. Each dataset is divided into a training and test set. The training sets are publicly released in a standardized format to foster algorithm development. Automatic evaluation of developed solutions on the test data is offered on the grand-challenge.org platform to ensure fair comparisons between methods.

To validate the readiness of the database for algorithm development, a baseline system was trained and tested successfully across all datasets. The baseline codebase is made publicly available along with a detailed tutorial on using the database for development and utilizing the automatic evaluation.

Chapter 6 first summarizes strategies for automating classification and regression and highlights how its findings may aid clinical decision-making. It then reflects on the practices employed in this thesis that are relevant for speeding up progress in the field of computer-aided medical image analysis. Finally, it provides a future outlook for research opportunities related to this thesis.

Samenvatting

Hoofdstuk 1 introduceert de concepten dit centraal staan in dit proefschrift. Het biedt eerst een achtergrond over Deep Learning (DL) en automatische driedimensionale (3D) medische beeldverwerking. Vervolgens introduceert het de regressie- en classificatietoepassingen waarop deze thesis voornamelijk is gericht, namelijk het verwerken van thoracale computertomografie (CT) scans om automatisch Pulmonale Functie Test (PFT) uitkomsten te bepalen en om COVID-19 patienten te classificeren. Tenslotte legt het kort de rol uit van competities in onderzoek over medische beeldanalyse.

In **Hoofdstuk 2** wordt een DL methode beschreven die zowel globale maten uit een afbeelding kan bepalen, als de bijdragen van individuele delen van de afbeelding aan deze globale maat kan bepalen.

Dit hoofdstuk toont eerst aan dat een model dat getraind om de som te bepalen van verzameling van afbeeldingen van cijfers, impliciet leert om correcte waarden toe te wijzen aan individuele cijfers. Vervolgens toont het aan dat het mogelijk is om van CT op kwab-niveau kwantiteiten te bepalen, zoals de ernst van een COVID-19 infectie, longvolume en functioneel longvolume van CT, terwijl alleen patiënt-niveau kwantiteiten worden gebruikt tijdens het trainen.

Ten slotte wordt aangetoond dat de geïntroduceerde DL methode kan worden gebruikt voor het bepalen van spirometrie en diffusiecapaciteit van koolmonoxide (DLCO) resultaten van CT-scans en het onderscheiden van de individuele bijdrage van longkwabben aan de longfunctie van een patiënt. De bevindingen die in dit werk worden gepresenteerd, kunnen het gebruik van CT bij screening, diagnose en stadiëring van restrictieve longziekten bevorderen, evenals bij risicobeoordeling voor longresectiechirurgie en bronchoscopische longvolumereductie.

In **Hoofdstuk 3** wordt waargenomen dat toegepast kunstmatige intelligentie (AI) onderzoek onevenredig veel aandacht besteedt aan nieuwe architectuurwijzigingen die niet noodzakelijkerwijs generaliseren naar andere datasets, terwijl systematische vergelijkingen tussen veelgebruikte algoritmecomponenten worden verwaarloosd. Dit probleem was vooral prevalent in onderzoek naar geautomatiseerde toepassingen van AI voor COVID-19-classificatie en -gradatie van CT-beelden met convolutionele neurale netwerken (CNNs).

Hoofdstuk 3 pakt dit probleem aan middels een systematisch onderzoek van COVID-19 gradatie algoritmecomponenten met behulp van een grote publiek beschikbare dataset. De resultaten van dit hoofdstuk zijn gepubliceerd in een online competitie. Deze bijdragen versnellen de ontwikkeling van AI-toepassingen voor COVID-19 gradatie door inzichten te verschaffen in de componenten van dergelijke

toepassingen en door toepassingen geproduceerd door toekomstig onderzoek op een eerlijke manier te kunnen vergelijken. Dit hoofdstuk gebruikt een gestandaardiseerd COVID-19 gradatiesysteem om de compatibiliteit tussen AI en klinische workflow vergroten.

Hoofdstuk 4 implementeert het Type Drie (T3) competitieformaat voor medische beeldanalyse uitdagingen. Dit formaat maakt training van oplossingen op privégegevens mogelijk en garandeert dat de trainingsmethoden ontwikkeld voor de competitie direct kunnen worden gebruikt door derden.

Het T3-formaat werd geïmplementeerd in de STOIC2021 competitie, om te voorspellen of proefpersonen een ernstige COVID-19 infectie hadden (patienten die binnen een maand moesten worden geintubeerd of overleden). De STOIC2021 competitie werd in twee fasen geïmplementeerd. Het bestond uit een Kwalificatiefase, waarbij competitiedeelnemers modellen ontwikkelden met behulp van 2.000 publiekelijk beschikbare CT-scans, en een Finale fase, waarbij deelnemers hun trainingsmethodologieën indienden waarmee modellen werden getraind met behulp van CT-scans van 9.724 patienten.

Zes van de acht inzendingen uit de Finale fase konden met succes worden getraind door de organisatoren van de competitie. De ingediende codebases voor training en het toepassen van de inzendingen op nieuwe CT scans werden publiekelijk vrijgegeven. De best presterende oplossingen waren algemeen toepasbare DL methoden die gebruik maakten van automatisch gegenereerde segmentaties.

Hoofdstuk 5 bouwt voort op bevinding uit de voorgaande hoofdstukken dat specifiek voor medische beeldclassificatie ontwikkelde oplossingen vaak worden overtroffen door algemeen toepasbare methoden. Om de ontwikkeling van universeel toepasbare en publiekelijk toegankelijke medische beeldclassificatiemiddelen te stimuleren, presenteert dit hoofdstuk een uitgebreide database voor de ontwikkeling van toegankelijke en universeel toepasbare 3D medische beeldclassificatie algoritmen. Deze datasets beschrijven een breed spectrum van classificatietaken met afbeeldingen van verschillende modaliteiten en lichaamsdelen. Elke dataset is verdeeld in een training- en testset. De trainingsets worden publiekelijk vrijgegeven in een gestandaardiseerd formaat om de ontwikkeling van algoritmen te bevorderen. Automatische evaluatie van ontwikkelde oplossingen op de evaluatie data wordt aangeboden op het grand-challenge.org platform om eerlijke vergelijkingen tussen methoden te verzekeren.

Om de gereedheid van de database voor algoritmeontwikkeling te valideren, werd een basissysteem succesvol getraind en getest op alle datasets. De basis-

systeemcode is publiekelijk beschikbaar gesteld samen met een gedetailleerde handleiding over het gebruik van de database voor ontwikkeling en het benutten van de automatische evaluatie.

Hoofdstuk 6 vat eerst strategieën samen voor het automatiseren van classificatie en regressie en belicht hoe de bevindingen kunnen helpen bij klinische besluitvorming. Vervolgens reflecteert het op de praktijken die in deze thesis zijn toegepast die relevant zijn voor het versnellen van de vooruitgang in het veld van computerondersteunde medische beeldanalyse. Tot slot biedt het een perspectief voor toekomstig onderzoek gerelateerd aan dit proefschrift.

Publications

Papers in international journals

N. Lessmann, C. I. Sanchez, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke, B. Geurts, H. A. Gietema, M. Groeneveld, L. van Harten, N. Hendrix, W. Hendrix, H. J. Huisman, I. Isgum, C. Jacobs, R. Kluge, M. Kok, J. Krdzalic, B. Lassen-Schmidt, K. van Leeuwen, J. Meakin, M. Overkamp, T. van Rees Vellinga, E. M. van Rikxoort, R. Samperna, C. Schaefer-Prokop, S. Schalekamp, E. T. Scholten, C. Sital, L. Stöger, J. Teuwen, K. Vaidhya Venkadesh, C. de Vente, M. Vermaat, W. Xie, B. de Wilde, M. Prokop, and B. van Ginneken. "Automated Assessment of CO-RADS and Chest CT Severity Scores in Patients with Suspected COVID-19 Using Artificial Intelligence". In: *Radiology* 298.1 (2021), E18–E28.

C. de Vente, L. H. Boulogne, K. V. Venkadesh, C. Sital, N. Lessmann, C. Jacobs, C. I. Sánchez, and B. van Ginneken. "Automated COVID-19 grading with convolutional neural networks in computed tomography scans: A systematic comparison". In: *IEEE transactions on artificial intelligence* 3.2 (2021), pp. 129–138.

E. Sogancioglu, K. Murphy, E. Th. Scholten, L. H. Boulogne, M. Prokop, and B. van Ginneken. "Automated estimation of total lung volume using chest radiographs and deep learning". In: *Medical Physics* 49.7 (2022), pp. 4466–4477.

L. H. Boulogne, J.-P. Charbonnier, C. Jacobs, E. H. van der Heijden, and B. van Ginneken. "Estimating lung function from computed tomography at the patient and lobe level using machine learning". In: *Medical Physics* 51.4 (2024), pp. 2834–2845.

L. H. Boulogne, J. Lorenz, D. Kienzle, R. Schon, K. Ludwig, R. Lienhart, S. Jegou, G. Li, C. Chen, Q. Wang, D. Shi, M. Maniparambil, D. Muller, S. Mertes, N. Schroter, F. Hellmann, M. Elia, I. Dirks, M. Nicolas Bossa, A. Diaz Berenguer, T. Mukherjee, J. Vandemeulebroucke, H. Sahli, N. Deligiannis, P. Gonidakis, N. Dung Huynh, I. Razzak, R. Bouadjenek, M. Verdicchio, P. Borrelli, M. Aiello, J. A. Meakin, A. Lemm, C. Russ, R. Ionasec, N. Paragios, B. van Ginneken, and M.-P. Revel. "The STOIC2021 COVID-19 AI challenge: applying reusable training methodologies to private data". In: *Medical Image Analysis* (2024), p. 103230.

Abstract in conference proceeding

L. H. Boulogne and B. van Ginneken. "Automatically Generated CT Severity Scores for COVID-19 Predict Death or Intubation at 1-Month Follow-Up". In: *Proceedings of the Annual Meeting of the Radiological Society of North America (RSNA)*. 2022.

Bibliography

[1] *Trends in aantal CT-onderzoeken - Trends en stand van zaken*. Accessed: 2024-03-24. National Institute for Public Health and the Environment (RIVM), 2023.

[2] *Echografie en MRI - Trends en stand van zaken*. Accessed: 2024-03-24. National Institute for Public Health and the Environment (RIVM), 2023.

[3] E. H. Dibble, E. Rubin, R. Duszak Jr, D. Parris, M. J. Tarrant, and J. R. Parikh. "The 2021 ACR/RBMA Workforce Survey: practice types, employment trends, and hiring needs". In: *Journal of the American College of Radiology* (2023).

[4] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[5] K. He, X. Zhang, S. Ren, and J. Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *International Conference on Computer Vision*. 2015, pp. 1026–1034.

[6] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Strom, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. van Boven, R. Vink, C. Hulsbergen-van de Kaa, J. van der Laak, M. B. Amin, A. J. Evans, T. van der Kwast, R. Allan, P. A. Humphrey, H. Gronberg, H. Samaratunga, B. Delahunt, T. Tsuzuki, T. Hakkinen, L. Egevad, M. Demkin, S. Dane, F. Tan, M. Valkonen, G. S. Corrado, L. Peng, C. H. Mermel, P. Ruusuvuori, G. Litjens, M. Eklund, A. Brilhante, A. Cakir, X. Farre, K. Geronatsiou, V. Molinie, G. Pereira, P. Roy, G. Saile, P. G. O. Salles, E. Schaafsma, J. Tschui, J. Billoch-Lima, E. M. Pereira, M. Zhou, S. He, S. Song, Q. Sun, H. Yoshihara, T. Yamaguchi, K. Ono, T. Shen, J. Ji, A. Roussel, K. Zhou, T. Chai, N. Weng, D. Grechka, M. V. Shugaev, R. Kiminya, V. Kovalev, D. Voynov, V. Malyshev, E. Lapo, M. Campos, N. Ota, S. Yamaoka, Y. Fujimoto, K. Yoshioka, J. Juvonen, M. Tukiainen, A. Karlsson, R. Guo, C.-L. Hsieh, I. Zubarev, H. S. T. Bukhar, W. Li, J. Li, W. Speier, C. Arnold, K. Kim, B. Bae, Y. W. Kim, H.-S. Lee, J. Park, and the PANDA challenge consortium. "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge". In: *Nature Medicine* (Jan. 2022).

[7] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225* (2017).

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86 (1998), pp. 2278–2324.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *arXiv:1512.03385* (2015).

[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 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[13] V. Cheplygina, M. de Bruijne, and J. P. Pluim. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis". In: *Medical image analysis* 54 (2019), pp. 280–296.

[14] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, and J. Xia. "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT". In: *Radiology* 296.2 (2020), pp. 65–72.

[15] M. Prokop, W. van Everdingen, T. van Rees Vellinga, J. Quarles van Ufford, L. Stoger, L. Beenen, B. Geurts, H. Gietema, J. Krdzalic, C. Schaefer-Prokop, B. van Ginneken, M. Brink, and COVID-19 Standardized Reporting Working Group of the Dutch Radiological Society. "CO-RADS - A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation". In: *Radiology* 296.2 (2020), E97–E104.

[16] J. Carreira and A. Zisserman. "Quo vadis, action recognition? a new model and the Kinetics Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[17] L. H. Boulogne, J. Lorenz, D. Kienzle, R. Schon, K. Ludwig, R. Lienhart, S. Jegou, G. Li, C. Chen, Q. Wang, D. Shi, M. Maniparambil, D. Muller, S. Mertes, N. Schroter, F. Hellmann, M. Elia, I. Dirks, M. Nicolas Bossa, A. Diaz Berenguer, T. Mukherjee, J. Vandemeulebroucke, H. Sahli, N. Deligiannis, P. Gonidakis, N. Dung Huynh, I. Razzak, R. Bouadjenek, M. Verdicchio, P. Borrelli, M. Aiello, J. A. Meakin, A. Lemm, C. Russ, R. Ionasec, N. Paragios, B. van Ginneken, and M.-P. Revel. "The STOIC2021 COVID-19 AI challenge: applying reusable training methodologies to private data". In: *Medical Image Analysis* (2024), p. 103230.

[18] E. Derom, C. Van Weel, G. Liistro, J. Buffels, T. Schermer, E. Lammers, E. Wouters, and M. Decramer. "Primary care spirometry". In: *European Respiratory Journal* 31.1 (2008), pp. 197–203.

[19] M. R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C. Van Der Grinten, P. Gustafsson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O. F. Pedersen, R. Pellegrino, R. Viegi, and J. Wanger. "Standardisation of spirometry". In: *European respiratory journal* 26.2 (2005), pp. 319–338.

[20] N. Macintyre, R. Crapo, G. Viegi, D. Johnson, C. Van der Grinten, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, P. Enright, P. Gustafsson, J. Hankinson, R. Jensen, R. McKay, M. R. Miller, D. Navajas, O. F. Pedersen, R. Pellegrino, and W. J. "Standardisation of the single-breath determination of carbon monoxide uptake in the lung". In: *European Respiratory Journal* 26.4 (2005), pp. 720–735.

[21] A. Brunelli, A. W. Kim, K. I. Berger, and D. J. Addrizzo-Harris. "Physiologic evaluation of the patient with lung cancer being considered for resectional surgery: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines". In: *Chest* 143.5 (2013), e166S–e190S.

[22] J. Flandes, F. J. Soto, R. Cordovilla, E. Cases, and J. Alfayate. "Bronchoscopic lung volume reduction". In: *Clinics in chest medicine* 39.1 (2018), pp. 169–180.

[23] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Zha, and Y. Yang. "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021). early access.

[24] B. Wang, S. Jin, Q. Yan, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng, W. Sun, L. Lan, W. Zhang, X. Mu, C. Shi, Z. Wang, J. Lee, Z. Jin, M. Lin, H. Jin, L. Zhang, J. Guo, B. Zhao, Z. Ren, S. Wang, W. Xu, X. Wang, J. Wang, Z. You, and J. Dong. "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system". In: *Applied Soft Computing* 98 (2021), p. 106897. ISSN: 1568-4946.

[25] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung, A. Bernheim, V. Mani, C. Calcagno, K. Li, S. Li, H. Shan, J. Lv, T. Zhao, J. Xia, and Y. Yang. "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19". In: *Nature Medicine* 26.8 (2020), pp. 1224–1228.

[26] O. Gozes, M. Frid-Adar, H. Greenspan, P. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel. "Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis". In: *Engineering* 6.10 (2020), pp. 1122–1129.

[27] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie. "COVID-CT-Dataset:a CT Scan Dataset about COVID-19". In: *arXiv preprint arXiv:2003.13865* (2020).

[28] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan. "Coronavirus Detection and Analysis on Chest CT with Deep Learning". In: *arXiv preprint arXiv:2004.02640* (2020).

[29] D. Singh, V. Kumar, and M. Kaur. "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks". In: *European Journal of Clinical Microbiology & Infectious Diseases* 39.7 (2020), pp. 1–11.

[30] X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi. "COVID-AL: The diagnosis of COVID-19 with deep active learning". In: *Medical Image Analysis* 68 (2020), p. 101913.

[31] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu, W. Gong, Y. Bai, S. Li, Z. Yongbei, L. Wang, and J. Tian. "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis". In: *European Respiratory Journal* 56.2 (2020).

[32] S. A. Harmon, T. H. Sanford, S. Xu, E. B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, M. Blain, M. Kassin, D. Long, N. Varble, S. Walker, A. Ierardi, E. Stellato, G. Plensich, and B. Turkbey. "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets". In: *Nature communications* 11.1 (2020), pp. 1–7.

[33] N. Lessmann, C. I. Sanchez, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke, B. Geurts, H. A. Gietema, M. Groeneveld, L. van Harten, N. Hendrix, W. Hendrix, H. J. Huisman, I. Isgum, C. Jacobs, R. Kluge, M. Kok, J. Krdzalic, B. Lassen-Schmidt, K. van Leeuwen, J. Meakin, M. Overkamp, T. van Rees Vellinga, E. M. van Rikxoort, R. Samperna, C. Schaefer-Prokop, S. Schalekamp, E. T. Scholten, C. Sital, L. Stöger, J. Teuwen, K. Vaidhya Venkadesh, C. de Vente, M. Vermaat, W. Xie, B. de Wilde, M. Prokop, and B. van Ginneken. "Automated Assessment of CO-RADS and Chest CT Severity Scores in Patients with Suspected COVID-19 Using Artificial Intelligence". In: *Radiology* 298.1 (2021), E18–E28.

[34] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng. "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT". In: *IEEE Transactions on Medical Imaging* 39.8 (2020), pp. 2615–2625.

[35] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song, F. Shi, H. Yuan, Y. Wei, X. Cao, Y. Gao, D. Wu, and Q. Wang. "Dual-Sampling Attention Network for Diagnosis of COVID-19 from Community Acquired Pneumonia". In: *IEEE Transactions on Medical Imaging* (2020).

[36] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, and D. Qian. "Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images". In: *IEEE Transactions on Medical Imaging* (2020).

[37] Y. Li, D. Wei, J. Chen, S. Cao, H. Zhou, Y. Zhu, J. Wu, L. Lan, W. Sun, T. Qian, K. Ma, H. Xu, and Y. Zheng. "Efficient and Effective Training of COVID-19 Classification Networks with Self-Supervised Dual-Track Learning to Rank". In: *IEEE Journal of Biomedical and Health Informatics* 24.10 (2020), pp. 2787–2797.

[38] Y.-M. Xu, T. Zhang, H. Xu, L. Qi, W. Zhang, Y.-D. Zhang, D.-S. Gao, M. Yuan, and T.-F. Yu. "Deep Learning in CT Images: Automated Pulmonary Nodule Detection for Subsequent Management Using Convolutional Neural Network". In: *Cancer Management and Research* 12 (2020), p. 2979.

[39] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai. "Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19". In: *Pattern Recognition* 114 (2021), p. 107848.

[40] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, and B. Xu. "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)". In: *European radiology* 31.8 (2021), pp. 1–9.

[41] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D.-C. Wang, L.-B. Shi, J. Mei, X.-L. Jiang, I. Pan, Q.-H. Zeng, P.-F. Hu, Y.-H. Li, F.-X. Fu, R. Huang, R. Sebro, and W.-H. Liao. "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT". In: *Radiology* 296.3 (2020), E156–E165.

[42] S. A. A. Ahmed, M. C. Yavuz, M. U. Sen, F. Gulsen, O. Tutar, B. Korkmazer, C. Samanci, S. Sirolu, R. Hamid, A. E. Eryurekli, T. Mammadov, and B. Yanikoglu. "COVID-19 Detection in Computed Tomography Images with 2D and 3D Approaches". In: *arXiv preprint arXiv:2105.08506* (2021).

[43] P. K. Chaudhary and R. B. Pachori. "FBSED based automatic diagnosis of COVID-19 using X-ray and CT images". In: *Computers in biology and medicine* 134 (2021), p. 104454.

[44] Z. Li, J. Zhang, B. Li, X. Gu, and X. Luo. "COVID-19 Diagnosis on CT Scan Images Using a Generative Adversarial Network and Concatenated Feature Pyramid Network with an Attention Mechanism". In: *Medical Physics* 48.8 (2021), pp. 4334–4349.

[45] A. Saha, J. J. Twilt, J. S. Bosma, B. van Ginneken, D. Yakar, M. Elschot, J. Veltman, J. Fütterer, M. de Rooij, and H. Huisman. *Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge*. 2022.

[46] L. Jin, J. Yang, K. Kuang, B. Ni, Y. Gao, Y. Sun, P. Gao, W. Ma, M. Tan, H. Kang, J. Chen, and M. Li. "Deep-Learning-Assisted Detection and Segmentation of Rib Fractures from CT Scans: Development and Validation of FracNet". In: *EBioMedicine* (2020).

[47] A. Saha, M. R. Harowicz, L. J. Grimm, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski. "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features". In: *British journal of cancer* 119.4 (2018), pp. 508–516.

[48] m.-p. Revel, S. Boussouar, C. Margerie-Mellon, I. Saab, T. Lapotre, D. Mompoint, G. Chassagnon, A. Milon, M. Lederlin, S. Bennani, S. Moliere, M.-P. Debray, F. Bompard, S. Dangeard, C. Hani, M. Ohana, S. Bommart, C. Jalaber, M. Hajjam, and H. Abdoul. "Study of thoracic CT in COVID-19: the STOIC project". In: *Radiology* 301.1 (2021), E361–E370.

[49] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. Gugten, P. Heng, B. Jansen, M. Kaste, V. Kotov, J. Lin, J. Manders, A. Sónora-Mengana, J. C. Naranjo, and E. Papavasileiou. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge". In: *Medical image analysis* 42 (2017), pp. 1–13.

[50] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, and R. Garnavi. "A feature agnostic approach for glaucoma detection in OCT volumes". In: *PloS one* 14.7 (2019), e0219126.

[51] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, C. Zhong, and C. Weight. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge". In: *Medical image analysis* 67 (2021), p. 101821.

[52] P. Bándi, O. Geessink, Q. Manson, M. van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Cetin, E. Halici, H. Jackson, R. Chen, F. Both, J. Franke, H. Kusters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens. "From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge". In: *IEEE Transactions on Medical Imaging* 38 (2 Feb. 2018), pp. 550–560.

[53] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34 (2015), pp. 1993–2024.

[54] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, AnnetteKopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, H. Huisman, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbelaez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, N. Kim, I. Kim, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng,

A. L. Simpson, L. Maier-Hein, and M. J. Cardoso. "The Medical Segmentation Decathlon". In: *arXiv preprint arXiv:2106.05735* (June 2021).

[55] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[56] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[57] M. Baumgartner, P. F. Jäger, F. Isensee, and K. H. Maier-Hein. "nnDetection: a self-configuring method for medical object detection". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 530–539.

[58] N. J. Gross. "The GOLD standard for chronic obstructive pulmonary disease". In: *American journal of respiratory and critical care medicine* 163.5 (2001), pp. 1047–1048.

[59] A. Brunelli, A. Charloux, C. T. Bolliger, G. Rocco, J.-P. Sculier, G. Varela, M. Licker, M. Ferguson, C. Faivre-Finn, R. M. Huber, E. M. Clini, T. Win, D. De Ruysscher, and L. Goldman. "ERS/ESTS clinical guidelines on fitness for radical therapy in lung cancer patients (surgery and chemo-radiotherapy)". In: *European Respiratory Journal* 34.1 (2009), pp. 17–41.

[60] A. T. Sengul, B. Sahin, C. Celenk, and A. Basoglu. "Postoperative lung volume change depending on the resected lobe". In: *The Thoracic and Cardiovascular Surgeon* 61.02 (2013), pp. 131–137.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[62] W. Brendel and M. Bethge. "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet". In: *International Conference on Learning Representations* (2019).

[63] N. Pawlowski and B. Glocker. "Is texture predictive for age and sex in brain MRI?" In: *arXiv preprint arXiv:1907.10961* (2019).

[64] I. Ilanchezian, D. Kobak, H. Faber, F. Ziemssen, P. Berens, and M. S. Ayhan. "Interpretable gender classification from retinal fundus images using BagNets". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 477–487.

[65] O. N. Hassan, M. J. Menten, H. Bogunovic, U. Schmidt-Erfurth, A. Lotery, and D. Rueckert. "Deep Learning Prediction Of Age And Sex From Optical Coherence Tomography". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 238–242.

[66] M. Paschali, M. F. Naeem, W. Simson, K. Steiger, M. Mollenhauer, and N. Navab. "Deep Learning Under the Microscope: Improving the Interpretability of Medical Imaging Neural Networks". In: *arXiv preprint arXiv:1904.03127* (2019).

[67] E. Sogancioglu, K. Murphy, E. Th. Scholten, L. H. Boulogne, M. Prokop, and B. van Ginneken. "Automated estimation of total lung volume using chest radiographs and deep learning". In: *Medical Physics* 49.7 (2022), pp. 4466–4477.

[68] M. Argus, C. Schaefer-Prokop, D. A. Lynch, and B. van Ginneken. "Function Follows Form: Regression from Complete Thoracic Computed Tomography Scans". In: *arXiv preprint arXiv:1909.12047* (2019).

[69] L. Fernández-Rodríguez, I. Torres, D. Romera, R. Galera, R. Casitas, E. Martínez-Cerón, P. Díaz-Agero, C. Utrilla, and F. García-Río. "Prediction of postoperative lung function after major lung resection for lung cancer using volumetric computed tomography". In: *The Journal of thoracic and cardiovascular surgery* 156.6 (2018), pp. 2297–2308.

[70] F. Liu, P. Han, G.-s. Feng, B. Liang, J. Xiao, Z.-l. Tian, and Z.-q. Lei. "Using quantitative CT to predict postoperative pulmonary function in patients with lung cancer." In: *Chinese medical journal* 118.9 (2005), pp. 742–746.

[71] H. Yabuuchi, S. Kawanami, T. Kamitani, M. Yonezawa, Y. Yamasaki, T. Yamanouchi, M. Nagao, T. Okamoto, and H. Honda. "Prediction of post-operative pulmonary function after lobectomy for primary lung cancer: a comparison among counting method, effective lobar volume, and lobar collapsibility using inspiratory/expiratory CT". In: *European journal of radiology* 85.11 (2016), pp. 1956–1962.

[72] C. T. Bolliger, C. Gückel, H. Engel, S. Stöhr, C. P. Wyser, A. Schoetzau, J. Habicht, M. Solèr, M. Tamm, and A. P. Perruchoud. "Prediction of functional reserves after lung resection: comparison between quantitative computed tomography, scintigraphy, and anatomy". In: *Respiration* 69.6 (2002), pp. 482–489.

[73] N. Sverzellati, A. Chetta, E. Calabrò, P. Carbognani, E. Internullo, D. Olivieri, and M. Zompatori. "Reliability of quantitative computed tomography to predict postoperative lung function in patients with chronic obstructive pulmonary disease having a lobectomy". In: *Journal of computer assisted tomography* 29.6 (2005), pp. 819–824.

[74] M.-T. Wu, H.-B. Pan, A. A. Chiang, H.-K. Hsu, H.-C. Chang, N.-J. Peng, P.-H. Lai, H.-L. Liang, and C.-F. Yang. "Prediction of postoperative lung function in patients with lung cancer: comparison of quantitative CT with perfusion scintigraphy". In: *American Journal of Roentgenology* 178.3 (2002), pp. 667–672.

[75] H. Nomori, Y. Cong, and H. Sugimura. "Systemic and regional pulmonary function after segmentectomy". In: *The Journal of thoracic and cardiovascular surgery* 152.3 (2016), pp. 747–753.

[76] Y.-C. Chang, C.-J. Yu, S.-C. Chang, J. R. Galvin, H.-M. Liu, C.-H. Hsiao, P.-H. Kuo, K.-Y. Chen, T. J. Franks, K.-M. Huang, and P.-C. Yang. "Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT". In: *Radiology* 236.3 (2005), pp. 1067–1075.

[77] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. "Genetic epidemiology of COPD (COPDGene) study design". eng. In: *COPD* 7 (2010), pp. 32–43.

[78] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken. "Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans". In: *IEEE Transactions on Medical Imaging* 39 (8 2020), pp. 2664–2675.

[79] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. "Transfusion: Understanding transfer learning for medical imaging". In: *arXiv preprint arXiv:1902.07208* (2019).

[80] D. S. Moore, G. P. McCabe, and B. A. Craig. *Introduction to the Practice of Statistics*. 6th ed. W. H. Freeman and Company, 2009.

[81] W. Yang, A. Sirajuddin, X. Zhang, G. Liu, Z. Teng, S. Zhao, and M. Lu. "The role of imaging in 2019 novel coronavirus pneumonia (COVID-19)". In: *European Radiology* 30.9 (2020), pp. 1–9.

[82] M. Barstugan, U. Ozkaya, and S. Ozturk. "Coronavirus (COVID-19) classification using CT images by machine learning methods". In: *arXiv:2003.09424* (2020).

[83] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, M. Zhang, L. Yu, F. Cheng, and H. Yu. "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography". In: *Scientific Reports* 10.1 (Nov. 2020), p. 19196. ISSN: 2045-2322.

[84] W. Ning, S. Lei, J. Yang, Y. Cao, P. Jiang, Q. Yang, J. Zhang, X. Wang, F. Chen, Z. Geng, L. Xiong, H. Zhou, Y. Guo, Y. Zeng, H. Shi, L. Wang, Y. Xue, and Z. Wang. "iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia". In: *Research Square* (2020).

[85] D. Di, F. Shi, F. Yan, L. Xia, Z. Mo, Z. Ding, F. Shan, B. Song, S. Li, Y. Wei, Y. Shao, M. Han, Y. Gao, H. Sui, and Y. Gao. "Hypergraph learning for identification of COVID-19 with CT imaging". In: *Medical Image Analysis* 68 (2020), p. 101910.

[86] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems*. 1990, pp. 396–404.

[87] S. Ji, W. Xu, M. Yang, and K. Yu. "3D convolutional neural networks for human action recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013), pp. 221–231.

[88] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems 25*. 2012, pp. 1097–1105.

[89] B. van Ginneken, S. G. Armato, B. de Hoop, S. van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. M. R. Schilham, A. Retico, M. E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Belloti, F. D. Carlo, R. Megna, S. Tangaro, L. Bolanos, P. Cerello, S. C. Cheran, E. L. Torres, and M. Prokop. "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study". In: *Medical Image Analysis* 14 (Dec. 2010), pp. 707–722.

[90] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud. "A Comprehensive Analysis of Deep Regression". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (2020), pp. 2065–2081.

[91] W. Ning, Y. Xue, S. Lei, L. Wang, Z. Wang, J. Yang, and Y. Zheng. *CT images and clinical features for COVID-19*. `http://ictcf.biocuckoo.cn/HUST-19.php`. Accessed: 2020-05-20.

[92] C. de Vente and L. H. Boulogne. *Grand Challenge - COVID-19 CT Classification challenge*. `https://covid19.grand-challenge.org/`. Accessed: 2020-09-01.

[93] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." In: *Nature medicine* 25 (6 2019), pp. 954–961. ISSN: 1546-170X.

[94] I. W. Harsono, S. Liawatimena, and T. W. Cenggoro. "Lung nodule detection and classification from thorax CT-scan using RetinaNet with transfer learning". In: *Journal of King Saud University-Computer and Information Sciences* (2020). to be published.

[95] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[96] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[97] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[98] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Thirty-first AAAI conference on artificial intelligence*. 2017, pp. 4278–4284.

[99] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. "Learning transferable architectures for scalable image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8697–8710.

[100] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[101] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in Neural Information Processing Systems*. 2019, pp. 3342–3352.

[102] A. Shoeibi, M. Khodatars, R. Alizadehsani, N. Ghassemi, M. Jafari, P. Moridian, A. Khadem, D. Sadeghi, S. Hussain, A. Zare, Z. Alizadehsani, J. Bazeli, F. Khozeimeh, A. Khosravi, S. Nahavandi, U. Acharya, and D. Srinivasan. "Automated Detection and Forecasting of COVID-19 Using Deep Learning Techniques: A Review". In: *arXiv preprint arXiv:2007.10785* (2020).

[103] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[104] C. de Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta. "Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-parametric MRI". In: *IEEE Transactions on Biomedical Engineering* (2020).

[105] T. Chen, B. Xu, C. Zhang, and C. Guestrin. "Training deep nets with sublinear memory cost". In: *arXiv preprint arXiv:1604.06174* (2016).

[106] C. M. Rutter. "Bootstrap estimation of diagnostic accuracy with patient-clustered data". In: *Academic Radiology* 7 (2000), pp. 413–419.

[107] F. Samuelson, N. Petrick, and S. Paquerault. "Advantages and examples of resampling for CAD evaluation". In: *IEEE International Symposium on Biomedical Imaging*. 2007, pp. 492–495.

[108] M. Aubreville, N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. Ter Hoeve, F. Ciompi, F. Wilm, C. Marzahl, T. A. Donovan, A. Maier, J. Breen, N. Ravikumar, Y. Chung, J. Park, R. Nateghi, F. Pourakpour, R. H. J. Fick, S. Ben Hadj, M. Jahanifar, A. Shephard, J. Dexl, T. Wittenberg, S. Kondo, M. W. Lafarge, V. H. Koelzer, J. Liang, Y. Wang, X. Long, J. Liu, S. Razavi, A. Khademi, S. Yang, X. Wang, R. Erber, A. Klang, K. Lipnik, P. Bolfa, M. J. Dark, G. Wasinger, M. Veta, and K. Breininger. "Mitosis domain generalization in histopathology images - The MIDOG challenge." In: *Medical Image Analysis* 84 (2022), p. 102699. ISSN: 1361-8423. aheadofprint.

[109] M. Schirmer, A. Venkataraman, I. Rekik, M. Kim, S. Mostofsky, M. B. Nebel, K. Rosch, K. Seymour, D. Crocetti, H. Irzan, M. Hütel, S. Ourselin, N. Marlow, A. Melbourne, E. Levchenko, S. Zhou, M. Kunda, H. Lu, N. Dvornek, and A. W. Chung. "Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge". In: *Medical image analysis* 70 (2021), p. 101972.

[110] Q. Da, X. Huang, Z. Li, Y. Zuo, C. Zhang, J. Liu, W. Chen, J. Li, D. Xu, Z. Hu, H. Yi, Y. Guo, Z. Wang, L. Chen, L. Zhang, X. He, X. Zhang, K. Mei, C. Zhu, and S. Zhang. "DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system". In: *Medical Image Analysis* 80 (2022), p. 102485.

[111] W. Ouyang, C. Winsnes, M. Hjelmare, A. Cesnik, L. Åkesson, H. Xu, D. Sullivan, S. Dai, J. Lan, P. Jinmo, S. M. Galib, C. Henkel, K. Hwang, D. Poplavskiy, B. Tunguz, R. Wolfinger, Y. Gu, C. Li, J. Xie, and E. Lundberg. "Analysis of the human protein atlas image classification competition". In: 16.12 (2019), pp. 1254–1261.

[112] H. Hassan, Z. Ren, H. Zhao, S. Huang, D. Li, S. Xiang, Y. Kang, S. Chen, and B. Huang. "Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks". In: *Computers in biology and medicine* 141 (2022), p. 105123.

[113] B. Ehteshami Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, the CAMELYON16 Consortium, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. Ahmady Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio. "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". In: *Journal of the American Medical Association* 318 (22 Dec. 2017), pp. 2199–2210.

[114] N. Lassau, I. Bousaid, E. Chouzenoux, J. Lamarque, B. Charmettant, M. Azoulay, F. Cotton, A. Khalil, O. Lucidarme, F. Pigneur, Y. Benaceur, A. Sadate, M. Lederlin, F. Laurent, G. Chassagnon, O. Ernst, G. Ferreti, Y. Diascorn, P. Brillet, and J. Meder. "Three artificial intelligence data challenges based on CT and MRI". In: *Diagnostic and Interventional Imaging* 101.12 (2020), pp. 783–788.

[115] H. Choi, H. Kim, K. Jin, Y. Jeong, K. Chae, K. Lee, H. S. Yong, B. Gil, H.-J. Lee, K. Lee, K.-N. Jeon, J. Yi, S. Seo, C. Ahn, J. Lee, K. Oh, and J. M. Goo. "A challenge for emphysema quantification using a deep learning algorithm with low-dose chest computed tomography". In: *Journal of Thoracic Imaging* 37.4 (2022), pp. 253–261.

[116] S. Halabi, L. Prevedello, J. Kalpathy-Cramer, A. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. Pereira, R. Sousa, N. Abdala, F. Kitamura, H. Thodberg, L. Chen, G. Shih, K. Andriole, M. Kohli, B. Erickson, and A. Flanders. "The RSNA pediatric bone age machine learning challenge". In: *Radiology* 290.2 (2019), pp. 498–503.

[117] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. Guo, B. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N. Nguyen, D. Tran, L. Huynh, N. Boutry, and J. Rittscher. "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy". In: *Medical image analysis* 70 (2021), p. 102002.

[118] F. Knoll, T. Murrell, A. Sriram, N. Yakubova, J. Zbontar, M. Rabbat, A. Defazio, M. Muckley, D. Sodickson, C. Zitnick, and M. Recht. "Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge". In: *Magnetic resonance in medicine* 84.6 (2020), pp. 3054–3070.

[119] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, L. Xinhui, L. Gao, T. Wu, J. Xiao, F. Wang, B. Yin, Y. Wang, G. Danala, L. He, Y. Choi, Y. C. Lee, and F. Meriaudeau. "Idrid: Diabetic retinopathy–segmentation and grading challenge". In: *Medical image analysis* 59 (2020), p. 101561.

[120] Y. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. Park, K. Lee, J. Kim, W. Hong, H. Jung, Y. Liu, H. Rajkumar, M. Khened, G. Krishnamurthi, S. Yang, X. Wang, C. Han, J. T. Kwak, and J. Choi. "PAIP 2019: Liver cancer segmentation challenge". In: *Medical Image Analysis* 67 (2021), p. 101854.

[121] H. Fang, F. Li, H. Fu, X. Sun, X. Cao, F. Lin, J. Son, S. Kim, G. Quellec, S. Matta, S. M S, Y.-T. Chen, C.-h. Wang, N. Shah, C.-Y. Lee, C.-C. Hsu, H. Xie, B. Lei, U. Baid, and Y. Xu. "ADAM challenge: detecting age-related macular degeneration from fundus images". In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2828–2847.

[122] Y. Sun, K. Gao, Z. Wu, G. Li, X. Zong, Z. Lei, Y. Wei, J. Ma, X. Yang, X. Feng, L. Zhao, T. Phan, J. Shin, T. Zhong, Y. Zhang, L. Yu, C. Li, R. Basnet, M. O. Ahmad, and L. Wang. "Multi-site infant brain segmentation algorithms: the iSeg-2019 challenge". In: *IEEE Transactions on Medical Imaging* 40.5 (2021), pp. 1363–1376.

[123] N. Sathianathen, N. Heller, R. Tejpaul, B. Stai, A. Kalapara, J. Rickman, J. Dean, M. Oestreich, P. Blake, H. Kaluzniak, S. Raza, J. Rosenberg, K. Moore, E. Walczak, Z. Rengel, Z. Edgerton, R. Vasdev, M. Peterson, S. McSweeney, and C. Weight. "Automatic Segmentation of Kidneys and Kidney Tumors: The KiTS19 International Challenge". In: *Frontiers in Digital Health* 3 (2022), p. 797607.

[124] M. Combalia, N. Codella, V. Rotemberg, C. Carrera, S. Dusza, D. Gutman, B. Helba, H. Kittler, N. Kurtansky, K. Liopyris, M. Marchetti, S. Podlipnik, S. Puig, C. Rinner, P. Tschandl, J. Weber, A. Halpern, and J. Malvehy. "Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge". In: *The Lancet Digital Health* 4.5 (2022), e330–e339.

[125] A. Kavur, N. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. Pham, S. Chatterjee, P. Ernst, S. Ozkan, B. Baydar, D. Lachinovk, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, and M. Selver. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation". In: *Medical Image Analysis* 69 (2021), p. 101950.

[126] A. Hakim, S. Christensen, S. Winzeck, M. G. Lansberg, M. W. Parsons, C. Lucas, D. Robben, R. Wiest, M. Reyes, and G. Zaharchuk. "Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the ISLES challenge". In: *Stroke* 52.7 (2021), pp. 2328–2337.

[127] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, F. Hou, J. Yang, G. Xiong, J. Tian, C. Zhong, J. Ma, J. Rickman, J. Dean, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, H. Kaluzniak, S. Raza, J. Rosenberg, K. Moore, E. Walczak, Z. Rengel, Z. Edgerton, R. Vasdev, M. Peterson, S. McSweeney, S. Peterson, A. Kalapara, N. Sathianathen, C. Weight, and N. Papanikolopoulos. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge". In: (Dec. 2, 2019). arXiv: `http://arxiv.org/abs/1912.01054v1 [eess.IV]`.

[128] H. Bogunovic, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, K. Gopinath, A. K. Gostar, K. Jeon, Z. Ji, S. H. Kang, D. D. Koozekanani, D. Lu, D. Morley, K. K. Parhi, H. S. Park, A. Rashno, M. Sarunic, S. Shaikh, J. Sivaswamy, R. Tennakoon, S. Yadav, S. De Zanet, S. M. Waldstein, B. S. Gerendas, C. Klaver, C. I. Sánchez, and U. Schmidt-Erfurth. "RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge". In: *Transactions on Medical Imaging* 38 (8 Aug. 2019), pp. 1858–1874.

[129] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, J. Lee, X. Li, P. Liu, S. Lu, B. Murugesan, V. Naranjo, S. Phaye, S. M S, A. Sikka, and H. Bogunović. "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs". In: *Medical image analysis* 59 (2020), p. 101570.

[130] J. Yang, H. Veeraraghavan, S. Armato III, K. Farahani, J. Kirby, J. Kalpathy-Kramer, W. Elmpt, A. Dekker, X. Han, X. Feng, P. Aljabar, B. Oliveira, B. Van der Heyden, L. Zamdborg, D. Lam, M. Gooding, and G. Sharp. "Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017". In: *Medical Physics* 45.10 (2018), pp. 4568–4581.

[131] J. Hirvasniemi, J. Runhaar, R. van der Heijden, M. Zokaeinikoo, M. Yang, X. Li, J. Tan, H. Rajamohan, Y. Zhou, C. Deniz, F. Caliva, C. Iriondo, J. Lee, F. Liu, A. Morales Martinez, N. Namiri, V. Pedoia, E. Panfilov, N. Bayramoglu, and S. Klein. "The KNee OsteoArthritis Prediction (KNOAP2020) challenge: An image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images". In: *Osteoarthritis and Cartilage* 31.1 (2023), pp. 115–125.

[132] I. Arganda-Carreras, S. C. Turaga, D. R. Berger, D. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J. M. Buhmann, T. Liu, M. Seyedhosseini, T. Tasdizen, L. Kamentsky, R. Burget, V. Uher, X. Tan, C. Sun, T. Pham, and H. Seung. "Crowdsourcing the creation of image segmentation algorithms for connectomics". In: *Frontiers in neuroanatomy* 9 (2015), p. 142.

[133] M. Ivantsits, L. Goubergrits, J.-M. Kuhnigk, M. Huellebrand, J. Bruening, T. Kossen, B. Pfahringer, J. Schaller, A. Spuler, T. Kuehne, Y. Jia, X. Li, S. Shit, B. Menze, Z. Su, J. Ma, Z. Nie, K. Jain, Y. Liu, and A. Hennemuth. "Detection and analysis of cerebral aneurysms based on X-ray rotational angiography-the CADA 2020 challenge". In: *Medical image analysis* 77 (2022), p. 102333.

[134] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. H. Rohban, S. Singh, and A. Carpenter. "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl". In: *Nature methods* 16.12 (2019), pp. 1247–1253.

[135] M. Simões, D. Borra, E. Santamaría-Vázquez, GBT-UPM, M. Bittencourt-Villalpando, D. Krzemiński, A. Miladinović, Neural_Engineering_Group, T. Schmid, H. Zhao, C. Amaral, B. Direito, J. Henriques, P. de Carvalho, and M. Castelo-Branco. "BCIAUT-P300: A multi-session and multi-subject benchmark dataset on autism for P300-based brain-computer-interfaces". In: *Frontiers in Neuroscience* 14 (2020), p. 568104.

[136] M. Veta, Y. J. Heng, N. Stathonikos, B. E. Bejnordi, F. Beca, T. Wollmann, K. Rohr, M. A. Shah, D. Wang, M. Rousson, M. Hedlund, D. Tellez, F. Ciompi, E. Zerhouni, D. Lanyi, M. Viana, V. Kovalev, V. Liauchuk, H. A. Phoulady, T. Qaiser, S. Graham, N. Rajpoot, E. Sjoblom, J. Molin, K. Paeng, S. Hwang, S. Park, Z. Jia, E. I.-C. Chang, Y. Xu, A. H. Beck, P. J. van Diest, and J. P. W. Pluim. "Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge". In: *Medical Image Analysis* 54.5 (May 2019), pp. 111–121.

[137] S. Winzeck, A. Hakim, R. McKinley, J. A. Pinto, V. Alves, C. Silva, M. Pisov, E. Krivov, M. Belyaev, M. Monteiro, A. Oliveira, Y. Choi, M. Paik, Y. Kwon, H. Lee, B. Kim, J.-H. Won, M. Islam, and M. Reyes. "ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI". In: *Frontiers in neurology* 9 (2018), p. 679.

[138] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, P. Golland, S. Klein, and D. Alexander. "TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data". In: *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*. Springer. 2019, pp. 1–10.

[139] Y. Balagurunathan, A. Beers, M. Mcnitt-Gray, L. Hadjiiski, S. Napel, D. Goldgof, G. Perez, P. Arbelaez, A. Mehrtash, T. Kapur, E. Yang, J. Moon, G. Bernardino, R. Delgado, M. Farhangi, A. Amini, R. Ni, X. Feng, A. Bagari, and K. Farahani. "Lung nodule malignancy prediction in sequential ct scans: Summary of isbi 2018 challenge". In: *IEEE transactions on medical imaging* 40.12 (2021), pp. 3748–3761.

[140] A. De Luca, A. Ianus, A. Leemans, M. Palombo, N. Shemesh, H. Zhang, D. C. Alexander, M. Nilsson, M. Froeling, G.-J. Biessels, M. Zucchelli, M. Frigo, E. Albay, S. Sedlar, A. Alimi, S. Deslauriers-Gauthier, R. Deriche, R. Fick, M. Afzali, and K. Schilling. "On the generalizability of diffusion MRI signal representations across acquisition parameters, sequences and tissue types: Chronicles of the MEMENTO challenge". In: *NeuroImage* 240 (2021), p. 118367.

[141] L. A. Bratholm, W. Gerrard, B. Anderson, S. Bai, S. Choi, L. Dang, P. Hanchar, A. Howard, S. Kim, Z. Kolter, R. Kondor, M. Kornbluth, Y. Lee, Y. Lee, J. Mailoa, T. Nguyen, M. Popovic, G. Rakocevic, W. Reade, and D. Glowacki. "A community-powered search of machine learning strategy space to find NMR property prediction models". In: *Plos one* 16.7 (2021), e0253612.

[142] E. E. Bron, M. Smits, W. M. Van Der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. M. Orellana, R. Meijboom, M. Pinto, J. Meireles, C. Garrett, A. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cardenas, and S. Klein. "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge". In: *NeuroImage* 111 (2015), pp. 562–579.

[143] A. A. A. Setio, A. Traverso, T. de Bel, M. S. N. Berens, C. v. d. Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. v. d. Gugten, P. A. Heng, B. Jansen, M. M. J. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. M. C. Manders, A. Sonora-Mengana, J. C. Garcia-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. A. Zuidhof, B. v. Ginneken, and C. Jacobs. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge". In: *Medical Image Analysis* 42 (Dec. 2017), pp. 1–13.

[144] I. Pan, H. H. Thodberg, S. S. Halabi, J. Kalpathy-Cramer, and D. B. Larson. "Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge". In: *Radiology: Artificial Intelligence* 1.6 (2019), e190053.

[145] D. M. Cash, C. Frost, L. O. Iheme, D. Ünay, M. Kandemir, J. Fripp, O. Salvado, P. Bourgeat, M. Reuter, B. Fischl, M. Lorenzi, G. Frisoni, X. Pennec, R. Pierson, J. Gunter, M. Senjem, C. Jack, N. Guizard, V. Fonov, and S. Ourselin. "Assessing atrophy measurement techniques in dementia: Results from the MIRIAD atrophy challenge". In: *Neuroimage* 123 (2015), pp. 149–164.

[146] Y.-G. Kim, I. H. Song, H. Lee, S. Kim, D. H. Yang, N. Kim, D. Shin, Y. Yoo, K. Lee, D. Kim, H. Jung, H. Cho, H. Lee, T. Kim, J. Choi, C. Seo, S. Han, Y. Lee, Y. Lee, and G. Gong. "Challenge for diagnostic assessment of deep learning algorithm for metastases classification in sentinel lymph nodes on frozen tissue section digital slides in women with breast cancer". In: *Cancer Research and Treatment: Official Journal of Korean Cancer Association* 52.4 (2020), pp. 1103–1111.

[147] Q. C. 2. O. Committee, B. Bilgic, C. Langkammer, J. P. Marques, J. Meineke, C. Milovic, and F. Schweser. "QSM reconstruction challenge 2.0: Design and report of results". In: *Magnetic Resonance in Medicine* 86.3 (2021), pp. 1241–1255.

[148] H. Fu, F. Li, X. Sun, X. Cao, J. Liao, J. I. Orlando, X. Tao, Y. Li, S. Zhang, M. Tan, C. Yuan, C. Bian, R. Xie, J. Li, X. Li, J. Wang, L. Geng, P. Li, H. Hao, and Y. Xu. "Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography". In: *Medical Image Analysis* 66 (2020), p. 101798.

[149] A. Babier, B. Zhang, R. Mahmood, K. L. Moore, T. G. Purdie, A. L. McNiven, and T. C. Chan. "OpenKBP: the open-access knowledge-based planning grand challenge and dataset". In: *Medical Physics* 48.9 (2021), pp. 5549–5561.

[150] D. Sun, T. M. Nguyen, R. J. Allaway, J. Wang, V. Chung, V. Y. Thomas, M. Mason, I. Dimitrovsky, L. Ericson, H. Li, Y. Guan, A. Israel, A. Olar, B. Pataki, G. Stolovitzky, J. Guinney, P. Gulko, M. Frazier, J. Chen, and X. He. "A Crowdsourcing Approach to Develop Machine Learning Models to Quantify Radiographic Joint Damage in Rheumatoid Arthritis". In: *JAMA network open* 5.8 (2022), e2227423–e2227423.

[151] M. Hatt, B. Laurent, A. Ouahabi, H. Fayad, S. Tan, L. Li, W. Lu, V. Jaouen, C. Tauber, J. Czakon, F. Drapejkowski, W. Dyrka, S. Camarasu-Pop, F. Cervenansky, P. Girard, T. Glatard, M. Kain, Y. Yao, C. Barillot, and D. Visvikis. "The first MICCAI challenge on PET tumor segmentation". In: *Medical image analysis* 44 (2018), pp. 177–195.

[152] T. Schaffter, D. S. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, W. Lotter, Z. Jie, H. Du, S. Wang, J. Feng, M. Feng, H.-E. Kim, K. Albiol, A. Albiol, S. Morrell, Z. Wojna, M. Ahsen, U. Asif, and H. Jung. "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms". In: *JAMA network open* 3.3 (2020), e200265–e200265.

[153] A. Reinke, M. D. Tizabi, C. H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. Blaschko, F. Buettner, M. J. Cardoso, J. Chen, V. Cheplygina, E. Christodoulou, B. Cimini, G. S. Collins, S. Engelhardt, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, B. Glocker, P. Godau, R. Haase, F. Hamprecht, D. A. Hashimoto, D. Heckmann-Nötzel, P. Hirsch, M. M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, A. E. Kavur, H. Kenngott, J. Kleesiek, A. Kleppe, S. Kohler, F. Kofler, A. Kopp-Schneider, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M. A. Noyan, J. Petersen, G. Polat, S. M. Rafelski, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C. I. Sánchez, J. Schroeter, A. Saha, M. A. Selver, L. Sharan, S. Shetty, M. van Smeden, B. Stieltjes, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. V. Calster, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, P. Jäger, and L. Maier-Hein. "Common limitations of image processing metrics: A picture story". In: *arXiv preprint arXiv:2104.05642* (2021).

[154] D. Merkel. *Docker: Lightweight Linux containers for consistent development and deployment*. `https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf`. Accessed 2022 Dec 5. 2014.

[155] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." eng. In: *Biometrics* 44 (1988), pp. 837–845.

[156] X. Sun and W. Xu. "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves". In: *IEEE Signal Processing Letters* 21.11 (2014), pp. 1389–1393.

[157] D. S. Moore and G. P. McCabe. *Introduction to the Practice of Statistics.* WH Freeman/Times Books/Henry Holt & Co, 1989.

[158] I. Loshchilov and F. Hutter. "Fixing weight decay regularization in Adam". In: *Proceedings of the ICLR 2018 Conference Blind*. CoRR, abs/1711.05101. 2018.

[159] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs. "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem". In: *European Radiology Experimental* 4.1 (2020), pp. 1–13.

[160] D. Müller, I. Soto-Rey, and F. Kramer. "Robust chest CT image segmentation of COVID-19 lung infection based on limited data". In: *Informatics in medicine unlocked* 25 (2021), p. 100681.

[161] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

[162] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. Le, and H. Adam. "Searching for mobilenetv3". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[163] NVIDIA NGC Catalog. *Clara_train_covid19_ct_lesion_seg*. `https://catalog.ngc.nvidia.com/orgs/nvidia/models/clara_train_covid19_ct_lesion_seg`. 2023.

[164] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. "Imagenet large scale visual recognition challenge". In: *International Journal of Computer Vision* 115.3 (2014), pp. 1–42.

[165] H. R. Roth, Z. Xu, C. Tor-Díez, R. S. Jacob, J. Zember, J. Molto, W. Li, S. Xu, B. Turkbey, E. Turkbey, D. Yang, A. Harouni, N. Rieke, S. Hu, F. Isensee, C. Tang, Q. Yu, J. Sölter, T. Zheng, and M. G. Linguraru. "Rapid artificial intelligence solutions in a pandemic—The COVID-19-20 Lung CT Lesion Segmentation Challenge". In: *Medical image analysis* 82 (2022), p. 102605.

[166] P. An, S. Xu, S. Harmon, E. Turkbey, T. Sanford, A. Amalou, M. Kassin, N. Varble, M. Blain, V. Anderson, F. Patella, G. Carrafiello, B. T. Turkbey, and B. J. Wood. *Ct images in covid-19. the cancer imaging archive*. 2020.

[167] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". In: *Journal of Digital Imaging* 26.6 (July 2013), pp. 1045–1057.

[168] S. P. Morozov, A. Andreychenko, N. Pavlov, A. Vladzymyrskyy, N. Ledikhova, V. Gombolevskiy, I. A. Blokhin, P. Gelezhe, A. Gonchar, and V. Y. Chernina. "Mosmeddata: Chest ct scans with covid-19 related findings dataset". In: *arXiv preprint arXiv:2005.06465* (2020).

[169] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. "Unified perceptual parsing for scene understanding". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 418–434.

[170] D. Kienzle, J. Lorenz, R. Schön, K. Ludwig, and R. Lienhart. "COVID detection and severity prediction with 3D-ConvNeXt and custom pretrainings". In: *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer. 2023, pp. 500–516.

[171] N. Lassau, S. Ammari, E. Chouzenoux, H. Gortais, P. Herent, M. Devilder, S. Soliman, O. Meyrignac, M.-P. Talabard, J.-P. Lamarque, R. Dubois, N. Loiseau, P. Trichelair, E. Bendjebbar, G. Garcia, C. Balleyguier, M. Merad, A. Stoclin, S. Jegou, and M. Blum. "Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients". In: *Nature communications* 12.1 (2021), pp. 1–11.

[172] S. Jégou. *Scancovia repository*. Accessed: 2022-12-20. 2022.

[173] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. "ibot: Image bert pre-training with online tokenizer". In: *arXiv preprint arXiv:2111.07832* (2021).

[174] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. "Models genesis: Generic autodidactic models for 3d medical image analysis". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 384–393.

[175] D. Müller and F. Kramer. *AUCMEDI - A Framework for Automated Classification of Medical Images*. Accessed: May 12th, 2022. 2022.

[176] D. Yamada, S. Ohde, R. Imai, K. Ikejima, M. Matsusako, and Y. Kurihara. "Visual classification of three computed tomography lung patterns to predict prognosis of COVID-19: a retrospective study". In: *BMC Pulmonary Medicine* 22 (2022), pp. 1–9.

[177] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal Loss for Dense Object Detection". In: (Aug. 7, 2017). arXiv: `http://arxiv.org/abs/1708.02002v2 [cs.CV]`.

[178] D. Müller and F. Kramer. "MIScnn: a framework for medical image segmentation with convolutional neural networks and deep learning". In: *BMC medical imaging* 21.1 (2021), pp. 1–11.

[179] *MONAI Documentation - DynUNet*. Accessed: May 12th, 2022. 2022.

[180] F. Isensee, J. Petersen, S. A. A. Kohl, P. F. Jäger, and K. H. Maier-Hein. "nnU-Net: Breaking the Spell on Successful Medical Image Segmentation". In: *arXiv:1904.08128* (Apr. 17, 2019). arXiv: `http://arxiv.org/abs/1904.08128v1 [cs.CV]`.

[181] *COVID-19 Lung CT Lesion Segmentation Grand Challenge*. Accessed: May 12th, 2022. 2022.

[182] *iCOVID AI*. Accessed: May 12th, 2022. 2022.

[183] *COPLE-Net*. Accessed: May 12th, 2022. 2022.

[184] J. Ma, C. Ge, Y. Wang, X. An, J. Gao, Z. Yu, M. Zhang, X. Liu, X. Deng, S. Cao, H. Wei, S. Mei, X. Yang, Z. Nie, C. Li, L. Tian, Y. Zhu, Q. Zhu, G. Dong, and J. He. *COVID-19 CT Lung and Infection Segmentation Dataset*. Version 1.0. 2020.

[185] N. N. Y. Tsang, H. C. So, K. Y. Ng, B. J. Cowling, G. M. Leung, and D. K. M. Ip. "Diagnostic performance of different sampling approaches for SARS-CoV-2 RT-PCR testing: a systematic review and meta-analysis". In: *The Lancet Infectious Diseases* 21.9 (2021), pp. 1233–1245.

[186] G. E. Humpire-Mamani, L. Builtjes, C. Jacobs, B. van Ginneken, M. Prokop, and E. T. Scholten. *Dataset for: Kidney abnormality segmentation in thorax-abdomen CT scans*. Zenodo, June 2023.

[187] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. "Big transfer (bit): General visual representation learning". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer. 2020, pp. 491–507.

[188] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. Ribeiro, and Y. Zhang. "Sparks of artificial general intelligence: Early experiments with gpt-4". In: *arXiv preprint arXiv:2303.12712* (2023).

[189] 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". In: *Nature methods* 18.2 (Dec. 7, 2020), pp. 203–211. ISSN: 1548-7091.

[190] J. Yang, R. Shi, and B. Ni. "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 191–195.

[191] 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. Fayad, and Y. Yang. "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning". In: *Radiology: Artificial Intelligence* 4.5 (2022), e210315.

[192] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng. "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis". In: *Ophthalmology* 121.11 (2014), pp. 2081–2090.

[193] P. Mokhles, J. S. Schouten, H. J. Beckers, A. Azuara-Blanco, A. Tuulonen, and C. A. Webers. "A systematic review of end-of-life visual impairment in open-angle glaucoma: an epidemiological autopsy". In: *Journal of Glaucoma* 25.7 (2016), pp. 623–628.

[194] A. R. Ran, C. C. Tham, P. P. Chan, C.-Y. Cheng, Y.-C. Tham, T. H. Rim, and C. Y. Cheung. "Deep learning in glaucoma with optical coherence tomography: a review". In: *Eye* 35.1 (2021), pp. 188–201.

[195] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, and R. Garnavi. "A feature agnostic approach for glaucoma detection in OCT volumes". In: *PloS one* 14.7 (2019), e0219126.

[196] D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks. "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening". In: *New England Journal of Medicine* 365 (2011), pp. 395–409.

[197] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W. P. Mali, F. A. A. Mohamed Hoesein, P. M. A. van Ooijen, J. G. J. V. Aerts, M. A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J. E. Walter, K. Ten Haaf, H. J. M. Groen, and M. Oudkerk. "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial". In: *New England Journal of Medicine* 382 (6 Feb. 2020), pp. 503–513.

[198] K. V. Venkadesh, A. A. A. Setio, A. Schreuder, E. T. Scholten, K. Chung, M. M. W Wille, Z. Saghir, B. van Ginneken, M. Prokop, and C. Jacobs. "Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT." In: *Radiology* 300 (2 Aug. 2021), pp. 438–447. ISSN: 1527-1315. ppublish.

[199] W. Siegel Miller and Jemal. "Cancer statistics, 2023". In: *CA Cancer Journal for Clinicians* 73 (2023), pp. 17–48.

[200] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". en. In: *CA Cancer J Clin* 71.3 (Feb. 2021), pp. 209–249.

[201] N. Alves and L. Boulogne. *Breast MRI molecular cancer subtype*. Zenodo, May 2023. DOI: `https://doi.org/10.5281/zenodo.7956360`.

[202] N. Alves and L. Boulogne. *Retina OCT glaucoma*. Zenodo, May 2023. DOI: `https://doi.org/10.5281/zenodo.7957454`.

[203] N. Alves and L. Boulogne. *Lung nodule CT false positive reduction*. Zenodo, May 2023. DOI: https://doi.org/10.5281/zenodo.8042115.

[204] N. Alves and L. Boulogne. *Rib CT fracture*. Zenodo, May 2023. DOI: https://doi.org/10.5281/zenodo.8054740.

[205] N. Alves and L. Boulogne. *Prostate MRI clinically significant cancer*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8042132.

[206] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 1*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.7969800.

[207] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 2*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8042589.

[208] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 3*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8042817.

[209] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 4*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8043089.

[210] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 5*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8043216.

[211] N. Alves and L. Boulogne. *Lung CT COVID-19 batch 6*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8043218.

[212] N. Alves and L. Boulogne. *Kidney CT Abnormality*. Zenodo, June 2023. DOI: https://doi.org/10.5281/zenodo.8043408.

[213] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention*. Vol. 9351. Lecture Notes in Computer Science. 2015, pp. 234–241.

[214] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[215] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[216] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[217] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815* (2017).

[218] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

Acknowledgements

I would like to express my heartfelt gratitude to **Bram**. Thank you for sharing your knowledge on medical image analysis, for providing actionable feedback on my research and writing, and for making time for me in your busy schedule. Access to your network was invaluable for our research. Your continued support until the end of my PhD means a lot to me.

To **Erik**, thank you for your feedback and the understanding that you provided me with about the clinical side of our research. Thank you for the opportunity of observing relevant medical procedures. These experiences provided me with insights that proved useful throughout my PhD and beyond.

Colin, I highly appreciate the feedback you provided on various projects. Thank you for providing an interactive environment in the BodyCT group, where we learned a great deal from each other.

Jean-Paul, thank you for our frequent meetings and stimulating discussions early during my PhD, and for your kindness and mentorship during those times.

I would like to thank **Nikolas** and **Clarisa** for leading the inspiring CORADS-AI effort that I was lucky to be a part of. A special thanks to **Kiran**, **Coen**, and **Cheryl** for our collective work on the COVID-19 grading model in CORADS-AI, which was foundational for much of my further research. Amid the pandemic, our collaboration was inspiring and our discussions contained the first spark for Chapter 5 of this thesis. I would also like to thank Kiran, Coen and Clarisa for providing your expertise and time for that chapter.

I am grateful to **James**, the **RSE team**, and the **grand-challenge team** for building and maintaining the grand-challenge platform, which facilitated much of the work in this thesis. Thank you for your availability and help, particularly during the STOIC2021 challenge.

Marie-Pierre, thank you for reaching out to collaborate on AI research using the STOIC dataset. Your feedback was invaluable throughout our research together. **Yannick**, thank you for your work at the AP-HP side on the STOIC dataset and its integration into the grand-challenge platform. Our meetings were always productive and enjoyable.

Alex, **Chris**, and **Raz**, thank you for sharing your expertise on scalable machine learning, as well as your support of grand-challenge, and of the STOIC2021 challenge in particular.

To all **participants of the STOIC2021 challenge**, thank you for your contributions and efforts that made it a success.

A heartfelt thanks to everyone at Radboudumc who helped with data collection, especially the **data team**, **Karlijn**, and **Joep**, whose support was crucial for my research.

Ferdi and **Roel**, thank you for your help with data collection and for showing me the clinical reality behind the data.

Cristina, thank you for giving me the opportunity to host the DDH. It was a wonderful chance to engage with DIAG members and external researchers. I learned a great deal in the process. I also appreciate the **ThiraLab members** for the enriching presentations, meetings, and discussions.

Thank you to **Natália**, **Luc**, **Khrystyna**, **Alessa**, and **Anindo** and all others who have provided invaluable contributions to Chapter 5 of this thesis. Thank you for lending your time and your expertise on the data sets, classification tasks, and data processing.

My gratitude also goes out to the **cluster team** and in particular **Stephan** for maintaining the cluster on which many of the models described in this thesis were trained.

My heartfelt thanks to all the wonderful people I had the pleasure of meeting through this research and the joy of spending time with outside of work. A special thanks to **Anton** and **Nicole** for sharing a mutual love of heavy board games. And also to **Riccardo** and **Eva**, **Thomas** and **Sanne**, and **Kiran** and **Ann** for the nice dinners and game nights. Kiran and Ann, I'm grateful to you for looking after Valiente and making him feel secure. A big thanks to the **jamming crew** at Anindo's for the great music and spicy food.

Thank you to all **DIAG and AXTI members** I've interacted with, whether through fun activities or insightful discussions. A special thanks to my **office mates in room 23**, who provided a great working atmosphere.

I would like to thank all my previous teachers. Your support has been essential in shaping my path.

To **Valiente**, for his companionship and for reminding me of life's simplicity.

Thank you to **my family** for always being there for me. Thank you for the healthy competition between brothers and for the continued, selfless, hands-on support from my parents.

Thank you **Marenthe** for helping me handle the stress, for looking out for my health, for the counterbalance to work, for your love, and for the room I am writing this in.

Curriculum Vitae

Luuk Boulogne was born on May 12, 1994, in Enschede, the Netherlands. Luuk studied Artificial Intelligence at the University of Groningen. In 2016, he obtained his Bachelor's degree, during which he gained a broad understanding of the field. He obtained his Master's degree in 2018, specializing in autonomous perceptive systems and deep learning. In 2019, he began his PhD in the Diagnostic Image Analysis Group under supervision of Bram van Ginneken, Erik van der Heijden, and Colin Jacobs. His research focused on 3D medical image classification and regression, with an emphasis on processing lung CT scans. The results of his research are described in this thesis.

PhD Portfolio

Name: Luuk Boulogne
Graduate school: Radboudumc Research Institute for Medical Innovation
Department: Medical Imaging, DIAG
PhD period: 01-04-2019 until 31-03-2023
PhD Supervisors: Prof. dr. B. van Ginneken, Prof. dr. H.F.M. van der Heijden
PhD Co-supervisor(s): Dr. ir. C. Jacobs

Courses	Year(s)	Hours
Introduction day Radboudumc	2019	6
RIHS introduction course for PhD students	2019	15
Achieving your goals and performing more successfully in your PhD	2019	28
Scientific integrity	2021	20
The next step in my career	2022	28
Seminars		
Radiology research meeting	2019	12
Deep learning Nijmegen meetup	2019	14
ThiraLab	2019-2020	28
BodyCT meeting	2019-2023	84
DIAG discussion hour	2019-2023	112
Deep learning journal club	2019-2023	28
Symposia & conferences		
RIHS PhD retreat	2021	32
Colourful AI: Current advances in Medical Imaging	2022	6
Radiological Society of North America	2022	56
Teaching activities		
ICAI Lecture	2019	14
Artificial Intelligence for Health	2020	112
Supervision of master student	2021	112
Total		707

Research Data Management

Ethics and privacy

This thesis is based on the results of research involving human participants and existing data from published papers, which were conducted in accordance with relevant national and international legislation and regulations, guidelines, codes of conduct and Radboudumc policy.

The institutional ethical review committee CMO Radboudumc, Nijmegen, the Netherlands has given approval to conduct the studies for which data was collected from Radboudumc (CMO2016-3045: Project 19049 for Chapter 2, Project 20027 for Chapter 3, Project 19010 for chapter 5). The protocol for the collection of the data used in Chapter 4 can be accessed through ClinicalTrials.gov with identifier NCT04355507. The rest of the data used to conduct the studies presented in this thesis were collected from publicly available sources.

The privacy of all participants in these studies was warranted by the use of either pseudonymization or full anonymization. In case of pseudonymization, the key was stored on a secured network drive that was only accessible to members of the project who needed access to it because of their role within the project. The pseudonymization key was stored separately from the research data.

Data collection and storage

Diagnostic Image Analysis Group (DIAG) data managers and the DIAG data team collected the Radboudumc data used for the studies described in Chapters 2, 3, and 5 from PACS and pseudonomyzed these data. The data from the COPDGene study for Chapter 2 was obtained through a formal application process. The data for Chapter 4 was collected and pseudonymized by Assistance Publique – Hôpitaux de Paris (AP-HP). All other data was collected from publicly available sources.

All data used for Chapters 2, 3, and 5 is securely stored within the Radboudumc storage system. The data for Chapter 4 is stored at AP-HP. All scientific experiments within the context of this thesis conducted on data that is not publicly available have been executed exclusively either within the Radboudumc IT infrastructure or on secure cloud computing platforms hosted by Amazon Web Services. These secure storage options safeguard the availability, integrity and confidentiality of the data.

Data sharing according to the FAIR principles

Training set A described in chapter 4 has been made publicly available under a CC-BY-NC 4.0 licence on the AWS registry of open data at `https://registry.opendata.`

aws/stoic2021-training/. The training data for chapter 5 has been made publicly available on Zenodo. The corresponding licences and URLs with DOIs can be found in table 5.1. The interoperable MetaImage Header Archive and Comma Separated Values file formats were used for sharing these data.

Radboudumc is the legal owner of the Radboudumc data used for the studies described in Chapters 2, 3, and 5. These data are stored within the Radboudumc IT infrastrucure. AP-HP de Paris is the legal owner of the STOIC data used for Chapter 4. These data are stored at AP-HP. Any collaborative research project led by an academic partner who requires access to the STOIC data shall be analyzed, validated, and authorized by the Steering Committee of STOIC. To this end, the academic partner shall send a document describing the research project to the Stoic Steering Committee at the following email address: marie-pierre.revel@aphp.fr, with the following subject: STOIC DATA ACCESS PERMISSION. After acceptance by the Steering Committee, the academic partner shall sign a specific agreement (Data Transfer Agreement - DTA) with AP-HP, who is legally responsible for the STOIC data as Sponsor of the STOIC research. Refer to the STOIC princeps paper [48] for more information.

The model for patient and lobe level lung function estimation presented in chapter 2 is publicly available for use on https://grand-challenge.org/algorithms/lobe-wise-lung-function-estimation/.

For the purpose of fair comparison to new research, solutions to the challenges presented in Chapters 3, 4, and 5 can be submitted to https://covid19.grand-challenge.org/, https://stoic2021.grand-challenge.org/ and https://auc23.grand-challenge.org/ respectively.

Baseline codebases with submission tutorials for the challenge presented in Chapter 4 were published at https://github.com/luukboulogne/stoic2021-baseline and https://github.com/luukboulogne/stoic2021-baseline-finalphase under the MIT license. Links to the codebases of the finalist solutions to this challenge and their corresponding licences can be found in Table 4.2

A baseline codebase with a submission tutorial for the challenge presented in Chapter 5 was published at https://github.com/DIAGNijmegen/auc23-baseline under the Apache-2.0 license.

9 789465 150277

Radboudumc university medical center

IN·DEI·NOMINE·FELICITER

Radboud University